ZE3RA: The ZEPLIN-III Reduction and Analysis Package

F. Neves[*,a,b], D.Yu. Akimov[c], H.M. Araújo[b], E.J. Barnes[d], V.A. Belov[c],
A.A. Burenkov[c], V. Chepel[a], A. Currie[b], L. DeViveiros[a], B. Edwards[e],
C. Ghag[d], A. Hollingsworth[d], M. Horn[b], G.E. Kalmus[e], A.S. Kobyakin[c],
A.G. Kovalenko[c], V.N. Lebedenko[b], A. Lindote[a,e], M.I. Lopes[a], R. Lüscher[e],
P. Majewski[e], A.St J. Murphy[d], S.M. Paling[e], J. Pinto da Cunha[a], R. Preece[e],
J.J. Quenby[b], L. Reichhart[d], S. Rodrigues[a], P.R. Scovell[d], C. Silva[a],
V.N. Solovov[a], N.J.T. Smith[e], P.F. Smith[e], V.N. Stekhanov[c], T.J. Sumner[b],
C. Thorne[b], R.J. Walker[b]

[a]*LIP–Coimbra & Department of Physics of the University of Coimbra, Portugal*
[b]*High Energy Physics group, Blackett Laboratory, Imperial College London, UK*
[c]*Institute for Theoretical and Experimental Physics, Moscow, Russia*
[d]*School of Physics & Astronomy, University of Edinburgh, UK*
[e]*Particle Physics Department, STFC Rutherford Appleton Laboratory, Chilton, UK*

Abstract

ZE3RA is the software package responsible for processing the raw data from the ZEPLIN-III dark matter experiment and its reduction into a set of parameters used in all subsequent analyses. The detector is a liquid xenon time projection chamber with scintillation and electroluminescence signals read out by an array of 31 photomultipliers. The dual range 62-channel data stream is optimised for the detection of scintillation pulses down to a single photoelectron and of ionisation signals as small as those produced by single electrons. We discuss in particular several strategies related to data filtering, pulse finding and pulse clustering which are tuned to recover the best electron/nuclear recoil discrimination near the detection threshold, where most dark matter elastic scattering signatures are expected. The software was designed assuming only minimal knowledge of the physics underlying the detection principle, allowing an unbiased analysis of the experimental results and easy extension to other detectors with similar requirements.

Key words: ZEPLIN-III, liquid xenon detectors, dark matter, signal analyses, data reduction

---

[*]Corresponding author:
   Email address: neves@coimbra.lip.pt (F. Neves)

1. Introduction

The present work evolved within the ZEPLIN-III experiment, a two-phase (liquid/gas) xenon detector aiming to measure very low energy nuclear recoils produced by the interaction of dark matter WIMPs (Weakly Interactive Massive Particles) [1, 2, 3, 4]. Searches for rare particle interactions in low-background physics experiments are inherently difficult, and any data reduction and analysis software must address two main challenges. Firstly, unusual event topologies will almost inevitably appear in the long exposures required for the science acquisitions (lasting typically for many months or even years). These are not exercised by the various calibration runs and may result from localized instabilities or unexpected backgrounds, for example. Secondly, WIMP-nucleus interactions are expected to result in very small energy transfers to xenon atoms, and so the search for scintillation and ionization signatures from WIMPs goes down to the quantum of response in these channels, i.e. to the photoelectron level in scintillation and to single electrons in ionization. At this level, signals become sparser in time and therefore less recognizable; inevitably, statistically-motivated pulse-finding techniques are required in order to separate real pulses from noise or to distinguish them from unrelated pulses. When signals are not only rare but also extremely small, the potential for their mis-parametrization (and consequent detection inefficiency) is high.

The ZEPLIN-III data Reduction and Analysis (ZE3RA) package must therefore be very accurate at separating small signals from the noise, thus maximizing sensitivity to WIMPs. It must deal with unexpected rare artefacts arising in very long measurements; these include internal effects (e.g. occasional micro-discharges), changes in the local underground environment (pressure, temperature, structural movements), long-term electronic drifts and occasional electromagnetic pick-up, and many others. It must also be very flexible since our knowledge of these effects improves as more data are accumulated. For these reasons we opted for deferring the physics interpretation of the detector response to a later stage in the data analysis. ZE3RA provides robust reduction of the raw data acquisition output in the form of full pulse parametrization. Other requirements include a powerful event display and the possibility to deal with blind analyses, whereby specific events or datasets present in the data repository should not be displayed or reduced until the final analysis stage so as not to bias the signal estimation.

In the context of direct WIMP searches, the ultimate benchmark of such a package is the level of electron/nuclear recoil discrimination. This is quantified by the relative number of events belonging to the heavily populated electron-recoil background population (mainly gamma-rays, with high ionization/scintillation ratios) that leak into the acceptance region of the signal (mimicking nuclear recoils with low such ratios). ZEPLIN-III achieved the best discrimination of any two-phase xenon detector reported so far and, in spite of a relatively high gamma-ray background, achieved a sensitivity for WIMP signals amongst the best in the world [4, 5, 6]. This relied in part on the ability of ZE3RA to parametrize events correctly. Some of the algorithms contained

within it can be applied to other detector arrays, not only in the context of underground experiments.

2. Setup and Data Processing

ZEPLIN-III is a two-phase (liquid/gas) xenon time projection chamber containing ≈ 12 kg of liquid xenon above a compact hexagonal array of 31 2-inch photomultipliers (ETL D766QA) [2, 3, 4]. The photomultipliers (PMTs) are immersed directly in the liquid at a temperature of -105^oC and record both the rapid scintillation signal (S1) and a delayed second signal (S2) produced by proportional electroluminescence in the gas phase from charge drifted out of the liquid [7]. The electric field in the active xenon volume is defined by a cathode wire grid 36 mm below the liquid surface and an anode plate 4 mm above the surface in the gas phase. These two electrodes define a drift field in the liquid of ≈ 4 kV/cm and an electroluminescence field in the gas of ≈ 8 kV/cm. A second wire grid is located 5 mm below the cathode grid just above the PMT array. This grid defines a reverse field region which suppresses the collection of ionization charge for events just above the array and helps to isolate the PMT input optics from the strong external electric field.

