ABSTRACT

HUFFER, CRAIG REEVES. Results and Systematic Studies of the UCN Lifetime Experiment at NIST. (Under the direction of Paul Huffman.)

The neutron β-decay lifetime is important in understanding weak interactions in the framework of the Standard Model, and it is an input to nuclear astrophysics and Big Bang Nucleosynthesis. Current measurements of the neutron β-decay lifetime disagree, which has motivated additional experiments that are sensitive to different sets of systematic effects. An effort continues at the NIST Center for Neutron Research (NCNR) to improve the statistical and systematic limitations of an experiment to measure the neutron β-decay lifetime using magnetically trapped UCN. In the experiment, a monoenergetic 0.89 nm cold neutron is incident on a superfluid ^{4}He target within the minimum field region of an Ioffe type magnetic trap. Some of the neutrons are subsequently downscattered by single phonons in the helium to low energies (≈ 200 neV), and those in the appropriate spin state become trapped. The inverse process, upscattering of UCN, is suppressed by the low phonon density in the < 300 mK helium. When the neutron decays, the energetic electron creates EUV scintillation light, which is down-converted and transported out of the cell to PMTs operated at room temperature. With this method, the decay of the UCN population can be monitored in situ. The apparatus, analysis, data, and systematics will be discussed. After accounting for the systematic effects the measured lifetime disagrees with the current PDG mean neutron β-decay lifetime by about 9 of our standard deviations, which is a strong indication of unaccounted for systematic effects. Additional ^{3}He contamination will be shown to be the most likely candidate for the additional systematic shift, which motivated the commissioning and initial operation of a heat flush purifier for purifying additional ^{4}He. This work ends with a description of the ^{4}He purifier and its performance.

Results and Systematic Studies of the UCN Lifetime Experiment at NIST

by
Craig Reeves Huffer

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Physics

Raleigh, North Carolina

2017

APPROVED BY:

David Haase	Andrew Taylor
John Thomas	Albert Young

Paul Huffman
Chair of Advisory Committee

BIOGRAPHY

Craig Huffer was born on October 24th, 1985 in Southbend IN. He was raised a bit outside of a small town. In high school, he got a job at a horse ranch. He learned a lot there, but maybe the most important lesson was that if you were not careful, you might spend the rest of your life sweeping the same barn every day for minimum wage. He went to Indiana University, Bloomington. While a student there he got his first job in nuclear physics working for Professor Mike Snow, which proved a rewarding and challenging experience. It was rewarding to do work that mattered, which seemed in stark contrast to some of his previous jobs. He continued in experimental nuclear physics following an opportunity to work on an exciting project for Professor Chen-Yu Liu at Indiana Unversity. During his last summer before graduating undergrad, Craig was part of the NIST SURF program. Finally, he attended graduate school at NCSU in the fall of 2008. In the summer of 2009, he had the opportunity to work on the UCN Lifetime Experiment at NIST. He very quickly became enamored with the project, and it did not take long before he joined the project full time.

ACKNOWLEDGEMENTS

We acknowledge the support of the NIST, US Department of Commerce, in providing support, including the neutron facilities used in this work. This work is also supported in part by the US National Science Foundation under Grant No. PHY-0855593 and the US Department of Energy under Grant No. DE-FG02-97ER41042.

I would like to thank my advisor Prof. Paul Huffman. Throughout my graduate career, his reassuring presence has always been there when I have needed it. In our conversations about design issues, hitting obstacles in the lab, or life in general, he has always provided thoughtful contributions that have been invaluable. Whenever I wanted help or advice, he was there, and for that, I am grateful.

I would also like to thank the staff, postdocs, and students at the NCNR at NIST. In particular, Pieter Mumm, the PI of the experiment and our fearless leader. Without the help of the group at NIST, this experiment would not have been possible, and their contributions are too many to list. The collaborative environment at NIST fostered discussions that have helped me develop my passion for science and research, and for that, you have my thanks. I look back fondly on my time at NIST, and that was in large part due to my coworkers there.

I also had the opportunity, while working in the lab at NCSU to meet and interact with some of the most intelligent and hardworking people I have ever met. Many of the friendships I have developed there have been continuing sources of inspiration for me on this project and in all aspects of my life.

I would also like to thank my friends and family. Your continued support has been essential in keeping me motivated, happy, and healthy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

This work describes many of the systems developed and operated as part of the UCN Lifetime Experiment at NIST. It is the third generation of an experiment that uses three-dimensional magnetic confinement to store neutrons for measuring the neutron β-decay lifetime. The method was designed, in part, to be sensitive to an independent set of systematic effects when compared with the rest of the field of neutron lifetime measurements. Both the experimental apparatus and the analysis software, which was developed in-house, are described in detail along with a description of the data and data quality. The trap lifetime is presented along with estimates of the systematic effects. Finally, a purifier for isotopically pure ^{4}He is discussed that was built at North Carolina State University (NC State) and operated at the National Institute of Standards and Technology (NIST) to produce new samples of isotopically pure ^{4}He in an attempt to verify the purification method. The goal of the experiment is to measure the neutron β-decay lifetime, which is discussed briefly below.

A neutron that is not bound in a nucleus can β-decay into a proton, electron, and an anti-electron neutrino via the weak interaction. Measuring the time constant associated with this process, the neutron β-decay lifetime, and the angular correlations between the spins and momenta of the particles taking place in the interaction has been an active area of research for over fifty years[1]. First, we will discuss neutron β-decay in general. This is followed by a discussion of some of the other experimental methods to provide a context of how our method differs from other current methods. The chapter continues with a description of the current state of the measured values of the neutron β-decay lifetime and how the disagreement in the best measurements motivates the use of new techniques that are subject to different sets of systematic effects to help resolve the disagreement in the current experimental values. Finally, a brief description of the experimental method used in this experiment is presented.

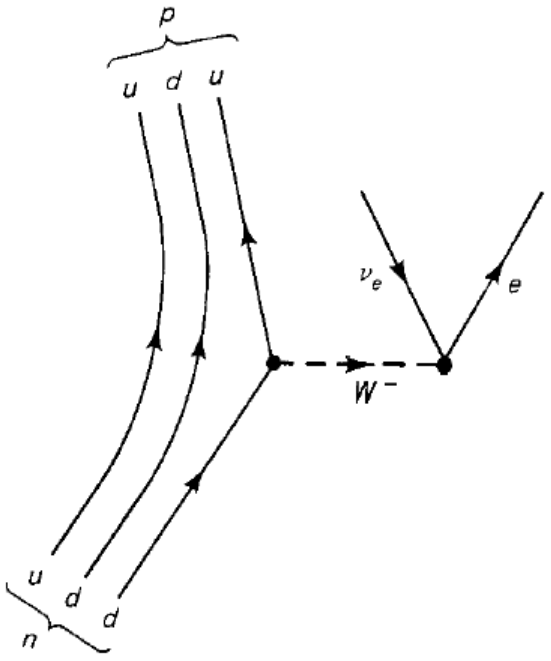
1.1 Neutron β-decay

As mentioned previously, neutrons β-decay by the process

$$n \rightarrow p + e^- + \bar{\nu_e} + 782.6 \text{ keV},$$

where 782.6 keV is the Q value for the interaction, which is the difference in mass-energy before and after the decay. Because the proton is much more massive than the other decay products, the kinematics result in the electron and neutrino splitting the majority of the kinetic energy from the reaction. A Feynman diagram of the process is included in Figure 1.1. It indicates the quark structure of both the neutron and proton and illustrates that neutron β-decay is the process where one of the down quarks in the neutron converts into an up quark that causes the transition between these two baryons. This interaction is mediated by a virtual W^- boson, which decays into an electron and an electron antineutrino. This is a weak interaction process and is described by the electroweak theory of the standard model. The neutron β-decay lifetime is ≈ 15 min.

Figure 1.1: A Feynman diagram of neutron beta decay that was borrowed from David Griffith's *Introduction to Particle Physics*[2].



1.2 Motivation

An extensive review of the experimental status of the neutron lifetime and its impact can be found in "The neutron lifetime" by F. E. Wietfeldt and G.L. Greene[1]. The Particle Data Group reviews[3] also highlight the current status of the neutron lifetime and a few of the motivations for lowering its uncertainty including Big Bang Nucleosynthesis and unitarity of the Cabibbo-Kobayashi-Maskawa (CKM) matrix through the V_{ud} matrix element, both of which are briefly discussed below.

1.2.1 Big Bang Nucleosynthesis

Big Bang Nucleosynthesis describes how the light elements were formed in the early universe. In the epoch of nuclear physics from ≈ 1 s to ≈ 200 s after the Big Bang, the neutron β-decay lifetime constrains the ratio of neutrons to protons, n/p, and consequently the abundances of the light elements. In particular, the neutron lifetime contributes the largest uncertainty in the theoretical ^{4}He abundance, which can be approximated by the relation[4]

$$Y_p = 2(n/p)/(n/p + 1),$$

where Y_p is the ^{4}He abundance, n is the number density of neutrons, and p is the number density of protons. The neutron lifetime impacts the final n/p ratio in two primary ways. It determines the temperature at which the charge-coupled weak interactions freeze out of thermal equilibrium, and neutron β-decay is the dominant process in determining the evolution of n/p between the freeze-out time and when the neutrons are absorbed into light nuclei. In the following, both of these effects are described in more detail.

When the charge-coupled weak interactions are in thermal equilibrium, n/p evolves according to the Boltzmann factor

$$n/p = e^{-\Delta_m/T},$$

where Δ_m is the mass difference of the proton and neutron, and T is the temperature of the universe. When the charge-coupled weak interaction rates decreases sufficiently, the system will leave thermal equilibrium, and the evolution of n/p will change. The reaction rates depend on the temperature and they depend on the same weak interaction vertex, which allows the interaction rate to be evaluated with the neutron β-decay lifetime[4]

$$\Gamma \propto T^5/\tau_n,$$

where Γ is the reaction rate, τ_n is the neutron β-decay lifetime, and T is the temperature of the universe. When the average duration between interactions falls below the age of the universe, a

reaction is said to freeze out because the reactions effectively cease. Because the reaction rate depends on the neutron lifetime, it can be used to determine the freeze-out time and therefore n/p at that time according to[3]

$$n/p = e^{-\Delta_m/T_F} \approx 1/6,$$

where T_F is the temperature of the universe at freeze-out.

At this point, ≈ 1 s after the Big Bang, all of the other charge-coupled weak interactions have effectively ceased, and the neutron population continues to decay away by neutron β-decay. During this period, the neutron lifetime plays a direct role in determining the evolution of n/p. It ends when the neutrons are stabilized by being absorbed into light nuclei, which occurs when the deuterium bottleneck is broken ≈ 200 s after the Big Bang with a final ratio of $n/p \approx 1/7$[3].

1.2.2 CKM Unitarity Tests

In the standard model, the CKM matrix, V, is the rotation matrix between the mass eigenstates and the weak interaction eigenstates. For the three generation quark model, this is a 3 x 3 matrix and therefore has 9 elements, where each element is related to the probability of an up-type quark (up, charm, or top) converting into a down-type quark (down, strange, bottom)[5].

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} = V \begin{pmatrix} d \\ s \\ b \end{pmatrix}$$

The standard model requires the matrix elements of the CKM matrix as inputs, and therefore they must be determined experimentally. The CKM matrix must be unitary, which reduces the number of free parameters from 9 to 4 in the case of the three generation quark model. These four free parameters can be written in terms of three generalized Cabibbo angles θ_1, θ_2, and θ_3 and a phase factor δ,

$$V = \begin{pmatrix} c_1 & -s_1c_3 & -s_1s_3 \\ s_1c_2 & c_1c_2c_3 - s_2s_3e^{i\delta} & c_1c_2s + s_2c_3e^{i\delta} \\ s_1s_2 & c_1s_2c_3 + c_2s_3e^{i\delta} & c_1s_2s_3 - c_2c_3e^{i\delta} \end{pmatrix},$$

where c_n is the cosine of θ_n and s_n is the sine of θ_n. The current experimental values of the

CKM matrix are[3]

$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} = \begin{pmatrix} 0.97417(21) & 0.2248(6) & 0.00409(39) \\ 0.220(5) & 0.995(16) & 0.0405(15) \\ 0.0082(6) & 0.0400(27) & 1.009(31) \end{pmatrix}.$$

From this, it is evident that the quarks strongly favor staying in the same generation by the near unity values of the diagonal elements.

The unitarity of the CKM matrix is a requirement that the probabilities are conserved. As an example, if the u quark decays, it must do so into either a d, s, or b quark and the sum of the probabilities of each of these interactions add to one. The probability of a given transition is related to the matrix element squared. Therefore, this particular example is given by the equation

$$V_{ud}^2 + V_{us}^2 + V_{ub}^2 = 1.$$

Unitarity tests, like this one, can be performed on any row or column of the CKM matrix. However, the one that is given above is the most stringent test to date. A deviation from unitarity would be an indication of either errors in the measurements if the standard model is to hold in its current state or physics beyond the standard model.

Currently, the most accurate method for determining V_{ud} is via $0^+ \rightarrow 0^+$ decays[3],

$$V_{ud} = 0.97417(5)ex.(9)nucl.dep.(18)RC,$$

where the uncertainties have been broken up into the contributions from the experimental uncertainty, nucleus-dependent corrections, and radiative corrections. This evaluation includes the contribution from thirteen unique transitions that are in good statistical agreement giving confidence in the corrections that have been applied. However, because the uncertainty is dominated by theoretical contributions, it will likely decrease slowly. Alternatively, V_{ud} can be determined with charged pion decays, nuclear mirror decays, or neutron β-decay. Given that the pion decays are strongly limited by the experimental precision of the decay rate, the neutron decay is an attractive system to determine V_{ud} because it does not require any nuclear structure related corrections. For a detailed review, see [6, 7].

Alternatively, V_{ud} can be evaluated using a measurement of the neutron β-decay lifetime and at least one neutron β-decay asymmetry parameter. The current best limit with this method uses A the β-asymmetry parameter[3],

$$V_{ud} = 0.9758(6)\tau_n(15)g_A(2)RC,$$

where the dominant uncertainties come from the experimental determinations of τ_n and g_A. At the moment, this method is not competitive, but the reduced theoretical uncertainties will make this method very attractive as the uncertainty of the neutron β-decay lifetime and β-asymmetry coefficients is reduced.

1.3 Lifetime Experiments

A variety of experimental methods have been used to measure the neutron β-decay lifetime[1]. The current generation of experiments can be classified into two general types of measurements that use different methods and consequently are sensitive to different systematic effects. The two classifications are beam-type and bottle-type experiments. Many of these experiments employ ingenious methods for estimating and reducing these systematic effects, which is what has allowed the uncertainty in the neutron β-decay lifetime to be pushed as low as it has. However, there is substantial disagreement in the results of the current generation of experiments. The reported uncertainties make it unlikely that this spread of values would occur due to random fluctuations, see Figure 1.2. Both of these types of measurements will be discussed in the following.

1.3.1 Beam-Type Experiments

Beam-type experiments pass a cold neutron beam of known fluence through a detection volume and detect the number of neutrons that decay in that detection volume. The neutron β-decay lifetime can be extracted if both the total amount of time that the neutrons spent in the decay region and the total number of neutrons that decayed in the decay region are determined.

The beam lifetime experiment at NIST will be used as a representative example of the beam-type experiments. In this experiment, a cold neutron beam of known flux, $\Phi(v)$, is passed through a well-defined volume of length, L, and area, A, where the flux depends on the neutron's velocity, v. To the extent that the neutron flux is constant in time, the expected number of neutrons in the decay volume will also be constant. To calculate the number of neutrons in the decay volume, one must account for the time that the neutrons spend inside the decay volume. The number of neutrons in the trap is

$$N_n = \int_A \int_v \Phi(v) \frac{L}{v} \partial v \partial A,$$

where N_n is the number of neutrons, v is the velocity of the neutron, L is the length of the trap, and A is the area of the trap.

Neutron β-decay is an exponential process in that the population decays away according to

the relation

$$N_n(t) = N_0 e^{-t/\tau_n},$$

where N_0 determines the initial population at time $t = 0$ and τ_n is the neutron β-decay lifetime. The differential equation that describes the β-decay rate is obtained by taking the derivative.

$$\dot{N}_n(t) = -\frac{1}{\tau_n} N_n(t)$$

In the beam lifetime experiment at NIST, the neutron decays are measured by trapping the protons from neutron β-decay in an electrostatic trap. They are stored in the trap volume until a gate potential is lowered allowing the protons to empty from the trap following field lines to a surface barrier detector. The detection efficiency of the protons, ϵ_p, relates the rate of neutron β-decays to the proton detection rate, which can be rewritten as a function of the number of neutrons in the trap,

$$\dot{N}_n = \frac{\dot{N}_p}{\epsilon_p} \Rightarrow \dot{N}_p = \frac{\epsilon_p}{\tau_n} N_n.$$

Combining these equations, the proton detection rate can be written in terms of the neutron lifetime and a set of parameters that characterize the neutron beam. The length of the trap can be pulled outside of the integral.

$$\dot{N}_p = \frac{\epsilon_p}{\tau_n} N_n = \frac{\epsilon_p}{\tau_n} L \int_A \int_v \Phi(v) \frac{1}{v} \partial v \partial A$$

A neutron fluence monitor was used to measure this integral. It measures the neutron fluence by shining the neutron beam on a thin lithium foil and detecting the decay products from the neutron absorption with surface barrier detectors with well characterized solid angles. The rate in the neutron monitor depends on its detection efficiency and this same integral.

$$\int_A \int_v \Phi(v) \frac{1}{v} \partial v \partial A = \frac{\dot{N}_\alpha}{\epsilon_0 v_0},$$

where ϵ_0 is the detection efficiency of the neutron monitor scaled to the thermal neutron velocity, v_0, and $\dot{N}_\alpha$ is the detection rate in the neutron monitor.

Taking the ratio of $\dot{N}_p$ and $\dot{N}_\alpha$ and solving for the neutron lifetime, it can be written in terms of the length of the trap, two measured rates, a detection efficiency for the protons, and the detection efficiency of the neutron monitor. The detection efficiency of the protons is simulated. The detection efficiency of the neutron fluence monitor must be determined.

$$\tau_n = \frac{L \dot{N}_\alpha \epsilon_p}{\dot{N}_p \epsilon_0 v_0}$$

A primary difficulty in this type of experiment comes from attempting to perform an absolute fluence measurement of the neutron beam[8]. It requires a detailed characterization of the solid angle of the surface barrier detectors, the shape of the lithium deposit, and an evaluation of the cross section of the ^{6}Li(n,t)^{4}He reaction.

Additional work was done to improve the calibration of the neutron monitor after the operation of the beam lifetime experiment. As a result of this work, the systematic effect due to measuring the neutron fluence was reduced to 0.5 s[9], a factor of 5 reduction in the uncertainty, and the dependence of their lifetime result on the ENDF evaluation of the ^{6}Li cross section was removed. Other major challenges in this experiment come from characterizing the trap volume, 0.8 s, and estimating the proton detection efficiency, which is dominated by proton backscattering from the surface barrier detectors, 0.4 s. This experiment is currently limited by statistics after the improvements in the calibration of neutron monitor. This is part of the motivation for the ongoing effort to upgrade the beam lifetime experiment and to take more data for a 1 s measurement of the neutron β-decay lifetime.

1.3.2 Bottle-Type Experiments

The bottle-type experiments take advantage of the properties of very low-energy neutrons, termed ultracold neutrons (UCN)s, which have sufficiently little kinetic energy ($\lesssim$ 300 neV) that they are unable to overcome the effective potential barrier of common trap materials. This makes it likely that an UCN will be reflected during a wall interaction, which allows UCNs to be stored in a material bottle for an extended period.

Alternatively, the UCNs can be manipulated with either magnetic fields through their magnetic dipole moment, 60 neV/T, or by gravitational fields, 100 neV/m. Clever trap designs can be used that incorporate a combination of these three potentials to improve the experimental accuracy. As an example of a commonly used trap, a material bottle with an open top is a combination of the nuclear potential and the gravitational potential.

There are a variety of recent bottle-type experiments. In most of these experiments, the basic concept is the same. A storage volume is developed that limits the wall losses substantially. Neutrons are loaded into the trap through a fill port from an external source. The fill port is closed, and the population begins to decay away according to

$$N(t) = N_0 e^{-t/\tau},$$

where N_0 is the number of neutrons in the trap when the fill port is closed, and τ is the trap lifetime. As the situation approaches the case where neutron β-decay is the only non-negligible loss mechanism, the lifetime of neutrons in the trap, τ, approaches the neutron β-decay lifetime, τ_n. After some predetermined amount of time, an exit port is opened, and the UCNs are emptied

from the trap to be counted using external neutron detectors. The number of neutrons that are counted is a lower limit of the number of neutrons that survived the storage period and constitutes a single measurement.

By completing a pair of measurements with different storage times but the same initial number of neutrons, the trap lifetime can be extracted. This is done by taking the ratio of the number of neutrons at the two times,

$$\frac{N_1}{N_2} = e^{-(t_1 - t_2)/\tau_n},$$

which can be rewritten to give the neutron lifetime,

$$\tau_n = (t_1 - t_2) \,/\, \ln\left(\frac{N_2}{N_1}\right).$$

Here t_1 and t_2 are the times at which the exit was opened for the first and second measurements; $N_1 = N(t = t_1)$ and $N_2 = N(t = t_2)$ are the number of neutrons detected in the two measurements.

The dominant systematic effects in bottle-type experiments are typically due to wall losses. Some of the bottle experiments vary the surface area to volume ratio in their trap to measure the effect of wall interactions. Additionally, ^{3}He detectors can be used to measure the inelastic scattering of UCNs from the walls. Our experiment is an example of a magnetic bottle-type experiment, which use magnetic fields to confine the UCNs to mitigate this systematic effect. Additionally, if neutrons can get into stable orbits inside of the trap they will not empty effectively, which can cause systematic errors in the measured trap lifetime. Recent traps have introduced asymmetry in the storage volume to help mix the phase space and prevent stable orbits of this kind[10]. Our experiment creates and stores the UCNs directly in the detection volume, which removes these types of systematic effects. It also detects the neutron β-decay's instead of the number of neutrons that survive until a given time, which allows the decay rate to be measured directly and gives a strong handle on time-dependent systematic effects that is unique to our method.

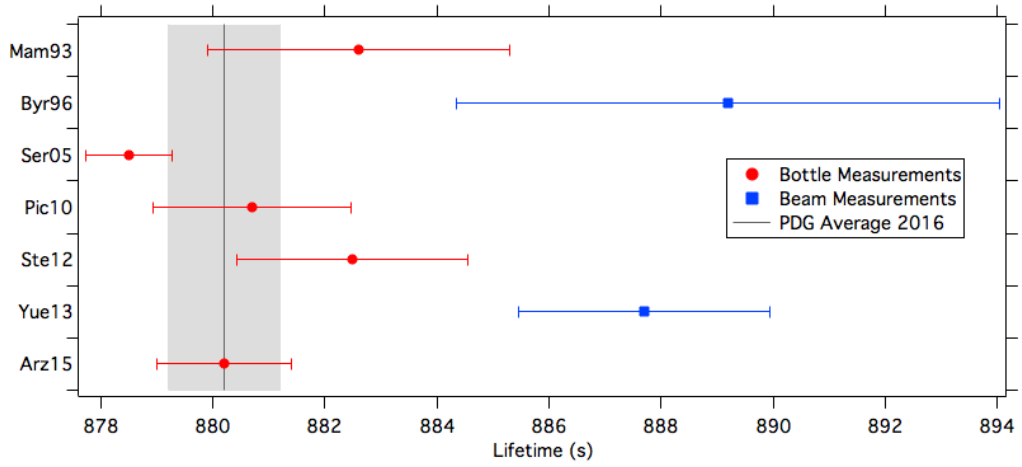
1.4 PDG Summary of the Published Neutron Lifetime Measurements

Before describing our experimental method in more detail, let's first set the stage by looking at the world average of the neutron lifetime measurement. This will help explain and quantify the disagreements in the current measurements that have already been mentioned. The Particle Data Group (PDG) summary includes a description of which experiments are included in the

current PDG average and why. In the case of the neutron β-decay lifetime, their data for the current summary is included in Figure 1.2.

The spread in the measured values is statistically improbable when the size of the error bars are taken into account. It is still important to have a current best estimate of the neutron lifetime to help constrain the theory. Therefore, the PDG averages the best values and inflates the error bars, by a factor of 1.9, to account for the disagreement in the measured values. This results in their published value, $\tau_n = 880.2 \pm 1.0$ s[3].

Figure 1.2: Summary of the current experimental status of the neutron lifetime as determined by the Particle Data Group (PDG). The individual measurements are included and designated by the first three letters of the lead authors last name and the publication year[3].



This disagreement in the measured neutron β-decay lifetime seems to be a clear indication that at least some of the current experiments are subject to systematic errors that are being improperly corrected or underestimated. This could be indicative of an unknown systematic effect. The knowledge that some of the experimental methods are subject to systematic effects that are not being properly accounted for is a strong motivation for running new experiments with improved capabilities for testing their systematic effects, that are designed to be less sensitive to particular systematic effects, or that use methods that are subject to completely different systematic effects. Over the last decade, multiple experiments have been constructed to address this problem, the UCN Lifetime Experiment at NIST is one such experiment.

1.5 The UCN Lifetime Experiment at NIST

In this experiment, a mono-energetic 0.89 nm cold neutron beam enters a $\leq$ 350 mK bath of isotopically pure ^{4}He. A small fraction of the cold neutrons interact by the superthermal process and are downscattered into UCNs. One of the two spin states of the UCNs is then attracted toward the center of the cell by the combined magnetic fields of a pair of solenoids providing longitudinal confinement and a quadrupole providing radial confinement. The magnets provide a magnetic potential well, and neutrons with energies below 139 neV do not have sufficient kinetic energy to overcome the magnetic potential and escape the trap. These UCNs are trapped. The stage of the experiment during which the neutron beam is on and a population of UCNs is produced in the trap is termed the filling stage. Approximately 40 min after the start of the filling stage, it is ended by closing the neutron beam. After this, no new UCNs are created. Therefore, the neutron population decays away by β-decay and any additional loss mechanisms.

The filling stage is followed by a brief delay during which the beam is closed, and the experiment is prepared for data collection. The duration of the delay ranges from roughly 142 s to 373 s depending on the type of data being taken. During this time, the detection system is powered up and, in the majority of the data types, the strength of the magnetic field is ramped. In the production data, the fields are ramped to preferentially purge marginally trapped neutrons (MTN)s to limit systematic effects due to this population of neutrons. After the delay, the observation stage begins. During the observation phase, a pair of photomultiplier tube (PMT)s, which are connected to the cell by a series of windows and acrylic light guides, detect light resulting from the scintillation of the energetic electron and proton created by β-decay inside the liquid helium (LHe) bath, which acts as a scintillator. These flashes of light indicate the decay of an UCN, and by recording these decays in real time, the neutron population can be observed decaying away.

Because the detection method detects light, other sources of light can introduce backgrounds. These background light sources include non-neutron β-decay induced charged particle scintillation and Compton scattering in the acrylic. A large fraction of background events is tagged and removed from the data set using pulse shape analysis. Additional background data is taken to determine the time-dependent background rate so that these backgrounds can be subtracted from the trapping data. In the background data, the cold neutron beam is emitted to the apparatus, but the quadrupole magnet is powered down. Therefore, UCNs are produced inside the cell, but without radial confinement, they are quickly lost from the trap. This preserves background sources that originate from the beam while eliminating the trapped neutrons. The response of the detection system depends on the magnetic field state. Therefore, the quadrupole magnet is ramped up before data collection begins. These background data sets are subtracted from the production data, which removes both time-dependent and constant backgrounds in a

model independent way.

Ultimately, the data is transformed into a time-dependent decay rate of neutron β-decay like events, which is fit to an exponential. The lifetime of the exponential fit is the extracted trap lifetime for that data set and is expected to be a lower limit for the neutron β-decay lifetime. A typical pair of trapping and background data files* takes a little less than 3 h to acquire, of this, the data acquisition phase is ≈ 1.3 h.

*Throughout this work, I will frequently use the term data file in place of data run when describing the process of taking data.

Chapter 2

Lifetime Apparatus

The apparatus for the UCN Lifetime Experiment at NIST has been previously discussed elsewhere. Its design, construction, and commissioning were almost entirely completed before I joined the project. Detailed information on a few of the systems has been published previously including the beam facility[8], the monochromator[11], the previous generation of the lifetime apparatus[12], and multiple status updates and other papers on the current experiment[13, 14, 15]. In addition, three theses contain design details about this version of the experiment and go into further depth than is needed here[16, 17, 18]. This chapter includes any information needed for calculating the trap lifetime of this apparatus, calculating systematic effects, or for a general understanding of the operating procedures.

The following sections follow the path of neutrons created at the NIST Center for Neutron Research (NCNR) that make it to the UCN Lifetime Experiment at NIST. This discussion will begin with the reactor, moderators, and neutron guides that create, down-convert to the desired energy range, and transport the cold neutrons to the experimental location. A discussion of the neutron entrance follows, which describes what the cold neutron beam "sees". The next section describes the experimental cell, where a fraction of the cold neutrons are energy down-converted into ultracold neutrons (UCNs), which is followed by a description of the magnetic trap that confines ultracold neutron (UCN)s in the cell for observation. A description of the light production from neutron β-decay is followed by a discussion of light transport and detection. This finishes the sequence of events that occur when a neutron β-decay is detected.

However, additional systems are required to operate the UCN Lifetime Experiment at NIST. The chapter continues with descriptions of the cryogenics infrastructure, gas handling system, slow control computer, and data acquisition system (DAQ). Finally, the auxiliary systems used to reduce systematic effects and backgrounds are described in detail. This includes the gain monitor, shielding, and the cosmic muon active veto system. The chapter ends with an analysis of the performance of various components of the system, suggested upgrades to both the

apparatus and methods, and additional data types that should be recorded.

2.1 Cold Neutron Production and Transport

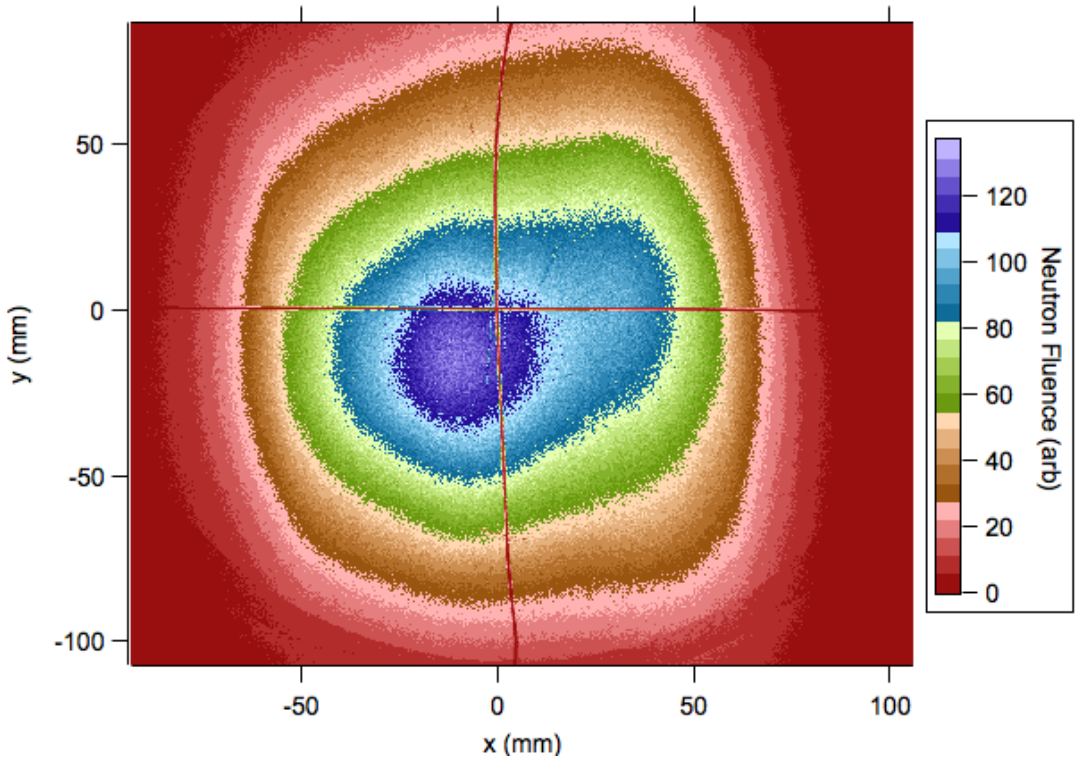
The NCNR went through a major facilities upgrade beginning April 2011 that included the addition of a new cold guide hall and a new cold source. These upgrades occurred after the UCN Lifetime Experiment at NIST finished taking data, and therefore the NCNR facility is described as it was before the upgrade to be representative of the state during data collection.

The experiment takes place at the NCNR building on the NIST campus in Gaithersburg, MD. Neutrons produced by the 20 MW research reactor are down-converted to thermal neutrons by a heavy water, D_2O, moderator. Nine thermal neutron beams emerge from the heavy water moderator. In addition, a liquid hydrogen cold source is employed to supply seven cold neutron beams. Neutrons interact frequently inside the liquid hydrogen cold source where they come to a quasi-thermal equilibrium and emerge with a Maxwellian distribution with a characteristic temperature that is slightly above the cold source temperature. The cold guide halls at the NCNR take advantage of evacuated, ^{58}Ni coated supermirror neutron guides to transport the neutrons from the cold source to the experimental stations[8]. The UCN Lifetime Experiment at NIST takes place on the NG6-U beamline. NG6-U utilizes a stage 2 potassium-intercalated graphite monochromator to select and reflect neutrons at 0.89 nm[11]. In addition to the 0.89 nm neutrons, the monochromator also reflects integer multiples of the selected frequency, with the largest contribution from neutrons of wavelength $\lambda/2$. This component is removed from the beam by reflection from a pyrolytic graphite crystal[8]. Finally, a beam shutter allows the NG6-U beamline to be opened and closed at the experimental station.

To evaluate the beam, neutron fluence measurements were performed, and beam images were taken. The neutron fluence was measured 1 m from the monochromator to be 4.7×10^6 cm^{-2}s^{-1}[8]. Beam images were taken, in August 2006, of the cold neutron beam before the construction of the apparatus was completed. An example of the beam image, before processing to remove the crosshair and to apply collimation, can be seen in Figure 2.1. The beam imaging technique takes advantage of photostimulated luminescence (PSL). The beam images are taken using FujiFilm BAS imaging plates, which contain a photo-stimulable phosphor that stores energy from neutron irradiation in color centers. This energy can then be released by photo-stimulation after the activated beam plate has been transported off site for analysis. The amount of energy released from the plate is related to the neutron fluence, which allows the spatial distribution of the neutron fluence to be estimated. The beam images are aligned using crosshairs constructed out of cadmium wire, which take advantage of the large neutron capture cross section of ^{113}Cd to absorb any neutrons that strike the crosshairs. This leaves a deficit of events in the beam image with the shape of the crosshairs. The cadmium crosshairs are then optically

illuminated with laser light from a theodolite to allow precise alignment of the crosshairs with respect to the apparatus, which allows the beam image to be aligned with respect to the apparatus. The beam image shows a smoothly varying spatial distribution, which is slightly off centered and has a cross section that roughly approximates a Gaussian distribution. There is additional structure at the perimeter of the beam images, which are edge effects associated with scattering from the crosshair support structure and the beam tube. Using the beam images, the neutron fluence is estimated to fall off by approximately 80% from the maximum of the beam to edge of the collimator. After additional processing, the beam images are used in the Monte Carlo simulation to estimate the initial spatial distribution of UCN inside the apparatus, see Section 4.1, which starts on page 113.