The PMT signals are digitized at 2 ns sampling over a time segment of 36 μs starting at -20 μs from the trigger point. Each PMT signal is fed into two 8-bit digitizers (ACQIRIS DC265) with a $\times 10$ gain difference between them provided by fast amplifiers (Phillips Scientific 770), to obtain both high (HS) and low (LS) sensitivity readout covering a wide dynamic range. The PMT array is operated from a common HV supply with attenuators (Phillips Scientific 804) used to normalize their individual gains. The trigger is generated using the shaped sum of the HS signals from all the PMTs. For the sake of illustration, Fig. 1 shows the sums of all HS and LS channels for a low energy multiple scattering event triggered by the first S2 signal.

2.1. Software architecture

The efficient analysis of the huge amount of data produced by the acquisition system (DAQ) demands its previous reduction to a set of relevant physical parameters (pulse timing estimators, height, area, etc). In the ZEPLIN-III experiment this task is performed by the ZE3RA software. ZE3RA is implemented in C++ following a comprehensive class-oriented architecture mimicking the various DAQ functional stages (run, event, channel, etc). The software design allows easy plug-in of new tools and to build different reduction templates targeting specific analyses. These features are of key importance for the optimization of the analysis which should produce consistent results over a large variety of acquisition scenarios lasting from a few hours to several months (different field configurations, calibration sources, etc). The ZE3RA architecture, illustrated schematically in Fig. 2, includes the skeleton classes:

- The *ZDisplay* class layers the end user interaction with the classes managing the analysis and holding both the raw and the reduced data structures.
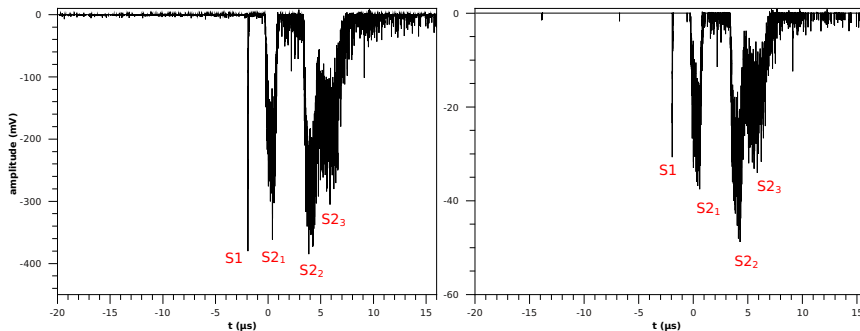
3

Figure 1: Sum of all HS (left) and LS (right) channels for a low energy event having 3 energy depositions in the liquid xenon sensitive volume. The distance between the fast S1 signals (all overlapping in time) and the corresponding S2 is a measure of the depth of the interaction. For each S2 pulse the light distribution over the PMT array allows to reconstruct the position of the interaction in the (x,y)-plane.

- The *ZRun* class inherits from *ZEvent* and manages all the settings and the different reduction templates.

- The *ZEvent* class stores and manages the access to individual events and folded data structures (i.e. *ZChannel*). It inherits from *ZRawFileHandler* and *ZNtupleFileHandler* which layer, respectively, the input from the raw DAQ data files and the output to the databases containning the reduced quantities. The *ZRawFileHandler* class inherits from *ZBlindManager* which implements the access policy to the raw data used in the WIMP search .

- The *ZChannel* class stores and manages the data from individual channels and all contained structures (i.g. *ZPulse*).

- The *ZPulse* and *ZCluster* classes store all information related with individual pulses and clusters of sequential pulses, respectively. It inherits from *ZStatistic* and *ZPhysics* which gather and maintain, respectively, statistical and physical data.

In the following subsections we present a short overview of the most relevant algorithms implemented and tested in the ZE3RA analysis framework for the ZEPLIN-III experiment. It should be noted that not all of those algorithms were used for the production of published results, either for the first or second science runs (in 2008 and 2010/11, respectively). In particular, the moving average algorithm (§2.3.2) was preferred to the more refined wavelet analysis (§2.3.4) as a trade-off between speed and performance. Rather than lessening the purpose of the software package, this puts the emphasis on its design and architecture as a general framework easily extensible to different acquisition scenarios and similar experiments with specific needs.
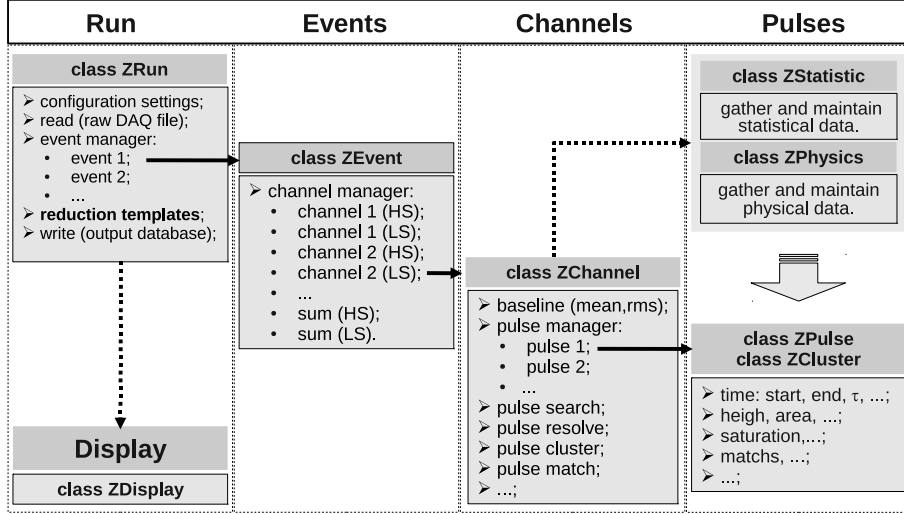
Run	Events	Channels	Pulses

class ZRun

- ➢ configuration settings;
- ➢ read (raw DAQ file);
- ➢ event manager:
  - • event 1;
  - • event 2;
  - • ...
- ➢ **reduction templates**;
- ➢ write (output database);

class ZEvent

- ➢ channel manager:
  - • channel 1 (HS);
  - • channel 1 (LS);
  - • channel 2 (HS);
  - • channel 2 (LS);
  - • ...
  - • sum (HS);
  - • sum (LS).

class ZChannel

- ➢ baseline (mean,rms);
- ➢ pulse manager:
  - • pulse 1;
  - • pulse 2;
  - • ...
- ➢ pulse search;
- ➢ pulse resolve;
- ➢ pulse cluster;
- ➢ pulse match;
- ➢ ...;

class ZStatistic

gather and maintain statistical data.

class ZPhysics

gather and maintain physical data.

class ZPulse
class ZCluster

- ➢ time: start, end, τ, ...;
- ➢ heigh, area, ...;
- ➢ saturation,...;
- ➢ matchs, ...;
- ➢ ...;

Display

class ZDisplay

Figure 2: Schematic representation of the ZE3RA software architecture.

2.2. Baseline characterization.

The baseline is parametrized using the waveforms containing the actual PMT signals. To avoid any bias due to the occurrence of transients or small spurious signals, the parametrization method relies on a consistency check of the noise distribution variance during a sufficiently large time window. For that purpose, the DAQ *pre*-trigger region is divided into $i = 1..M_0$ consecutive regions containing m samples each. For each of these regions, the variances $\left\{\sigma_i^2,\ i = 1..M_0\right\}$ of the signal amplitude distribution are calculated. The *F-distribution* probability function (Q) is used to check if the variances are statistically consistent:

$$Q = \frac{\Gamma(2\nu)}{2\Gamma(\nu)} \int_0^{(1+F)^{-1}} (t - t^2)^{\nu-1} dt \ , \tag{1}$$

where $\nu = (m - 1)/2$ and

$$\left\{ \begin{array}{ll} F = \sigma_i^2/\sigma_{i+1}^2, & \sigma_i > \sigma_{i+1} \\ F = \sigma_{i+1}^2/\sigma_i^2, & \sigma_i \leq \sigma_{i+1} \end{array} \right. \ .$$

Q is therefore the significance level at which that hypothesis ($\sigma_i^2 \equiv \sigma_{i+1}^2$, $i = 1..M_0$) can be rejected [8]. For each of the M_0 regions, the means $\left\{\mu_i^2,\ i = 1..M_0\right\}$ of the signal amplitude distribution are also calculated. The noise (σ_{bas}) and mean (μ_{bas}) characterizing each waveform are then defined as

$$\sigma_{bas} = \langle\sigma_i\rangle,\ i = 1..M \ , \tag{2}$$

$$\mu_{bas} = \langle\mu_i\rangle,\ i = 1..M \ , \tag{3}$$

5

for those M regions satisfying $Q < Q_{crit}$. For the ZEPLIN-III analysis the values of $Q_{crit} = 0.0001$ and $m = 25$ (50 ns) were used. The maximum length of the total sampled waveform was 2 μs ($M_0 = 40$). For each event the σ_{bas}, μ_{bas} and M values are stored for all channels and can be used, for example, to identify misbehaving baselines.

2.3. Raw data filtering.

In order to enhance the signal-to-noise ratio and help with the identification of relevant pulse structures, a set of general filter algorithms are available in ZE3RA. Besides the built-in filters, which are briefly described below, the software framework allows easy plug-in of new algorithms and their use at any defined configuration. It should be noted that the DAQ raw data is never modified during the analysis; instead, an auxiliary buffer containing the filtered data is maintained for every channel.

2.3.1. Moving average.

The implemented moving average algorithm is a simple low-pass filter defined as

$$\overline{y}_i = \frac{\sum_{k=i-m}^{i+m} y_k}{2m+1} \, , \tag{4}$$

where $\overline{y}$ and y represent, respectively, the filtered and raw data buffers. Although appealing because of its speed, the moving average produces a significant loss of information when pulses are narrow compared with the filter width ($2m+1$). The loss of information or even distortion of the filtered signal ($\overline{y}$) is an effect of the weight given to the k-th data point in Eq. 4 being independent of its distance to i. In ZEPLIN-III this presents a problem since the S1 and S2 signals have very different time constants, respectively, ~ 30 ns and ~ 0.6 μs (parametrized by the signal mean arrival time). If m is tuned for S2 then S1 information is lost. Conversely, setting m to preserve S1 does not help with the detection of S2 pulses. These issues related with the time scale at which one is filtering or looking for a signal are addressed by both the methods described in §2.3.2 and §2.3.4.

2.3.2. Moving average with variable width.

A solution to the problem highlighted above, which retains most of the simplicity and speed of the moving average algorithm, is to adapt the width of the filter (Eq. 4) based on some characterization $Q_y(i, m_i)$ of the local data being analysed:

$$\overline{y}_i = \frac{\sum_{k=i-m_i}^{i+m_i} y_k}{2m_i+1} \, , \tag{5}$$

with

$$\begin{cases} m_{i+1} = m_i + 1, & Q_y(i, m_i) < Q_{thr} \bigwedge m_i < m_{max} \\ m_{i+1} = m_i - 1, & Q_y(i, m_i) \geq Q_{thr} \bigwedge m_i > m_{min} \end{cases},$$

where m_{min}, m_{max} and Q_{thr} are user-defined parameters controlling the limits and the adaptation of the filter width. In the ZEPLIN-III analysis $Q_y(i, m_i)$ was set to the variance of y calculated in the $[i - m_i, i + m_i]$ interval. Typical values of $m_{min} \sim 35$ (70 ns), $m_{max} \sim 100$ (0.2 μs) and $Q_{thr}^{1/2} = 3\sigma_{bas}$ (Eq. 2) show good performance in terms of both S1 and S2 pulse detection. Figures 3 and 4 show examples of applying this filter to typical ZEPLIN-III events and it comparison with the results obtained using the algorithm described in §2.3.4.

2.3.3. Fourier analysis.

Besides the random noise intrinsic to any signal processing chain, in ZEPLIN-III one can also observe the occurrence of coherent noise. This noise is most often induced by electric equipment working in close vicinity to the detector or following the saturation of the amplifiers. In both cases the noise exhibits time periodicity and is composed of a small set of characteristic frequencies. These properties make the Fast Fourier Transform (FFT) analysis particularly suitable for the identification and removal of such occurrences.

The FFT is an efficient computational tool to calculate the *Fourier* transform of a function (y) sampled at a finite number of N points $\{y_n, n = 0..N - 1\}$ [8]. In the frequency domain the amplitude of the component $f_n = n/N\Delta$ is given by

$$H(f_n) = \Delta \sum_{k=0}^{N-1} y_k e^{i2\pi kn/N} = \Delta \sum_{k=0}^{N-1} \left[\cos\left(\frac{2\pi kn}{N}\right) - i\sin\left(\frac{2\pi kn}{N}\right) \right], \quad (6)$$

where Δ represents the sampling time interval. In ZE3RA the coefficients $H(f_n)$ are calculated using the FFTW library [9].