Figure 2.1: A fluence beam image of the NG6-U beamline that was taken upstream of the apparatus. The absorption of the cadmium crosshairs is evident along with edge effects associated with scattering from the crosshair support structure on the perimeter of the beam image. For reference, the collimator has an inner diameter of 7.12 cm.



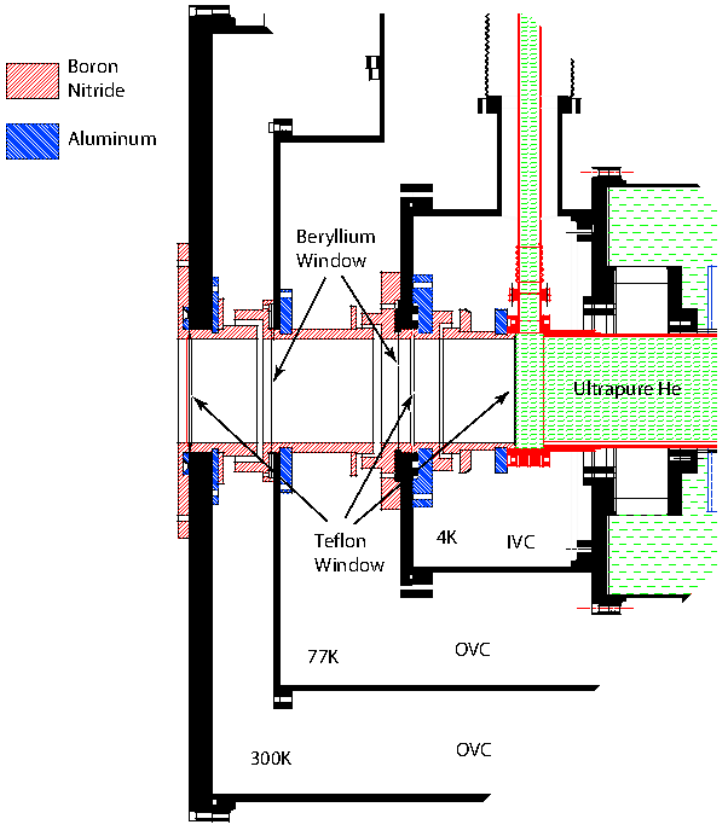
2.2 Neutron Entrance

The neutron entrance is the portion of the apparatus directly downstream the beam shutter. It is designed to transport neutrons into the experimental cell while minimizing backgrounds related to the cold neutron beam and neutron-induced activation. Shielding walls and boron doped materials are the primary methods for reducing background events associated with the cold neutron beam. They are discussed in Section 2.9, which starts on page 48. To further reduce backgrounds, a secondary shutter is implemented behind the main cold neutron shutter. It is designed to place either an additional lead wall or a boron nitride (BN) guide tube into the beam path depending on the stage of the data collection. During the data acquisition phase, the lead wall is placed in the beam to further attenuate any prompt gamma rays from the cold neutron beam. During the neutron filling stage, the BN beam tube is placed in the beam path. It allows the neutron beam into the cell while absorbing neutrons that scatter out of the beam path. The beam tube absorbs cold neutrons that scatter in the air, which will broaden the angular distribution of the cold neutron beam and activate parts of the apparatus outside the beam path. To further reduce neutron activation, all metal surfaces at the beam entrance are covered with neutron absorbing materials. The majority of the neutron shielding consists of custom machined BN covers that were augmented with additional boroflex shielding. BN spontaneously emits uncorrelated photons after neutron activation in a process referred to as neutron-induced luminescence. Our collaboration measured this process in an environment very similar to what exists inside our apparatus[19]. Because of this process, any BN shielding on the beam entrance that is within sight of the experimental cell is coated with graphite to act as an optical shield for these uncorrelated photons.

To make it into the cell, the cold neutrons travel through a series of vacuum gaps, Teflon vacuum windows, beryllium window radiation shields, and a collimator. A schematic of these components can be seen in Figure 2.2. Each Teflon window acts both as the vacuum window and as the gasket. They consist of a sheet of perfluoroalkoxy (PFA) Teflon that is 508 μm thick and have proven to be superfluid leak tight[20]. The beryllium windows act as radiation shields to limit the heat loads[17, 21]. The collimator has an internal diameter of 7.16 cm and is made of graphite coated BN to absorb any incident cold neutrons. It uses a knife edge design to limit the number of neutrons that are scattered into the experimental cell from the collimator surface. The collimator both reduces possible neutron activation of the cell materials and limits the total energy of UCN produced in the cell by constraining their initial positions to the center of the cell where the magnetic potential energy is lowest.

Neutrons first travel through a Teflon window, which is used to separate atmosphere from the outer vacuum chamber (OVC), the vacuum space that houses the 77 K and 4 K apparatus. The neutrons must then pass through two beryllium windows, which are attached to the 77 K and

Figure 2.2: Schematic of the neutron entrance showing the interlocking nature of the BN shielding and the locations of the vacuum (Teflon) and radiation shield (beryllium) windows. This image was borrowed from Chris O'Shaughnessy's thesis[17].



4 K radiation shields inside the OVC. A second Teflon window defines the interface between the OVC and the inner vacuum chamber (IVC), a separate vacuum space that houses the experimental cell at 350 mK. The beam passes through the collimator, which is attached to the experimental cell and is located just upstream of the final Teflon window, which defines the front surface of the experimental cell.

2.3 Experimental Cell

The experimental cell is the region of the apparatus where the UCN are produced and stored until their decay. The UCN are produced by the superthermal process where cold neutrons from the beam are down scattered by the creation of phonons in superfluid LHe. This process is allowed by energy and momentum conservation because of two intersections between the

energy-momentum curve of the free neutron, $KE = p^2/(2m)$, and the LHe dispersion curve at 0 meV and 1.03 meV. The higher energy crossing corresponds to a neutron wavelength of 0.89 nm. This allows 0.89 nm cold neutrons to interact with the LHe and, by creating a single phonon, deposit the vast majority of their energy into the helium bath. The resulting neutron has an energy that has been reduced by about four orders of magnitude and is called an ultracold neutron (UCN).

Once the UCN are produced, they reside in the helium in the cell. ^{3}He has an extremely large capture cross section; therefore, even a small amount of ^{3}He will quickly absorb any UCN that were produced in the cell. As a result, the LHe in the experimental cell must be isotopically pure ^{4}He. The purity of the isotopically pure ^{4}He, the size of the systematic error, and the purification method are discussed in subsequent chapters.

The experimental cell is a stainless steel tube. Its dimensions are included in Table 2.1. The dimensions of the structural tube for the cell were strongly constrained by the quadrupole magnet because the cell had to be slid inside quadrupole's inner bore. To allow a larger cell, only one end flange was soldered in place before assembly. The second end flange was soldered onto the tube after it was inserted through the quadrupole magnet.

To prevent neutron irradiation of the stainless steel tube, a set of interlocking BN sleeves were constructed to slide inside the structural tube in order to provide a hermetic neutron absorbing shield. These BN sleeves were made out of the hexagonal close-packed crystalline structure, h-BN. The interlocking nature of the sleeves was provided using overlapping steps of approximate length 1.27 cm on a sleeve of total length 15.24 cm. BN spontaneously emits photons after being exposed to neutrons in a process called neutron-induced luminescence. Because this experiment relies upon light detection as the method for detecting neutron β-decays, neutron-induced luminescence can result in both a substantial background and detector deadtime[16]. To reduce the number of neutron-induced luminescence photons that reach the detection system, optically opaque graphite sleeves line the inside of the BN. Flexible, approximately half-cylindrical Gore-Tex$^{\text{TM}}$, expanded Polytetrafluoroethylene (ePTFE), inserts line the inside of the graphite sleeves. The inner surface of the ePTFE inserts are coated with tetraphenyl butadiene (TPB). These last two materials are used for their light transport and wavelength shifting properties and are discussed in detail in Section 2.5, which starts on page 26.

All of the materials inside the BN sleeves are made from materials that can be produced with very limited and well-understood contamination. This was essential to prevent neutron activation of radioisotopes that could introduce substantial, time-dependent backgrounds in the data. Studies of the activation and neutron-induced luminescence in the cell materials can be found in reference[22].

This design limits both neutron-induced activation and neutron-induced luminescence from

Table 2.1: A list of interesting length scales in the experiment. * Indicates data that was borrowed from Chris O'Shaughnessy's thesis[17].

Property	Dimension (m)
Collimator Radius	7.16×10^{-2}
Cell Length	1.829
Cell ID	1.24×10^{-1}
BN Thickness	3×10^{-3}
BN OD	1.233×10^{-1}
Graphite Thickness	1×10^{-3}
Graphite OD	1.17×10^{-1}
ePTFE Thickness	2×10^{-3}
TPB Thickness	1.5×10^{-6}
Cell Acrylic Length	$6.9815\times10^{-1*}$
Cell Helium Length	$7.5\times10^{-1*}$
TPB Characteristic Roughness	$2-15\times10^{-6}$
Gore-Tex Fiber Width	0.2×10^{-6}
Gore-Tex Fiber Length	$1-5\times10^{-6}$
Gore-Tex Dimple Width	$5-7\times10^{-6}$
Gore-Tex Dimple Length	20×10^{-6}
Gore-Tex Node Width	$2-4\times10^{-6}$
λ of 50 neV UCN	1.3×10^{-7}
λ of 300 neV UCN	0.52×10^{-7}
λ of 1.03×10^{6} neV cold neutrons	8.9×10^{-10}

the cold neutron beam while providing a ≤ 350 mK isotopically pure ^{4}He bath in which the cold neutrons are down-converted into UCN and can be stored in an environment that minimizes loss mechanisms. Once the UCN are created, the next step is to confine the UCN for a substantial fraction of the neutron β-decay lifetime. In the UCN Lifetime Experiment at NIST, this is done with a magnetic trap, which is discussed in the following section.

2.4 Magnetic Trap

Neutrons interact with the magnetic field with their magnetic moment according to the potential, $V_B = -\vec{\mu}\cdot\vec{B}$. This allows the low-field-seeking state, with their magnetic moments aligned with the magnetic field, to be trapped in a conservative potential, while the high-field-seeking states are expelled from the trap. Using the dipole moment of the neutron, the conversion between the magnetic field and potential energy is 60 neV/T. In order for the low-field-seeking states to be trapped, a few requirements must be satisfied.

The direction of the magnetic field must vary slowly enough, from the reference frame of the

UCN, that the magnetic moment of the neutron can reorient itself to follow the magnetic field lines. This is called the adiabatic condition. The adiabatic condition fails as the angular velocity of the rotating magnetic fields that the neutron experiences approach the Larmor precession frequency of the neutron. The adiabatic condition is tested in Section 5.11 and found to be easily satisfied.

Any neutron with sufficient energy will be able to strike the cell wall. These neutrons are termed above-threshold neutrons. Marginally trapped neutrons (MTN)s are a subpopulation of the above-threshold neutrons, which are discussed in more detail in chapters on the Monte Carlo and systematics effects. Above-threshold neutrons experience an additional loss mechanism due to wall interactions and therefore can result in a systematic error in estimating the neutron β-decay lifetime.

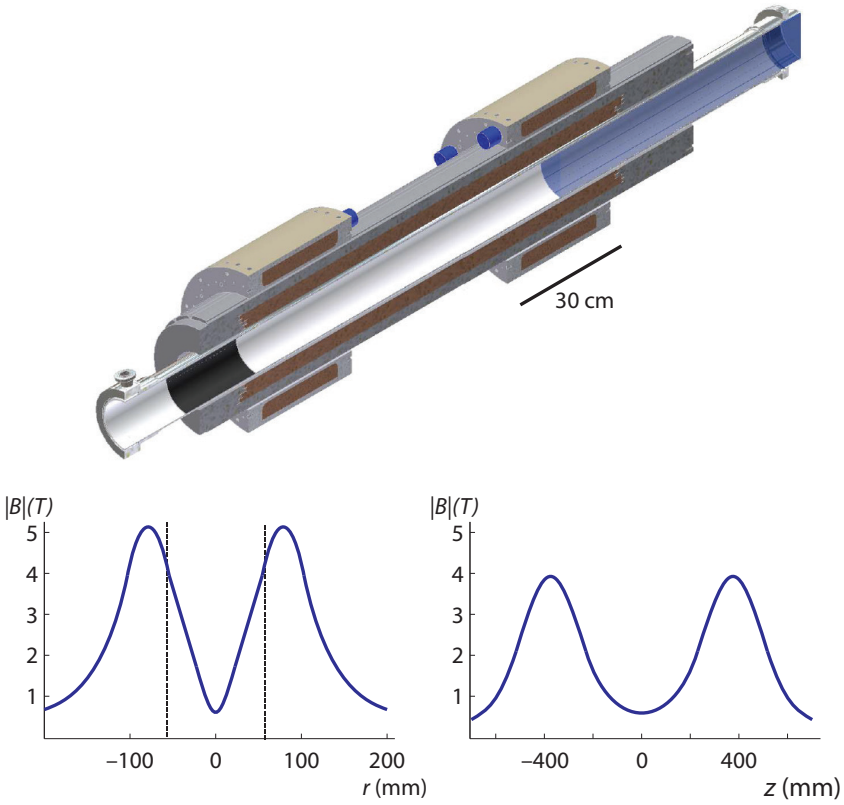
Finally, if the neutron experiences any low-field regions, it may randomly flip its spin orientation: a Majorana spin flip. In a quasi-classical picture, the UCN forgets its spin orientation therefore increasing the contribution to the wave function from the high-field-seeking spin state. When the UCN moves back into a strong magnetic field region, it is more likely to be in the high-field-seeking state and therefore ejected from the trap.

The magnetic trap was designed specifically to maximize the cell size and trap depth while both satisfying the adiabatic condition and maintaining a field minimum substantially above zero inside the cell.

The magnetic trap uses an Ioffe configuration, which consists of a quadrupole magnet that provides radial confinement and a pair of solenoids with an identical current direction that provide the longitudinal confinement. This trapping arrangement has no low-field regions that would allow Majorana spin flips. The field minimum inside the cell, which is located on the cell wall, is estimated to be 0.57 T. In addition, the small field gradients easily satisfy the adiabatic spin transport condition.

Both magnets are superconducting magnets, which provide the 3.4 T field while minimizing the ohmic heat loads. The trap potential along with a sketch of the magnets can be seen in Figure 2.3. The use of high currents requires substantial infrastructure for supplying current and controlling the temperature of the leads. In addition, systems are required to protect the magnetic trap in the event of a quench, the process of the magnetic trap rapidly leaving the superconducting state. This occurs when a small portion of the coil leaves the superconducting state because of mechanical vibrations, the temperature passing above the critical temperature, or due to the electromagnetic fields in the coils becoming too large. The portion of the coil that loses superconductivity begins to produce ohmic heat. Due to the minuscule specific heat of materials at cryogenic temperatures, the ohmic heating quickly warms the surrounding superconductor above its superconducting transition, which results in an extremely rapid warmup of the magnet as a substantial fraction of the magnet's energy is quickly deposited into the coils as

Figure 2.3: (Top) Cutaway diagram of the experimental cell and magnetic trap along with projections of the magnetic potential onto the (Bottom Left) radial and (Bottom Right) longitudinal axes.



heat. The localized heating from the tremendous currents can damage or destroy the magnet. Therefore, a quench detection circuit was implemented to detect a quench, immediately redirect as much energy as possible from the quadrupole magnets into a dump resistor, and to inform the slow control computer and thereby the operator. The magnets are submerged in a LHe volume to act as a thermal bath. The heat load deposited into the magnets during a quench, in addition to putting the magnetic trap in danger, also deposits a substantial amount of energy into the LHe bath causing violent boiling and expansion. If the cryogens were trapped, this would result in an explosion hazard. To prevent this, safety relief values are included in the apparatus to allow the trapped gas to escape in the event of a quench. By dumping some of the current into a dump resistor, the quench detection circuit also limits the amount of energy deposited into the LHe bath thereby reducing the amount of cryogens lost during a quench.

Quenches that occur during typical operation are thought to occur primarily due to reorientations of the magnetic coils. The reorientation is a result of the magnet attempting to

reduce its internal energy. Small perturbations in the coil locations, induced by the forces on the magnet, can result in improved magnet stability without being sufficiently powerful to destroy the superconducting state. The process of slowly improving the stability of the magnet through running it at consistently higher currents is referred to as "training" the magnet. By monitoring the voltage taps on the superconducting leads with an oscilloscope, it is possible to observe voltage spikes that are thought to be associated with perturbations of the coil orientations. As mentioned, short-term fluctuations in the voltage are indicative of reorientations of the magnetic coils. In contrast, a constant voltage is representative of stable operation. The majority of the reorientations take place while the magnet is being ramped. There is also an increase in the amount of activity when the magnet is ramped to a magnetic field strength that is higher than it has been run at previously. The increased instability decays away on the hour to day timescale. The frequency of the voltage jumps appears to be strongly correlated with how much stress the magnets are under and the likelihood of a quench. By "training" the magnet, the operational current can be slowly increased to much higher values than what would have been obtainable immediately after beginning to operate the magnet. A quench can also result in a reorientation of the coils and frequently results in a relatively long-term reduction of the stability of the magnet, which is attributed to overcorrecting the coil location. In addition, a quench can permanently damage the magnets. This motivates a very slow and steady approach to "training" the magnets to prevent any quenches as the trap strength is slowly increased.

A higher magnetic field strength is desired to increase the number of neutrons that can be trapped, which is expected to go as $N_n \propto B_{eff}^{3/2}$. Here B_{eff} is the effective trap depth and is defined as the difference in the magnetic potential between the minimum and maximum of the trap potential. In this experiment, the maximum magnetic field in the trap is 4 T at design current, which defines the minimum of the trap potential. The maximum potential of the trap is more challenging to estimate because the experiment does not have an effective method of cleaning neutrons above a well-defined energy. Therefore, a conservative estimate is to use the minimum field strength on the surface of the cell, 0.57 T, to calculate the maximum potential of the trap. Therefore, the effective field strength at design current is 3.4 T, which corresponds to 2.38 T at 70% of design current, the strength at which the trap was operated. Using this effective field strength, 2.38 T, the trap depth is estimated to be V_B =139 neV[23]. The minimum magnetic field inside the trapping volume was determined using magnetic field maps that were developed for use in the Monte Carlo simulation[17]. Additional details on calculating the magnetic potential can be found in the corresponding section of the chapter on the Monte Carlo, see Section 4.3.1. The size of the trap is estimated as the volume of the trap with a magnetic potential that is less than the minimum potential on the cell wall, which is estimated with the field maps to be 5.27 l. The size of the cell is substantially larger than this, however few UCNs exist in other portions of the cell.

Another factor that can substantially increase the number of neutrons detected is the ramping speed of the magnets. The data is taken in pairs of trapping and non-trapping data. The non-trapping data is used to assess the backgrounds so that they can be subtracted from the trapping data. The non-trapping data is taken by leaving the quadrupole magnet off during the neutron fill. This results in a substantial reduction in the radial confinement so that any UCN produced in the trap interacts with the wall without experiencing the slowing effect of the quadrupole's magnetic potential. As a result, the UCN are very quickly purged from the trap. After the end of the fill stage of the non-trapping file, the quadrupole magnet is quickly ramped up to the operational trap depth to mirror the state of the trapping file as closely as possible without trapping any of the UCN in the trap. The amount of time it takes the quadrupole magnet to ramp up to full current determines how soon data can be taken after the end of the fill stage. All loss mechanisms, including β-decay, diminish the neutron population during the field ramping and therefore decrease the effective number of trapped neutrons.

A quick ramping speed is desired to limit the number of neutrons that decay between the end of the neutron fill and the start of data acquisition. The solenoid magnet has a much higher inductance than the quadrupole which limits its ramping rate. If the trap was ramped at the maximum rate that the solenoid can be ramped, it would result in substantial neutron losses due to the increased delay. Therefore, an alternative ramping scheme was used where the solenoid current was constant, and only the quadrupole was ramped. This allowed the ramping rate to be much higher and is expected to increase the number of trapped neutrons. Additionally, it causes the direction of the magnetic fields changes as a function of time, not just the amplitude.

It was found that the magnets did not train as effectively or as quickly as expected. It is suspected that this new ramping scheme may have contributed to the reduced training of the magnets. Changing the direction of the fields changes the coil orientation that the magnet is attempting to settle into and could be the reason that the magnets were never able to be run at their design current. It is also possible that the sensitivity to quenches was instead caused by damage to the magnets from an early quench.

The quadrupole magnet is on loan from the High Energy Accelerator Research Organization (KEK). It is an electron focusing magnet used at the TRISTAN accelerator[24, 16]. The magnet uses NbTi superconductor embedded in a copper matrix with a copper to superconductor ratio of 1:8. The critical temperature and field for the magnet are approximately 10 K and 7.1 T respectively. The maximum operating current is 3400 A. The inductance of the quadrupole is approximately 58 mH. The quadrupole magnet is powered with a set of 4x Agilent (Hewlett Packard) 6680A power supplies operated in master/slave mode. The power supplies can each supply a maximum voltage of 5 V and a maximum current of 875 A[17].

A pair of solenoid magnets were designed in house and constructed by American Magnetics, Inc. The maximum operating trap strength for the solenoids occurs at a current of 225 A,

which is 75% of the loadline[16]. The inductance of the solenoids is 7 H, which is much larger than that of the quadrupoles. The current for the solenoid magnets was supplied by 2x Agilent (Hewlett Packard) 6681A power supplies[17]. They each supply a maximum voltage of 8 V and a maximum current of 580 A.

2.4.1 Magnet Infrastructure

In addition to the magnets themselves, a variety of components were required to operate the magnets. A set of compensation coils is used to cancel the effect of the solenoid magnets at locations far from the experiment. High-temperature superconducting leads limit the heat load associated with transporting the extremely large currents into the magnets at 4.2 K. In addition, room temperature leads supply the current to the high-temperature superconducting leads. Finally, the quench detection circuitry protected the quadrupole magnet in the case of a quench.

The quadrupole magnet uses an active quench detection circuit that, in the event of a quench, shorts the magnet coils dumping a large fraction of the energy from the magnet into a dump resistor. The quench detection circuitry monitors the voltage across adjacent coils with opposite current orientations, which cancels out fluctuations due to ramping the magnets while leaving the contribution from ohmic heating. During normal operation, four silicon-controlled rectifier (SCR)s, which are in parallel with the coils of the quadrupole, are in the powered on, closed state. When a quench is detected, the quench detection circuitry engages a fifth SCR that flips all of the quadrupole SCRs into the powered down, open state. This shorts the quadrupole allowing a large fraction of its internal energy to be dumped to ground.

The solenoids use a passive quench protection circuit because their large inductance limits the effectiveness of an active quench detection circuit. Instead, the solenoid coils are broken up into six sections, and each section has a large diode in parallel with the coil. In a quench, the voltage in the coils will increase above the diode threshold allowing a fraction of the current to bypass the coils to ground. The rest of the energy in the solenoid magnets is deposited into the LHe in the magnet bath. The quench detection circuit was designed in part by Chris O'Shaughnessy, Sergei Dzhosyuk, and Liang Yang. Detailed information on the quench detection circuitry can be found in their theses[17, 25, 16].

The compensation coils were required to minimize the magnetic field at the location of the Spin-Echo experiment, another experiment in the cold guidehall at the NCNR that is extremely sensitive to magnetic fields. A requirement for running the UCN Lifetime Experiment at NIST was that the magnetic field from the experiment must be below the limit of 3 mG at the location of the Spin-Echo experiment. This was only possible with active compensation. A set of four rectangular compensation coils were constructed from copper bars with a 0.635 cm by 3.175 cm cross section[17]. Each coil resembles a helix with the copper bars stacked on each other in a

spiral pattern. Two of the coils were smaller allowing them to fit inside the larger coils. The small coils were connected in series with the larger coils. Each of the combined compensation coils is about 280 cm tall and 140 cm wide. Each coil consisted of 39 copper bars that were operated at 540 A from an EMHP 40-600-4111 power supply. Adjacent bars were either electrically isolated with Kapton tape, a flexible tape with very high electrical resistance, or connected with copper grease. The coils were oriented so that their normal vector pointed along the direction of the neutron beam. They were connected to the outside of the support frame of the cryostat, one on the outside of each tower of the cryostat. A plexiglass shield encased the compensation coils to prevent them from shorting to the outside world. The relatively large current draw resulted in a substantial heat load. Fans were installed at the top of the compensation coils to provide forced air cooling through vents near the bottom of the compensation coils. Thermocouples were embedded in the compensation coils and monitored by the slow control computer to notify the user after a first temperature threshold was crossed and then to shut down the experiment if the temperature surpassed a higher threshold, as a safety precaution.

The quadrupole current leads, which were donated by Fermilab, are a set of prototype leads designed for an upgrade to the Fermilab Tevatron. They consist of three sections, a low-temperature superconductor section; high-temperature superconducting section; and a normal state section. Each of the sections is connected by copper jumpers that are soldered to the leads. The low-temperature superconductor is submerged in LHe. The high-temperature superconductor alloy that was used was Bismuth strontium calcium copper oxide, which is referred to as BSCCO-2232. It has a superconducting transition temperature of 108 K. It is actively cooled by a liquid nitrogen bath, the 50 K stage of the magnet tower cryocooler, and vapor cooling from both the nitrogen and helium baths. It is maintained below about 80 K during typical operation. Finally, the normal state conductor is made of copper. Thermometers are embedded in the leads to facilitate monitoring and controlling the temperatures of the leads in order to keep them below the superconducting transition. The temperature of the leads can be adjusted by varying the flow of liquid nitrogen (LN) boiloff through the high-temperature superconducting section of the leads. These leads are estimated to introduce a heat load that is a factor of six smaller than what could be achieved with conventional vapor cooled leads[26, 15].

The leads for the solenoid magnets were manufactured in house[15, 16, 17]. Similar to the case of the quadrupole leads, it was determined that a substantial reduction in the heat load could be obtained using high-temperature superconducting leads as opposed to traditional vapor cooled leads. The smaller current requirements made it possible to construct the majority of the leads out of high-temperature superconducting tape, which is then soldered to rigid copper bar leads, which in turn are mechanically connected to large multicore copper cables outside of the cryostat. The operational temperature of the solenoid leads was 64 K, which was cooled by the 50 K stage of the magnet tower cryocooler.

2.5 Detection System

Once UCNs are trapped in the apparatus, the final step is to detect the rate of neutron β-decay events. The energetic electron from a β-decay will ionize the LHe, which results in the emission of extreme ultraviolet (EUV) photons. These EUV photons traverse the experimental cell until they strike the cell wall where they are absorbed by the TPB, which down-converts the light to blue photons. The blue photons have a longer mean free path in acrylic, which allows them to be extracted from the cell. The blue photons are emitted from the TPB back into the cell with a random direction. Some fraction of the emitted photons make it into the acrylic light guide at the downstream side of the cell, through multiple optical windows, and are guided to a pair of PMTs operating at room temperature. The two main detection PMTs are operated in coincidence, and if the signal in each passes a hardware threshold, in addition to some other requirements, a trigger signal is sent to the DAQ cards and the voltage trace in the two main detection channels, the active veto channel, and the reference channel are all recorded along with a timestamp. The following sections describe each of the systems that participate in this process.

2.5.1 Light Production

The process for light production from a neutron β-decay in the UCN Lifetime Experiment at NIST is relatively complicated. Both the energetic electron and the proton have enough energy to create scintillation light. However, the majority of the light is expected to come from the electron, so it will be the focus of this section. The passage of charged particles through LHe will result in the creation of elementary excitations in the liquid, helium ionic states, and helium molecular states. Some fraction of the ionic helium decays into the excited helium molecular states. The excited helium molecular states are created in both the triplet and singlet states, both of which can radiatively decay resulting in scintillation light.

Many of the reaction rates in these processes depend on the density of the states and their mobility. The states will tend to diffuse after creation resulting in a complicated timing scheme for the various reactions. For example, the fraction of the ions that make it into the molecular states is strongly dependent on the rate of recombination of the ions with electrons in the liquid and therefore on the separation distance between individual ionization events. The average distance between ionization events is strongly dependent on the energy deposited per unit path length, which varies strongly for different radiation types. This manifests as a variation in the light production efficiency for different types of radiation. The ratio of the light yield of an α particle and an electron of the same energy has been shown to be 0.182 in LHe[27].

The singlet and triplet states can both decay radiatively emitting EUV photons that have a maximum in their energy distribution of $\approx$ 15.5 eV and a full width at half maximum of

$\approx$ 3 eV[28]. The singlet state has a much shorter lifetime than the triplet state and has been shown to be less than 10 ns[29, 30]. The triplet state, in contrast, decays with a 13 ± 2 s lifetime[30], which is larger by 9 orders of magnitude. Other time dependencies exist in the scintillation light. Both β and α particles exhibit a component that decays with a 1.7 μs lifetime. The α particles also show a $1/t$ dependence that is much smaller for β radiation[31, 29]. This structure highlights the complicated nature of LHe scintillation. Ultimately the UCN Lifetime Experiment at NIST attempts to measure the prompt signal from the singlet decay, neglecting the longer lifetime processes. The other time dependencies in the scintillation light can create afterpulsing, which is a potential background for the experiment.

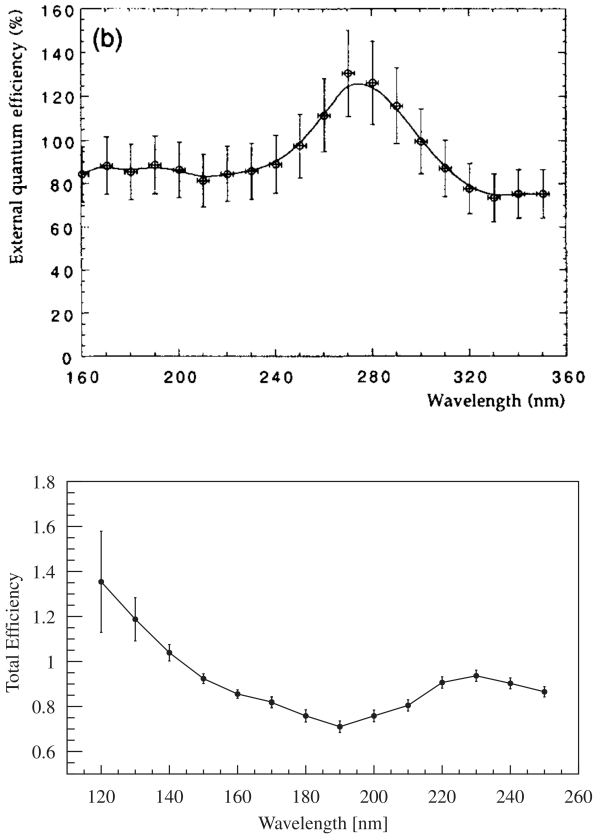
The fraction of the electron energy that is emitted as prompt EUV light has been measured to be 24%[32]. This reference was later cited along with a citation for personal communications with one of the authors of the original paper to be 35%[29]. This is an important value for estimating the number of photoelectrons (PE) expected in the neutron β-decay signal and the disagreement between these two numbers without justification is troubling. The value 35% along with an average photon energy of 16 eV results in 22 γ/keV, which is the typical value that is used in the field. Using the electron endpoint energy for neutron β-decay, 782.6 keV, and the average EUV photon energy, 16 eV, the number of EUV photons from a maximally energitic electron is 1.17×10^4 or 1.71×10^4 EUV photons respectively for 24% and 35%.

The scintillation yield has also been shown to be a function of the temperature of the LHe. The temperature dependence of the scintillation differs strongly for β and α scintillation[31, 29]. Quantitative descriptions of the temperature dependence for β scintillation in our temperature range are unavailable, however extrapolating from the higher temperatures in [31] the effective gain difference is estimated to be $< 1\%$ between our warmest data (1 K) and production data (250 mK to 400 mK). Furthermore, this would only affect the data by an effective gain difference, which would be introduced through changes in the location of the pulse shape cut thresholds. This gain drift is unmeasurable with the gain monitor because the photons from the reference LED will not be subject to this effect. Since the temperature variation inside and between the production data files is more than an order of magnitude smaller than this difference, this effect is expected to be small and is not considered further.

The LHe is transparent to the EUV photons, which is allowed in part because the energy of the first excited atomic state of helium, 20 eV, is greater than the photon energy. This allows any EUV light from the radiative decay to traverse the cell unhindered. The inside of the cell is covered by ePTFE inserts that are coated in TPB. TPB is an organic fluor. There is disagreement in measurements of TPB's quantum efficiency as is demonstrated in Figure 2.4. Overall, the quantum efficiency is relatively constant throughout the measurement range, although the measurements were completed at longer wavelengths than the scintillation spectrum from radiative decay of the singlet states in LHe. The quantum efficiency was measured

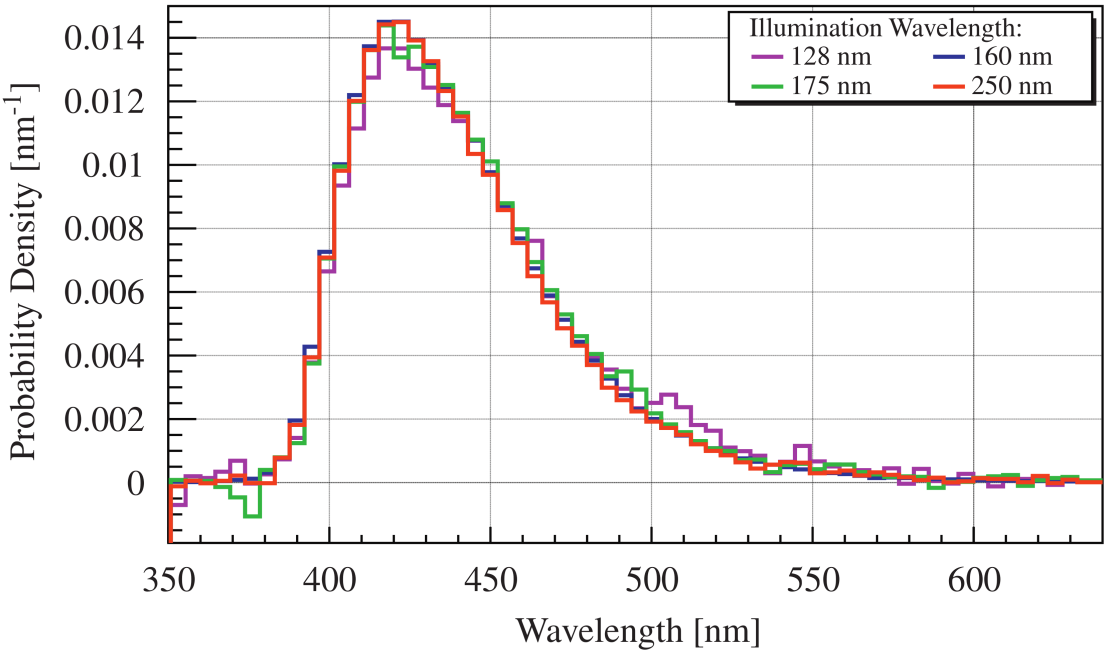
by our collaboration in the previous apparatus to be 1.35±0.1[16]. It is expected to be similar in the current apparatus. The wavelength spectrum of the emitted fluorescent photons is relatively independent of the incident photon energy, see Figure 2.5. TPB fluoresces with a lifetime that has been measured to be 0.64 ns and 1.68 ns when dissolved in the solvents pseudocumene and cyclohexane respectively[33]. It is not clear how representative these lifetimes are for the TPB evaporated on an ePTFE substrate at 350 mK.