This FFT tool was used only on a dedicated dataset acquired to study single electron emission from the liquid into the gas phase [10]. For the acquisition of that dataset the DAQ was triggered externally with a pulse generator which accidentally induced coherent noise into the HS channels. After calculating the FFT coefficients for each timeline, a 10-*pole Butterworth* filter was used to attenuate $H(f_n)$ (Eq. 6) corresponding to the set of noise frequencies $\{f_{noise}\}$ contained in each channel. This procedure allowed successful recovery of the signal as shown by the comparison of the single electron emission results obtained using this particular dataset with the results from another method described in Ref. 10 using WIMP search data sets.

2.3.4. Wavelet analysis.

Resulting from being encoded both in terms of amplitude and phase of sines or cosines (Eq. 6), the FFT analysis gives no direct information about the time occurrence of transients [8]. One solution to this problem would be to divide the

entire domain into small regions and analyse them separately. Nevertheless, this method would imply the loss of information at lower frequencies when compared with the duration of the region being analysed. Another solution allowing to preserve both time and frequency information is to use the Discrete Wavelet Transform (DWT) analysis [8].

The DWT scales (s) and shifts (τ) a mother function $\mathcal{F}$ along the time domain while recording its level of correlation with the signal into a set of coefficients $w(s, \tau)$:

$$w(s,\tau) = \sum_{i=-\infty}^{+\infty} y_i * \mathcal{F}_{(s,\tau),i} \,, \tag{7}$$

where y represents the raw data stream and $\mathcal{F}_{(s,\tau)}$ are the set of basis functions obtained from scaling and translating $\mathcal{F}$ at fixed (s, τ) steps. Unlike the sinusoidal functions which define a unique FFT of the signal, there are many possibilities for $\mathcal{F}$ producing different DWT coefficients. The choice of $\mathcal{F}$ is based on the trade-off between localization/smoothing (time/frequency domain) required for a particular application. The ZE3RA framework simply wraps the DWT code available from GSL [11] and expands its functionality to facilitate the manipulation of the coefficients $w(s, \tau)$ (Eq. 7). The available $\mathcal{F}$ family functions are: Haar, Daubechies and biorthogonal b-spline [11].

Currently ZE3RA provides two different noise removal and smoothing algorithms based on DWT decomposition. The first of these algorithms implements the soft threshold technique described in Ref. 12 for a uniformly distributed *Gaussian* noise: for a given scale s the N_S wavelet coefficients are translated towards 0 by an amount

$$\delta_s = \frac{\sqrt{2 \ln N_s}}{0.6745} MAD \,, \tag{8}$$

where MAD represents the median absolute deviation of $\{w_j(s,\tau) \, j = 1..N_s\}$. The second algorithm combines data smoothing with the edge detection and preservation method described in [13]. The edge detection is based on the multiscale behavior of the local maxima of the wavelet coefficients moduli (Eq. 7). The value of $|w(s,\tau)|$ measures the derivative of the smoothed signal at the scale s and a signal sharp variation (e.g. S1-like pulses) produces moduli maxima at different scales [13]. After calculating the moduli of the wavelet coefficients and finding local maxima, the implemented algorithm maps and stores the interscale evolution of the $|w(s,\tau)|$ values. This information can be used to study the Lipschitz regularity of the signal and further selection of the edge/pulse types to retain [13]. The signal coefficients $w(s,\tau)$ which do not belong to a valid edge structure can either be smoothed using the soft threshold method described above (Eq. 8) or simply reset to 0 at selected higher order scales (higher frequencies). An additional benefit from using the second algorithm to smooth the data is the intrinsic availability of the time position of the edges even in the occurrence of a baseline drift. This information can later be matched against the algorithm described in §2.5 to enhance the pulse finding efficiency.

Figures 3 and 4 show the comparison between the results of the wavelet analysis with edge detection and of the moving average with adaptive width (§2.3.2) when applied to two typical events. The DWT decomposition was done into 14 scales using a bi-orthogonal b-spline mother function ($\mathcal{F}$) of order $(1, 3)$ [11]. For the sake of illustration, both figures show the result of clearing all coefficients above the 8th scale ($w(s \geq 8, \tau) = 0$) together with the effect of keeping those up to $s = 10$ when belonging to an identified S1-like edge. Besides improving the sensitivity to the general shape and start time of the fast S1 signals (Fig. 4), smoothing the data using the edge detection algorithm also improves the discrimination between S2 and S1 when they are very close (Fig. 3). The latter class of events corresponds to energy depositions near the surface of the liquid.
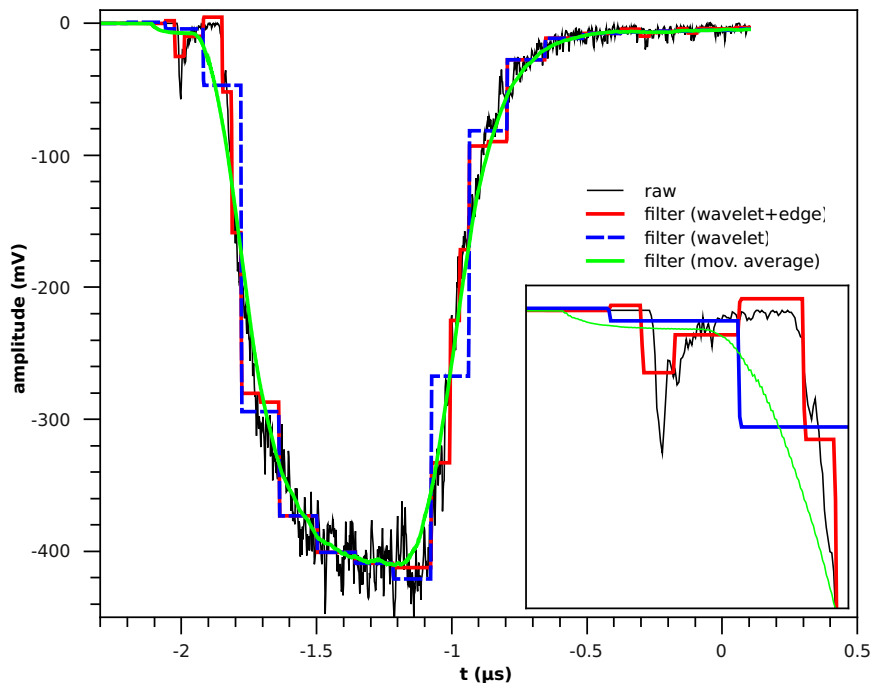
Figure 3: Comparison of the smoothing results obtained using the moving average with variable width (green line) and a DWT decomposition keeping (red line) or ignoring (blue line) the edge information. This event corresponds to an interaction just below the liquid xenon surface. In this instance the small S1 pulse, shown also in the inset, proceeds a much larger S2 by a very short time.

2.4. Channel delays.

Due to differences in the signal processing chain (PMTs, cabling, amplifiers, etc), all ZEPLIN-III channels exhibit relative delays ranging typically from ±10 ns. The correct alignment in time of all channels is crucial for the
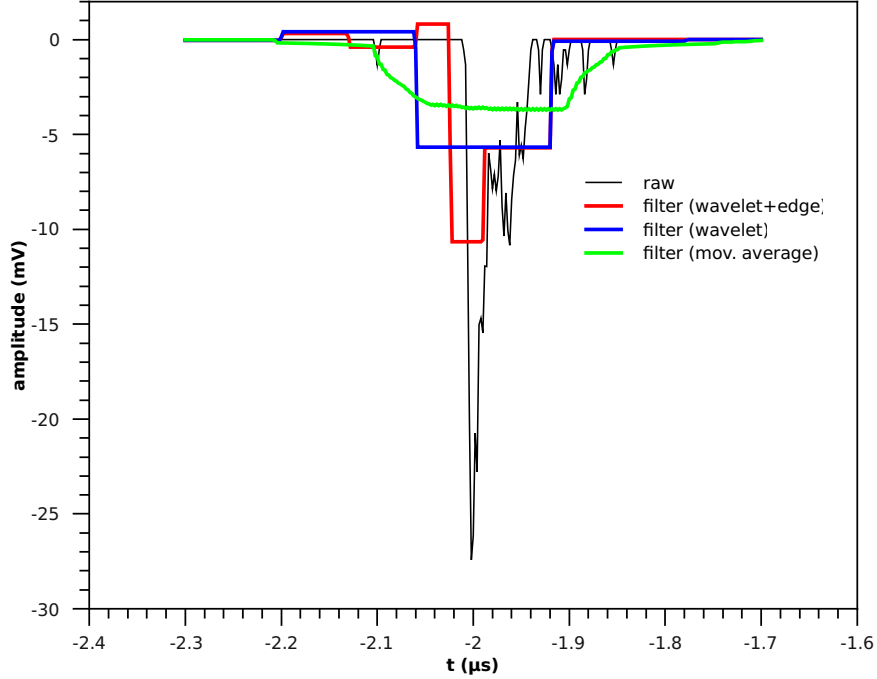
Figure 4: Comparison of the smoothing results for a typical S1 pulse obtained using the moving average with variable width (green line) and a DWT decomposition keeping (red line) or ignoring (blue line) the edge information.

performance of both the pulse finding and matching algorithms (§2.5 and §2.8, respectively). This is specially relevant for signals corresponding to low energy deposits, which constitute the region of interest in ZEPLIN-III.

The individual channels are realigned in ZE3RA by defining the beginning of the raw data buffer $y^{(i)}$ to point at the k-th element of the buffer $Y^{(i)}$ containing the original data read from the DAQ,

$$ k_Y^{(i)} = \delta^{(i)} - \min\left\{\delta^{(j)}, \, j = 1..J\right\} \, , \tag{9} $$

where $\delta^{(i)}$ is the individual delay for channel i and J represents the total number of channels. The size S_y for all y buffers is calculated using

$$ S_y = S_Y + \min\left\{\delta^{(j)}, \, 1..J\right\} - \max\left\{\delta^{(j)}, \, j = 1..J\right\} \, , \tag{10} $$

where S_Y is the size of the original Y buffers. The very simple operations defined in Eq. 9 and Eq. 10 avoid any extra time- or memory-consuming manipulations aside from reading the DAQ files into buffers $\left\{Y^{(j)}, \, j = 1..J\right\}$.

10

2.5. Pulse finding.

The pulse finding procedure implemented in ZE3RA consists of searching for excursions of the signal amplitude above a defined threshold V_{thr}. However, in the ZEPLIN-III detector both S1 and S2 signals can contain several of these excursions depending on the distance from a particular PMT to the interaction point and on the energy deposited. In particular, an S2 signal can consist of a few tens of photoelectrons spread over a period of time of ~ 1 μs. Taking advantage of the underlying structure revealed from filtering (§2.3), this problem was partially solved in ZE3RA by first searching for pulses on the smooth data buffer ($\overline{y}$). One must keep in mind though that the effective enhancement of smoothing depends on the applied filter and how sparsely the individual data excursions occur (§2.3). Regardless of any loss of information in $\overline{y}$, ZE3RA keeps the sensitivity to the smallest structures by also searching for pulses in the original data buffer (y). The final set of pulses available for all subsequent analysis is the union of pulses collected from both $\overline{y}$ and y. For the ZEPLIN-III analysis V_{thr} was chosen to be

$$V_{thr} = \mu_{bas} + 3\sigma_{bas} \,, \tag{11}$$

where σ_{bas} is the noise (Eq. 2) and μ_{bas} the mean (Eq. 3) values characterizing each raw waveform (§2.2). This software threshold is nominally equivalent to an energy threshold of only 1.67 keV for electron recoils detected through S1. As an example, Figure 5 shows a larger pulse which is chosen from $\overline{y}$ with the remaining fastest pulses being picked from the unmodified y buffer.