Figure 2.4: The quantum efficiency of TPB as a function of the wavelength of incident photons from two publications showing the substantial disagreement in the published spectral shape (Top [34], Bottom [35].)



The visible light that is produced by the TPB has a large mean free path in both helium and acrylic. The down conversion is the last process in preparing the scintillation light to be extracted to the PMTs at room temperature.

Figure 2.5: TPB's fluorescence spectrum[35].



2.5.2 Light Transport

Some fraction of the blue photons that are emitted from the TPB inserts is transported through the acrylic light guides, vacuum gaps, a compound parabolic concentrator, and split with a custom acrylic splitter between two PMTs operating at room temperature. The following section describes the light transport and parameters that relate to the light transport efficiency.

The first step for a blue photon to be detected is to make it into the acrylic light guide, which is located on the downstream side of the cell. The light is expected to be emitted isotropically from the cell wall. Most of the blue photons reflect from the cell wall a few times before making it to the light guide. The absorption of the wall materials is estimated to be $\approx 10\%$. Therefore, it is improbable for photons to make it out of the cell if they take more than a handful of bounces to enter the light guide. ePTFE was chosen as the cell wall material in part because it is a diffuse scatterer of light because of its very irregular porous structure (Figure 4.7.)

The 350 mK light guide sits inside the BN and graphite inserts on the downstream side of the cell. It extends almost all of the way to the end of the cell. Light passing out of the far end of the 350 mK light guide traverses a small LHe filled gap before reaching a 0.48 mm acrylic window[17] that is part of the cell's end cap. The positioning of the cell is aligned by a Vespel 22

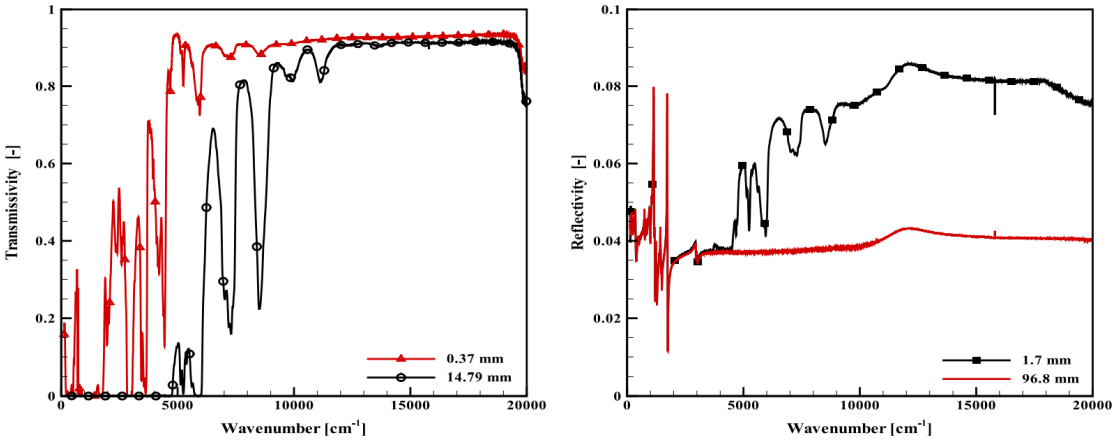
pin, chosen in part for its low thermal conduction, that allows the cell to be positioned in a very reproducible manner at a position approximately 0.1 mm from the 4 K light guide[17]. The close positioning limits the light losses through the gap while the reproducibility of the positioning limits variations in the light detection efficiency each time the cell is aligned. The pin shape was chosen to limit the conductive heat transfer from the 4 K surface to the cell at 350 mK. The gap between the acrylic window and the 4 K light guide is part of the IVC, and therefore is under vacuum. The light passes through a second vacuum gap, this time in the OVC, through a sapphire window at 77 K to limit infrared heating to 4 K, through a third vacuum gap again in the OVC, into the 300 K light guide. Once in the 300 K light guide, it passes through an optical grease joint, and into the compound parabolic concentrator. The compound parabolic concentrator acts both as a focusing element and as an adapter that downsizes the acrylic light guide to match the size of the PMT tubes. After the compound parabolic concentrator, the light encounters a splitter that splits the light through the vertical plane between two PMTs operated at room temperature.

To estimate the number of photons expected from a neutron β-decay event, an accurate estimate of the light detection efficiency is needed. This has been estimated with a GuideIt simulation that models the photon transport through materials. The details of this simulation can be found in Chris O'Shaughnessy's thesis[17]. The detection efficiency is a weak function of the location of the β-decay along the plane perpendicular to the longitudinal axis. The light detection efficiency is, in contrast, a strong function of the location along the longitudinal axis. It is almost linear at distances greater than roughly 200 mm from the end of the light guide. There is also a slight reduction in the light collection efficiency very close to the acrylic light guide, which is thought to be due to the reduced solid angle coverage of TPB near the light guide because the TPB coating ends at the start of the light guide. The maximum of the light collection efficiency can be estimated as a percentage using Figure 2.7. By dividing the maximum of the expected number of photons for a 364 keV event by the estimate of 22 γ/keV, the light collection efficiency is estimated to be 0.69%.

The range of the electrons in the LHe could also influence the light detection efficiency. If the path length is sufficiently large, the detection efficiency as a function of the location could couple with the direction of the electron momentum when the UCN β-decays. Electrons lose about 30 keV/mm as they pass through LHe[32]. For the endpoint energy and the maximum of the energy distribution, this corresponds to a path length of 26 mm and 12 mm, respectively. The Guideit simulation suggests that the light detection efficiency does not vary much on this length scale, see Figure 2.7. This effect is slightly exacerbated by the neutron β-decay asymmetry coefficient that determines the correlation between the neutron spin and the electron momentum, the big A correlation coefficient[36]. The A coefficient is the subject of continued investigation. The 2016 PDG review lists a relatively weak anti-correlation between the two

vectors, $A = -0.1184 \pm 0.0010$[3]. The fact that the adiabatic condition is easily satisfied in the trap suggests that the neutrons are strongly polarized by the magnetic trap. However, the direction of the magnetic field at any given location is not particularly straight forward. Along the circular cross section of the trap, the magnetic field is expected to be roughly radially inward in one pair of opposing quadrants and radially outward in the other two. This, in addition to the cylindrical symmetry of the cell, does not seem to have an effect on the light detection efficiency. The solenoid fields, on the other hand, create a magnetic field that points toward the front or back of the cell. This causes an asymmetry in the electron trajectories. Considering the weak correlation of A and the relatively short electron path lengths, the effect on the detection efficiency is still expected to be negligible. Unfortunately, this effect was not considered until after the apparatus was disassembled and since no effects were thought of that coupled with the absolute orientation of the magnetic fields, the current sense in the magnets was not recorded. This effect is expected to result in a slight offset in the mean of the light detection efficiency curve in addition to slight changes in the spectral shape. However, it is not known which direction the shift is in. The dominant effect is expected to be the part that is direction independent and will result in a broadening of the neutron pulse area distribution. As mentioned previously, the effect on the light detection efficiency is expected to be negligible.

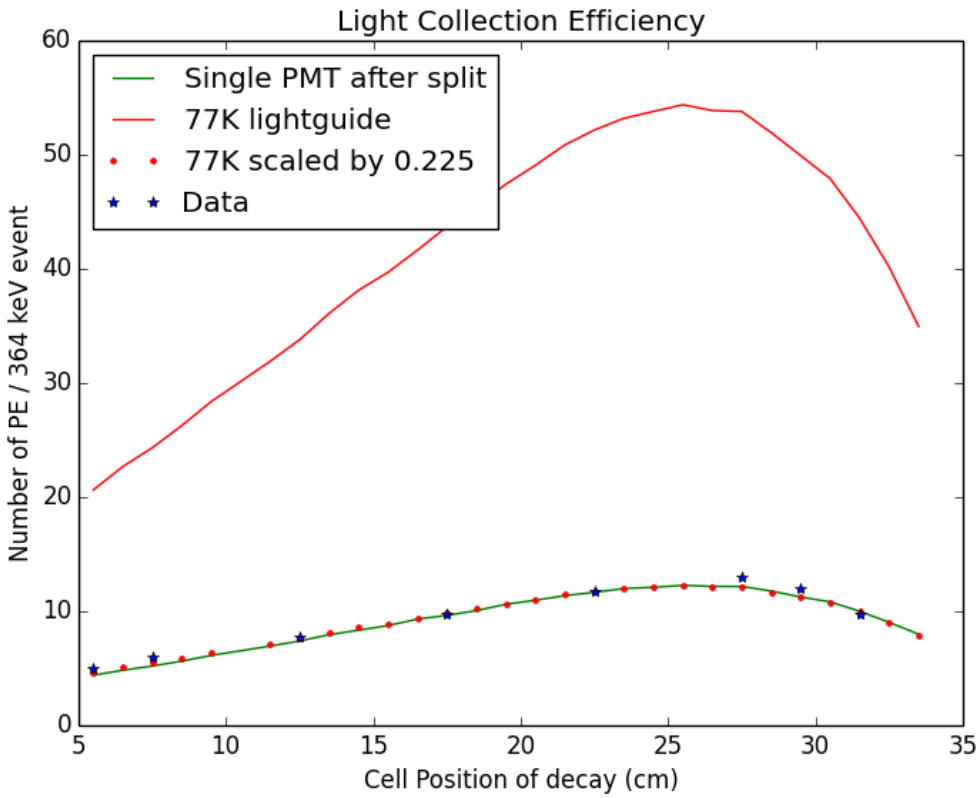
Figure 2.6: PMMA transmission and reflectance to visible light[37].



(a) Transmissivity (b) Reflectivity

The light collection efficiency simulation requires accurate inputs for the optical properties of all of the materials used in the light transport system. The published data on some of these materials are not in particularly good agreement. For example, the transmittance through poly-

Figure 2.7: The light transport efficiency as a function of the location along the longitudinal axis of the cell. This figure is an updated version of Figure 3.21 from Chris O'Shaughnessy's thesis[17].



methyl methacrylate (PMMA) acrylic seems to vary between publications. One paper reports the transmittance of light through PMMA as being roughly constant at approximately 90% from about 1600 nm to 525 nm with a kink resulting in a transmission of roughly 80% at 500 nm where their measurements were performed on a 14.9 cm PMMA slab[37]. Other sources have varying results for the transmission, some showing 90% transmission all the way to 375 nm and 10% transmission above 300 nm. The transmission was also measured through 0.1 mm to 0.2 mm PMMA films to be between 80% to 85% in a wavelength range of 375 nm to 500 nm[38]. All reported values suggest a short mean free path for UV photons in PMMA with an almost constant transmission for the majority of the visible spectrum. This substantial disagreement in the measured transmission through PMMA and in all of the other materials will introduce additional uncertainty in an absolute determination of the light collection efficiency. However, it is not expected to affect the qualitative features of the light transport simulation.

2.5.3 Light Detection

When a neutron β-decay occurs in the cell, a large number of photons are produced. Some fraction of those photons transverse the light guides and shine on one or the other of the main detection PMTs. The PMTs convert the photons into an electric signal, which is highly correlated with the number of photons that struck the PMT. The trigger logic of the detection circuitry analyzes the signal in the two main detection channels to determine if the event should be recorded. When the trigger logic is satisfied a trigger is sent to the DAQ which then records the voltage trace in the two main detection channels as well as the active veto and reference PMT channels. The active veto PMT measures the signal in the active veto system, which is discussed in Section 2.12. The reference PMT is part of the gain monitor, which is discussed in Section 2.11.

The PMTs and the trigger logic both leave features in the data that constrain the way that the analysis can be performed. Therefore, each of these systems is described in substantial detail in the following sections. Special attention is given to features in the light detection system that introduce corresponding features in the data in an attempt to improve our understanding of both the data and the apparatus. At some level, the DAQ cards will also introduce some features in the trace in ways that are similar to the PMTs and trigger logic. However, the size of these effects are smaller, and the DAQ is sufficiently complicated to warrant its own section, see Section 2.6, which starts on page 44 for additional information on the DAQ and the DAQ cards specifically.

Photomultiplier Tubes (PMTs)

The main detection system uses photomultiplier tubes (PMTs) to detect the light from neutron β-decay. The following paragraphs describe the basic operating principles of PMTs, factors that affect the linearity of the detector, and the physical mechanisms that result in some of the inherent features in the data. Hamamatsu's PMT handbook[39] is an excellent resource for additional information including basics about how PMTs work, proper operation, and design characteristics.

PMTs consists of a photocathode, dynodes, and an anode. The operational voltage varies drastically with both the PMT and socket design. However, a typical operational voltage is a few kV. The photocathode is made of a material that has a relatively large probability of emitting electrons when it is struck by a photon. An electron emitted from the photocathode is termed a primary electron. The probability of a primary electron being emitted for a single photon for a typical PMT photocathode is approximately 20% for photons in the wavelength range where the photocathode is most senstive. In a typical event, many photons strike the photocathode in a small period of time resulting in a large number of primary electrons. These primary electrons

are then accelerated in an electric field toward a set of electrodes, at incrementally increasing voltages, called dynodes. Each primary electron that strikes the first dynode releases a large number of secondary electrons, which are then accelerated by the electric potential toward the second dynode. At each dynode stage, the number of electrons grows geometrically, which results in the amplification of the signal. The resulting cascade of electrons is finally deposited on the anode resulting in the output signal of the PMT, the anode current. The contribution to the anode current resulting from an individual primary electron is termed a PE, which is a term that will be used frequently throughout this work.

The response of PMTs is very linear during proper operation, however, in a few circumstances, the linearity of the PMT's response can be affected. In the following paragraphs, we consider mechanisms that are expected to introduce gain drifts in this experiment due to details of the methodology used in this experiment.

Gain Drift Mechanisms

The known mechanisms that can introduce gain drifts in this experiment can be broken into three basic types. Temperature-induced gain drifts are caused when the PMTs are turned on just before data collection. The increased currents cause additional ohmic heating, which results in the PMTs warming up. The properties of the electronics in the PMTs change as the temperature changes, which causes a gain drift. The second type of gain drift mechanisms are rate-dependent effects. Changing the photon flux on the photocathode is known to introduce hysteretic gain drifts. This could be caused by mechanisms associated with the photocathode or with the dynode chain. Magnetic effects are the final gain drift mechanisms. The gain of PMTs is sensitive to magnetic fields, and time-dependent magnetic fields can cause gain drifts. To account for these gain drifts, a gain monitor was developed to calculate and correct the gain as a function of time. For information on the system used to monitor the gain drifts, see Section 2.11, which starts on page 52. A discussion of how the gain is calculated and corrected can be found in Section 3.2.4, which starts on page 90. In the following paragraphs, each of the gain drift mechanisms will be discussed and, when possible, characterized using a combination of production data and diagnostics data. It is important to keep in mind that the gain drifts that are described here are before the gain correction. The gain correction removes these gain drifts to a high level before applying the cuts, which will reduce the size of systematic effects due to gain drifts. The accuracy of the gain calculation is assessed in Section 5.4, which starts on page 143 and is used to place limits on the size of the systematic effect introduced by inaccuracies in the gain calculation.

First, let us consider temperature effects in the PMTs and the PMT bases. As mentioned previously, the PMTs are powered down to a lower voltage during the neutron filling stage.

This protects them from the large amount of light that is created by the cold neutron beam as it is dumped into the helium bath. When the neutron filling stage ends, the beam closes, and the PMTs are ramped back up to operational voltage before data collection begins. In all non-trapping files and any other files where the magnetic trap is flushed, the quadrupole magnet is also ramped before the start of data collection. The duration that the quadrupole is ramped gives the PMTs a little bit of time to warm up after being set to their operating voltage. Data files from the same series use the same duration between ramping the PMTs and the start of data acquisition for both the trapping and non-trapping files to make the files as similar as possible. The data can be divided into two general classifications with different delay times. In the static production data, the quadrupole field is held constant throughout both the filling stage and the data acquisition stage. In contrast, the ramping production data lowers the quadrupole field briefly between the filling stage and the data acquisition phase to purge higher energy neutrons from the trap, which reduced the effect of wall losses during the data acquisition stage; this topic is described in more detail in Section 5.3. In the static production data, this duration between the PMTs being ramped and the start of data acquisition ranges from 60 s to 80 s*. For flushing production data, the duration is 200 s to 275 s. As a result of the shortness of this duration, the components in the PMT bases are still warming up during data collection in many of the data files, which is known to cause gain drifts in PMTs[39]. For reference, when using PMTs general practice is to leave the PMT running for a few hours before data collection so that the temperatures can stabilize properly. That is not possible in this experiment.

The temperature dependent gain drifts can be isolated by taking data where during data collection the PMTs are powered down, the system is left sitting for a well-defined duration, and then the PMTs are powered back up. When the power is turned back up, the location of the reference pulser is monitored. This process was done twice. The first time the PMTs were powered down for 5 min. The second time they were powered down for 30 min. The temperature-induced gain drifts were a known effect in the previous generation of the experiment. Therefore, the channel 1 PMT base was redesigned and commissioned by Karl Schelhammer to take advantage of resistors that are insensitive to temperature changes. In the upgraded PMT base, the data shows no evidence of temperature-dependent gain drifts. The gain in the traditional PMT base (channel 2) fits well to an exponential with a fractional amplitude of 10% (16%) for the 5 min (30 min) duration. In both cases, the time constant was consistent with a value of 52 s. The diagnostic data used in this analysis was taken by Karl Schelhammer and is shown in Figure 4.11 of his thesis[18]. Note that the scale on the x-axis of his thesis is too small by a factor of 10, which is why, by eye, the data in his thesis appears to be consistent with a 5 s

*Note, this is not the delay time mentioned in the analysis, it is the time between powering up the PMTs and beginning data collection, which is somewhat shorter than the delay time.

time constant.

The gain of PMTs is also susceptible to changes in the magnetic field strength inside the PMT. This is because PMTs are very carefully designed to optimize the trajectories of the primary and secondary electrons inside the dynode chain thereby optimizing the amplification and timing of the signal on the anode. Because of this, the gains of PMTs are very sensitive to magnetic fields, which will alter the trajectories of the electrons inside the dynode chain. If the magnetic fields were static, this would result in an overall reduction in the gain that was constant in time. However, in this experiment where the magnetic fields are ramped, it results in a time dependence of the gain. The magnetic fields are not ramped during data collection. However, the support structure becomes magnetized, and the magnetic shielding saturates both with timescales that can extend into the data collection phase. PMTs can be designed to be insensitive to magnetic fields[39]. However, this typically comes with additional financial costs and reduced performance. Because traditional PMTs are used in this experiment, the PMTs are expected to experience gain drifts due to magnetic field effects. This has two contributions, one from the quadrupole magnet and the second from the solenoid magnets. The effects are separate because the two sets of magnets were not ramped together. They will be considered separately in the following paragraphs.

The gain drift from ramping the solenoids can be extracted from the data using diagnostic data sets that were taken after each cryogen fill*. The magnet ramp-up and overshoot script was used during production data collection to monitor the gain drift due to the solenoid ramp. These data files were used after a cryogen fill both to give the gains time to stabilize and to verify, from the shape of the pulse area spectra, that the gains were consistent with expectation before taking production data files. In these files, the gain drifts caused by ramping the solenoid can be observed by watching the location of the reference pulser as a function of time in the main detection channels. The gain drift due to solenoid ramping was estimated from both the magnet ramp-up and overshoot script, s16r31, and the following test run script, s16r32. There is a gap between these two files of about 2.5 min. The extracted fit parameters are in good agreement for each file individually as well as when a simultaneous fit is applied to both data files. The simultaneous fit takes advantage of the longer duration, and therefore it was used to characterize the gain drift. First, for channel 1 the fit returned $\tau = 1548.1$ s; $A = -4.5167 \times 10^4$; and $y_0 = 1.1599 \times 10^5$, i.e. the gain is increasing in time. This corresponds to the gain drifting by 2.1% of the steady state value during the data acquisition phase of the first data file after the cryogen fill. For channel 2, the fit parameters are $\tau = 1268.2$ s, $A = -2.0 \times 10^4$, and $y_0 = 9.5 \times 10^4$. This corresponds to a smaller relative gain drift with an increase of about 0.59% of the steady state value.

*The practice of doing test runs after cryogen fills was implemented on 1/20/2011 before file s11r100. Cryogen fills before that do not have this type of diagnostic data.

The solenoid gain drift seems to be well characterized by an exponential. It has a small fractional amplitude and a long time constant, which results in a small effect that primarily affects the first few data files after a cryogen fill. This is because the solenoid is only ramped during cryogen fills and therefore any files that are more than a few time constants after the cryogen fill are negligibly affected by the solenoid ramp. As mentioned elsewhere, production data can be taken for about 23 h before needing a cryogen fill. The first data file after a cryogen fill starts data acquisition about 1 h after the end of the solenoid ramp. After that, the duration between the start of data acquisition for consecutive files is about 1.5 h. Taking this into account, approximately 6% to 12% of the data is affected in a meaningful way by the solenoid gain drifts.

The solenoid gain drift, since it uses magnet ramp-up and overshoot script, is calculated using an entire cryogen fill worth of data. It was found that the fit parameters for different cryogen fills vary sufficiently that the average fit parameters very poorly describe the individual cryogen fills. This seems to suggest that there is some other mechanism that contributes on a similar level to the solenoid gain drift that is not properly being addressed. This additional factor could be a hysteretic effect or an environmental variable like the ambient temperature or humidity, for example. As a result, the numbers presented above should be taken as only a rough estimate of the size and timing distribution of the solenoid gain drift.

The quadrupole ramp is more complicated than the solenoid ramp. Whereas the solenoids are ramped up at the start of a cryogen fill and then left at operational strength, the quadrupole magnets are ramped frequently during normal operation. In addition, the way that the quadrupole is ramped varies between different data sets, which includes changes in the ramp speed, delay, flushing depth, and trap strength. Each of these parameters is expected to affect the gain drift. The quadrupole is ramped at times very close to the times at which other parameters are changing that could independently cause gain drifts, e.g. closing the beam and voltage ramps of the PMTs. All of these factors make it very challenging to separate out the quadrupole gain drift and will cause this component of the gain drift to vary from run type to run type. To be able to characterize the gain drift at all, it must be separated from the other gain drift mechanisms.

The best data for separating out the effect of the quadrupole ramp is the static, i.e. non-flushing, production data. In the static data, the trapping files do not have a quadrupole ramp before the start of data acquisition. The non-trapping files, in contrast, ramp the quadrupole magnet from no current to operational strength just before the start of data acquisition. The rest of the operating parameters for these files, namely the timing of the beam opening and the PMTs ramping up to full voltage, are the same. Unfortunately, all of the static production data at 60% field strength has gain jumps, a phenomenon that will be discussed later, which introduces uncertainty into our characterization of the quadrupole gain drift. A diagnostics data set was taken at a higher temperature that is a static 70% data and wasn't subject to

the gain jumps. Therefore, this data series was used to estimate the size of the quadrupole magnet ramping gain drift. The difference in the gains in channel 2 between the trapping and non-trapping data is consistent with a constant value, i.e. there appears to be no gain drift in channel 2 from this mechanism. Channel 1 shows a reduction in the gain by 2% of the steady state value with a time constant of 4.4×10^3 s.

This characterization is subject to all of the caveats mentioned previously. This was calculated at one operating strength, with one ramping speed, and it has a simple ramping scheme since the trap is not flushed. We do not have data that allows us to remove the effect of ramping the quadrupole in the rest of the running configurations and so we do not know how much the gain drift might differ from one data type to the next. Therefore, additional data should be taken to separate out the gain drift to the quadrupole ramp in each running configuration. A more detailed discussion on this topic can be found at the end of this chapter.

As mentioned previously, the gain of PMTs are known to be rate-dependent[40, 41]. The rate-dependent gain drifts are thought to be caused by electron and hole buildup on the last few dynodes and corresponding current discharges. This process has been shown to fit to the sum of three exponential populations decaying away with lifetimes that are shorter than a few hundred seconds in a manner that is strongly hysteretic[42]. During the neutron filling stage, the PMTs are exposed to a much larger incident photon flux than what is expected during the observation stage due to processes involving the large flux of cold neutrons being dumped into the apparatus. As mentioned previously, the PMTs are ramped down to a lower voltage setting during the neutron filling stage, which limits the amplification of the electron cascade in the dynode chain. This protects the later dynodes from the extremely large currents that they would otherwise experience, which could permanently damage the PMTs.

After the neutron fill stage ends and, consequently, the number of primary electrons emitted from the photocathode begins to decrease, the PMTs are powered back up, and the currents in the dynode chain increase to a much higher value. This introduces a gain drift as charge builds up on the later dynodes and they become susceptible to discharges on the dynode surface. Additionally, during the data acquisition stage, the data rate is expected to decrease as both the neutron activation and UCN populations decay away. This could introduce a second rate-dependent gain drift, although the derivative of the rate will be much smaller and therefore the gain drift should be as well. In summary, rate-dependent gain drifts are expected in this experiment. In addition, it seems possible that the gain drifts could have a complicated time dependence.

The PMT Handbook[39] mentions a similar effect. It mentions that if the PMT is operating and exposed to intense light, the voltage divider may be unable to supply the necessary current and charge may begin to build on dynodes. This is particularly important on the later dynodes, where the electron flux is much higher and therefore a large current is required to supply the

emitted electrons. This can result in a substantial reduction in the amplification of the PMT. It suggests that increasing the operating voltage of the PMT can increase the dynode current restoring the linearity, however, at some point, the electron current can become sufficiently high to damage the dynode. Additionally, it mentions that the linearity can also be improved at high currents by modifying the voltage divider and the PMT design.

It is possible that there are other gain drift mechanisms that are associated with when the beam is closed. For example, exposing the photocathode to a large light flux just before operation could affect the quantum efficiency of the photocathode. If this was hysteretic, it could extend into the data acquisition phase causing an additional gain drift. We refer to this potential mechanism as photocathode blinding*. Because this is an atypical way of operating PMTs, a lack of published data can not be taken as an indication that photocathode blinding does not occur. Unfortunately, we do not have data that can be used to place a limit on the size of this effect. Without the ability to constrain this effect experimentally and with no published evidence, I am forced to assume that photocathode blinding, if it exists at all, is negligible. At the end of this chapter, specific data sets are suggested which are designed to estimate the size and time dependence of photocathode blinding and to separate it out from rate-dependent gain drifts due to the amplification in the dynode chain.

Our experiment lacks data that isolates any remaining gain drift mechanisms. Therefore, the last step is to estimate the effect of any remaining gain drifts in the data. First, we subtract the estimates of the known gain drifts from the gain drift data leaving only the mechanisms that have not been accounted for. Then we fit what remains to a single exponential plus a y-offset to characterize the sum of any additional gain drift mechanisms, which is expected to include rate-dependent effects and potentially photocathode blinding. This has to be calculated independently for the trapping and non-trapping data in both the channel 1 and channel 2 main detection PMTs. The results can be found in Table 2.2. The remaining gain drifts are found to be relatively large, with a lifetime that is on the order of the neutron lifetime. These lifetimes are much longer than what is expected for rate-dependent gain drifts. We also find some disagreement between the values in the non-trapping and trapping data for the same channel. This is not expected for photocathode blinding, and it is hard to believe that the rate-dependent gain drifts would vary this much between the trapping and non-trapping data. This may be an indication that the uncertainties presented in the table, which are the outputs from the fit function, are underestimates. Despite this, this work has succeeded in calculating quantitative estimates of a few distinct gain drift mechanisms.

This concludes our attempt to characterize the gain drift due to the individual mechanisms listed above. We have found that the gain drifts affect the two main detection channels quite differently. The most extreme example is that the gain drifts are consistent with zero due to

*I have been unsuccessful in conducting a literature search on photocathode blinding

Table 2.2: The fit coefficients that characterize the residual gain drifts in the series 18 data runs after correcting for the known gain drift effects.

Data Type	Lifetime (s)	Amplitude (%)
Channel 1 Trapping	670 ± 40	5.7 ± 0.1
Channel 2 Trapping	820 ± 60	7.9 ± 0.2
Channel 1 Non-Trapping	560 ± 40	5.3 ± 0.2
Channel 2 Non-Trapping	680 ± 40	8.2 ± 0.2

temperature effects in channel 1 and the quadrupole ramp in channel 2. We have also found that the temperature effects and the rate-dependent effects have the largest amplitudes at roughly 16% in channel 2 and 5.5% (8%) in channel 1 (channel 2), respectively. Despite the large amplitude of the temperature effects, the short lifetime suggests that this effect is reduced to a 2% level after only ≈ 110 s. Therefore, in all of the production flushing data, which has a duration that is longer than this, the effect will be small. The static data has durations ranging from 60 s to 80 s*, therefore in these files, the temperature effects are expected to contribute. The solenoid and quadrupole ramp gain drifts have very small gain drift amplitudes and long lifetimes. Therefore, they are primarily a concern because they differ between the trapping and non-trapping data, which could affect the background subtraction. With this set of estimates, we gain a substantial understanding of the gain drift mechanisms that dominate in this experiment. All of these estimates are before the gain is corrected using the gain monitor. Therefore, the size of the gain drifts in the analysis will be drastically reduced.

This characterization of the gain drift could be used to attempt to limit the systematic uncertainty in the final result from the gain correction using a similar method to what was done here. However, as mentioned previously, we have found that attempting to break the gain drift up into the individual mechanisms and to account for each independently does not produce fits that accurately capture the trends in the data. It seemed that without fits that convincingly captured the trends in the data, any attempt to limit the systematic effect with this method would hold little water. Therefore, although this method has been used to estimate the gain drift parameters for the mechanisms previously mentioned, it will not be used to estimate the size of the systematic effect associated with the gain drift. Instead, gain drift fits will be performed on each file individually. When only including an individual file, the trend in the gain drift can be accurately captured with a simpler model because many of the gain drift mechanisms are either negligible or strongly correlated with other gain drift mechanisms in the majority of the data. A method, which compares two independent algorithms of calculating the gain drifts on

*Note, this is not the delay time mentioned in the analysis, it is the time between powering up the PMTs and beginning data collection, which is somewhat shorter than the delay time.

each individual file, was developed to estimate the sensitivity of the trap lifetime to the gain correction*.

Statistical Fluctuations in the Response of a PMT to Single Photons

In addition to mechanisms that introduce time dependence in the average response of the PMT, there are also statistical fluctuations in the signal. These effects will influence the accuracy with which the pulse area can be used to estimate the energy of the incident event.

These statistical fluctuations come about from the physics that is occurring inside the PMTs. The number of secondary electrons emitted from each dynode is expected to vary according to counting statistics, and therefore fluctuations in the integrated anode current are expected from each primary electron. The width of the signal comes about from the spread of the time of flight for the individual electron trajectories when they are quasi-randomly emitted from the dynode surfaces. Since the exit angle and momentum of the secondary electrons are randomly sampled from a broad distribution, fluctuations in the timing of the single PE pulses are expected. These two effects result in both the area and the width of peaks varying for every individual PE that makes up an event. Because of these effects, variations in the height of single PEs have been observed at the 20% level. When larger events are considered, a variety of additional parameters introduce variations in the peak size and shape. For example, the quantum efficiency and light transport efficiency add additional uncertainty into the voltage traces from a light source in the cell, even for a light source that produces a constant number of photons for each event.

The uncertainty in the voltage trace for each of the PEs that make up an event has additional implications for the structure of the data. This structure is useful when determining the energy calibration, between the area of an event and the number of PE. This topic is discussed in Appendix A, which starts on page 227.

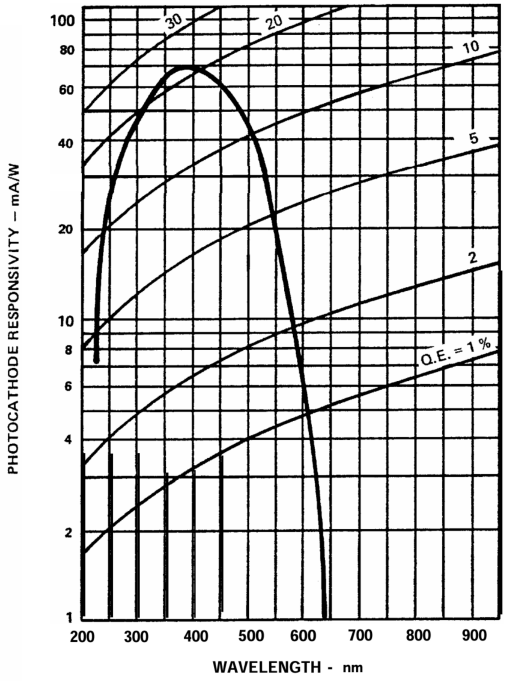
The main detection system uses two Burle 8854 PMTs. The responsivity, σ, which is proportional to the quantum efficiency of the tubes, η, is of particular interest, see Figure 2.8. The quantum efficiency is the average number of PE emitted from the photocathode per incident photon.

$$\eta = \frac{\sigma h c}{e \lambda}$$

Where h is Planck's constant, c is the speed of light, λ is the photon wavelength, and e is the charge of the electron. The quantum efficiency can be estimated from the responsivity curve. The peak of the quantum efficiency is 22.5% at 385 nm. The quantum efficiency falls below 5% at approximately 230 nm and 540 nm[43].

*Information on this method can be found in Section 5.4, which starts on page 143. A discussion of the system implemented to monitor the gain drifts can be found in Section 2.11, which starts on page 52. The gain correction is described in Section 3.2.4, which starts on page 90.

Figure 2.8: Burle 8854 responsivity curve[43].



Fundamentally, the quantum efficiency of the tubes, $\eta(\nu)$, can be understood by the expression,

$$\eta(\nu) = (1 - R)\frac{P_\nu}{k}\frac{1}{1 + 1/k_L}P_s,$$

where R is the reflection coefficient, k is the full absorption coefficient, P_ν is the probability that light absorption will excite the electron above the vacuum level, L is the mean escape length of excited electrons, P_s is the probability that an excited electron that reaches the surface of the photocathode is then emitted into the vacuum, and ν is the frequency of light[39].

The following paragraphs describe why the pulse area is used as the primary method for estimating the energy of an individual event. To do this, I will describe how the number of PE is actually the most accurate estimate of the event energy. I will then explain why the number of PE can not be accurately determined in many of the events in this experiment, which is motivation for using the less accurate but more consistent pulse area metric as the event energy parameter.