2.6. Pulse clustering.

Deciding when to stop accruing small excursions above threshold into pulse clusters is not straightforward and has important consequences for the detection efficiency of small signals. For S1 pulses, for example, a fixed-length integration, typically implemented as a coincidence window between channels, can lead to unnecessary inclusion of noise and unequal integration efficiency for different particle species (due to the different scintillation decay time constants). Alternatively, one may cluster subsequent candidates into the pulse based on time separation, although over-clustering can lead to run-away effects in this instance. The fraction of the pulse area integrated in each approach can be calculated analytically if the scintillation responses are known for each species, but this calculation fails for very low photoelectron numbers, when the start time of the pulse is not defined by the rise time of the scintillation signal but rather by the delayed arrival of the first photoelectron. A detailed comparison between the constant integration and the pulse gap methods was carried out with a toy *Monte Carlo* accounting for the DAQ sampling rate, the width of the single photoelectron response and the scintillation decay times and respective intensities for electron and nuclear recoils in liquid xenon [14] (however, noise or afterpulsing distributions were not included). This was used to calculate the
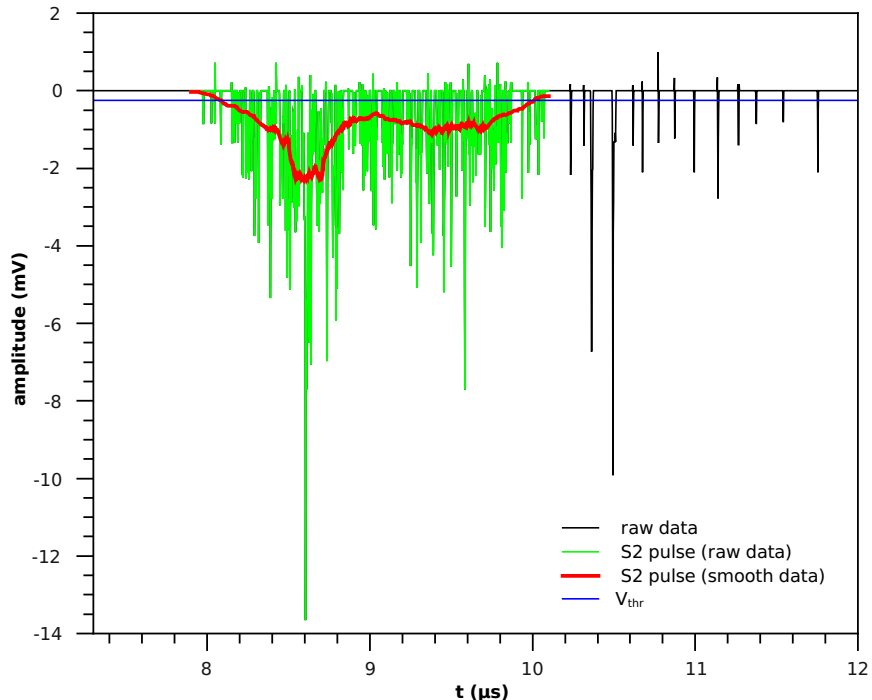
11

Figure 5: Example of a ZEPLIN-III event with a S2 pulse (green line) being selected from an excursion of $\overline{y}$ (red line) above V_{thr} (blue line). The fastest pulses on the right tail of the S2 are chosen from y (black line).

fraction (η) of S1 signals which is lost from the integrated pulse area. The integration of S2 pulses is not so critical as η is expected to be always very small in this case.

The simplest pulse clustering algorithm implemented in ZE3RA consists of merging all pulses found within a time window of constant width (t_{win}). The value of t_{win} is chosen according to the characteristics of the signal (i.e, S1 or S2). The respective *Monte Carlo* results are shown in Fig. 6 as a function of the sampled number of photoelectrons for window sizes of 50 ns and 100 ns. As expected, the missing area fraction η is smaller for nuclear recoils (due to the increase of the faster xenon scintillation component); significantly, η is not constant for small signals, but rather it decreases due to the delayed detection of the first photoelectron, which is a purely statistical effect.

The alternative clustering algorithm implemented uses the time distance between pulses to decide if they correspond to the same interaction in the liquid xenon. The algorithm iterates through pulses and recursively merges consecutive occurrences if the time elapsed between the end of the first and the begin of the next is smaller then a certain value t_{gap}. It should be noted that the saturation tails from the amplifiers and the existence of afterpulsing signals originated

12

in the PMTs [15] can bias the results towards excessive clustering. To mitigate these, one additional constraint is imposed in ZE3RA: that the clustering should not extend out to more then a user defined factor of the maximum mean arrival time of photoelectrons for the pulses being clustered. The *Monte Carlo* results are also shown in Fig. 6 for $t_{gap} = 25, 60$ ns. For pairs of values (t_{gap}, t_{win}) returning similar values at the lowest number of sampled photoelectrons, the performance of the gap method improves quickly with the magnitude of the signal and almost independently of the type of particle.

Using real detector data, we verified that the time gap algorithm was more robust than the constant time window option when dealing with different acquisition scenarios (different fields, detector tilt or changes in the gas gap, etc) and with the interaction of different particles in the liquid xenon (from gamma and neutron calibrations). We also found that the constant integration method, with t_{win} tuned for specific conditions and single scatters in the liquid, often clipped multiple overlapping pulses, therefore biasing their parametrization and correct identification. These reasons, and the potential for better performance on rare pulse topologies which may arise in very long exposures, led us to adopt the time difference method, with a simple heuristic adjustment of t_{gap} to 20 ns and 100 ns for S1- and S2-like pulses, respectively.
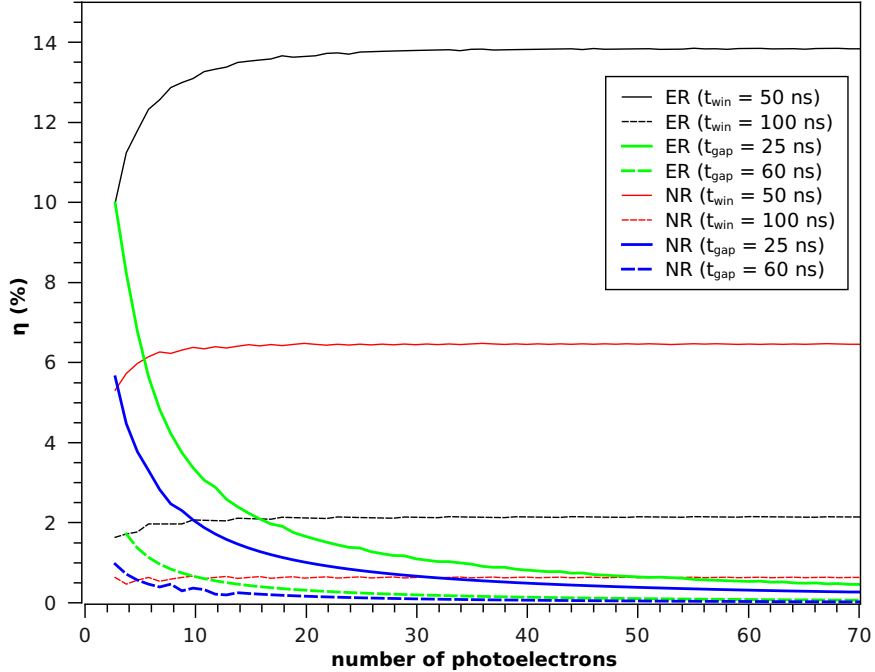

Figure 6: Comparison for both electron (ER) and nuclear (NR) recoils in the liquid xenon of the fraction of lost area of S1 signal (η) when clustering pulses using a constant time window ($t_{win} = 50, 100$ ns) or the time gap between pulses ($t_{gap} = 25, 60$ ns).