It should be noted that the amount of light is not related to the UCN's energy and instead primarily depends on how the kinetic energy is split amongst the decay daughters, the electron, proton, and neutrino. The kinetic energy in the neutrino is not detectable in this experiment, and the electron produces more light than the proton. Therefore, the amount of light produced

is strongly related to the amount of kinetic energy that is given to the electron. The reason that we care about the amount of light produced is that it may help us distinguish background events from neutron β-decay events. For example, if it could be determined that a particular event has more energy than the endpoint energy of the electron in the β-decay that event could be easily removed as a background event.

The number of PE in an event is the most accurate determination of the amount of energy of the incident particle that is possible with this detection method. This is because the number of PE, after taking into account the quantum efficiency of the tube, is directly correlated with the number of photons that strike the photocathode. This, in turn, only depends on the light detection efficiency and the amount of light produced in the event. For a given event mechanism, the amount of light produced is strongly correlated with the energy of the event. As I have indicated above, in multiple stages of this detection process there are mechanisms that introduce a lack of correlation in the signal. These mechanisms limit the accuracy with which the incident energy of a given event can be determined. The other primary candidate for estimating the incident energy of an event is the pulse area. If there were no fluctuations in the amplification or timing of a single PE event and no background voltage fluctuations in the trace, the pulse area and the number of PE would be equally accurate estimates of the event energy. However, this is not the case. In addition, the pulse area must be calculated in a window, which must be aligned about the event. Variations in the exact location of the window from event to event can introduce additional error into the pulse area calculation. These effects make a determination of the PE superior to a calculation of the pulse area of an event.

To maximize the accuracy of the PE calculation and thereby the determination of the energy, the main detection PMTs are operated in a single counting mode so that individual PE can be resolved in the voltage trace. As the timing resolution of the PMT becomes faster than the timing difference between the photons striking the photocathode, it becomes increasingly simple to extract the number of PE from the event in an unambiguous way.

In practice, extracting the locations of the individual PEs was not possible because the detection electronics were not fast enough to provide sufficient separation between the PEs so that each of the PEs could be resolved. By fitting the voltage trace to a combination of PE template of varying heights, similar information can be extracted, but some of the information in the trace will be lost. In addition, this method is extremely computationally demanding because it requires fits with a large number of free parameters, a height and location for each individual PEs. Furthermore, the method runs into convergence issues when two PE are very close to each other. Although attempts to fit to a combination of PE can further constrain our estimates of the energy from the pulse area, the increased complication and the substantial computational requirements prevent this from being feasible. The fast timing resolution does, however, allow better pulse shape discrimination, which is useful for distinguishing backgrounds

of various types in this experiment, even if counting the number of PE is preferable.

Trigger Logic

To be recorded, the signal in the PMTs must trigger the detection system. The hardware detection circuitry first compares both of the PMT anode current signals to a hardware pulse height discriminator threshold, which is set just above the single PE peak in the PE spectrum, using a Phillips Scientific 708 Leading Edge Discriminator. This limits the number of uncorrelated background photons. The logic signal from the pulse height discriminators is split for each channel. One part is sent to a Phillip Scientific 794 Quad Gate/Delay Generator, which creates a 258 ns gate. The other part is sent, along with the gate signal, to a 622 Quad 2-fold LOG Unit, which is a coincidence module. The gate turning on when the first of the two main detection channels surpasses the discriminator threshold. If the other main detection channel exceeds the discriminator threshold while the gate is still active, a logic pulse is sent from the coincidence module to the DAQ cards to record an event.

The trigger logic also contains a loop that inhibits an additional trigger for 4.6 μs after a trigger signal has been sent to the DAQ cards. This loop prevents any deadtime issues introduced by the DAQ cards take a varying amount of time to perform a software rearm after an event is detected.

2.6 Data Acquisition System

The DAQ code was written in-house by Scott Dewey of NIST. The DAQ software communicates with two GaGe CompuScope 12501/12502 DAQ cards, which digitizes and stores both the voltage traces and the corresponding timestamps for each event. The DAQ cards each have two input channels. The signals from the two main detection PMTs are connected to the two channels of the first card. The sum of the active veto paddles uses the first input of the second card (channel 3) and the reference PMT from the gain monitor uses the final input (channel 4).

The two DAQ cards are connected by a jumper cable supplied by the manufacturer, which syncs the two cards and forces them to use the same trigger and clock, thereby reducing any potential timing issues. In this experiment, the card's onboard clock was used to produce the timestamps, which is quoted to have a 1 sample "jitter" resulting in about a 2 ns uncertainty.

The DAQ software is also in charge of the settings of the DAQ cards including a variety of running parameters like the offset, range, and gain of the voltage values recorded by the channels; the trigger logic of the cards; the memory buffer; and the mode that the data is collected in. The memory buffer determines the location of the trigger signal timing in the voltage trace and therefore how far the voltage trace extends before the trigger signal. It is

set to 384 ns. Therefore, the voltage trace, which is 1408 ns long, extends 1024 ns beyond the trigger. The DAQ software also controls the conversion between the digitized number and the voltage. The main detection channels were operated in a ± 2 V range with an offset of 1.5 V to move zero voltage fluctuations toward one edge of the digitization range thereby increasing the effective digitization range. The reference channel used the same ± 2 V digitization range with no DC offset and the active veto system, due to its reduced gain, used a smaller ± 1 V digitization range, also with no DC offset. The voltage values are translated to a 12-bit value, which results in a voltage resolution of approximately 1 mV in the main detection and reference channels and 0.5 mV in the active veto channel. The voltage traces are written with 32-bit resolution, which results in a voltage to pulse height conversion of $\frac{\Delta V}{2^{32}} = 9.3 \times 10^{-10} V$.

The cards were run in the multiple record mode where the card fills its memory with events before writing the entire memory to disk. In this mode, the card performs a hardware rearm before the first event of the spill but only performs a fast, software rearm between the other events. The memory on the cards is sufficient to store 10000 events. In the rest of this document, the term spill will be used frequently to refer to either a set of 10000 events from a single memory dump or to the process of writing the data from the card's memory to disk.

The DAQ computer communicates with the slow control computer to allow the slow control computer, which controls the automation of the parts of the experiment that are automated, to determine when a data file should begin. Before a data file begins, the DAQ computer continuously checks for a data collection flag. When the signal to collect data is received, the DAQ software reads the initialization files to determine the operating parameters, informs the DAQ cards, creates a new data file, writes a header to the new data file describing the running parameters, and finally arms the DAQ cards as it prepares for the first trigger. As triggers come in, the DAQ cards write the voltage traces and timestamps to memory until the end of the first spill. The DAQ then writes the four voltage traces for each event in the trace to disk. Each voltage trace consists of 704$\times$ 32-bit word voltage records written in little-endian ordering with a record rate of 500 MHz, which corresponds to a 2 ns step size. After the last voltage trace of the spill is written to disk, the DAQ writes the timestamps to disk in two formats. The timestamps are single 64-bit IEEE floating point numbers, also written in little-endian ordering. After the spill finishes writing to disk, the DAQ cards rearm and wait for the next trigger. After a predetermined amount of time has passed, the slow control computer sends a stop signal to the DAQ computer. The DAQ computer waits to finish the current spill, writes the spill to disk, writes a brief footer to the file, and closes the file. It then prepares for the start of the next file.

The DAQ cards were found to have a firmware issue that results in incorrectly recording a small fraction of the timestamps. Communications with GaGe have indicated that the issue has since been fixed, but it is not possible to retroactively correct data that was already taken.

This issue is described in more detail along with the solution in Section 3.2.1, which starts on page 86.

2.7 Slow Control Computer

The slow control computer uses software written in-house by Alan Thompson of NIST to both periodically record a large number of diagnostic parameters related to the operation of the apparatus as well as to control various hardware components. The slow control computer uses a custom scripting language that allows the development of scripts to perform a variety of tasks for the experiment. This allows the execution of a particular type of data files to use the same script to ensure that all of the data files of that type are as similar as possible. The script controls ramping the magnetic fields and PMT power supplies, opening and closing the neutron beam, and coordinating with the DAQ computer in addition to a variety of other tasks.

The callable functions in the scripting language were designed to be sufficiently robust to perform complicated tasks like ramping up the quadrupole magnet, which requires switching the SCRs*. Each of the SCRs needs to be supplied a sufficient amount of voltage from the quadrupole power supplies to be able to switch properly. The amount of voltage varies somewhat. Because of this, it is possible that occasionally a SCR will fail to switch properly even in a script that has successfully run in the past. Therefore, before the quadrupole is ramped any further, it must be verified that the SCR latched properly by monitoring the voltage across the SCRs. The slow control computer takes care of ramping the magnets, switching the SCRs, and verifying that they latch properly before continuing to ramp the quadrupole. If the SCRs do not latch properly, it aborts the ramp and notifies the operator.

The slow control computer also monitors the status of various operational parameters including the readings from the thermometers inside the apparatus, the thermocouples on the compensation coils, and the voltage taps on the quench detection circuitry. For each of these parameters, alarm thresholds were developed to notify the operator. In addition, a server with phone access was implemented to alert additional team members via text message depending on the severity of the alert. In the event of a quench, the slow control computer puts the apparatus into a safe configuration automatically.

The slow control computer records a variety of messages, related to the commands in the script, recording what was done as well as the temperatures and other diagnostics as a function of time. Because it is entirely automated, this information is not subject to the same types of record errors that a handwritten lab notebook is and therefore in some ways this is the most reliable, if somewhat incomplete, log of the operation of the apparatus.

*The SCRs are described in more detail in Section 2.4.1, which starts on page 24, which describes the quadrupole quench detection circuitry.

2.8 Cryogenics

There are a variety of requirements that must be met by the cryogenics systems in order for the experiment to run successfully. The cell requires temperatures below roughly 350 mK to limit the phonon population in the LHe. This limits the rate at which the reverse superthermal effect can inelastically scatter UCN in the trap to energies well above the trap depth, thereby depopulating the trap. The size of this effect is a strong function of the temperature, $\tau_{up} \propto T^{-7}$, and is discussed in more detail in Section 5.1. In addition, the magnets use low-temperature superconductor that is submerged in a LHe bath to ensure that the magnets remain below their superconducting phase transition. The magnet leads use a combination of low and high-temperature superconductor, which has a critical temperature that is high enough that a LN bath is sufficient to keep it below its superconducting phase transition. The apparatus is also subjected to substantial heat loads during typical operation. Some of the more dominant heat loads come from eddy current heating from ramping the magnetic fields and thermal conduction down the acrylic light guide. Cryogenic systems were developed to meet these constraints in light of the large heat loads. This was the subject of a recent paper, which describes the cryogenic systems in substantial detail[15]. Additional information can be found in two theses on the experiment[17, 16]. The following paragraphs contain a very brief description of the cryogenics system followed by a few performance characteristics that constrain data collection.

The cryostat consists of four layers at 300 K, 77 K, 4.2 K, and 350 mK respectively. The outer cryostat sits at 300 K and serves as both the primary vacuum enclosure of the apparatus as well as the safety envelope. LN reservoirs maintain the temperature of the 77 K parts of the apparatus, which is primarily a thermal bath and radiation shield to reduce the heat load on the 4.2 K surfaces. The 4.2 K parts of the apparatus use LHe baths to maintain the temperature. Similarly, the 4.2 K regions are a radiation shield for the 350 mK cell while providing the separation between the OVC and IVC vacuum volumes. The quadrupole and solenoids are also components of the 4.2 K cryostat. Finally, the cell sits at approximately 350 mK. The additional cooling to the cell, to bring it below the boiling temperature of helium, is supplied by an Oxford Instruments Kelvinox 400 dilution refrigerator, which is located in the fridge tower, the upstream of the two towers. Thermal contact to the dilution refrigerator is supplied by a vertical column of superfluid LHe through a tube on the upstream side of the cell that connects to the mixing chamber of the dilution refrigerator. Additional heat loads are removed from the LHe volume by two Gifford-McMahon cryocoolers. These cryocoolers have two stages; one stage provides 25 W of cooling power at 50 K, and the other stage provides 1.5 W at 4 K. One of the cryocoolers is connected to the magnet leads at the top of the leads tower. The other is located underneath the cell and connects to the 77 K radiation shield and a copper bar on the bottom of the 4.2 K apparatus, which limits the thermal gradients.

With this configuration, production data can be taken for approximately 22 h between cryogen fills. This duration is a key parameter in determining the rate at which data can be acquired because the pauses in data collection for cryogen fills were a non-negligible contribution to the experiment's downtime. An estimate of the uptime of the experiment, including estimates of the duration of cryogen fills, can be found in Section 2.13.2, which starts on page 60.

The cryogen consumption rate is a substantial contribution to the operational costs of the experiment. Keeping the apparatus cold without the additional heat loads associated with taking data uses about 80 l/day of LHe. The helium consumption rate while taking data is 100 l/day[15]. The difference between these two numbers is primarily due to the additional heat loads associated with operating the magnetic trap.

2.9 Shielding

The larger size of the magnetic trap in the current generation of the UCN Lifetime Experiment at NIST has increased the number of trapped neutrons. However, many of the background types also scale with the size of the apparatus. As a result, reducing the backgrounds has become even more important. The strategy for reducing the backgrounds is four-fold. First, the background rates are reduced by shielding the apparatus from external background sources. Second, the cell wall materials, neutron windows, and light guide materials are carefully chosen to limit activation and luminescence backgrounds. Third, pulse shape analysis is performed to remove a large fraction of the background. Finally, the background subtraction further reduces any backgrounds that make it into the data. Of these methods, the first two are the most powerful because they remove events before the analysis and therefore do not introduce uncertainty into the timestamp histograms. They are the topic of the following section.

First, the backgrounds can be broken into two general types based on their origin, external and internal backgrounds. Because external backgrounds emanate from outside the experiment, they pass through at least a few cm of solid matter before they create light in either the LHe or acrylic. Alternatively, they can interact directly with the photocathode of the PMTs. However, the coincidence requirement between the two main detection PMTs single-handedly removes background sources that interact directly with the photocathode. For the background to pass through the few to many cm of matter before getting to acrylic or LHe, it must have substantial penetrating power. This excludes backgrounds from low-energy charged particles, for example. Instead, external backgrounds will tend to be either high-energy charged particles, e.g. cosmic muons, or high-energy photons, i.e. gamma rays. Energetic charged particles can directly create scintillation light in the LHe in a process that is very similar to the energetic electron from the neutron β-decay. γ radiation can create light by either Compton scattering in the LHe or the acrylic light guide. In the LHe, the liberated electron can create scintillation light; In the acrylic,

the electron instead will create Cherenkov radiation. Additionally, when neutrons interact with matter, they can create either prompt gamma rays or high-energy charged particles that can penetrate into the experiment. Because the experiment takes place in the experimental hall of a nuclear reactor, there is a high rate of neutron-induced backgrounds.

Internal activation can manifest as neutron activation of the various radionuclides in the experiment or as neutron-induced luminescence of the BN shielding. Additionally, natural occurring radioactivity from materials that make up the apparatus can create backgrounds. However, this is expected to be a small contribution to the total background rate. Activation or luminescence with lifetimes that are shorter than a few seconds will decay away before the start of data collection and therefore will provide negligible contributions to the data rate during data collection. Mechanisms with lifetimes longer than a few hours will contribute an almost constant background rate and are expected to have a negligible effect on the trap lifetime. The primary components for neutron activation, within a relevant time frame for the experiment, are the aluminum of the cryostat and support structures, the stainless steel of the cell, and the fluorine in the Teflon vacuum windows. Each of these materials is discussed below.

Early in this generation of the experiment, a relatively strong aluminum activation signal was observed. This background source was reduced by increasing the amount of neutron shielding at the neutron entrance of the experiment. Fluorine, which is present in the Teflon window, has a short lifetime of approximately 11.16 s* as a result of the decay chain $^{19}\mathrm{F}(n,)^{20}\mathrm{F} \rightarrow^{20}\mathrm{F}(,\gamma e^{-})^{20}\mathrm{Ne}$. Because of the short lifetime associated with the $^{20}\mathrm{F}$ β-decay, very few fluorine events make it into the data acquisition phase which begins more than 143 s after the neutron beam is closed. The stainless steel of the cell and the aluminum infrastructure are both hermetically shielded with neutron absorbers, and therefore the background rates due to activation of these materials are expected to be greatly reduced. The neutron-induced luminescence and the graphite shielding that is used to limit this background is discussed in the section on the experimental cell, for more details see Section 2.3, which starts on page 17.

The following paragraphs describe each of the shielding materials and how they were used. The experiment was surrounded by shield walls, sheets of shielding material, and custom shielding covers that were strategically placed around the experiment to reduce the background rates. A variety of shielding materials were used including polyethylene, boroflex, lithium doped plastics, lead, steel shot, and concrete. These shielding materials, as well as their advantages and weaknesses, are described in the following paragraphs.

Polyethylene is a class of common plastics that subscribe to the chemical equation $(\mathrm{C_2H_2})_\mathrm{n}\mathrm{H_2}$, where the variations between the polyethylene classes are due to the selection of n. Polyethylene is very effective at inelastically scattering neutrons because of its large hydrogen content. This makes polyethylene an effective moderator for thermal neutrons thereby down-scattering them

*https://t2.lanl.gov/nis/data/ndbrowse.php, ENDF/B-VII.1, Material 106, 4/28/2015

into energy ranges where typical neutron absorbers, like boron, have a larger absorption cross section.

Materials that are rich in boron can be used as neutron shielding. Boroflex is an example of a rubber with high boron content. Boron nitride, BN, is a rigid ceramic material that is used for shielding in this experiment. Both of these shielding materials take advantage of the $^{10}\mathrm{B}(\mathrm{n},\gamma)^{11}\mathrm{B}$ interaction to absorb neutrons in the process emitting a prompt gamma ray. This experiment does not take data while the neutron beam is open. Therefore, the prompt gamma rays from neutrons that originate from the beam and that strike the boron-rich shielding materials do not introduce a background. Other neutron sources that continue after the beam is closed can be absorbed on the neutron shielding during data collection and, consequently, emit a prompt gamma ray that can contribute to the backgrounds. The neutron absorption cross section follows a $1/t$ dependence due to the time that the neutron spends in the nuclear potential. As a result, boroflex is a much more efficient absorber for lower energy neutrons. Therefore using polyethylene to moderate thermal neutrons before they are exposed to either the boroflex or BN can substantially increase the shielding efficiency.

The stopping power for gamma rays is proportional to the charge density of the material. Lead has a very high charge density, which makes it very effective at attenuating gamma rays. Concrete and steel can also be used to effectively attenuating gamma radiation.

The NCNR uses modular shield walls throughout the facility. These shield walls are filled with polyethylene encased steel shot and take advantage of both the properties of having high density in the steel shot as well as the neutron absorption provided by polyethylene. Therefore, these are a very convenient choice of shielding when full sized shield walls can be used. Placing high Z materials inside the boron based neutron absorbers can help reduce the potential backgrounds from the prompt gamma rays.

Using these shielding materials, the following shielding strategy was developed. The secondary beam shutter, the guide tube, and BN covers around the beam entrance have already been discussed*. In addition, sheets of boroflex shielding have been placed both on the outside of the BN shielding and coating most of the surface on the outside of the apparatus. This was done to reduce activation from neutron background sources other than the NG6-u beam shutter. The guide hall has a large constant neutron background that behaves like a diffuse gas, in that the neutrons bounce repeatedly uniformly filling phase space instead of following linear trajectories. Coating the outside of the cryostat with layers of boroflex and polyflex sheeting converts a large fraction of the neutron background into a prompt gamma ray signal that can be blocked by high Z materials.

To reduce the prompt gamma rays from the boroflex shielding, in addition to other sources of gamma rays, shield walls have been constructed between the cryostat and the boroflex shielding.

*Section 2.2, which starts on page 16.

Logistical and support constraints prevented the construction of a hermetic lead shield, instead individual shield walls were put in key locations to shield from either specific experiments in the guide hall or regions that were found to have a large gamma ray component from radiation surveys in the experimental location.

The typical lead shield walls were constructed from stacked lead bricks and were 10.16 cm thick. The primary exception is on the top of the apparatus where lead sheets that were ≈ 0.6 cm in thickness were laid over the support frame of the cryostat. These sheets were sufficiently pliant that they could be bent to cover the contours of the cryostat providing a substantial solid angle coverage of the cell.

2.10 Gas Handling System

The UCN Lifetime Experiment at NIST requires an extremely high level of purity of ^{4}He in the experimental cell to limit ^{3}He absorption, an expected loss mechanism that will introduce a systematic error in the extracted trap lifetime. An isotopic purity of $R_{34} = \frac{N_{^3\mathrm{He}}}{N_{^4\mathrm{He}}} < 1 \times 10^{-15}$ is desired to limit this systematic effect to the level required for a 1 s measurement of the neutron β-decay lifetime; see Section 5.2, which starts on page 136 for more details on the ^{3}He systematic effect. This requires the production and long-term storage of isotopically pure ^{4}He. Because of the high level of purity desired, the helium is very sensitive to ^{3}He contamination. Even exposure to the trace amounts of ^{3}He in atmospheric helium could be enough to contaminate the entire 21 l (liquid) supply of isotopically pure ^{4}He. To combat this, a gas handling system was designed at NIST to store and manipulate the isotopically pure ^{4}He. The system is described in greater detail in Section 6.2.3, which starts on page 202 because the same gas handling system was used for the purifier with some additions, which is described in that section. The following is a very brief description of the gas handling system.

The primary design constraint for the system was to eliminate any rubber or plastic components, which could allow helium diffusion into the gas handling system. Therefore, the gas handling system uses all-metal seals and all-metal pressure gauges. It contains high-pressure storage bottles for long term storage of the isotopically pure ^{4}He, a custom built compressor for compressing the helium into the high-pressure bottles, intermediate-pressure storage volumes colloquially referred to as the dumps, a small all-metal seal diaphragm pump to manipulate the isotopically pure ^{4}He, a LHe cold trap to clean the helium, a capillary fill port for condensing helium into the cell, and finally a larger pumpout port for extracting isotopically pure ^{4}He from the cell. The gas handling system also contains a few additional pumps, ports for extracting gas samples and leak checking, and both one-way and pneumatic valves to protect the system.

2.11 Gain Monitor

The gain of a detector is a term that is used to describe the amount that the detector amplifies the signal. A gain drift, therefore, is a time dependence in the amplification, which can introduce an additional time dependence into the data. Gain drifts, when coupled with either hardware or software pulse shape cuts, can introduce a systematic effect into trap lifetime by causing a time-dependent cut efficiency, i.e. changing the fraction of both background and neutron events that are allowed through the cuts as a function of time. The size of the systematic error depends both on the size and time dependence of the gain drift. Gain drifts were estimated to cause a 8 s[16] systematic effect in the previous generation of this experiment. The design goal of the current apparatus was a 1 s statistical uncertainty which motivated the design and implementation of a gain monitor to measure the gain as a function of time and correct for any gain drifts. This system was designed by Karl Schelhammer and is described in substantial detail in his thesis[18].

The gain monitor uses a reference pulser operating at 100 Hz with a very stable peak size to monitor the gain of the main detection PMTs. This is done by splitting the output from the reference pulser into two parts. The first part is sent to the main detection PMTs while the second part is directed to a reference PMT, which is located far from the apparatus where it is less susceptible to the magnetic fields. The reference PMT is used to tag the reference events with near unity efficiency so that they can be selected in the analysis to calculate the gain of the main detection PMTs using just the reference pulser events. The reference events have a large signal in the reference PMT allowing for very clear separation between the reference events and the zero peak, which is why the system is capable of tagging reference events with near unity detection efficiency. Because of this large separation, extremely large gain drifts would be required in the reference PMT in order for the tagging efficiency of reference events to be affected. In contrast, the gain drift in the reference PMT is found to be very stable in time. The reference pulser events make up about 22% of the data. The gain monitor allows a $\approx 0.5\%$ estimate of the gain with 15 s timing resolution*.

2.12 Active Veto System

The active veto system consists of a set of scintillating paddles, monitored by PMTs, strategically located around the apparatus to detect charged particles passing through the system

*The method of calculating and correcting the gain is discussed in the analysis chapter, see Section 3.2.4, which starts on page 90. An estimate of the systematic effect from residual gain drifts after the gain correction can be found in Section 5.4, which starts on page 143. Additional information related to the performance of the gain monitor and an attempt to break down the contribution from the different mechanisms that cause the gain drifts was discussed previously in this chapter, see Section 2.5.3, which starts on page 34.

from the outside. The majority of the events tagged by the active veto system are expected to be cosmic ray muons. The following sections describe some general properties of cosmic ray muons, which are a relatively large background source for this experiment, the design of the active veto system, and finally its performance*.

2.12.1 Cosmic Ray Muons

The particle data group has a great summary on cosmic rays and cosmic ray muons[44]. The following borrows a couple of the numbers that they give to help evaluate the expected signal from cosmic muons in this system.

Cosmic rays interacting with the upper atmosphere result in high-energy muons that bombard the earth. The cosmic muon rate at sea level is roughly 1 $\text{cm}^{-2}\text{min}^{-1}$ and follows an angular distribution of $\cos^2\theta$. Muons are a heavier cousin of electrons in the lepton family and as such are electrically charged and primarily interact with matter by interacting with the electric field of the electrons in matter. This frequently results in the muons ionizing any matter they travel through by liberating many of the electrons they interact with. The muons have a large amount of energy, on the order of 6 GeV, and deposit approximately 2 $\frac{\text{MeV}}{\text{g/cm}^2}$ as they pass through matter via ionization. This means that the muons can easily pass through the atmosphere and deep into the earth before coming to a halt as they run out of kinetic energy. The UCN Lifetime Experiment at NIST expects a large muon flux throughout the apparatus. The detection method of neutron β-decay takes advantage of the ionization of LHe by the energetic electron; the signal from cosmic muons in the LHe uses the same mechanism, and therefore cosmic muons are expected to produce background events that are similar to neutron β-decay events in this experiment.

The expected cosmic muon rate in the data can be estimated from the numbers presented above in addition to information about the geometry of the apparatus. The cell is a cylinder with an inner radius of about 6.1 cm and a length of 500 cm. Using the estimated muon flux of 1 $\text{cm}^{-2}\text{min}^{-1}$ and the cross-sectional area of the cell, the expected muon rate can be estimated to be roughly 40 s^{-1}. Some of these muons will only skim the edge of the cylinder of helium and will not deposit very much energy. To provide some context, the mean path length of cosmic muons that traverse across the cross section of the cell, i.e. no component of velocity parallel to the axis of the cylinder, is 9.7 cm resulting in roughly 2.9 MeV of energy deposition in the helium. This is more than three times the endpoint energy of the electrons from neutron β-decay. Including a component of the velocity parallel to the axis of the cylinder will further increase the energy deposited. This demonstrates that cosmic muons will result in large pulses

*Additional information on the active veto system, specifically regarding analysis pertaining to the system, can be found in the corresponding sections in the chapter on the analysis, see Section 3.2.2, which starts on page 87.

in this experiment.

Some fraction of the muons will result in a response in our detectors that is more similar to the neutron β-decay. A subpopulation of the muons will have a sufficiently short path in the LHe that they deposit less energy than the maximally energetic electron from neutron β-decay. This fraction can be estimated using the muon energy deposition rate, 2 $\frac{\text{MeV}}{\text{g/cm}^2}$, along with an approximate density of superfluid helium, 0.15 g/cm^3, the path length comes out to 2.6 cm. In the two-dimensional circular cross-sectional calculation, less than 2% of the muons will produce less energy than an electron with the maximal electron energy from neutron β-decay. A much smaller fraction of muons will produce less energy than the electrons in three dimensions. Therefore, although the muon events will extend all the way to zero energy, the vast majority of the muon events are expected to be well above the energy of neutron events.

A couple of factors have been neglected in the previous discussion. For one, the detection efficiency in the cell varies as a function of location. This is expected to impact the neutron decays and cosmic muons differently because their spatial distributions in the cell are different. The neutrons are confined by the magnetic trap so that they do not spend an equal amount of time in all locations of the cell, the muons, on the other hand, will pass through all portions of the cell uniformly. The neutrons, by design, are held in portions of the trap that have the highest light detection efficiency. Therefore, in comparison, some fraction of the muons are expected to experience a substantially lower light detection efficiency, which will push more of these events into the portions of the pulse shape phase space that the neutrons inhabit.

All of the arguments so far have only considered cosmic muons, but the cosmic muons have enough energy to create secondary particles, which can also be detected. These daughter particles are expected to cascade in conical showers from the muon's trajectory. Therefore, in some of the cosmic muon events, the muon will miss the cell, but the secondary and later generations of particles may interact with the cell. It is much more challenging to estimate the size of the signal from all possible events that could be associated with a cosmic muon and is beyond the scope of this work. That said, cosmic muons are the most abundant particle at the surface of the earth resulting from cosmic rays and are therefore expected to be the dominant contribution to cosmic ray backgrounds in the experiment.

2.12.2 Active Veto Design

To combat this source of background, an active veto system was developed. It consists of six scintillating paddles with independent PMTs strategically placed around the apparatus. When a muon passes through the scintillating paddles, it creates scintillation light that is detected. A muon that passes through both an active veto paddle and the cell will result in a concurrent signal in both the main detection and the active veto PMTs. Muon like events are tagged by

comparing the pulse area in the active veto channel to a threshold. For additional details about how the threshold is selected see Section 3.2.2, which starts on page 87. This allows cosmic muon like events to both be tagged in the analysis for further study and to be removed from the data as background events. The active veto paddles will also pick up any other sufficiently energetic charged particle that passes through the active veto paddles. However, due to the geometry, the coincidence requirement, and the known background sources, the vast majority of the events tagged by the active veto system are expected to be muons.

The six scintillating paddles are broken up into three pairs of paddles that mirror each other across the central vertical plane of the experiment. Two large, square paddles straddle the top of the cryostat between the two towers and provide the majority of the vertical solid angle coverage of the cell. The remaining paddles are all of the same type and consist of a right hexahedron of scintillating plastic with dimensions of $146 \times 32 \times 2.5$ cm. The scintillator is coupled to a PMT via a triangular section of clear light guide of length 36 cm, with a final glue joint converter to a circular cross section that is glued directly to the PMT. The second pair of paddles sit on either side of the cryostat with their central axis parallel to the cell axis and their wide axis oriented vertically. The final pair of paddles rest horizontally under the apparatus and are centered longitudinally along the light guides. This third pair of paddles provides the majority of the solid angle coverage of the acrylic light guide.

The active veto PMTs are close to the cryostat and therefore are exposed to large magnetic fields from the magnetic trap. This results in a poor gain in the active veto PMTs. A layer of Netic shielding was fabricated to provide passive magnetic shielding. Additionally, each PMT was placed inside a solenoid which was manually adjusted to improve the gain. They were adjusted before operation to maximize the gain in each of the active veto PMTs. This process was very time-consuming, and so the optimization of the gain in the active veto system was only performed at 60% and 70% of the design current of the magnets. For each data set, the solenoid currents were adjusted to the more appropriate of these two settings.

The output from each of the active veto PMTs is summed and sent to the third channel of the DAQ cards. This provides the information needed to tag active veto events so that they can be removed from the data set before calculating the trap lifetime.

2.12.3 Performance

The solid angle coverage of the active veto paddles is estimated to be $\approx 58\%$ for the helium in the cell and $\approx 30\%$ for the acrylic light guide. The majority of the coverage of helium comes from the two active veto paddles between the towers, and the light guide is primarily covered by the paddles underneath the apparatus. The cross-sectional area of the helium in the cell is estimated to be 0.15 m^2. The light guide has a cross-sectional area of 0.14 m^2. Assuming a muon

rate of 1 $cm^{-2}min^{-1}$, the expected muon tag rate is 22.1 s^{-1}. Of this, about 30% is expected to be in the acrylic and is therefore expected to have a fast voltage structure as determined by the kurtosis pulse shape metric, see Section 3.1.5, which starts on page 81 for information about how the kurtosis selects on the timing distribution of the events.

The values in the data files vary slightly between the running configuration, which presumably is an indication that the active veto PMTs have a lower gain in some of the running configurations, which results in fewer active veto events passing above the active veto threshold. The situation is complicated further because the gains in the individual active veto PMTs are not expected to change by the same amount. Therefore in different running configurations, the sensitivity to charged particles coming from different directions can change. The active veto pulse area threshold is determined for the entire active veto pulse area spectrum, therefore in particular running configurations, the veto efficiency in the acrylic and helium may vary independently. For the muon rates of a typical 70 → 50 → 70 non-trapping file, s16r33b1.itx, see Table 2.3.

Table 2.3: Expected and measured tag rates of the active veto system. Data is from a 70 → 50 → 70 non-trapping data file, s16r33b1.itx.

Type	Expected Tag Rate	Expected	Measured Tag Rate	Measured
	s^{-1}	%	s^{-1}	%
All	22.1	-	16.9	-
Fast	7.1	32	9.0	53
Slow	15.0	68	7.9	47

Because the relative tagging efficiency of the muons in the acrylic and helium can vary depending on magnet field state, the measured percentages in Table 2.3 are not particularly meaningful. Despite this, a constraint can be placed on the measured rates for each type of event. If the measured rate exceeds the expected tagging rate, it is an indication that the active veto cut is removing non-muon events. As is evident in s16r33b1.itx, the measured tagging rate for fast muon events does exceed the expected tag rate. This is an indication of an overzealous active veto cut. However, the fast events will be removed by the kurtosis cut in the analysis*. The slow active muon events are the important parameter here because a large fraction of those events will not be removed by the other cuts. In the case of the slow muon events, the measured tagging rate is below the expected rate as desired, which is an indication that the active veto threshold appears to be reasonable.

*See Section 3.1.5 for a description of the kurtosis pulse shape metric.

One aspect related to the performance of the active veto system is the frequency of false positive muon tags. These can be caused by accidental coincidence where a particle passes through an active veto paddle at the same time as any other type of event triggers the main detection system. Alternatively, random voltage fluctuations can cause the active veto pulse area to exceed the threshold. The first of these can be estimated by looking at the accidental coincidence rate between the active veto system and the reference pulser. Using the well-defined rate of the reference pulser events, it is possible to extrapolate from the number of random coincidence events between the active veto system and the reference pulser to the total data rate. This estimates the total accidental coincidence rate with the active veto system. The rate in the active veto system is sufficient that $5 \times 10^{-2}\%$ of all events are expected to experience accidental coincidence with the active veto system. This accidental coincidence in addition to the time dependence of both the data rate and the detection efficiency in the active veto system can introduce a systematic effect. As will be discussed below, the muon rate shows no evidence of a time dependence. Therefore, this systematic effect is expected to be negligible.

The pulse area in the active veto PMT is referred to as *muon_pulse_area*. An independent study verified that the probability of random voltage fluctuations causing the *muon_pulse_area* to pass above the active veto threshold is extremely low and can be neglected. In the study, the zero peak fluctuations in the active veto channel were fit to a Gaussian distribution. The distance from the center of the Gaussian fit to the threshold was determined in units of the standard deviation of the gaussian fit. The mean of the zero peak was found to be greater than 5 σ from the cut threshold.

The main detection pulse area has been found to have poor correlation with the signal in the active veto system. Equivalently, this means that the energy deposited in the main helium bath is poorly correlated with the energy deposited in the active veto system. This is believed to be due to geometric effects. A small change in the angle of the muon through the rectangular cross section of the active veto paddles only slightly changes the path length inside the scintillation paddles. That same small angle, when extrapolated to the cell location can result in a much larger change in the path length due to the circular cross section of the cell. In addition, the light detection efficiency in the veto paddles is expected to be a strong function of the location, which could further reduce the correlation in the two signals. This is not thought to affect the analysis in any meaningful way and is presented as supplemental information.