2.7. Multiple pulse resolution.

In the ZEPLIN-III experiment only events with one S1 and one S2 are considered for most analysis (e.g. WIMPs search). Events corresponding to multiple scatters can be promptly identified using the S2 channel if the individual energy deposits occur at different depths in the LXe active volume. The time separation between the S1 and S2 signals is equal to time taken by the ionization electrons to travel from the interaction site to the liquid surface along the drift field. This mechanism is independent of the position of the interactions in the (x, y)-plane which, in any case, can only be obtained using position reconstruction after the correct identification and parametrization of the pulses.

The algorithm implemented in ZE3RA to identify and resolve multiple scatter events with overlapping S2 signals consists simply of reusing the pulse finding and clustering algorithms described in §2.5 and §2.6 for higher thresholds. For each S2 a list of thresholds is obtained by scanning the smooth buffer ($\overline{y}$) and accumulating the values $V_l = \overline{y}_l + V_{thr}$ (Eq. 11) obeying:

$$\overline{y}_k - \overline{y}_l > V_{thr} \wedge \overline{y}_m - \overline{y}_l > V_{thr},\, k < l < m\,, \tag{12}$$

where k and m are constrained by the candidate pulse start and end times (t_{start}, t_{end}). The pulse finding and clustering algorithms are then applied to the data sub-domain defined by (t_{start}, t_{end}) using the lowest value in $\{V_l\}$. The algorithm recursively applies this method to every resolved S2 pulse using the remaining threshold values in the V_l list. To illustrate the procedure, Figure 7 shows the set of test thresholds $\{V_{11.7\,\mu s}, V_{7.2\,\mu s}, V_{9.9\,\mu s}\}$ for a ZEPLIN-III event with 4 overlapping S2 pulses.

2.8. Pulse matching.

One frequent requirement of experiments using an array of photodetectors is to order in time and match signals from all DAQ channels. The simplest approach would be to have the ordering and matching function O depending both on the start and end times (t_{start}, t_{end}) of any two pulses $\{A, B\}$ occurring in different channels $\{a, b\}$,

$$O(A, B) = \begin{cases} A\,\text{precedes}\,B & ,\, t_{end}^{(A)} < t_{start}^{(B)} \\ A\,\text{follows}\,B & ,\, t_{start}^{(A)} > t_{end}^{(B)} \\ A\,\text{matches}\,B & ,\, \text{otherwise} \end{cases}. \tag{13}$$

An obvious scenario where the above method fails is when the pulse limits are extended out due to the occurrence of some sort of noise, amplifier saturation tail, etc. The possibility of having multiple interactions with partially overlapping S2s can also drive Eq. 13 to faulty results (§2.7). In order to solve these problems, an algorithm was implemented which takes into account a relative quantification W of the matching between pulses,

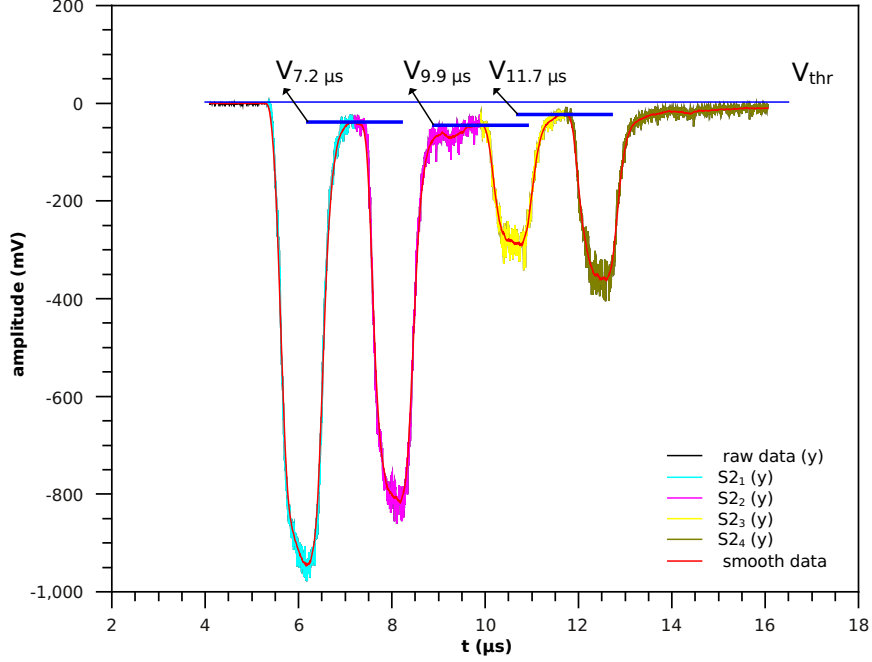$$W(A, B) = \sum_{i \in \{A \cap B\}} y_i^{(a)} * y_i^{(b)}\,, \tag{14}$$

14

Figure 7: ZEPLIN-III event with 4 overlapping S2 pulses (light blue, magenta, yellow and olive). The figure also shows the set of thresholds $\{V_{11.7\,\mu s}, V_{7.2\,\mu s}, V_{9.9\,\mu s}\}$ gathered from scanning $\overline{y}$ (red line) and used to resolve the original pulse found from an excursion of the data past V_{thr}.

where $y^{(a)}$ and $y^{(b)}$ represent, respectively, the data streams containing pulses A and B. The algorithm starts by using Eq. 13 to order and group overlapping pulses from the two channels $\{a, b\}$. For each one of these groups, eq. 14 is used to generate an overall matching quantity W defined as,

$$W = \prod_{k,m} W(A_k, B_m)\delta_{km} \, . \tag{15}$$

The values of δ_{km} are initially set to 1 for all pairs of (k, m) pulses. A fast *best-survival* algorithm was then implemented which maximizes the value of W while setting δ_{km} to 0 (Eq. 15) for all but the best degenerated matches $W(A_k, B_m)$ (Eq. 14). The application of this method in ZE3RA introduces a flavour of most probable match, even if the true pulse shape is not considered, which increased significantly the robustness of the pulse finding analysis.