Finally, the *muon_pulse_area* pulse shape metric uses a very small fixed integration window to calculate the area, which is then compared to the *muon_pulse_area* threshold to tag cosmic muons. The small integration window was a reason for concern in the analysis. If the timing correlation for all cosmic muon events is not very high, some fraction of the muon events will be missed because the corresponding pulses will fall outside of the integration window. To verify that this was a negligible effect, the size of the region of interest was varied. Plotting

the *muon_pulse_area* metric for a variety of integration window sizes as a function of each other gives a strong indication of how the integration window effects the number of events that are removed by the cut. A larger integration window allows the very small number of events with poor timing correlation to be cut. However, it also introduces larger uncertainty in the *muon_pulse_area* calculation due to the random voltage fluctuations in the trace. This larger uncertainty is expected to introduce additional false positive and false negative active veto tags. It would be possible to include two active veto thresholds, one for two distinct regions of interest sizes. This would take advantage of the relatively small uncertainty in the smaller region of interest to define the cut accurately. The larger window could then be utilized with a more conservative threshold to cut additional large active veto events with less timing correlation. Instead, to simplify the analysis, the single small region of interest was chosen, and this choice is expected to have a negligible effect on the trap lifetime. The fact that the majority of the muon like events are captured by a 20 ns integration window is strong evidence that these events, which are highly correlated in time with the signal in the main detection PMTs are in fact cosmic muons.

The active veto system does not have an external gain monitor system like the main detection PMTs. Unfortunately, this makes estimating the gain drift in the active veto channel challenging. The magnetic shielding is expected to slowly saturate after the magnets are ramped up and, as a result, a gain drift is expected. To estimate this, a typical test run data file was split into 1000 event bins. The mean and uncertainty of the mean were calculated for each bin. The resulting curve was fit to a line of the form $a + bx$ where x is the bin index, $a = 690.51 \pm 30.2$, and $b = -4.0667 \times 10^{-4} \pm 7.56 \times 10^{-4}$. This indicates that the gain in the active veto channel is constant throughout the file. This estimate was calculated with data file s13r1b1m0.dat. Including both cosmic muon events and non-muon events in this analysis caused the data to be sufficiently non-gaussian that estimating the uncertainty was not straight-forward. A more detailed analysis, including an accurate uncertainty estimate, was not warranted because the effect is sufficiently small.

In summary, an active veto system is in place to tag a large fraction of the cosmic muon events. The cosmic muon events make up a small fraction of the data rate and are therefore expected to have a small effect on the extracted trap lifetime, however, due to the strong physical motivation for the active veto cut, it is still performed.

2.13 Performance

In addition to describing the apparatus, it is worthwhile to spend some effort benchmarking the performance of the various systems. This effort is meant to inform future generations of experiments about how successful various systems were and what must be improved upon in

future designs. The performance of a few of the systems has already been described in the corresponding sections describing that system. This section contains the performance estimates that do not have a more logical home elsewhere.

2.13.1 Ultracold Neutron (UCN) Production

The UCN density in the trap is a useful parameter for determining the rate at which an experiment can acquire statistics. Although this experiment is not designed so that the UCN can be extracted or used for other measurements, the UCN density is still a good parameter for comparing our source production to other state-of-the-art UCN sources.

The Mark II apparatus consisted of a 1.1 T magnetic trap with a volume of 1.5 L. In comparison, the KEK trap was designed for a maximal trap strength of 3.4 T with a volume of 5.27 L. The Mark II apparatus estimated 4000 UCN per fill and a reduction by a factor of two due to the flushing[16].

Assuming that the UCN evenly fill momentum phase space, the number of trapped neutrons, N, is found to be proportional to the trap depth to the three-halves power and the trap volume. Using these relations, the UCN density in the KEK trap can be estimated from the Mark II apparatus. In practice the magnet was run at or below 70% of the design current; the improvement in the number of trapped neutrons from the trap depth can be estimated from the ratio $\left(\frac{B_{\mathrm{KEK}}}{B_{\mathrm{MarkII}}}\right)^{3/2} = \left(\frac{0.7\times 3.4\mathrm{T}}{1.1\mathrm{T}}\right)^{3/2}$ suggesting a multiplicative improvement by a factor of 2.2. The volume of the trap has also increased. The number of trapped UCN can be scaled linearly with the fractional increase in the volume, an additional factor of 3.5. Combining these two effects the KEK trap is estimated to contain approximately 8 times as many UCN per fill as the Mark II apparatus in static data sets. This suggests that the KEK trap should be able to fill 32000 or 16000 UCN for the static and flushing cases respectively, assuming the same fractional loss of below threshold UCN due to flushing the magnets.

$$N \propto \int_0^{v_{\max}} v^2 dv \propto PE^{\frac{3}{2}} \propto B^{\frac{3}{2}}$$

$$N_{\mathrm{KEK}} = \left(\frac{B_{\mathrm{KEK}}}{B_{\mathrm{MarkII}}}\right)^{3/2} \frac{V_{\mathrm{KEK}}}{V_{\mathrm{MarkII}}} N_{\mathrm{MarkII}}$$

Where N is the number of trapped neutrons, v is the neutron velocity, PE is the trap depth, B is the magnetic strength of the trap, and V is the volume of the trap.

The number of trapped neutrons can also be extracted directly from the fit to the data. The final stages of the analysis produce a timestamp histogram of the number of neutron-like events that decay as a function of time throughout the observation stage. This curve is fit to an exponential with a y-offset to account for any constant backgrounds that make it through the

background subtraction. This is a fit to the decay rate, not the neutron population. However, the neutron population can easily be extracted by integrating the decay rate.

$$\dot{N}\,(t) \;= A_0 \exp\{(\,-t/\tau)\,\} + y_0$$

$$N\,(t) \;= A_0 \tau \exp\{(\,-t/\tau)\,\}$$

Where the constant term was discarded before integration and the integration constant was also discarded. This shows that multiplying the amplitude of the fit by the extracted lifetime gives an estimate of the number of neutrons at the time $t = 0$, the start of data collection. To extrapolate to the number of neutrons at the end of the filling stage $N(t = -delay)$ is used, where *delay* is the duration of the delay between the end of the neutron fill and the start of data collection.

In the flushing data, $N(t = -delay)$ is an underestimate of the number of neutrons in the trap when the beam is closed because it does not take into account the additional losses from the wall interactions that occur during the ramp.

Finally, the UCN density can be estimated by dividing the number of neutrons at the end of the fill by the volume of the trap, 5.27 L. This is a rough estimate, and is only expected to be representative of the average UCN density in the cell. In the magnetic trap, the UCN density is expected to be highly non-uniform because the magnetic potential pushes the UCN toward the center of the trap. As a result the maximum UCN density is expected to be higher at the minimum of the trap and substantially lower toward the Teflon window and any other higher potential regions of the cell volume. The estimated number of trapped neutrons for many of the data series are presented in Table 2.4. The estimates from the data are consistently lower than what was estimated from the previous experiment. In the best data sets, the UCN density is about 0.9 UCN/cc.

2.13.2 Operational Uptime

With the limited beam time available to the experiment, it is important to take full advantage of the available time. One key factor that limits the rate at which statistics can be acquired is any downtime when the apparatus is not taking data or conversely the operational uptime. Some down periods are required for safety, e.g. ramping the magnets down during cryogen fills. Other down periods are required to prepare the apparatus for taking data, e.g. the neutron filling stage to load the trap. Any ability to remove unnecessary downtime and to limit necessary downtime will improve the rate at which statistics can be acquired.

A typical run cycle consisted of pairs of trapping and non-trapping data files. Each individual run consisted of a neutron filling stage typically followed by ramping the quadrupole magnet and

Table 2.4: The number of neutrons and estimated UCN density, estimated from the data, for each of the major data sets. The delay, i.e. the time between closing the neutron beam and starting data collection is also listed.

Series	Type	N_0	ρ_N	delay
		n	n/cc	s
8	60% static cold	5900	1.1	181
13	60% static cold	8000	1.5	145
18.1	60% static cold	8700	1.7	149
11.1	$70 > 35 > 70$ cold	2800	0.53	373
11.2	$70 > 35 > 70$ cold	4300	0.82	349
11.4	$70 > 35 > 70$ cold	3700	0.79	349
12.1	$70 > 35 > 70$ cold	4300	0.81	349
12.2	$70 > 35 > 70$ cold	7200	1.4	349
14.1	$70 > 35 > 70$ cold	2900	0.55	349
14.2	$70 > 50 > 70$ cold	6600	1.3	276
16	$70 > 50 > 70$ abbreviated cold	8500	1.6	174
17	$70 > 50 > 70$ cold	8000	1.5	174

then the data collection phase in which the DAQ actively monitored the population of neutrons as they decayed away. For different running configurations, the duration of the magnet ramping varied depending on the trap depth, the flushing depth, and the ramping rate. These factors introduced some variations in the fraction of time that data could be collected. The filling stage was open was 2500 s in duration. The delay between the beam closing and the start of data collection, the delay time, varied from 142 s to 373 s for typical data. The longer delay times corresponded to data with a deeper trap depth, slower ramping speeds, and a deeper flush depth. Finally, the duration of data collection was 2560 s to 2620 s. The width in the data collection time allowed the DAQ to continue collecting data until an integer multiple of 10000 events were collected, which corresponds to a spill of the DAQ card's memory.

There were additional contributions to the downtime to prepare the apparatus for the next file. In the case of a non-trapping file followed by a trapping file, the magnetic fields were already at the operational depth. Therefore, the trapping run was started immediately. For a non-trapping file following a trapping file, the quadrupole magnet must first be ramped down. This took 150 s for a 60% running configuration or 175 s for a 70% running configuration.

The final, necessary factor that contributed to the downtime is due to cryogen fills. To do cryogen fills safely, both magnets had to be ramped down before the fill could begin and then ramped back up after the cryogen fill completed. The ramp rate of the solenoid magnets and the compensation coils was quite slow at approximately 1% every 12 s. This corresponds to 720 s (840 s) for a 60% (70%) trap depth. In addition, ramping the solenoid introduced large gain drifts

as the magnetic shielding became saturated and the support infrastructure became magnetized. To monitor the gain drifts while allowing them to stabilize, diagnostic data sets were taken after the cryogen fill. These files were used to verify that the apparatus was behaving as expected with special attention being paid to the gains of the main detection PMTs. To improve the rate at which the gains stabilized, a specialized startup script was developed, which was referred to as the magnet ramp-up and overshoot script. In this script, the solenoids were ramped above the operational strength for 10 min before ramping down to operational strength during the diagnostic data. This increased the rate at which the gains stabilized after a cryogen fill. As an example, there was a typical cryogenic fill before file, s18r46. It took 17 min to ramp the magnets down and then another 43 min before the cryogenic fill was completed. The magnet ramp-up and overshoot script took 44 min followed by an additional diagnostic file that took roughly 8 min. Therefore the cryogen fill including all magnet ramping and diagnostics took a total of 112 min.

During typical data collection, the apparatus could be run for approximately 22 h on a single cryogen fill. Therefore, the downtime due to cryogen fills was about 2 h/day or 8.3%. Using the estimates listed above, livetime associated with data collection during typical operation was about 46.4% neglecting cryogen fills or 41.5% including cryogen fills. This is estimated from ideal data collection, which does not include downtime associated with sharing beam time with the aCORN experiment or other logistical issues like reactor downtime, additional diagnostic data, etc.

Additional work could be done to optimize the duration of the neutron fill and data collection to maximize statistics. It is not expected that adjusting these parameters will be able to improve the rate of statistics acquisition substantially.

2.13.3 Drift in the Data Acquisition (DAQ) System Clock

The accuracy of the DAQ clock is vital for the experiment because the DAQ clock is the clock to which the trap lifetime is compared. Therefore, a systematic error in the timestamps recorded by the DAQ can result in a systematic error in the extracted trap lifetime. We used a comparison between the DAQ clock and a well characterized reference pulser to investigate the accuracy of the DAQ timestamps.

The comparison is performed by tagging reference events and calculating the time between adjacent reference pulser events. The timing difference across spills are removed from the investigation. There is a population of missing reference events, which are discussed in more detail in Appendix B.1, which starts on page 230. In order to account for this, the next reference pulser event is removed if it is greater than 15 ms from the previous reference pulser event. Histograms are calculated for the 2nd, 5th, 10th and 20th reference event after a given refer-

ence event, which can be seen in Figure 2.9. This shows a complicated functional form, which appears to suggest that the accuracy of the two clocks is not the same. Also, the center of the distribution is offset from 10 ms, which indicates a linear drift between the two clocks.

The linear drift between the two clocks is calculated using the mean of the duration between reference pulser events that are 20 reference pulser events apart if the events fall within a 10 μs window of the expected timing. Unfortunately, the mean is sensitive to false positive reference pulser events, which will tend to drag the mean down. The 10 μs threshold window was chosen to eliminate any false positive reference events that are far enough from the expected value that they will have a noticeable effect on the mean. The mode and median were also considered. The mode is sensitive to random fluctuations due to the relatively broad peak of the distribution. The median is relatively insensitive to false positive muons and random fluctuations, however, it continually under corrected the linear drift. Therefore, the mean was chosen as the best metric.

If each time measurement introduces additional uncorrelated error, the timing histogram will spread out as the duration between the reference pulser events increases. However, this is not apparent in the data as can be seen in the lower graph of Figure 2.9 because this analysis is done on time scales that were too short. Larger time scales can not be used to show the spread of the distribution because the adjacent reference pulser events must be inside the same spill and the rate of spills is too high. Instead, this effect can be shown by correcting for the linear drift and then displaying the cumulative sum of the reference timestamps. The result can be seen in Figure 2.10. We find that the drift between the two clocks is minuscule when compared to the neutron β-decay lifetime, and therefore we expect the systematic correction to be insignificant.

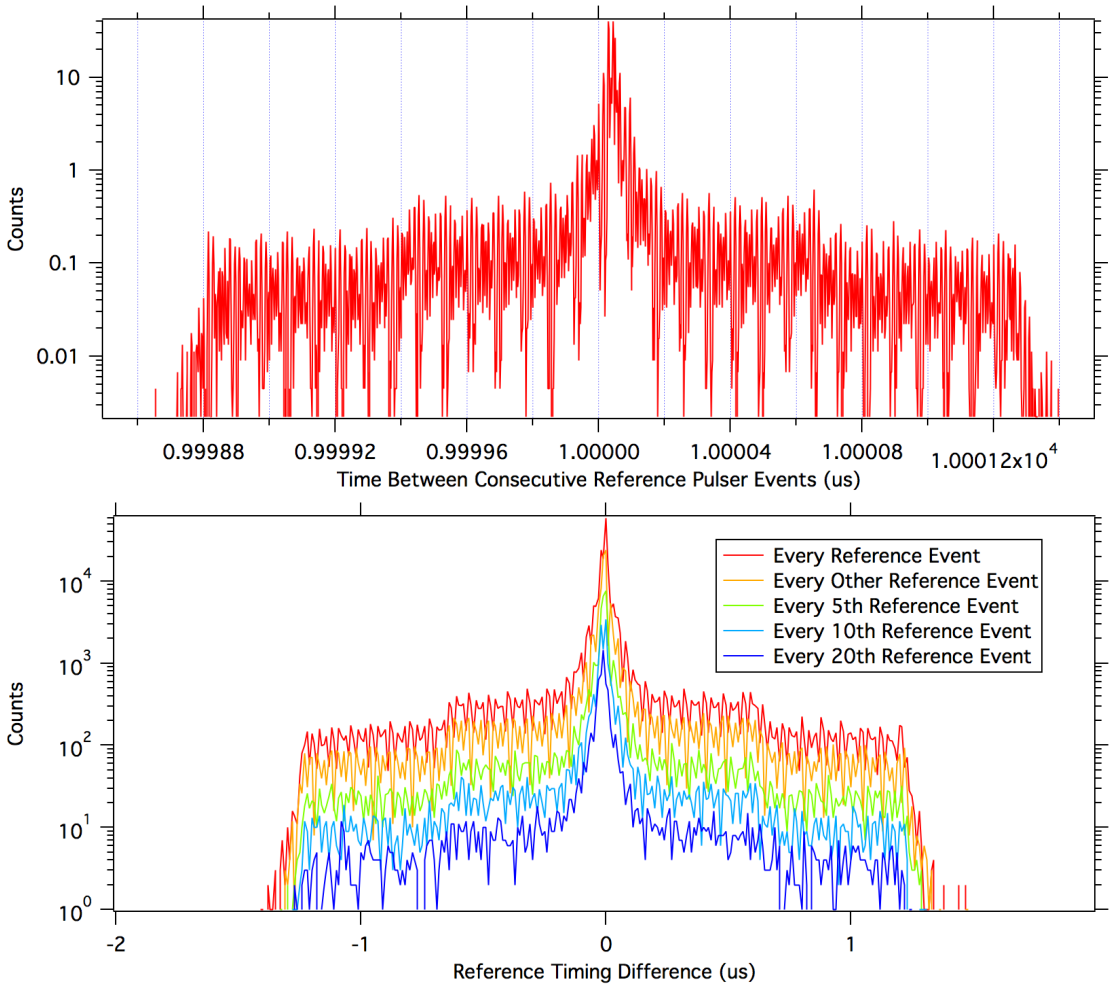
To quantify the systematic correction, two methods are used to estimate the random jitter between the two clocks. The first metric is the the range of the random jitter, and the second is the total distance traveled by the random jitter, i.e. the sum of the absolute value of the difference between the two clocks for each reference pulser event. These two values are calculated for a sampling of 94 data files pulled from three data series. A histogram of the range and total distance traveled can be seen in Figure 2.11. As is expected, the total distance traveled is much larger than the range of the random walk. The systematic correction can be quantified in a simple way by calculating the fractional change in the neutron β-decay lifetime, σ_τ/τ based on a lifetime of $\tau = 880$ s. The range is expected to be more characteristic of the actual size of the systematic effect. The total distance travel is included in an abundance of caution as it is the worst that the random jitter could be if it was modeled as a random walk. For the largest range value in the files analyzed, the correction is $\sigma_\tau/\tau = 7 \times 10^{-7}$. For the largest total distance traveled, the fractional correction comes out to $\sigma_\tau/\tau = 3 \times 10^{-5}$. Therefore, this systematic effect is well below the current statistical uncertainty of this experiment.

This investigation has been calculating the difference between the DAQ clock and the ref-

erence pulser timing. This systematic error calculation is based on the assumption that the reference pulser has the correct timing and the DAQ clock is drifting. Any fraction of the drift between these two clocks that is caused by the reference pulser will not introduce a systematic correction. In fact, it seems more likely that it is the reference pulser that is drifting in time. Therefore this is an extremely conservative estimate.

In the future, if the statistical uncertainty of this measurement is pushed into the realm where this is a potential systematic effect, it will motivate a characterization of the DAQ clock

Figure 2.9: A histogram of the duration between reference pulser events. (Top) The duration between consecutive reference pulser events showing a complicated structure with an approximate width of 3 μs and with a center of the distribution that is slightly offset from the expected value of 10 ms. (Bottom) The duration between reference pulser events after a removing the 10 ms delay and the linear drift. No deviation between the center of the distribution and zero is observed, which suggests that the drift in the DAQ clock has been estimated accurately.



using a calibrated clock. In an experiment like this that is completely reliant on the accuracy of a clock, it is essential that the clock is characterized before operation.

2.13.4 Detection System Timing Resolution

Two other parameters are important for determining the timing distribution of the signal from neutron β-decay like events. They are the path length of the photon trajectories and the timing resolution of the detection system.

Figure 2.10: A graph of the drift between the DAQ clock and the reference pulser.

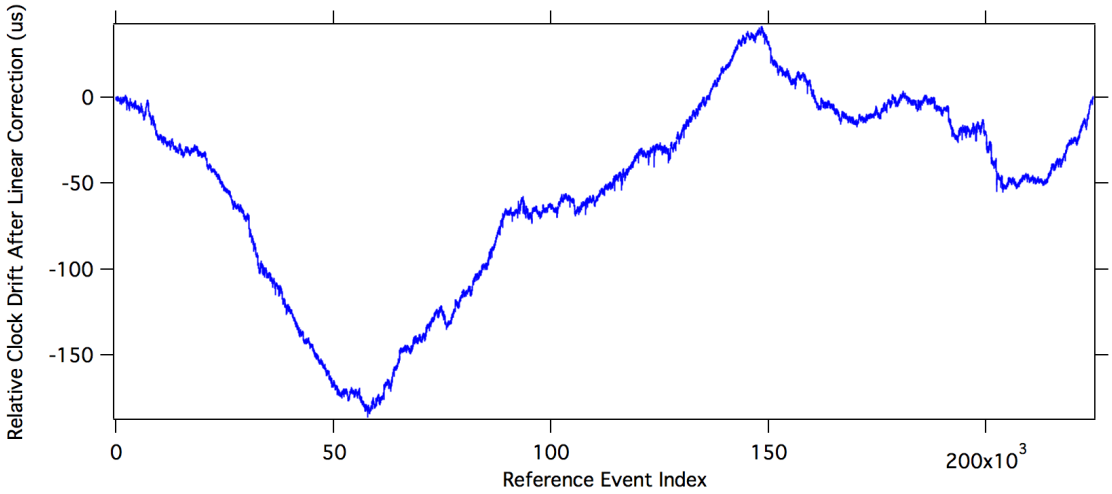
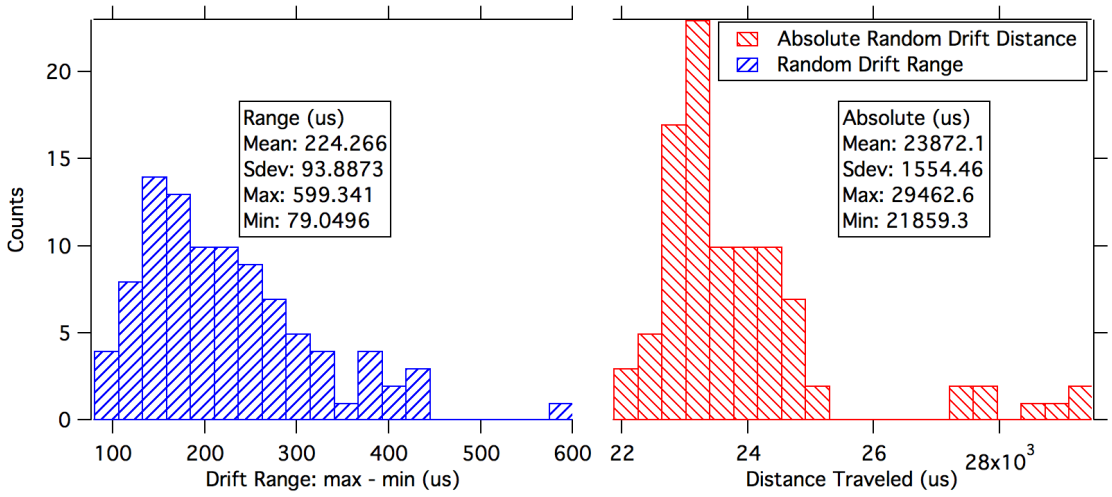
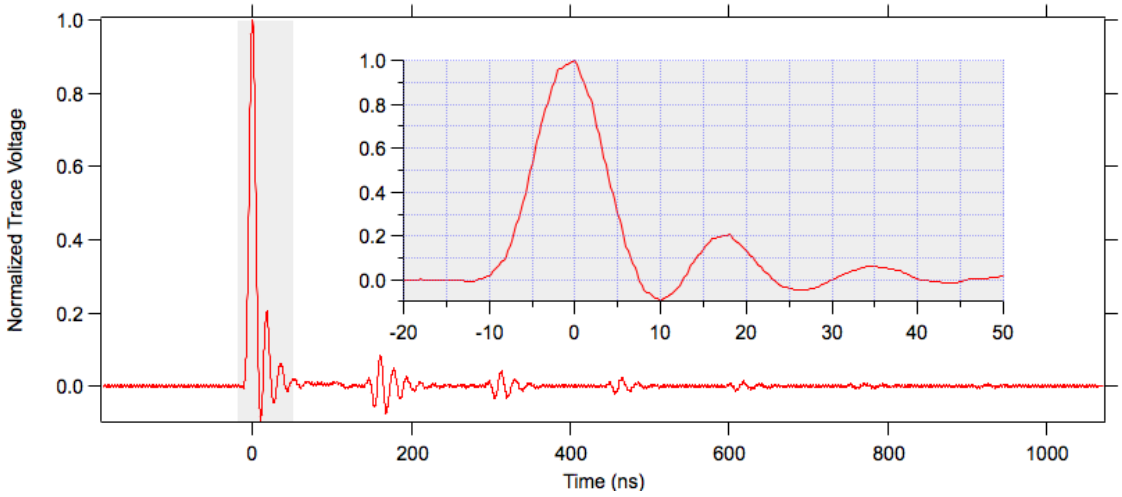


Figure 2.11: A histogram of the range and integrated duration drifted between the DAQ clock and the reference pulser. This is the drift between the clocks during an entire data file, and therefore is over a duration of ≈ 45 min.



The timing resolution of the detection system can be estimated from the shape of a single PE event. To estimate the shape of single PE events, all of the single PE events in a file are averaged to form an average single PE template. The risetime and falltime of the average single PE template is calculated between 10% and 90% to be 6 ns and 5 ns respectively. This risetime can be compared to the quoted risetime of the Burle 8854 PMTs of 2.9 ns at 3000 V. The PMTs in the experiment were operated at a lower voltage of 2400 V, which could affect the risetime and falltime. In addition, the single PE events include broadening from the timing resolution of the DAQ cards, which quote a typical risetime of 1 ns. This risetime appears to be in satisfactory agreement with the data after accounting for these effects and inaccuracies in the risetime measurement due to ringing and overshoot in the trace, which are obvious in the template see Figure 2.12.

Figure 2.12: Average pulse shape for a normalized single photoelectron event. The trace shows substantial ringing, and overshoot in the single PE peak.



Finally, the variation in path lengths of the trajectories of both the EUV and visible photons will introduce an additional spread in the arrival time on the photocathode. Although the process of simulating the photon arrival times is a relatively straightforward one, it is very labor intensive to do accurately. Instead, consider that it takes a photon about 0.5 ns to traverse the width of the cell. Due to the rough structure of the TPB coated ePTFE on the cell wall, photons are expected to scatter diffusely. Therefore, the photons in the cell will experience a random walk where some are scattered toward and some away from the light guide on any given bounce. Some small fraction of the photons will enter the light guide after just a few bounces, but the majority are expected to take many bounces. This is limited by the probability of absorption

when a photon strikes the cell wall. The absorption per bounce has been estimated to be 10%. Due to the nature of the Lambertian wall model, most of the light scattering from the surface will traverse the width of the cell, whereas a purely diffuse model would have a larger fraction of scattered photons that skim the cell wall. This is expected to further increase the spread in the timing distribution as compared to the purely diffuse model. Because of the 10% absorption probability, many of the detected photons will have traversed a sufficient distance that the dispersion due to the varying path lengths will contribute at or above the level of the TPB fluorescence lifetime and the singlet state lifetime. As a result, this mechanism is expected to be a non-negligible contribution to the timing distribution of photons striking the photocathode.

The photon timing spectrum can be estimated by inspecting the pulse shape of average traces. To limit systematic effects in the calculation, only a select set of events were included in the average trace. By selecting relatively large events, it is easier to pinpoint the peak of the trace. This is because a sufficiently large number of photons are included to prevent the single PE fluctuations, due to counting statistics, from dominating the pulse shape. This is important because it allows a more accurate determination of the maximum pulse height and therefore also the thresholds of 10% and 90% of the maximum. Events that are too large will be clipped by the linear fan because the voltage extends beyond the range where the linear fan's response is linear. Because these events do not have a linear response and the timing parameters require an accurate calculation of the peak height they can not be used to determine the timing distribution of photons without introducing systematic effects. Therefore, the photon timing distribution was estimated by calculated the risetime and falltime of the average trace of events with pulse heights between 20×10^3 and 25×10^3. In the rest of this paragraph, these events will be referred to as medium events. In addition, the medium events were classified by their pulse shape using the kurtosis metric. Fast events are those with a pulse kurtosis $\geq 70^*$; the remaining medium events are termed slow, medium events. Slow, medium events have a risetime of 9.4 ns and falltime of 49.7 ns, where once again these values are calculated from 10% to 90% of the maximal value. Figure 2.13 shows the average trace of slow, medium events used in this calculation in blue. An average 1 PE pulse and a fast, medium event, i.e. with low kurtosis, are shown for reference. The similarity in shape between the 1PE template and the fast, medium events suggests that the fast, medium events, like 1PE events, are dominated by the timing response of the detection system, and not from the physics of the light production mechanisms or the light transport.

In conclusion, this suggests that the timing response of the combined light detection system is less than about 5 ns. This is clear because the fastest events show a timing response on this order, which contains the contribution of the light detection system response and the

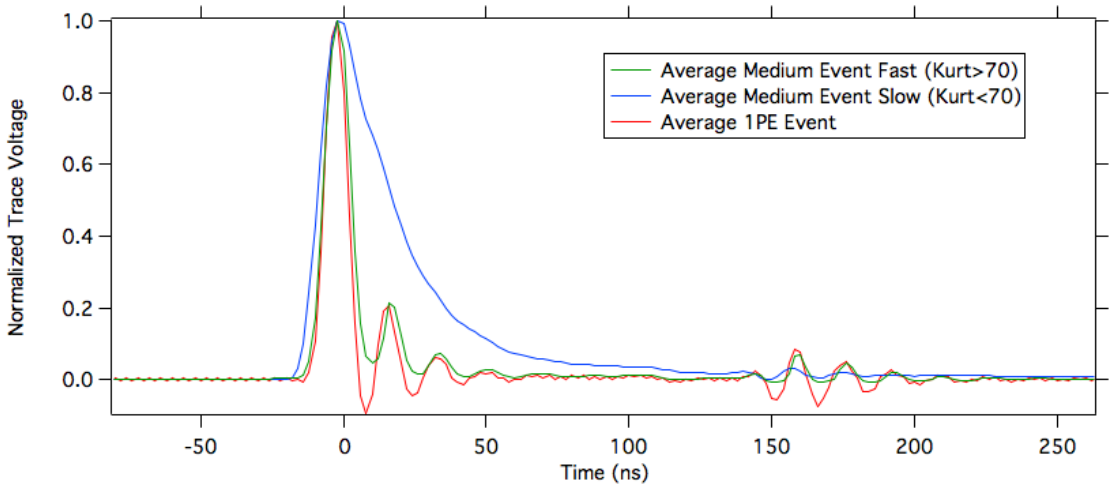
*This kurtosis metric was calculated over the entire voltage trace and therefore differs from the kurtosis calculated in the analysis chapter, which only calculates the kurtosis in the region of interest.

physical mechanisms involved in the light creation of those events. In comparison, the slow, medium events have a much longer characteristic falltime. The falltime of the slow, medium events is expected to be a combination of the LHe singlet decay lifetimes, the TPB fluorescence lifetime, the light detection system response, and the spread in time of flights of the photons from the original scintillation until striking the photocathode. The TPB fluorescence lifetime is quite short and can be neglected. An upper limit of 10 ns has been put on the singlet decay lifetime[29, 30]. Taking these two lifetimes into account along with the upper limit of 5 ns for the characteristic lifetime of the detection system, it is clear that the light transport contributes a sizeable fraction of the characteristic timing width of the slow, medium events. Unfortunately, this prevents this apparatus from putting a more stringent limit on the lifetime of the singlet state in LHe.

2.14 Future Work and Suggested Upgrades

During the process of running the experiment, finishing the analysis, and estimating the systematic errors, a variety of possible upgrades came to light. Methodology or analysis changes that were thought of while the experiment was running were implemented as quickly as possible. However, some of the ideas required substantial upgrades to the apparatus which prevented immediate implementation. During the analysis and systematics estimates, a few additional

Figure 2.13: The average of traces for medium sized, slow events and medium sized, fast events separately. The average trace of single photoelectron events is included for comparison. Medium events, in this case, are events with pulse height in channel 1 between 20×10^3 and 25×10^3. The kurtosis mentioned here was calculated on the entire voltage trace, not just the region of interest as is done throughout the rest of this work.



types of data were thought of to either improve our understanding of the system or to constrain some of the systematic effects. In practice, many of these upgrades could be implemented immediately, if additional data was taken, with little additional work.

2.14.1 Active Veto System

The active veto system could use an independent method of monitoring the gains in the active veto PMTs and for verifying that the system is working properly. It is very challenging, using the data that has already been collected, to determine how effectively the active veto system is running. To combat this, the voltage pulse sent from the signal generator to the reference pulser could be split and sent to LEDs in each of the active veto paddles. The timing of the signals to the different paddles should be adjusted so that they come in after the main event. This will prevent the event from being tagged as a muon. Each peak should have its own offset so that the gains of the individual active veto PMTs can be monitored independently. For example, if the main detection trigger is set to occur at 320 ns after the start of the trace window with a region of interest of 270 ns to 470 ns, the first muon paddle reference peak could be placed at 490 ns, the second at 530 ns, the third at 570 ns, etc. This will work best if the response in the active veto PMTs from the LED recovers more quickly than 40 ns, which must be verified. By including the LED signal in the reference events in this way no additional deadtime is introduced. These events are already tagged by the reference PMT so pulling them out of the data is simple. It would make adjusting the active compensation coils for each running configuration much simpler and would make it easier to match the gains of the different active veto PMTs. It also allows the gain of each of the active veto PMTs to be assessed independently despite the channels being added together.

2.14.2 Livetime Hardware Improvements

With the DAQ cards used in this experiment, it has been shown that it takes a varying amount of time for the system to rearm after writing events from the cards internal memory to disk. In future experiments, the DAQ and trigger systems should be set up so that the first event of the spill is taken as soon as the card is armed. This event, which would be ignored in the analysis, tells exactly when the DAQ card was rearmed. This allows the livetime to be calculated more accurately. Ultimately, this experiment attained an accuracy that closely approached this using a software veto. However, it did so at the expense of throwing out more events. In the case where the rearm duration of the cards on future experiments do not fluctuate, this will not provide any real advantage, however in this experiment, it would have guaranteed that a very small fraction of the events, 1:10000, was thrown out, but the software veto would no longer be necessary.

2.14.3 Upgrading the Detection Scheme

Another upgrade, which has been discussed previously in the theses, is to modify the detection scheme. In particular, moving the detectors inside the cryostat to 77 K, for example, would substantially improve the performance characteristics of the apparatus. First of all, removing the 300 K light guide would decrease the heat load from thermal conduction down the light guide. This would result in a lower operating temperature and a reduced systematic effect from thermal upscattering of the UCN. Additionally, this would reduce the volume of acrylic and thereby a fraction of the background associated with Cherenkov radiation in the apparatus. Finally, by moving the detectors closer to the cell, the light collection efficiency should be improved, which will increase the statistical sensitivity of the pulse shape parameters.

In addition to these general upgrades by moving the detectors inside the cryostat, changing the type of detector could also result in substantial improvements to the data quality. A discussion of the current detection types that could be used is beyond the scope of this work, but with current state-of-the-art detectors, substantial improvements to the data quality could be expected.