The relation between pulses from different channels generated by the matching algorithm was used in several of the reduction stages in order to:

- ignore LS pulses lacking a match in the respective HS channel (§2.5);

- cluster groups of scattered pulses in individual channels matching only one

15

pulse in the correspondent HS or LS sum channels (§2.7);

- disentangle the correspondence between pulses in individual channels with overlapping pulses in the sum channels for multiple interactions at different (x,y)-positions (§2.8);

- avoid misinterpretation of amplifier saturation tails at individual channels as multiple pulses in the sum channels;

- map the correspondence between pulses or groups of pulses (clusters) in all channels for the output databases containing the reduced quantities (§2.9);

2.9. Pulse parametrization.

Once the pulse start and end times are set resulting from the operations described in sections 2.2 to 2.8, a number of parameters are extracted from the raw data buffers. The actual list of parameters can depend on the pulse context information (e.g. whether the pulse has a saturation or ringing tail, etc) and is extensible to support any subsequent analysis. For the ZEPLIN-III analysis the reduced databases include:

- pulse start time, width (between threshold crossings, at 10% and at 50% height), amplitude, area, signal mean arrival time, pulse symmetry, etc;

- pulse matching mapping over different acquisition channels, etc;

During any hypothetical step of the analysis it may be convenient to access some characteristic of an individual pulse prior to its formal and complete parametrization: consider, for example, that one requires the area of a pulse before its clustering (§2.6). For that purpose, any parameter operator α that applied to pulses A and B obeys

$$\alpha(A) + \alpha(B) = \alpha(A \cup B) \tag{16}$$

is updated every time the pulse start or end times changes. This feature of the framework increases significantly the performance of the analysis since it avoids redundant loops over the data buffers for α.

3. Interface.

The ZE3RA human interface layers the end-user interaction with the base analysis framework and can be used either in graphical or batch operation mode.

The graphical interface mode was designed to allow easy navigation through events and channels while providing an inline tuning of the relevant analysis parameters and visualization of their output in terms of pulse identification and parametrization. To help understanding of the detector output, the interface also incorporates a *2D* plot of the relative response of the array, the possibility to plot together any set of high/low sensitivity channels in any temporal scale

and mouse context information on pulse properties. The interface was coded using the cross-platform GUI toolkit FLTK (Fast Light Toolkit). However, it is worth noting that the core analysis classes are self-contained (class *ZRun*) and independent of the graphical interface. This provides the ability to promptly reuse the developed analysis framework together with any available graphical tool or within the DAQ environment for online monitoring of the detector. Significantly, it makes the porting of this framework to other detector systems quite straightforward.

The batch mode consists of a simple command line input designed mainly to perform the mass reduction of any set of data files with minimum user interaction (for example, the first science run data consisted of 13,234 binary files containing over 23 million triggered events). The analysis input parameters are fed as a configuration file which can easily be created from the graphical interface.

An important feature affecting both interface modes is the built-in blind manager. This manager allows a *super-user* to set, based on a simple set of rules, which events or datasets can be visualized and reduced at a specific step of the WIMP search analysis. Such signal-blind analyses are an essential tool of rare rare event searches.

4. Conclusions

A robust and versatile software package was developed for the analysis and reduction of the raw data from the ZEPLIN-III experiment. The framework allows the easy plug-in of new tools and building of different reduction templates targeting specific analyses. The very high electron/nuclear recoil discrimination achieved in the WIMP search carried out in the first science run (>99.98%) benchmarks these algorithms. These techniques should find application in data reduction from detectors with a large number of channels, beyond the field of rare event searches.

5. acknowledgments

References

[1] T. J. Sumner, Proc. 3rd Int. Workshop on the Identification of Dark Matter, ed. N. J. C. Spooner and V. Kudryavtsev, Singapore: World Scientific (2001) 452.

[2] H. M. Araújo, et al., The ZEPLIN-III dark matter detector: performance study using an end-to-end simulation tool, Astroparticle Physics 26 (2006) 140.

[3] D. Akimov, et al., The ZEPLIN-III dark matter detector: Instrument design, manufacture and commissioning, Astroparticle Physics 27 (1) (2007) 46 – 60.

[4] V. N. Lebedenko, et al., Results from the first science run of the ZEPLIN-III dark matter search experiment, Phys. Rev. D (Particles and Fields) 80 (5) (2009) 052010.

[5] V. N. Lebedenko, et al., Limits on the Spin-Dependent WIMP-Nucleon Cross Sections from the First Science Run of the ZEPLIN-III Experiment, Phys. Rev. Lett. 103 (15) (2009) 151302.

[6] D. Akimov, et al., Limits on inelastic dark matter from ZEPLIN-III, Physics Letters B 692 (3) (2010) 180 – 183.

[7] B. A. Dolgoshein, V. N. Lebedenko, B. U. Rodionov, new method of registration of ionizing-particle tracks in condensed matter, JETP Lett. 11 (1970) 351.

[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, Numerical recipes in C : The art of scientific computing, Cambridge University Press, 2002.

[9] M. Frigo, S. G. Johnson, The design and implementation of fftw3, in: Proceedings of the IEEE, no. 93 (2), 2005, pp. 216–231.

[10] ZEPLIN-III, Single electron emission in the zeplin-iii two-phase xenon detector, In Preparation.

[11] M. Galassi, et al., GNU Scientific Library Reference Manual (3rd Ed.), ISBN 0954612078.

[12] D. Donoho, De-noising by soft-thresholding, Information Theory, IEEE Transactions on 41 (3) (1995) 613 –627.

[13] S. Mallat, S. Zhong, Characterization of signals from multiscale edges, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (1992) 710–732.

[14] J. Kwong, P. Brusov, T. Shutt, C. Dahl, A. Bolozdynya, A. Bradley, Scintillation pulse shape discrimination in a two-phase xenon time projection chamber, Nucl. Inst. and Meth. A 612 (2) (2010) 328 – 333.

[15] P. B. Coates, The origins of afterpulses in photomultipliers, Journal of Physics D: Applied Physics 6 (10) (1973) 1159.