2.14.4 Upgrading to a Non-Metallic Cell

Another improvement that has been discussed in previous theses is to switch from the stainless steel cell to a non-metallic cell. The primary goal is to reduce the heat load due to eddy current heating, which from the temperature profile observed in the data, is the largest contribution to time-dependent temperature drifts during the data acquisition phase. This will both reduce the base operating temperature and reduce time-dependent drifts in the temperature. The result will be a simpler calculation of the systematic effect from superthermal upscattering and a smaller systematic correction. This does add some complications related to maintaining the structural integrity of the cell and added difficulty in ensuring good thermal contact between the thermometry and the LHe.

2.14.5 Additional Systematics and Performance Data to Take

The following paragraphs present suggestions for additional types of data to be taken in the experiment to improve the performance estimate of the apparatus.

Specific PMT blinding data should be taken to evaluate if the quantum efficiency of the photocathode of the PMTs is affected by exposing the photocathodes to intense light while the PMT is in a low power state. The magnets should be powered up for a long time before this data is taken to prevent any time-dependent magnetic field contributions. Then the PMTs are ramped down, the beam is opened, a delay occurs, the beam is closed, and the PMTs are ramped back up. This is compared to almost identical data where the PMTs are ramped down

at the same rate and for the same duration and then are ramped back up without opening the neutron beam. This will allow a direct measurement of the time constant and amplitude of the gain drift due to photocathode blinding. It should be done at two different magnetic field strengths to limit the extent to which the effect depends on the magnetic field state. Similar data could also be taken with the magnetic fields off, but the gains in the two main detection PMTs are drastically reduced in this state, which may reduce the accuracy of the results.

Data should be taken to separate the rate-dependent gain drifts from the temperature effects. A radioactive source could be used to change the data rate while the PMTs are operated at full power. Placing a source near the acrylic splitter will allow the detection Cherenkov radiation from Compton scattering in the acrylic. By characterizing the rate-dependent gain drifts, it will be possible to determine if the gain drifts after the PMTs are ramped up are due to an increase in the data rate or if they are truly due to temperature effects. It is strongly expected that the gain drifts are due to the temperature effects because they are only seen in the main detection PMT that is not using resistors that are insensitive to temperature changes. This would provide an independent method of quantifying the rate-dependent gain drift.

Specific quadrupole ramping, gain drift data should be taken. In this case, the solenoid field should be powered up for a long time to prevent gain drifts due to the solenoid ramp from contributing to the data. The PMTs should not be ramped down, to prevent temperature effects, and the neutron beam should be left closed, to prevent photocathode blinding. The DAQ should collect data while the quadrupole magnet is ramped to operational strength, and then the reference pulser should continue to be monitored for 15 min to 60 min after the quadrupole reaches its operational current. This should be done multiple times to verify the repeatability. It will give independent data for isolating and quantifying the gain drift that results from ramping the quadrupole field.

Additional data should be taken to combine the gain drifts due to each of the magnets while still excluding gain drifts due to temperature effects and PMT blinding. This should be done by modifying the production data scripts so that the beam is not opened and the PMTs are not ramped down. Data collection should begin at the start of the file instead of after the end of ramping. This will allow the gain to be assessed during the filling stage and observation stage, which will be a stronger constraint on the quadrupole and solenoid gain drifts. This data should be able to demonstrate if the combination of these two effects in the data is just the sum of the gain drifts from the quadrupole and solenoid magnet ramps. It also lets these two gain drifts to be assessed before including the PMT blinding and the temperature effects in the gain drift, which will hopefully give additional clues about how the gain drifts combine in the data files.

At least four data files with a much longer data acquisition phase should be taken to put stronger constraints on the constant offset in the exponential fit to the background subtracted timestamp histograms. This will help determine if the constant offset can be set to zero. This

data will also constrain whether slow, time-dependent effects could be affecting the data through the background subtraction. Two pairs of data files should be taken including a non-trapping file followed by a trapping file.

Additional data should also be taken to measure the effect of neutron-induced luminescence of the BN shielding. This background has been verified to be pinned by magnetic fields, and due to the complicated ramping profile in our data, taking data to understand this background source inside our apparatus is motivated. For more information about the neutron-induced luminescence see Section 5.7.1, which starts on page 157. This background source can be measured by taking warm data or ^{3}He data to reduce the population of UCN inside the experiment without changing the neutron-induced luminescence from the cold beam. Additionally, tests could be run using the same experimental scripts as the final analysis, but taking data before the cell is filled with isotopically pure ^{4}He, which will prevent the down conversion of the cold neutrons into UCN.

A set of measurements should be performed to verify the calibration of the thermometers used in this experiment before additional data is taken. The size of the systematic effect due to thermal upscattering and its uncertainty could both be reduced by taking additional calibration data. In particular, a secondary calibration that could be performed near the operating temperature range of the trap in the experiment before running would result in a substantial improvement in my confidence of the absolute temperature measurements. This step was overlooked during the commissioning of the apparatus for the data described in this work.

An additional set of measurements should be performed to push both the temperature dependence of the TPB quantum efficiency measurements and LHe scintillation yields to the temperature range of this experiment. These values, which are temperature dependent, have to be extrapolated outside of the temperature ranges of the published literature in order to estimate the light production and light collection efficiency in our experiment. Additional measurements, either in our apparatus or in a separate measurement cryostat, should be taken to extend these measurements to our temperature range, which would improve our confidence in these parameters.

Chapter 3

Analysis

The analysis takes the digitized voltage data from the experiment and extracts a trap lifetime. It is broken into three different tiers of analysis, for ease of use, which was done to allow smaller portions of the code to be rerun when changes were made to the higher level code. First, the low-level analysis condenses the information stored in the voltage traces into pulse shape parameters. The pulse shape parameters are designed to extract as much useful information from the voltage traces as possible and to store that data in an efficient format. The mid-level analysis applies corrections for effects like gain drifts, firmware errors, and the deadtime. It also applies cuts to reduce the background rates in the data. The upper-level analysis takes the timing distribution of events that passed the cuts, combines it with data of the same type, performs the livetime correction and background subtraction, and extracts the trap lifetime. Additional analysis has been performed to estimate systematic effects. However, this work was done separately and will be discussed in its own chapter.

A guiding design principle when developing the analysis was to use methods that were straightforward and, consequently, easily understood. The primary motivation for this was to make the systematic effects simpler to understand and estimate.

The following sections describe each of the parts of the analysis. They include the methodology used, motivation, and a visual representation of the data as applicable. This is meant to be sufficient information to evaluate the effectiveness of the analysis code and to estimate the systematic effects, which are discussed in Chapter 5, which starts on page 132.

3.1 Low-Level Analysis

The low-level analysis operates on the raw binary files. It converts the binary data into the voltage traces and timestamps. It then distills the voltage data into a more compact form by calculating pulse shape parameters to extract information on the size, shape, and timing of

the events. These pulse shape parameters, along with the timestamps, are written to an Igor text format. A representative binary data file is 4.7 GB and is compressed to 282 MB after the low-level analysis.

The structure of the binary data is a result of the DAQ system. As mentioned previously, when the DAQ cards receive a trigger they record an event, which consists of a voltage trace for each of the four channels and a corresponding timestamp. The first two voltage traces are the main detection PMTs, the third is the sum of the active veto PMTs, and the final channel is used for the reference PMT. Each voltage trace consists of 704× 32-bit word voltage records written in little-endian ordering with a record rate of 500 MHz, which corresponds to a 2 ns step size. The timestamps are single, 64-bit, IEEE floating point numbers, which are also written in little-endian ordering. This data is stored in local memory on the DAQ cards until 10000 events are recorded. After the 10000th data point, the stored voltage traces and timestamps are written to disk, and the DAQ card is rearmed.

The binary files share this structure. A brief header of 5000 Bytes contains running information. Then the data starts with the voltage trace for event 1, channel 1. This is followed by the voltage traces for event 1 in channels 2, 3, and 4. Event 2 follows event 1. This pattern continues until the 10000th event. After the last voltage trace of the 10000th event, the corresponding 10000 timestamps are written to disk in 2 formats. Once the data is written to disk and the card finishes rearming, the DAQ system starts collecting data for the next spill. When the stop signal is sent to the DAQ system, it finishes the current spill, writes it to disk, adds a brief footer, and closes the file. The low-level analysis is designed to read in this data format and calculate useful pulse shape parameters from the individual traces.

In the analysis, the three dominant pulse shape parameters that are used are the pulse area, pulse kurtosis, and pulse height. The pulse area is most strongly correlated with the energy of the incident event. This is, therefore, a useful parameter for limiting low-energy backgrounds and excluding events that are too energetic to be neutron β-decays. The timing distribution of an event, i.e. the width of the pulse, is strongly dependent on the physical mechanism that produces the light that is detected from the event. The risetime and falltime are typical pulse shape parameters for extracting the timing distribution. Alternatively, the pulse height could also be used, once the pulse area is taken into account. In this analysis, an alternative pulse shape metric, the pulse kurtosis, is used instead. To the best of our knowledge, this is the first time that the pulse kurtosis has been used as a pulse shape metric for estimating the timing distribution of events in a low-energy particle physics experiment. The pulse height remains an important parameter in the analysis because components of the detection electronics, like the pulse height discriminator, interact with the height of the peak in the voltage trace. Therefore, the pulse height can be used to study how these components affect the data. Using these parameters, a substantial amount of information can be extracted from the data and used to

tag and remove background events of different types.

Before the pulse area, pulse timing, and pulse height pulse shape parameters can be calculated, some supplemental calculations are required. Most of these additional parameters are also saved during the low-level analysis for diagnostic purposes. The most obvious of these secondary pulse shape parameters include the baseline of the voltage trace and the location of the peak inside the voltage trace window. The following sections discuss the voltage data, how it is prepared for pulse shape analysis, and finally, the pulse shape calculations.

A continuing trend in the analysis, of which the region of interest is an example, was to use pulse shape algorithms that treat all events the same way. This was done to make the analysis more transparent and to simplify the calculation of systematic effects. It results in a reduction in the resolution of the pulse shape parameters that is caused by systematic differences in the pulse shape parameters of events that are very different. However, due to the simplicity of the algorithms, when and how strongly these effects occur can be more easily calculated.

Another goal of the analysis, which is particularly important for the low-level analysis, was the computational efficiency. Just reading the binary files into Igor was a very time-consuming process and with $\approx 3.2 \times 10^5$ voltage traces in a typical data file even simple calculations on the voltage traces can very quickly bog down the analysis code. As mentioned previously, this was the primary motivation for separating the different tiers of the analysis, but it was also an important design constraint in calculating the pulse shape parameters. Frequently throughout the low-level analysis, the pulse shape parameters were chosen because they could be calculated much more quickly than more general and accurate approaches. An illustrative example of this was an attempt to write the code so that the voltage traces were fit to a combination of single PE pulses. By recording the timing of the individual PE peaks, the number of photons could be determined extremely accurately. It is probably the most accurate determination of the incident energy and timing distribution of events with this detection method. Although the storage of the timing distribution of single PE pulses was not as efficient as the pulse shape parameters, the main issue was the computational difficulty of fitting an unknown number of PE to a voltage trace. This process was inherently quite slow even for tiny events. If the PE fitting was performed on just the small events, the computation times for a data file could already become orders of magnitude longer than the data collection process. Performing diagnostics and exploring the data would take so long with this method that it was discarded, not due to lack of merit, instead due to logistical costs. Most of the design choices that were based on using faster methods had much smaller effects, but in the case where two methods were in good agreement, the faster method was frequently chosen.

3.1.1 Raw Traces and the Baseline Calculation

The voltage traces consist of negative pulses on a large positive offset, for an example of a raw voltage trace see Figure 3.2. The large positive offset is referred to as the baseline. To calculate the pulse area, the baseline must be subtracted. After that, the trace is flipped, although this step is purely cosmetic.

An accurate baseline calculation is important because an offset in the baseline results in a systematic uncertainty in the pulse area and pulse height. The data is insensitive to an offset in the entire pulse area spectrum. However, it is sensitive to a time dependence in the pulse area metric. It is therefore expected to be more sensitive to a time dependence in the baseline calculation. It is hard to imagine how the baseline calculation could couple with a time-dependent process that would allow it to introduce time dependence in the pulse area metric. If this does occur at some level, it will introduce a systematic effect in the trap lifetime.

The baseline calculation method that was used scans through the trace with a small window that looks for regions of the trace that are characteristic of the baseline. This is done by stepping a 50 ns window through the trace with a 10 ns step size. In each window, the standard deviation is calculated. The standard deviation is expected to increase if a pulse overlaps with the window. Therefore, windows with small standard deviations tend to be more characteristic of the baseline. The region of the trace with the smallest standard deviation is selected, and the mean of that region is used as the baseline value.

This method has been found to be very robust. However, a few failure modes have been discovered. The baseline calculation is sensitive to features in the trace that have a small standard deviation that is roughly the size of the window. As an example, for very large events that extend beyond the voltage range where the linear fan is linear, the tops of the events are underestimated. In these regions, the fluctuations in the voltage trace are also underestimated resulting in a reduced standard deviation. Despite this, the standard deviation in these regions is still much larger than in regions with no pulses.

Another type of baseline calculation failure occurs when events are so large that substantial afterpulsing occurs for multiple seconds, after the initial event. This afterpulsing can repeatedly trigger the DAQ system and result in a voltage trace that is almost full of uniformly distributed single PE events. If the singles rate is high enough, it is possible for there to be no 50 ns portions of the trace that are characteristic of the baseline. For these events, this baseline method is likely to fail. However, these are not desired events, and they are quite rare.

There is an additional set of rare events where the voltage reads an identical value for many time bins in a row at a random location in the voltage trace and therefore a random height. This flat portion of the trace is thought to be an electronics error. It can results in a region that has a very small standard deviation, and this baseline calculation method is likely to return

an unrepresentative value on these events. Once again, this type of failure is extremely rare, at most a few events per file.

The final error mode of the baseline calculation is due to the detection electronics. The PMTs and bases were selected for fast timing, which typically results in more ringing in the anode current. This is particularly evident in events with a very quick timing distribution as can be seen in the red and green traces in Figure 2.13. In a small subset of the data, the baseline is sufficiently large that the ringing of the pulses passes beyond the digitization range of the DAQ cards, when this happens the maximum of the range is stored. This can result in a very small standard deviation for a limited width of the pulse, which is limited by the frequency of the ringing to be less than about 10 ns. This effect was seen in the first half of s15 and was due to a broken input of the linear fan in the detection electronics. For this data set, failure in the baseline calculation resulted in a substantial error in the pulse area metric. As a result, this data was not used in the final analysis.

The baseline calculation is insensitive to the size of the baseline calculation window. A width of 50 ns was chosen as a compromise between a small window, which has a larger uncertainty in the mean, and a larger window, which is more likely to be contaminated by portions of the voltage trace that are not consistent with background fluctuations.

3.1.2 Region of Interest

To maximize the signal to noise of the pulse area and the pulse kurtosis metrics, they are calculated over a limited range of the trace about the location of the peak. The region of interest is defined to be the range of the voltage trace in which the pulse shape metrics are calculated. Both the size and location of the region of interest are tailored to the characteristics of the signals in each of the channels independently. The following paragraphs describe aspects of the detection system that drove the selection of the region of interest, and then the corresponding region of interest is described for each channel.

Before we begin, it will be useful to discuss the hardware trigger logic briefly to give some context. The first of the main detection channels to trigger its discriminator opens a gate to look for coincidence between the main detection PMTs. If a pulse triggers the other channel while the gate is still active, a coincidence signal is sent to trigger the DAQ cards to record the event. Therefore, the DAQ cards are instructed to record the event when the second of the main detection channels triggers its discriminator. This results in the second of the two pulses coming in at a fixed time relative to the start of the voltage window. However, for a given event, either of the PMTs could have triggered first. Therefore, there is some jitter in the location of the two peaks depending on which came first and the timing correlation between them. To allow the analysis to account for the variable location of the peaks, the location of the

region of interest in the two main detection channels is dynamic. It is determined by finding the peak of a smoothed version of the trace. The smoothing is to limit the fluctuations caused by single PE pulses in small events. The smoothing is done with a five pass binomial smoothing algorithm that is built into Igor. The peak location is selected as the location of the maximum of the smoothed trace in the window from 200 ns to 360 ns after the start of the trace window. The region of interest starts 50 ns before the peak location and extends 150 ns after the peak location. An example of the region of interest for traces of three different sizes can be seen in Figure 3.1.

If the region of interest is too large for an event of a given size, there will be additional uncertainty in the pulse shape metrics due to the inclusion of zero value regions in the region of interest. This results in a decrease in the resolution of the pulse shape metrics. On the other hand, if the region of interest is too small for an event, the pulse will extend outside of the region of interest resulting in a systematic error in the pulse shape metrics, e.g. the pulse area would be underestimated. The size of the region of interest was chosen as a compromise between these two effects.

The signal in the reference PMT is very large for reference events resulting in a clear separation of the pulse area in the reference PMT between reference events and non-reference events. Therefore, a large static window is sufficient to tag these events. The region of interest was chosen to extend from 260 ns to 560 ns from the start of the trace window.

The active veto system has a much lower gain than the reference pulser system. Also, the signal in the active veto channel is the sum from six different PMTs, which results in much larger background fluctuations when compared to the signal from a single PMT. The combination of these two effects can result in very small pulses that are largely obscured by the background noise. It was discovered that there was a large number of very small events with very high timing correlation with the main detection system. To capture these events and to maximize the signal to background ratio, a very small region of interest was selected in the location of these peaks. The region of interest in the active veto channel extends from 320 ns to 340 ns after the start of the trace window. The small size of the static region of interest is discussed in more detail elsewhere, see Section 2.12.3, which starts on page 55.

3.1.3 Pulse Area

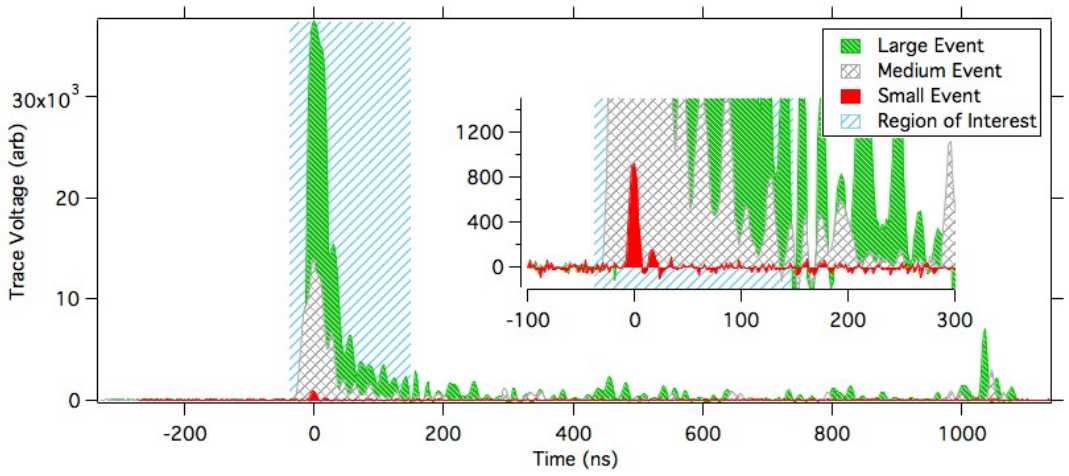
Fitting the voltage trace to a set of single PE peaks would result in the most accurate determination of the incident particle energy that is possible with our data; this topic has already been discussed in detail, see Section 2.5.3, which starts on page 41. Here, a brief refresher will be provided before discussing the details of the pulse area calculation in the analysis.

Charged particles that ionize the LHe as they pass through the cell result in scintillation

light. The energy deposited per unit path length is approximately constant. Therefore, the amount of scintillation produced is proportional to the energy deposited in the helium. If the particle is stopped in the helium, this is related to the total energy of the particle when it entered the helium. If the particle traverses the cell, it is an indication of the particle's path length in the cell. In this case, the amount of scintillation produced is less than what would be expected from the energy of the incident particle, which will smear the features in the pulse area histogram. The scintillation is quantized by the fact that an integer number of photons are produced by the incident particle. Some fraction of these primary photons will ultimately result in photons striking the photocathode of our main detection PMTs and producing a photoelectron.

A PE is the response of the PMT to a single photon striking the photocathode and emitting a single electron. Since the number of PEs is an indication of the energy of the incident particle, a single PE is the signal corresponding to the smallest detectable event in our detection method. However, the size of the signal in the anode current for each PE will vary. Due to the variation in the anode current from the individual PE, the pulse area metric, which integrates the anode current, is unable to distinguish a smaller number of large PEs from a larger number of small PEs. This can introduce substantial uncertainty in determining the number of PEs that are present in an event and hence in determining the initial particle energy. A large contribution to this variation is due to counting statistics of the number of secondary electrons emitted on each dynode of the PMT during the amplification process. These variations for each PE are sufficient that, although the number of primary electrons is a discrete function, the pulse area

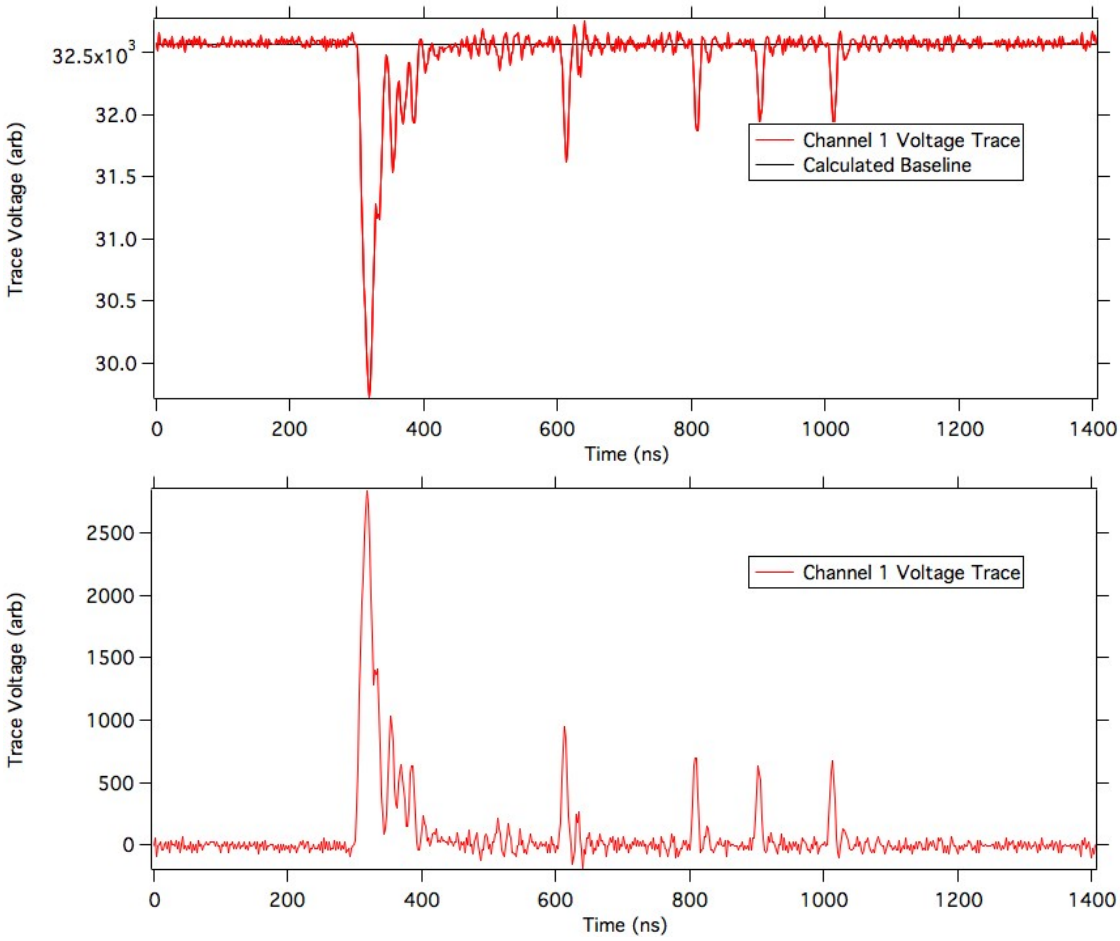
Figure 3.1: Graphical representation of the region of interest, which shows three traces that span the spectrum of peak sizes. The traces have been aligned such that the regions of interest match, which is shown in blue. The larger events can be seen extending outside of the region of interest. The inset image is zoomed in on the small trace to show additional detail.



histogram is continuous.

Despite this, at low energies, the contribution from each of the individual PE can be seen in the pulse area histogram. The location of these low-energy PE peaks can be used to estimate the conversion between the pulse area and the number of PEs. Using this conversion, an approximate number of PE in each event can be estimated. This is only a rough estimate of the number of PEs because more accurate information about the number of PEs in the event was thrown out during the integration process. The hardware pulse height discriminator affects the apparent location of the low PE peaks in the pulse area histogram and, therefore, adds a systematic error in the pulse area to PE conversion. However, if the shape of the pulse height discriminator can

Figure 3.2: Image of a raw voltage trace before and after the baseline subtraction. (Top) An image of a raw voltage trace (red) with the calculated baseline value (black). (Bottom) The same trace is shown after subtracting the calculated baseline and inverting the trace.



be determined, this effect can be accounted for*.

As mentioned previously, the pulse area is the sum of the voltage trace in the region of interest. Throughout this work and the analysis code, the pulse area in main detection channel 1(2) will be referred to as $pulse1\_area(pulse2\_area)$, the active veto pulse area as $muon\_pulse\_area$, and the reference PMT pulse area as ref_pulse_area.

3.1.4 Pulse Height

The pulse height is another straight forward and powerful pulse shape parameter. It can be used along with the pulse area to estimate the timing distribution of the events. It was found that the pulse kurtosis, which is described later, is a very clean metric for separating events by their timing distribution. As a result, the pulse height was not used as a method of classifying different event types in the main analysis. Instead, it was used to help understand the response of the detection system. Some of the events in our data are sufficiently large that they extend beyond the range where the response of our detection electronics is linear. This results in an underestimation of the pulse height and pulse area metrics. Therefore, an upper pulse height threshold was used to discard events that were sufficiently large that they were non-linear. This will be discussed in more detail in the section on the upper pulse height cuts. The pulse height is also the closest pulse shape parameter to what the hardware discriminator threshold cuts on, which makes the pulse height particularly useful when investigating the effect of the pulse height discriminator.

The pulse height metric was calculated as the mean of the unsmoothed trace in a 8 ns window starting 2 ns before the peak location. It was calculated only for the two main detection channels and was given the designation $pulse1\_height(pulse2\_height)$ for channel 1(2).

3.1.5 Pulse Kurtosis

The kurtosis is the pulse shape parameter that is used to determine the timing distribution of the events. First, this section will describe the rise and falltimes of the pulse, which are typical pulse shape parameters for determining the timing of an event. This is followed by a discussion of the inherent difficulties in calculating the rise and falltimes, which motivates the need for a more effective pulse shape parameter for determining the timing distribution of the events in this experiment. Next, the kurtosis metric is discussed as an alternative. Finally, the correlation between the kurtosis and the falltime of the data is presented with our data to demonstrate that the kurtosis is, in fact, selecting on the timing distribution.

The noisy nature of the voltage traces, which can be seen in Figure 3.2, introduces difficulties in calculating the rise and falltimes in a fast, reliable, simple way that is effective for all of the

*This PE to pulse area calibration is discussed in more detail in Appendix A, which starts on page 227.

events in our data. Typically, the risetime is calculated as the amount of time it takes the trace to rise from 10% to 90% of the difference between the baseline of the trace and its peak; correspondingly, the falltime is the time it takes to fall from 10% to 90%. Therefore, an accurate calculation of the rise and falltimes requires an accurate determination of both the baseline and peak height of the pulse. The neutron β-decay events in our data have a slow timing distribution, which is characterized by a risetime of ≈ 9 ns and a falltime of ≈ 50 ns. Single PE events have a characteristic rise and falltime of ≈ 65 ns. Because of the broadness of the neutron β-decay timing distribution, some of the smaller neutron β-decay events will have little or no overlap between the PEs that compose the pulse, which reduces the correlation between the pulse area and pulse height, but also limits the reliability of the risetime and falltime timing metrics.

Early on in the analysis, these uncertainties in estimating the timing of the pulse using the rise and falltimes motivated a search for an alternative pulse shape metric. This search culminated with the discovery of the kurtosis metric.

The kurtosis was found to be strongly correlated with the pulse timing. It can be calculated with the built-in function, *Wavestats*, in Igor. This function, which is optimized C code, is much faster than anything of comparable complication that can be written in the Igor scripting language. *Wavestats* also calculates a large quantity of other useful parameters including the location and height of the maximum of the trace, the average, the sum, and the standard deviation of the voltage values. The sum of the region of interest was already being calculated for the pulse area metric; by switching this calculation to *Wavestats*, both of these parameters could be calculated on the region of interest with a minor increase in the computation time. Due to the enormous throughput of data generated in this experiment, the speed with which the kurtosis could be calculated was its most influential selling point once its strong correlation with the pulse timing was discovered.

The kurtosis is a measure of the "peakedness" of the voltage trace. It can be written in terms of the fourth moment of a distribution and is calculated according to

$$\text{Kurtosis} = \frac{1}{N}\sum_{i}^{N}\left(\frac{y_i - \bar{y}}{\sigma}\right)^4 - 3,$$

where it is a function of the number of points, N; the individual heights, y_i; the average height, $\bar{y}$; and the standard deviation of the heights, σ.

The correlation between the risetime, falltime, and kurtosis can be used to demonstrate that the kurtosis is selecting on the timing distribution. As mentioned previously, it is difficult to calculate the fall and risetimes due to the noisy nature of the voltage traces. Therefore, the correlation between the risetime, the falltime, and the kurtosis is most easily demonstrated by calculating them on average traces that are composed of a large number of events. To do

this, the pulses were aligned and averaged together and then the risetime, the falltime, and the kurtosis were calculated. This was done on events that are sufficiently large that the peak lies in the center of the digitization range. Requiring relatively large events ensures that they were composed of a large number of PEs, which results in each PE having a small effect on the pulse shape. This results in a much smoother waveform that makes the calculation of the rise and falltimes more reliable. Events that were too large were excluded from this analysis to enforce that the response of DAQ system remains highly linear. For very large pulses, the peak extends beyond the region where the response of the linear fan is linear, this results in an underestimate of the pulse height for large pulses, and it would introduce error into all three of these parameters. Therefore, by selecting events in the middle of the digitization range, the timing pulse shape parameters can be determined more accurately.

The risetime of our events is much faster than the falltime. It is sufficiently fast that the risetime can not be estimated as accurately because the timing resolution of the digitization is too slow to accurately determine the shape of the peak in the regions where the voltage is increasing. Therefore, the kurtosis is instead compared to the falltime, which describes a portion of the trace that varies more slowly and, therefore, can be estimated more accurately. The correlation between the falltime and the kurtosis can be seen in Figure 3.3. The kurtosis is seen to be a smoothly varying monotonically decreasing function of the falltime with a discontinuity at approximately 10 in the kurtosis metric. This is strong supporting evidence that the kurtosis is an effective parameter for determining the shape of events in this experiment.

As mentioned previously, the kurtosis metric, when combined with the pulse area metric, contains the same features as the pulse area, pulse height phase space plots. This is evident in Figure 3.4, which shows scatter plots of the gain corrected pulse area as a function of both the pulse height and the pulse kurtosis. In both graphs, the third pulse shape parameter is indicated by the color. The banded structure in color is evidence of the correlation between the pulse shape parameters. It is clear from the top of Figure 3.4 that, above the reference pulser, which has a gain corrected pulse area of 1×10^6, the kurtosis of the two bands is different. Therefore, it can be used to discard events with a fast timing distribution, i.e. the events in the lower band. Keep in mind that the upper pulse area threshold is set to 115 PE.

The kurtosis pulse shape metric has also proven to be insensitive to gain drifts. Karl Schelhammer performed a series of simulations where he compared the kurtosis calculation for pulses before and after applying a linear scaling to the pulses. He found that the kurtosis calculation is completely invariant under a linear scaling in the voltage trace. This is strong supporting evidence that the kurtosis metric is insensitive to gain drifts of this type. Since gain drifts are such an important issue in our experiment, picking a pulse shape metric that is insensitive to gain drifts has merit.

The kurtosis histograms in the two main detection PMTs were found to be very similar.

However, channel 1 was found to have a smaller range of values. In order to allow a simpler comparison between the two channels, a multiplicative scaling of 18.35/14.3 was applied to the kurtosis in channel 1 to bring the two histograms into approximate agreement. The remainder of this work will use the scaled kurtosis metrics, which are $p1\_kurt = 18.35/14.3 \times pulse1\_kurtlocal$ and $p2\_kurt = pulse2\_kurtlocal$ for channels 1 and 2 respectively, where $pulse1(2)\_kurtlocal$ is the raw kurtosis of the region of interest of channel 1(2).

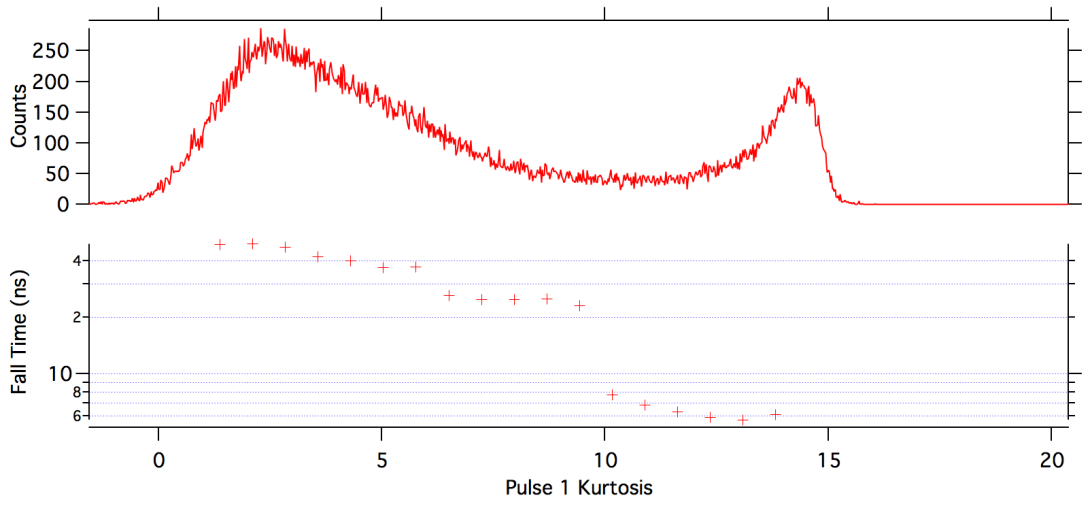
In conclusion, the kurtosis metric has been found to be an efficient and effective metric for characterizing the shape of events in this experiment. It has also been found to be invariant to gain drifts. Finally, a linear scaling is used to make the kurtosis in the two main detection channels more easily comparable*.

3.2 Mid-Level Analysis

The mid-level analysis takes the pulse shape metrics developed in the low-level analysis and, after applying a few corrections, uses them to calculate cut thresholds, apply cuts to remove background events, and to produce timestamp histograms of the remaining neutron-like events. In addition to producing the timestamp histograms, the majority of the diagnostics checks,

*Additional information about the kurtosis that specifically relates to the kurtosis cut threshold can be found in Section 3.2.7, which starts on page 93.

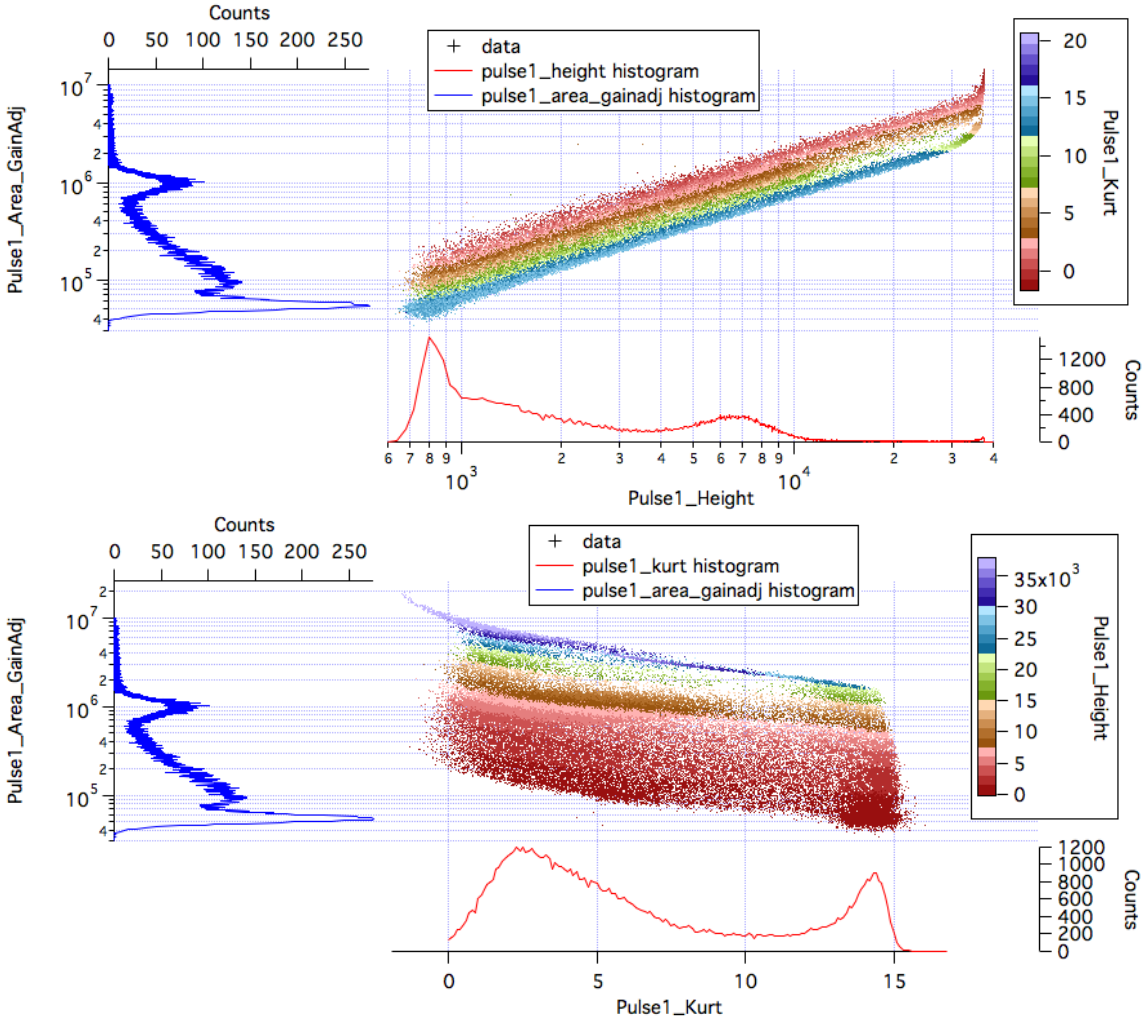
Figure 3.3: Comparison between the kurtosis and the falltime of medium events. (Top) A kurtosis histogram of channel 1. (Bottom) The falltime of the average trace for all medium events as a function of their kurtosis. Medium events in this case are events with pulse height in channel 1 between 20×10^3 and 25×10^3. There is a clear discontinuity in the falltime spectrum around 10 in kurtosis.



systematics investigations, and exploration work was done with the data from the mid-level analysis.

The background cuts include the lower and upper pulse area cuts, an upper pulse height cut, a kurtosis cut, an active veto cut, and a reference pulser cut. The data is corrected to account for the detector livetime, which includes a software veto that throws out events that fall within a fixed duration after a spill to ensure that the time when the detection system goes

Figure 3.4: Phase space plots comparing the kurtosis, area, and height pulse shape metrics. (Top) A figure of the gain corrected pulse area as a function of the pulse height colored by the pulse kurtosis. (Bottom) A corresponding image of the gain corrected pulse area as a function of the pulse kurtosis colored by the pulse height. Histograms are included to give a sense of the distribution of the data in the pulse shape phase space. This data is from channel 1 of file s16r2b1m0.dat.



live is known. Finally, additional events are removed from the analysis because of a firmware issue of the DAQ cards. Each of these topics is discussed in detail in this section.

The primary goal of the mid-level analysis is to reduce the background rate by tagging and removing events based on a set of cut thresholds. This is done by creating a second timestamp array in which the value is replaced with "Not a Number" (NaN). Igor automatically ignores NaN values when calculating histograms, which removes these events from the analysis. The timestamp histograms and the livetime arrays are the primary output of the mid-level analysis.

Naively, one might think that the background subtraction between the trapping and non-trapping data would remove the effect of background events in the experiment and as a result cuts would not be necessary. However, the uncertainty in the background-subtracted, timestamp histograms increases with the data rate of both the trapping and non-trapping data. Therefore, any background events that can be removed before the background subtraction will increase the statistical sensitivity of the experiment, which makes the cuts extremely important. Assuming that the cut efficiencies are the same for the trapping and non-trapping data, the cuts will not introduce a systematic effect in the lifetime. The following sections discuss the methodology used in performing the corrections to the data, the background cuts, and in calculating the timestamp histograms.

3.2.1 Bad Timestamp Cut

Two issues with the timestamps were discovered early on during the analysis. These issues were later verified with Gage, the manufacturer of the DAQ cards, to be due to a firmware issue that had since been fixed but could not be used to retroactively fix data that was already taken. Addressing these issues is the first task in the mid-level analysis.

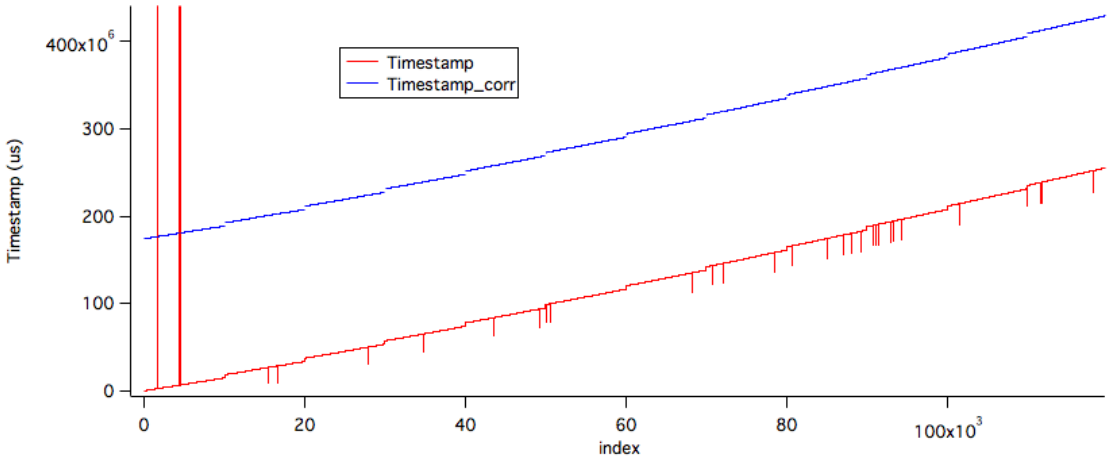
The timestamps are expected to be monotonically increasing as a result of the order in which the DAQ cards write the data to disk. Instead, as can be seen in Figure 3.5, there are two types of jumps that occur in the timestamp data. The first of these is large jumps into the future that occur near the start of the run. These jumps are on the order of about 40 min. The timestamp of the following event appears to be recorded correctly. Therefore, in this case, it is the event before the jump backward in time that is bad. The size of this 40 min jump is similar to the duration of the file. However, this is a coincidence. Using shorter diagnostics data files, the size of these jumps is found to be uncorrelated with the duration of the file and is always approximately this size.

The second type of faulty timestamp is seen throughout the timestamp array. It is a small jump into the past. The typical size of this jump is about -2.7 s. This type of failure occurs much more frequently in the data files. It is also worth noting that for both types of jumps the size of the jump varies randomly and can not be accounted for with a single bit being faulty in

the DAQ card assuming that the bits are always read in the same order.

The bad timestamps were removed with a very simple algorithm that calculates the difference in time between consecutive events and throws out both timestamps if the difference is negative. In both cases, this throws out one additional data point that is not a bad timestamp, but it is a very small number of events and was considered an acceptable loss.

Figure 3.5: A timestamp array before and after correcting the bad timestamp events. The corrected timestamp in blue also contains a y-offset to account for the delay time, the amount of time between closing the neutron beam and starting data collection. This image only shows the early events in the file to make the bad timestamps more apparent.



The reference pulser system proved instrumental in investigating the bad timestamps. The regular spacing of the reference pulser events allows an accurate prediction of the location of future reference pulser events. By looking at the timing of events that had bad timestamps that were also reference pulser events, it was possible to prove that the size of the effect on bad timestamps is random in nature. The spread in the distribution was sufficiently large that attempting to fix the bad timestamps would introduce substantial uncertainty in their corrected value. Because improperly correcting the timestamps could introduce a systematic effect, it was decided that it was better to remove these events instead of attempting to correct them.

3.2.2 Active Veto Cut

The active veto system, as described previously, is designed to tag charged particles passing through the cell so that they can be removed from the data to improve the signal to background

of the experiment*. A histogram of the *muon_pulse_area* can be seen in Figure 3.6 along with the *muon_pulse_area* threshold, the fit region, and the fit curve. The histogram consists of a roughly Gaussian zero peak in addition to a shoulder at larger *muon_pulse_area* that corresponds to events that triggered the active veto system. Due to the lack of separation between the zero peak and the muon signal, a clean cut that removes cosmic muons and no other events is not possible. In light of this, a conservative cut threshold that prefers missing cosmic muons over falsely tagging non-cosmic muon events was desired. Estimating the expected rate of type 2 errors (falsely tagging a non-muon event as a muon) requires an understanding of the distribution of the non-cosmic muon events in *muon_pulse_area*. Because muons will tend to be at higher *muon_pulse_area*, below the maximum of the distribution the spectra is expected to be dominated by non-muon events. The mean and standard deviation of the non-muon events was estimated by fitting the low *muon_pulse_area* side of the zero peak to a Gaussian. For the fit to converge, the fit region must extend beyond the maximum of the distribution so that the fitting algorithm can isolate the effect of the mean from the standard deviation of the distribution. Therefore, the fit region was chosen to extend from $-\infty$ to 32 *muon_pulse_area* greater than the maximum of the *muon_pulse_area* histogram. The threshold was chosen to be five standard deviations, 5 σ, from the x-offset of the fit.

As is evident from Figure 3.6, there are many large events in *muon_pulse_area* that are clearly not neutron events and can be easily removed from the data with little risk. Due to the weak signal to noise in the active veto system, there is a large population of events where it is less clear if the event is a muon. The 5 σ threshold was selected as a conservative estimate to limit the number of type 2 error tags. The cosmic muon rate is expected to sample from a Poisson distribution that is constant in time, and therefore, it should also be roughly constant in time. To the extent that the gains are stable, the detection rate of cosmic muons should also be constant. If the detection rate of cosmic muons is constant then the effect of type 2 error muon tags will decrease the signal to noise in the experiment. However, it is not expected to introduce a systematic effect. Type 1 error muon tags, on the other hand, will be proportional to the total detection rate, which is not expected to be constant and therefore would be expected to introduce a systematic effect. This is the motivation for using the conservative active veto threshold. The average fraction of events that failed the active veto cut was 3.2%, which corresponds to approximately 12 s^{-1}.

*The design of the active veto system along with a description of cosmic muons can be found in Section 2.12, which starts on page 52.

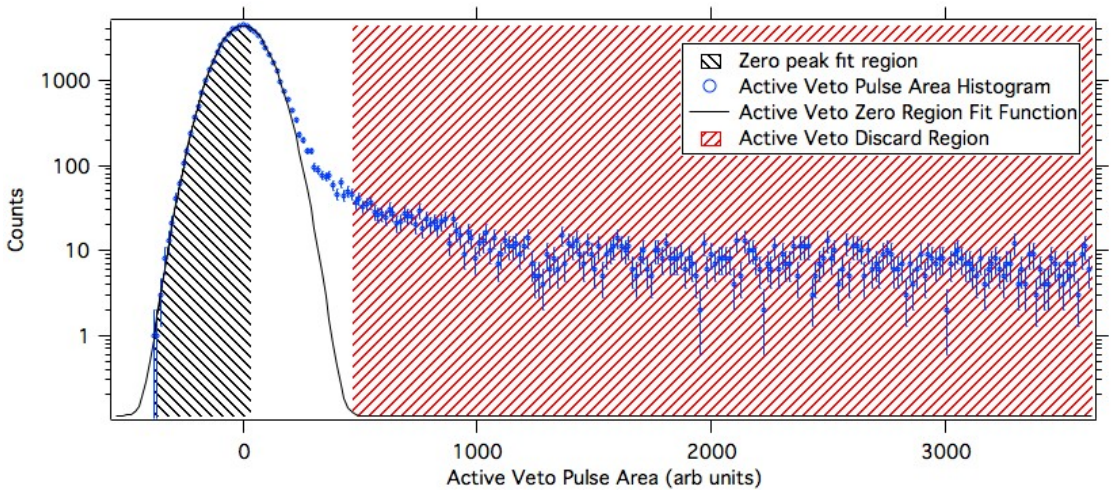
3.2.3 Reference Pulser Cut

The reference pulser events are included in the data to allow for the calculation and correction of gain drifts as a function of time in the main detection PMTs*. A graph of a ref_pulse_area histogram is shown in Figure 3.7. The zero peak is clearly separated from the reference events. A conservative static threshold of 5000 ref_pulse_area was chosen. The standard deviation of the zero peak of the ref_pulse_area histogram is determined to be ≈ 600 by fitting the zero peak to a Gaussian. This puts the ref_pulse_area threshold approximately eight standard deviations away from the center of the zero peak.

Because of the clear separation between the zero peak and the reference events, the tagging efficiency of reference events is expected to be effectively unity. There are some reference events that do not follow a perfectly Gaussian distribution, as can be seen in Figure 3.7, as the few events between 40×10^3 to 65×10^3 . Because of their small number, these events are ignored. For completeness sake, the rate of reference pulser tags should be calculated using the same method as the other cuts. The average fraction of events cut by the reference pulser cut is 27.5%, which corresponds to an approximate rate of 102 s^{-1}. The discrepancy between the expected rate of 100 s^{-1} and the measured average rate is an indication of the uncertainties in the calculation of the rate, which has been only approximately corrected for the detector livetime. Therefore, these estimates of the rates are expected to be accurate at the few percent level.

*The reference pulser system is described in Section 2.11, which starts on page 52. The correction of the gain drifts is described in the following section, Section 3.2.4, which starts on page 90

Figure 3.6: A pulse area histogram for the active veto system. A fit to the zero peak is included in black. The region of data used in the fit is shown in the black hash. The region excluded as muon like events is shown in red hash, which corresponds to a 5 σ cut on the zero peak.



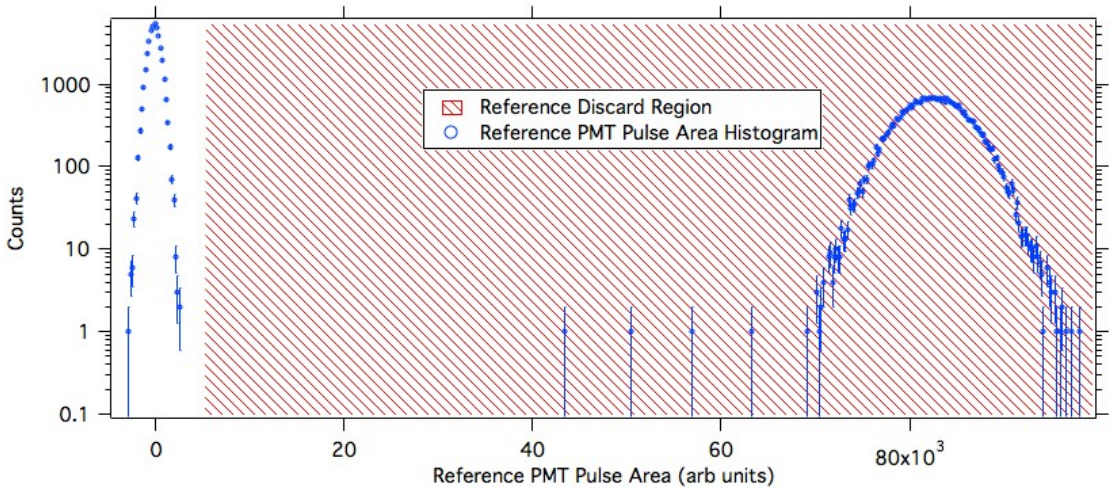
3.2.4 Gain Correction

A gain monitor is in place to allow the gain of the main detection PMTs to be estimated and corrected to remove any time dependence from the pulse area and pulse height cuts. This gain correction is done independently for the two main detection channels. The gain is estimated by monitoring the location of the reference pulser events in the main detection PMTs. The reference events are tagged using the reference PMT by comparing the ref_pulse_area to the ref_pulse_area threshold. Events that exceed the ref_pulse_area threshold will be referred to as reference events.

To calculate the gain drifts, the reference events are broken up into 15 s time bins. In each time bin, the mean and standard deviation of the main detection pulse area is calculated for the reference events. The mean is an estimate of the gain, and the standard deviation is an estimate of the uncertainty in the gain. The time bin width of 15 s results in a sufficient number of counts in each bin for Gaussian statistics to be satisfied while keeping the timing resolution high enough to constrain the timing distribution of the gain drifts.

A Loess time-averaging method is used smooth out the fluctuations in the gain calculation, and the gain is evaluated at arbitrary times by interpolating the smoothed data. The Loess smoothing, which is a locally weighted, regressive smoothing algorithm, used a smoothing window containing 11 consecutive data points. It was selected because it was built into Igor and

Figure 3.7: A pulse area histogram of the reference PMT including the cut threshold. The ref_pulse_area histogram is shown in blue. The reference threshold is 5000 in ref_pulse_area. Any events with a ref_pulse_area greater than the threshold are tagged as reference events, which is indicated by the red hashed region. These reference events are used to calculate the gain of the main detection PMTs as a function of the time. They are excluded from the final analysis.



not because of the specifics of the smoothing algorithm itself.

Once the gain drift has been determined, the gain correction is applied by multiplying the pulse area and pulse height by the ratio of the desired location of the reference pulser events, which we chose to be 1×10^6, and the gain evaluated at the time of the event. To preserve the ratio of the pulse area to pulse height from before the gain correction, the pulse height was corrected by an additional factor of the average pulse height divided by the average pulse area of the reference events.

Ultimately, the overall scaling of the gain adjusted pulse area is not important. This is because the signal in the reference PMT is found to be constant in time and consistent between different data types. This suggests that the light from the light emitting diode (LED) is extremely uniform, and therefore, any pulse area or pulse height cuts that are performed with respect to the time averaged location of the reference pulser, i.e. the gain corrected pulse area and pulse height, will be consistent for different running configurations. This is important for the background subtraction and is why all pulse area and pulse height cuts are performed on the gain adjusted pulse shape metrics. This argument is expected to be independent on how accurately we can calibrate the location of the reference pulser peak in PE in the main detection PMTs, which we estimate to be ≈ 29 PE.

If the gain correction does not properly correct the gain drifts it will result in an additional systematic effect in the trap lifetime extracted from the data. This is addressed in Section 5.4, which starts on page 143. Additional information related to the performance of the gain monitor and understanding the gain drifts can be found in Appendix B, which starts on page 230.

3.2.5 Lower Pulse Area Cut

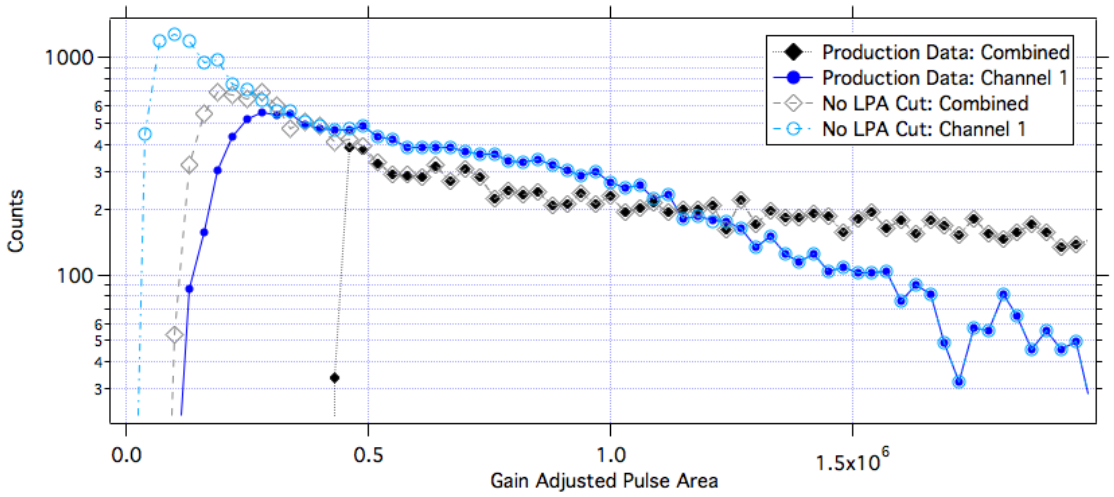
A lower pulse area cut is included in the analysis to remove the portion of the backgrounds that come from low-energy events, in particular, the random coincidence of uncorrelated photons. The primary, low-energy backgrounds in the experiment come from neutron-induced luminescence of the BN, spontaneous emission from the photocathode, and any other low-energy events in the detection system. These sources of background are a large fraction of the contribution to the constant background. Because these background events should be removed from the data, a higher threshold for the hardware pulse height discriminator would have been selected if a pulse area cut was not possible. The lower pulse area cut threshold is chosen to be sufficiently high that it is more stringent than the pulse height discriminator threshold throughout the data file. The lower pulse area cut is performed on the gain adjusted pulse area to prevent a time-dependent detection efficiency for the neutron β-decay events. Removing the majority of the low-energy backgrounds with a software pulse area cut has an added benefit. It allows some of the low-energy events to be digitized so that they can be studied during the data exploration

to better understand the data.

The lower pulse area cut will also remove some of the neutron β-decay events. Although an average neutron β-decay event is expected to produce a large number of photons in the helium, for some of the neutron β-decay events only a few photons will be detected. This is because the amount of energy radiated as photons is strongly correlated with the energy of the β particle emitted from the β-decay, which has a continuous energy spectrum that extends to zero energy. In addition, the light detection efficiency, which depends on the location in the cell, will further reduce the number of photons detected from an event. Therefore, the inclusion of a lower pulse area cut will remove some fraction of the neutron events from the data.

The lower pulse area cut threshold combines an individual cut in each channel of 3 PE and a cut on the sum of the two channels of 12 PE. On average, this combination of cuts tags 45% of all events. However, there is substantial correlation between the lower pulse area cuts and the kurtosis cut.

Figure 3.8: A gain-adjusted pulse area histogram.



3.2.6 Upper Pulse Area Cut

An upper pulse area cut is included to remove events that are sufficiently large that they could not have been a neutron β-decay event. In the following, the maximum number of PE from a neutron β-decay event is estimated. For a 364 keV electron in LHe, 24% of the energy is deposited as extreme ultraviolet (EUV) scintillation phonons[32]. Assuming that this is a good estimate of EUV production at the neutron β-decay electron endpoint energy, the amount of

energy produced in EUV photons is $\approx$ 190 keV. The average EUV photon energy is 15.5 eV which corresponds to $\approx 1.2 \times 10^4$ photons. Using a Guideit simulation that he developed, Chris O'Shaughnessy[17] estimated the effective photon detection efficiency, the probability of an EUV photon creating a PE in the detection system. The simulation predicts a maximum in the effective photon detection efficiency of $\approx 0.5\%$, which occurs about 10 cm from the end of the light guide. Taking these numbers into account, an average 786 keV electron is expected to produce $\lesssim 60$ PE in each main detection PMT with a variance according to counting statistics of ≈ 8 PE.

The upper pulse area threshold is selected to be 115 PE in each channel individually, which was chosen by estimating the location of where the neutron spectra approached zero counts. By selecting a threshold that is above the region where the neutrons exist, the risk of introducing a systematic effect into the lifetime is reduced, but it will allow the removal of a small fraction of the background events before the background subtraction. This threshold provides a safety factor of ≈ 2 over our estimate of the energy of a maximally energetic neutron β-decay event. On average this cut tags 6.6(3.2)% of all events in channel 1(2). However, there is substantial overlap between these two cuts as well as between these cuts and the upper pulse height cuts.

3.2.7 Pulse Kurtosis Cut

Two related types of kurtosis cuts have been considered, a combined kurtosis metric, which combines the kurtosis in the two channels as the square root of the sum of the squares, and a method using an independent cut in each of the main detection channels*. In this section, both of these methods will be discussed.

As was discussed previously, the top image in Figure 3.4 shows how the kurtosis is an effective metric for distinguishing between events in the two bands of the pulse area, pulse height phase space. In that figure and in Figure 3.9, kurtosis histograms are presented, which show that the separation between the two bands results in two peaks with a minimum between. The separation between these peaks is what allows the kurtosis to be an effective selection criterion for the pulse timing. The counts between the two peaks are the result of events with low pulse area that have sufficiency few PEs to prevent an accurate determination of the pulse shape with the kurtosis. Comparing these histograms to the lower graph in Figure 3.4, it is apparent that if the threshold for the lower pulse area cut is high enough, it can be used to remove these events where the kurtosis cut is not effective at selecting on the timing distribution.

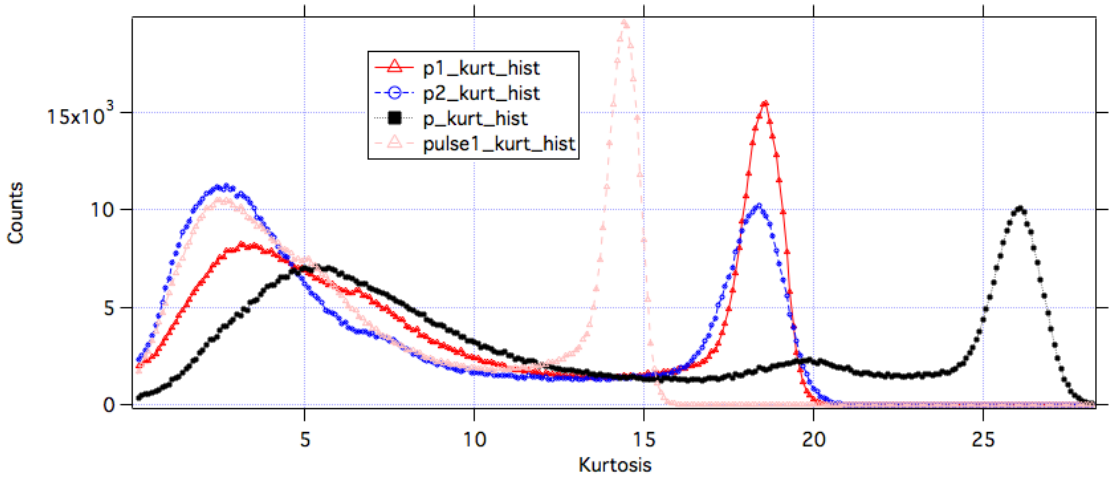
If the kurtosis is a powerful metric for selecting on the timing distribution of the events and the timing distribution is an indication of the physical process that creates the light, there should be a strong correlation between the kurtosis in the two channels. This correlation can

*A general description of the kurtosis, how it is calculated, and its significance can be found in Section 3.1.5, which starts on page 81.

be assessed visually in Figure 3.10. There is evidence of a strong correlation in the kurtosis at both low and high kurtosis. However, there are also a non-negligible number of events in the wings where the correlation is less strong. It was the shape of the data in this portion of the phase space that initially lead to the combined kurtosis metric. By cutting on the combined kurtosis, a single cut could be used to separate the slow events at low kurtosis from any events that were inconsistent with being slow. From this graph, it is also clear that the effect of a cut in the combined kurtosis, which would be a circular cut in this plot, is likely to have a similar effect as a cut in each of the channels, which would correspond to vertical and horizontal cuts. By comparing the extracted lifetime in the series 16 data with two independent kurtosis cuts and a single combined kurtosis cut, a systematic difference of -0.8 s or $\approx 0.1\%$ is obtained. This shows that our method is not sensitive to the choice of these two types of kurtosis cuts.

The kurtosis cut is dynamically determined in each channel to correspond to the minimum in the kurtosis histogram. Using the minimum results in the fewest number of events being affected by small changes in the cut threshold due to any uncertainty in the threshold determination. The kurtosis cut tags 44% of all events on average. However, there is substantial overlap between the kurtosis and the lower pulse area cut.

Figure 3.9: A representative example of pulse kurtosis histograms that shows the effect of the scaling in channel 1 and includes the combined pulse kurtosis metric for comparison. Histogram of the kurtosis in channel one before (pink, dashed, triangle) and after (red, solid triangle) the scaling are shown. Additionally, histograms of the kurtosis in channel 2 (blue, fine dashed, circle) and the combined kurtosis (black, dotted, square), which is the sum of the squares of the scaled kurtosis values, are included.

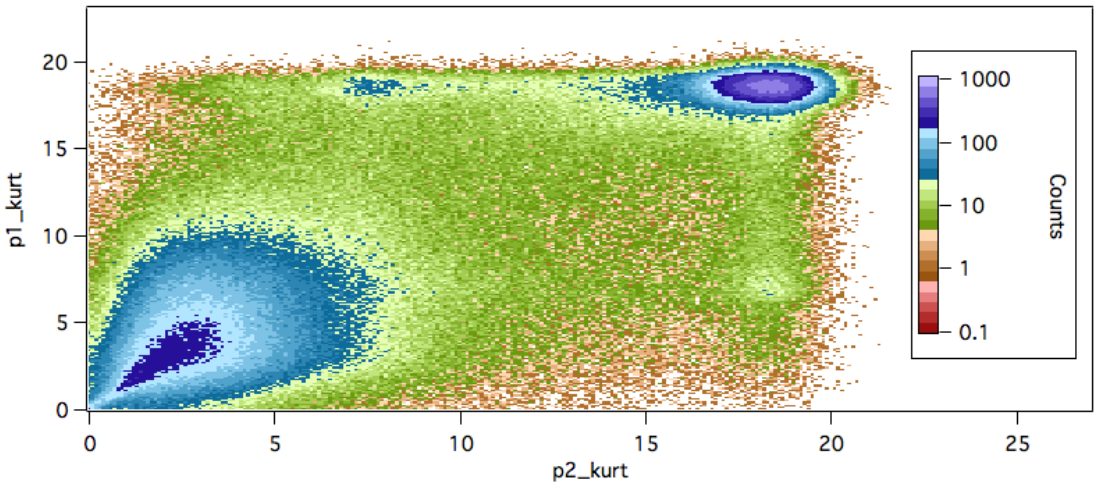


3.2.8 Upper Pulse Height Cut

The response of the detection system becomes non-linear for events that have an amplitude that is too large. An upper pulse height cut is used to remove these events. The non-linearity is caused by the linear fan, which has an effective voltage range in which its response is linear. Some of the events in our data pass outside of this range. The result is an underestimate of the height and correspondingly the pulse area of events these events with amplitudes that are too large. It is worth noting that this effect is smoothly varying, it does not cleanly chop off the parts of the voltage trace that are beyond the limits of the linear fan. A consequence of this effect on the data can be seen in Figure 3.4 where it manifests itself as curling up of the two bands structure at large pulse height. This also affects the pulse area, pulse kurtosis phase space by causing events in the lower band that have large pulse height to overlap with the bottom of the upper band. The effect of the pulse height cut on the pulse area, pulse kurtosis phase space can be seen in Figure 3.11.

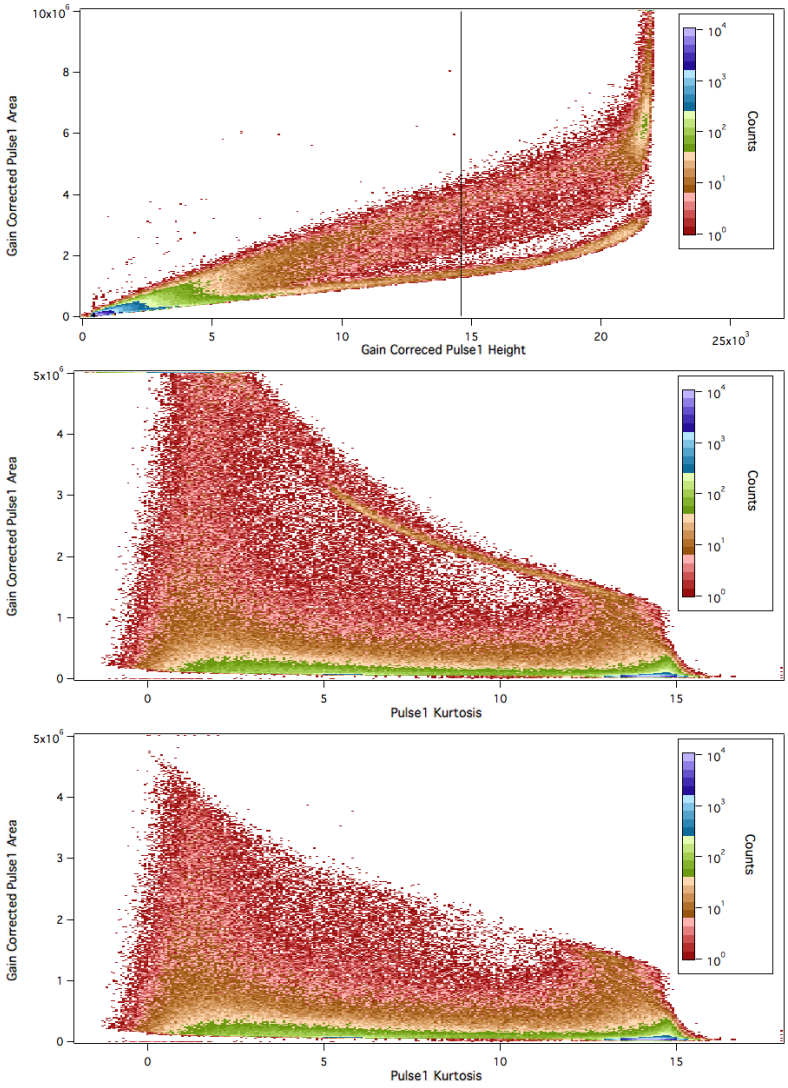
It is the height of the event, including the effect of the gain, that determines if the event is underestimated. Therefore, during periods when the gain is low, a higher energy event could be affected less strongly than a less energetic event when the gain is higher. Because of this, performing the upper pulse area cut on the non-gain corrected pulse height might seem to be the proper course of action. However, using the non-gain corrected pulse height would result in the

Figure 3.10: A visual representation of the correlation in the kurtosis of the two main detection channels. It shows a strong correlation between the two channels, with a non-negligible number of events in the wings of anti-correlation. The shape of the distribution seems to suggest that the uncorrelated events appear to be more strongly coupled to the high kurtosis correlated peaks than the low kurtosis correlated peaks.



cut threshold varying in time. Therefore, to maintain the validity of the background subtraction, it is essential that the cut is performed on the gain corrected pulse shape parameters. To take advantage of the strengths and avoid the weaknesses of these two approaches, a methodology was developed that combines both of them using cuts that are chosen so that all of the events that are distorted are removed. It is described below.

Figure 3.11: Graphical representation of the effect of the upper pulse height cut on the data. (Top) A graph of the pulse area as a function of the pulse height for the channel 1 main detection PMT with the cut threshold indicated. The pulse area as a function of kurtosis in the channel 1 main detection PMT is shown with (Bottom) and without (Middle) applying the upper pulse height cut. This shows how the pulse height cut removes the portion of the lower band that overlaps with the upper band in the pulse area, pulse kurtosis phase space.



First, the non-gain corrected pulse height histograms are analyzed to determine the minimum in each file, which is an estimate of where the bands begin to curl over in the non-gain corrected pulse shape arrays. The minimum is determined by taking the derivative of the pulse height histogram, applying a boxcar smoothing algorithm, and determining where the smoothed derivative crosses zero. The lowest threshold for all of the files in both the trapping and non-trapping data is selected as the non-gain corrected threshold. The same threshold is used for both the trapping and non-trapping data.

The second step is to calculate the gain corrected pulse height threshold. This threshold is calculated by applying the minimum of the gain correction to the non-gain corrected pulse height threshold that was determined previously. This determines the minimum value of the gain corrected pulse shape parameter that corresponds to the non-gain adjusted pulse height threshold. This is done for every file in the series, including both the trapping and non-trapping files. Once again, the gain corrected pulse height threshold from the file with the lowest threshold is used for all of the data files in the series.

This two-pass methodology has two advantages. Using the smallest non-gain adjusted pulse height thresholds results in a conservative cut threshold, which will minimize the number of events allowed through the cut for which the response of the detection system was non-linear. Transferring the threshold to the gain adjusted pulse height arrays ensures that the cut threshold is in the same location for all of the data files that are used in the background subtraction, which reduces any systematic effects that this cut could introduce through the background subtraction.

Using this cut, on average 5.6(6.2)% of all events are tagged in channel 1(2). However, these cuts and the upper pulse area cuts have a large amount of overlap. This cut is found to reduce a non-negligible fraction of the events. Additionally, it is found to complement the upper pulse area cuts.

3.2.9 The Relative Selective Power of the Cuts

The fraction of events removed from each of the cuts can be presented in different ways. The simplest method is to list the number of events that were tagged by each cut. Alternatively, the number of events that were tagged by a particular cut and no other cuts might be more representative of how strongly a particular cut will influence the final result. The result of both of these methods is provided for individual cuts as well as for some combined cuts, see Table 3.1 and Table 3.2.

These statistics show some interesting information about the pulse shape phase space. First of all, the reference pulser events are a large contribution to the data rate. However, they are easily tagged, and their rate was chosen. Therefore, the fact that they compose 28% of the data

Table 3.1: The fraction of events that were tagged by the indicated cut or combination of cuts. This data represents the events that are tagged by at least one of the indicated cuts and does not take into account if the event was tagged by any other cuts.

Cut type	Cut fraction
Lower pulse area (LPA)	0.45
Kurtosis (Kurt)	0.44
Combined LPA and Kurt	0.52
Reference pulser	0.28
Active veto	0.032
Upper pulse height channel 1	0.059
Upper pulse height channel 2	0.061
Combined, upper pulse height	0.088
Upper pulse area channel 1	0.068
Upper pulse area channel 2	0.034
Combined, upper pulse area	0.068
Combined, upper pulse area and height	0.12

is an overrepresentation of their importance when considering the cuts.

The lower pulse area and kurtosis cuts tag the largest fraction of the events by tagging 45(44)% of events, respectively. The combined lower pulse area and kurtosis cut demonstrates that these two populations of events have a very strong overlap. The other pulse shape cuts tag 3% to 9% of the events. The combination of the upper pulse area and upper pulse height cuts tags 12% of the data. This data also shows that about 85% of the events are removed by the cuts before the background subtraction. Of those, 54% are removed by the pulse shape cuts.

Table 3.2: The fraction of events that were tagged by the indicated cut or combined cuts but no other cuts. All non-pulse shape cuts is the combination of the reference pulser, active veto system, software veto, and bad timestamp cuts.

Cut type	Cut fraction
All cuts	0.85
Lower pulse area (LPA)	0.068
Kurtosis (Kurt)	0.041
LPA + Kurt	0.44
All non-pulse shape cuts	0.31
Combined, upper pulse height	0.02
Combined, upper pulse area	$\ll 0.001$

3.2.10 Sensitivity to the Pulse Shape Thresholds

Cut thresholds that strongly affect the trap lifetime must be very well understood to prevent the introduction of systematic effects. Therefore, it is important to investigate how sensitive the trap lifetime is to each of the cuts. This was done by varying the location of each of the cut thresholds by $\pm 5\%$ of the threshold's value. These diagnostics data sets were carried through the analysis and a trap lifetime was extracted and compared to the production trap lifetime.

Some of the data series showed large systematic effects, in particular, the early data series. These early data series also typically showed larger statistical uncertainties in the trap lifetime. However, a strong correlation between the size of the statistical uncertainty in the trap lifetime and the sensitivity to the cut threshold was not observed.

There are a few trends in the data that suggest the reason that the data appears to be sensitive to the cut thresholds is because of statistical uncertainties introduced into the calculation when the threshold values are changed. Most notably, the systematic correction associated with a particular change in the cut threshold is not correlated between the different data series. Additionally, the size of the systematic effect is not anti-correlated with the effect of moving the cut threshold in the other direction. Finally, the data sets with small uncertainties in their trap lifetimes show substantially less variation in their sensitivity to cut thresholds. Each of these observations is consistent with the estimated systematic effect being sensitive to statistical fluctuations.

If we assume that the apparent sensitivity to the threshold values is caused by statistical fluctuations, it is reasonable to look at the systematic difference in the lifetime normalized by the statistical uncertainty in the trap lifetime. In the 11 data series that are included in the final analysis, the normalized systematic difference is never larger than 0.3. The average normalized systematic effect is 0.011(0.032). The data series with the largest uncertainty, s11.4, also has the largest systematic difference, which is ≈ 60 s and is due to the lower pulse area cut. On the other end of the spectrum, for five of the last six production data series that were taken, the systematic effect was less than ≈ 5 s for each of the cut threshold.

Although there is evidence that the analysis is sensitive to the location of the cut thresholds, we believe that this apparent sensitivity is due to statistical fluctuations and that as additional data is taken the estimated sensitivity is likely to decrease.

3.2.11 Delay Time

The timestamps are recorded in microseconds from the start of data collection. Later in the analysis, it will be more useful for the timestamps to be relative to the close of the neutron beam. During operation, the DAQ writes a message to the log that records the amount of time between closing the beam and sending the signal to the DAQ cards to begin collecting data.

This duration is referred to as the delay time, which is entered into a database that was built into Igor from the ground up. It records the value and, at this point in the analysis, adds the delay time to each of the timestamp values. This aligns the timestamp histograms and livetime array for different data files, which makes it trivial to combine them together in the upper-level analysis. The delay time varies between 142 s to 373 s in the production data depending on the running configuration.

3.2.12 Timestamp Histograms

Timestamp histograms are the primary output of the mid-level analysis; they characterize the time dependence of both the neutron like and background events. They are calculated using the cut timestamp arrays. The timestamp histograms are constructed after the correction for the delay time, and therefore they measure time relative to the close of the neutron beam. Both the timestamp histograms and livetime arrays, which are discussed in the next section, use bins with a width of 15 s.

3.2.13 Livetime Calculation

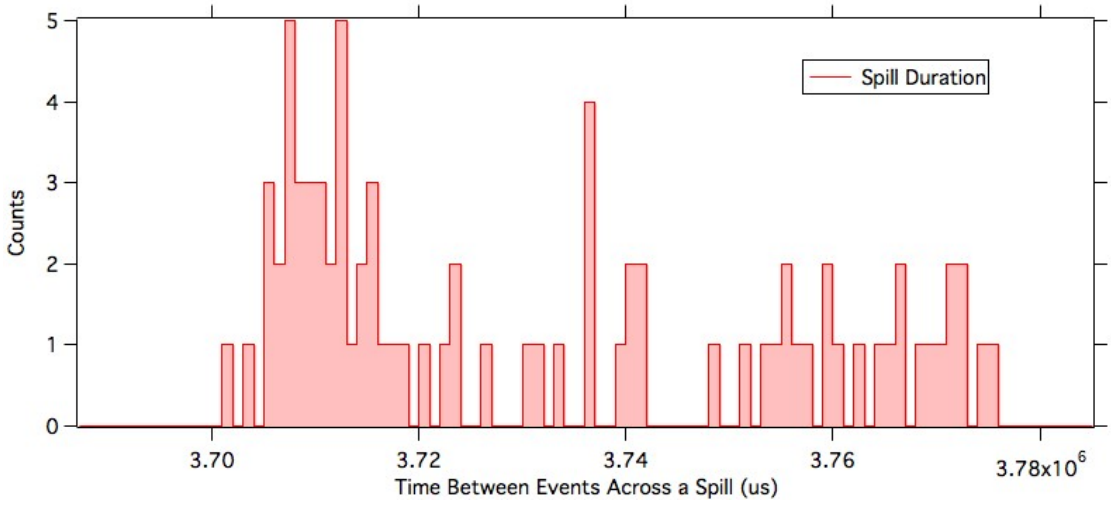
The imperfect livetime of the detection system was another effect that must be corrected for. The livetime is the fraction of the time that the detector is able to receive and recognize events. The dead time is the period of time directly proceeding an event during which the detection system is in the process of returning to its ready state, is writing data to memory or to disk, or is for any other reason unable to accept an additional event. In random processes, there is some probability that another event will come in before the detection system returns to its live state. This results in a missed event and an underestimate of the rate. The duration of the deadtime is expected to be independent of the data rate, therefore as the data rate increases the fractional deadtime of the detection system does as well. In the UCN Lifetime Experiment at NIST, this results in a lower detection efficiency at early times when the rate is higher than at later times and therefore results in a systematic effect in the extracted trap lifetime.

To account for this, the data is corrected for the deadtime of the detection system. The dominant mechanism for deadtime, in this experiment, is a result of the time it takes the DAQ cards to write memory to disk and perform the hardware rearm. This duration is referred to as the spill duration. The DAQ system is set up so that the data comes in in contiguous blocks of 10000 events, which are referred to as spills. The spill duration has been found to vary from spill to spill. At the end of a spill, the memory of the card is written to disk. Therefore, the deadtime starts immediately following the 10000th event in the spill, which allows a precise determination of the start of the dead period. In contrast, the time when the DAQ system returns to its live state is not known. The card must have gone live before the first event of

the next spill, but, the amount of time that the card is live before the first event of the spill is expected to vary as a result of the random timing of the events.

If the spill duration was constant, the time between two consecutive reference pulser events, 0.01 s, would be an upper bound for the duration between the last event of the previous spill and the first event of the current spill. At durations longer than this, a reference event would be guaranteed to come in and trigger the system. Instead, the variation in the spill durations is much larger than this, which can be seen in Figure 3.12. Some reference pulser events are missing, however not enough to account for the spread in the spill durations. Therefore, it is concluded that this additional spread is a manifestation of variations in the spill duration. If the spill duration was constant, the shortest observed value could be used as an estimate of the spill duration and specific data could be taken to measure it very accurately. Because the spill duration varies, a software veto was implemented to enforce a constant spill duration by throwing out any events that occur during a 3.8 s window that starts at the time of the last event of the previous spill. This duration was chosen because almost all spills have a spill duration that is shorter than this value. Implementing this software spill veto allows the livetime to be calculated precisely while only throwing out a handful of data points.

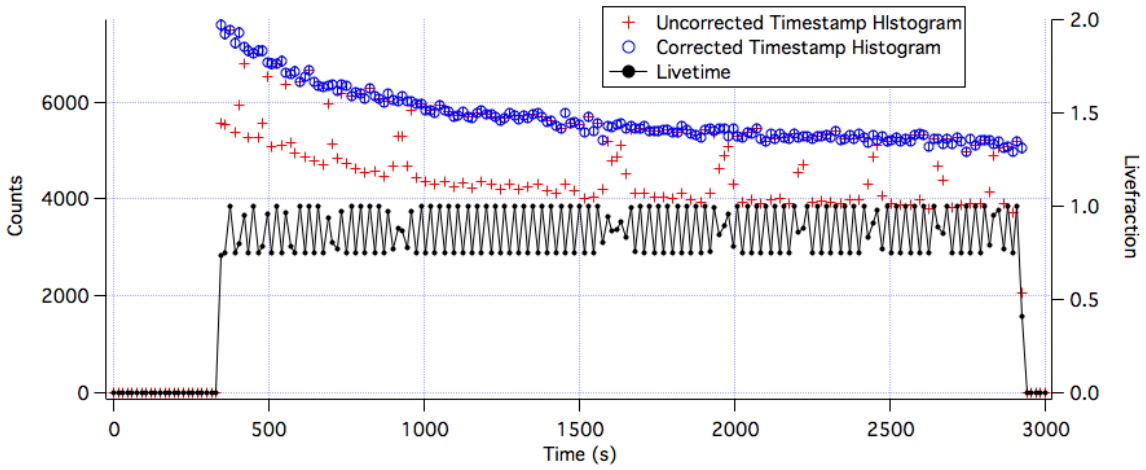
Figure 3.12: A histogram of the spill duration, the amount of time it takes the DAQ cards to write its memory to disk and perform a hardware rearm. In this case, it is measured as the time between the last event of the previous spill and the first event of the current spill, which is an overestimate of the spill duration.



The live fraction is calculated by finding the overlap between the software veto and the bins of the timestamp histograms. The livetime array has the same number of bins as the timestamp

histogram. It starts out with a value of one in each bin. For each spill, the analysis determines which bins overlap with the spill, and it appropriately reduces the livetime of the affected bins. Using the livetime array, the timestamp histograms can be livetime-corrected to account for diminished decay rate due to the deadtime. This is done by dividing the timestamp histograms by the live fraction. The main analysis uses 15 s time bins in the timestamp histograms and the livetime arrays. With this bin size, the majority of the bins have a livetime between 10.5 s to 15 s. A typical livefraction graph along with the corresponding uncorrected and livetime-corrected timestamp histograms is shown in Figure 3.13. The livetime correction is no applied to each individual file as was done in this example. Instead, a pooled livetime correction is applied to the pooled timestamp histograms for each data series in the upper-level analysis.

Figure 3.13: An example of a livefraction array as well as an uncorrected and livetime-corrected timestamp histogram. The count rate has been modified to account for the 15 s bins so that the count rate is in s^{-1}. The livetime correction is not applied to individual files in the data, it is done in this case just to show the effect of the livetime correction. The jagged feature is a beat frequency due to the relative frequency of the spills and the 15 s bins in the histogram. The uncertainty in the timestamp histograms is also carried through the livetime correction, which frequently results in a large uncertainty in the first or last data point in the file because it has a large livetime correction.



An additional, albeit much smaller contribution to the deadtime, occurs after every event and corresponds to the time it takes the DAQ card to perform a software rearm between events. To remove the possibility of variations in the software rearm timing, a hardware gate was put in place to prevent the DAQ system from triggering for 4.7 μs after an event. The average data rate in a typical file is in the range 250 s^{-1} to 350 s^{-1} resulting in a fractional deadtime

from this mechanism of less than 2×10^{-8}. This source of deadtime is sufficiently small that it is neglected in the analysis.

3.3 Upper-Level Analysis

The upper-level analysis is the final portion of the analysis, which extracts the trap lifetime. It takes the timestamp histograms and livetime arrays for each of the data series and combines them to create an average livetime-corrected, timestamp histogram for the trapping and non-trapping data separately. A background subtraction is performed by subtracting the non-trapping data from the trapping data. This removes a large fraction of the remaining background events in a model independent way. Finally, the upper-level analysis fits the background-subtracted, livetime-corrected, timestamp histogram to a single exponential with an offset to extract the trap lifetime. Each of these steps is described in more detail in the following sections.

3.3.1 Combining Data Sets

The data files have already been aligned in time by applying a delay time to the timestamp arrays during the mid-level analysis, which accounts for the amount of time that passed between closing the neutron beam and the start of data acquisition. The timestamp histograms are aligned so that the start of the first bin is at the time that the beam is closed. This means that the timestamp histograms can be combined as their sum, which results in a total number of detected events as a function of time after the start of data acquisition. The resulting histograms are referred to as the total timestamp histogram. This process is completed separately for the trapping and non-trapping data series. At this stage of the analysis, the uncertainties in each bin are still the square root of the number of counts in that bin.

3.3.2 Livetime Correction

The total timestamp histograms are the total number of detected events in each timing bin, but it is the total number of events that is desired. To obtain this value, the total timestamp histograms must be corrected for the deadtime. The first step in this process is to calculate the total effective livetime for each bin, which is the sum of the livetime arrays. Recall that the individual livetime arrays are an indication of the fraction of the time that the detection system was live during each bin of the timestamp histogram. Therefore, each entry is in the range $0 \leq livetime \leq 1$. After summing the livetime arrays, the total livetime array is equivalent to the fractional numbers of detectors that were live during the given bin and therefore is in the range $0 \leq livetime \leq n$, where n is the total number of files that were combined.

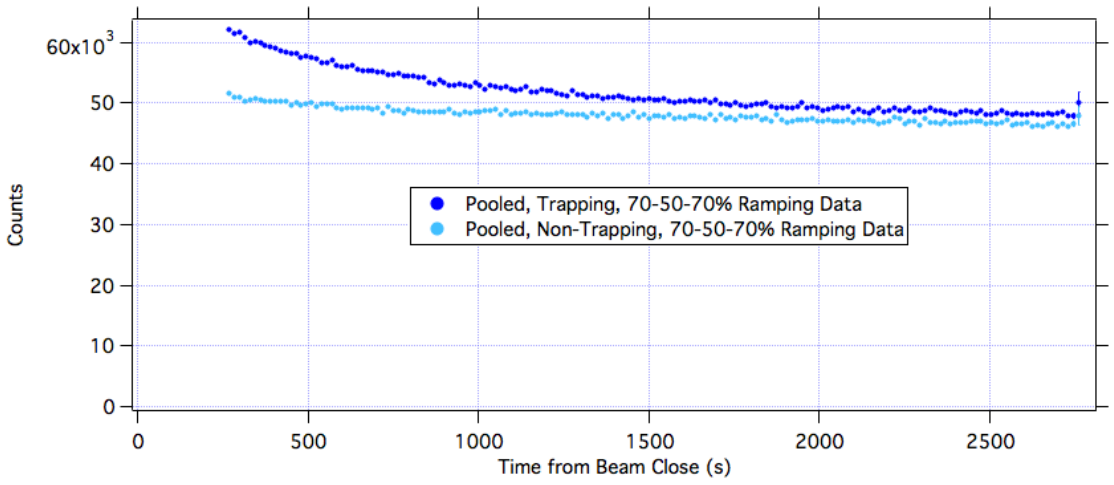
The livetime correction is applied by dividing the total timestamp histogram by the total livetime array. This results in the livetime-corrected timestamp histogram, which is the average number of events in a time bin inflated to account for the deadtime of the detection system. Similarly, the uncertainties are carried through the standard error propagation formalism, which inflates the uncertainties in bins with large livetime corrections.

All of the data experiences large fluctuations in the data rate after the livetime correction. This effect is caused by the timing structure of the files being very similar in all of the data files of the same type and therefore a strong correlation in the timing of the spills existing for similar data files. This results in what appears at first to be non-statistical fluctuations in the livetime-corrected timestamp histogram. Instead, this is an inherent property of the way that the DAQ system is set up.

3.3.3 Background Subtraction

The background subtraction takes the non-trapping and trapping data of a given running configuration and subtracts them to remove both constant backgrounds and time-dependent backgrounds from the trapping data. Plots of the pooled data before and after the background subtraction are included in Figure 3.14 and Figure 3.15. The difference in the background rate for the different data sets is the effect of relatively consistent backgrounds rates in the different data types, but varying numbers of data files of the various data types.

Figure 3.14: Histograms of the pooled, trapping and non-trapping data after the livetime correction.

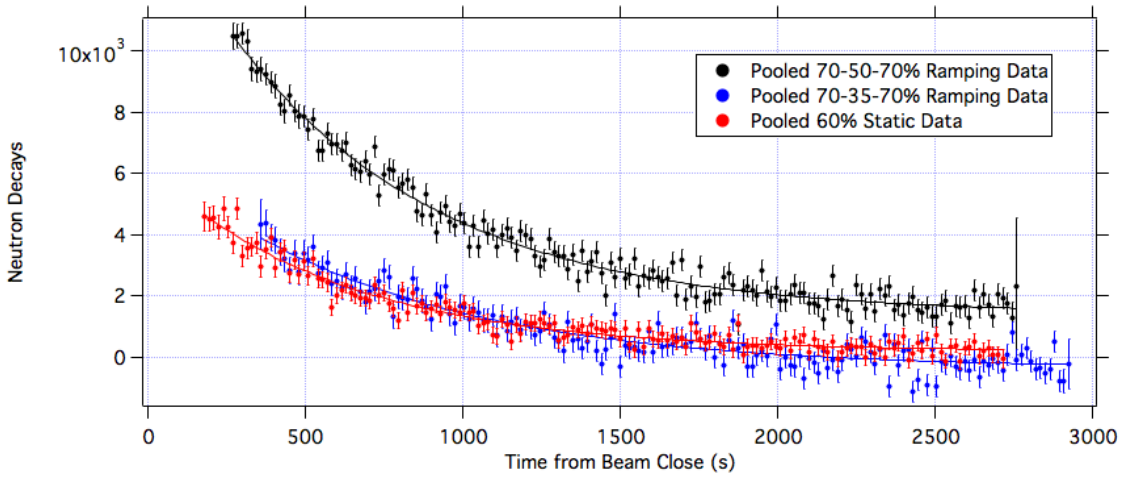


There are two main ways that the background subtraction can improve the data. First,

neutron-induced activation of radioisotopes can result in some background in the livetime-corrected timestamp histograms. It would take an unreasonable amount of data to be able to distinguish the contribution to the timestamp histograms from each potential radioisotope and would introduce substantial additional uncertainty in the extracted lifetime if these isotopes were included in the final fit to the timestamp histograms. With the background subtraction, this is not needed because the time dependence in the non-trapping data includes the effect of these radioisotopes and removes them from the data.

A strength of the background subtraction is its model independence. This is in contrast to effects like the gain drift, for which a functional form of the gain drift must be assumed. In the case of the gain drift, this is either the functional form that the gain drift is fit to or it is related to the algorithm that is used for a time-averaging method. The background subtraction takes advantage of data taken with a very similar running configuration, but with essentially no neutrons, to directly estimate the background rates as a function of time after the start of data acquisition. This allows the background subtraction to clean the data of systematic effects and can allow it to correct for effects that are sufficiently subtle that the other methods may not be sensitive to them. As an example, if the gain drifts in the files are not purely exponential or are improperly corrected, to the extent that the non-trapping data is identical to the trapping data sans neutrons, both the trapping data and non-trapping data will contain the incorrect gain drift, and the background subtraction will tend to remove any background events allowed through the cuts due to the imperfections in the calculation and thereby reduce the corresponding systematic effect. The background subtraction will, of course, not remove any time dependence of the cut thresholds on the neutron signal because they only are present in the trapping data.

Figure 3.15: Decay rate histograms of the pooled, background-subtracted data.



The greatest weakness of the background subtraction is due to the substantial uncertainty that it introduces into the livetime-corrected, background-subtracted, total, timestamp histograms. The event rate in the non-trapping data is a large fraction of that of the trapping data. Using standard error propagation, the uncertainty in the background-subtracted data goes as the square root of the sum of the squares of the uncertainties in the livetime-corrected total timestamp histograms for the trapping and non-trapping data. This results in a very large fractional error in the background-subtracted timestamp histograms and is the reason that removing as many background events as possible before the background subtraction is so important; It reduces the uncertainty in the background-subtracted livetime-corrected total timestamp histograms.

3.3.4 Extracting the Trap Lifetime

The trap lifetime is estimated by fitting the background-subtracted, livetime-corrected, total timestamp histogram to a single exponential with an offset. This is done with a built-in Igor function Curvefit, which uses a Levenberg-Marquardt least-squares method. It uses the point by point uncertainty and correlation between the fit parameters to estimate the uncertainty in each of the fit parameters. The lifetime from the fit and its uncertainty are taken as the trap lifetime and its statistical uncertainty. The amplitude and y-offset are left as free fit parameters, which increases the statistical uncertainty in the extracted trap lifetime. Specifically, we do not feel that we can justify setting the y-offset to zero, which would increase the statistical sensitivity of the result, but could introduce a systematic effect if the background subtraction is imperfect. An example of a fit to the $70 - 50 - 70\%$ ramping data is shown in Figure 3.16.

3.4 Trap Lifetimes by Data Series

Using the methodology described above, the trap lifetime has been determined for each of the primary data sets. The results can be seen two formats in Figure 3.17. In the first subplot, the data for the different ramping schemes are grouped together, and a weighted average for each ramping scheme is calculated and displayed. In the second subplot, the data is shown in the order that it was gathered.

In the series by series data, there is no clear evidence that different data series are returning inconsistent results. This suggests that none of the changes in the operating procedures that were implemented appear to be affecting the trap lifetime. However, because of the size of the statistical uncertainties only a weak limit can be placed.

The data shown in the second subplot of Figure 3.17 help illustrate the change in our strategy that occurred at the end of data series 12.2. Up to that point, the goal had been to

continue to reduce our backgrounds, increase our ramping rate and trap depth* to increase our statistics, and to, as quickly as possible, start taking production data with systematic effects that were controlled at the level required for a 1 s measurement. Early in our production data, we achieved this milestone for the first time in Series 11.1. Over the next few data series, the backgrounds and ramp rates were improved. This increased the rate at which the statistical uncertainty was decreased. By Series 12.2 it was clear that the difference between the PDG mean lifetime and our measured value was not likely a statistical anomaly and our systematic corrections did not appear to be large enough to account for the disagreement. It was also clear that at the rate we were acquiring statistical sensitivity, taking additional $70-35-70\%$ data might not be the best use of our remaining time before a long shutdown for a reactor facility upgrade. This motivated the change in strategy.

The new goal for the remaining data was to constrain the systematic effects to the extent possible with the remaining beam time. One of our most important secondary goals was to benchmark the Monte Carlo (MC) simulation, which is described in its own chapter. We had developed a detailed MC simulation and were in the process of upgrading it further. Using the MC we estimate the systematic effect due to a dominant systematic effect marginally trapped neutrons (MTN)s, but we wanted to benchmark the MC against the data to demonstrate that we understood the MTN systematic effect, that the MC correctly estimated the corresponding systematic correction, and thereby that it was extremely unlikely that the MTNs could be causing this discrepancy between our trap lifetime and the PDG mean lifetime.

*Recall that the ramping rate and trap depth were limited by the stability of the magnets to quenching. By running at progressively higher trap depths and faster ramp rates, the magnets were being trained.

Figure 3.16: Representative fit to the decay rate histogram of the pooled, background-subtracted $70-50-70\%$ ramping data.

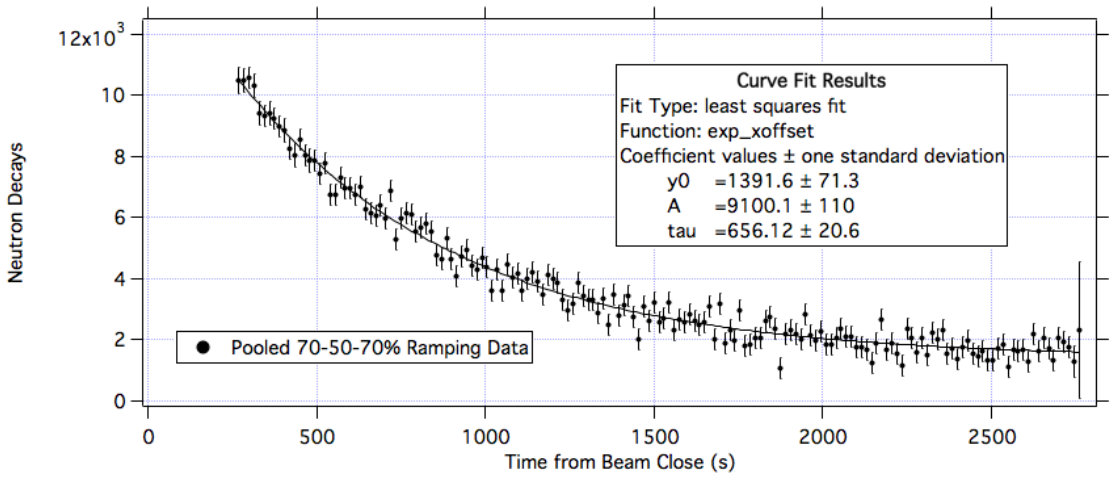
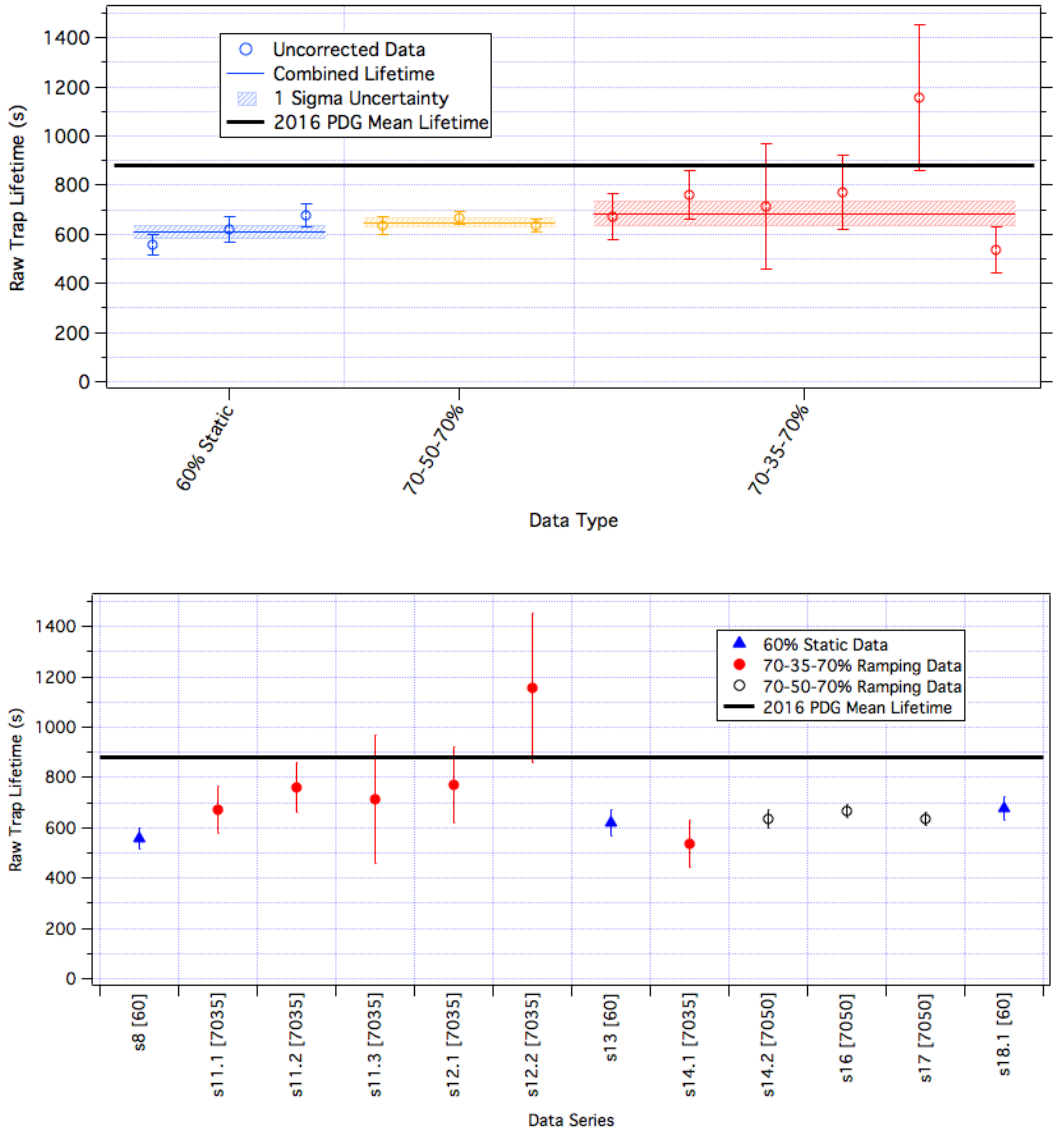


Figure 3.17: Raw trap lifetimes for each of the primary data series including the effect of all cuts. The Particle Data Group (PDG) mean lifetime from the 2016 review is included in both plots for reference. (Top) Raw trap lifetimes for the primary production data sets grouped by data with similar ramping schemes, which result in similar systematic corrections. (Bottom) The raw trap lifetimes for the data series sorted by the order that the data was taken. The square brackets indicate the ramping configuration. 60, 7050, and 7035 stand for data with the 60% static, $70-35-70\%$, and $70-50-70\%$ ramping schemes respectively.



This motivated taking additional data with different ramping schemes to vary the size of the MTN systematic effect. Then by comparing the data to the results of the simulation, the fidelity of the simulation could be tested. This motivated Series 13, which increased our statistical precision in the 60% static data. It also demonstrated that any changes to the apparatus since Series 8 had not had a large effect on the trap lifetime.

After a little bit more production data in Series 14.1, the first real offspring of our new motivation was born, the $70-50-70\%$ ramping data. In this data, the magnet is not ramped as deeply. This allows a larger population of MTNs to survive into the data collection phase, which increases the systematic correction. However, it also results in a substantial improvement in the number of fully trapped neutrons that survive into the data collection, and therefore, it improves our signal to noise in the experiment. In the end, because of this improvement in the signal to noise, the $70-50-70\%$ Ramping data had the largest contribution to our final lifetime, and it put our collaboration in a much better place to assess how large the systematic difference was between the PDG mean lifetime and our fully corrected lifetime.

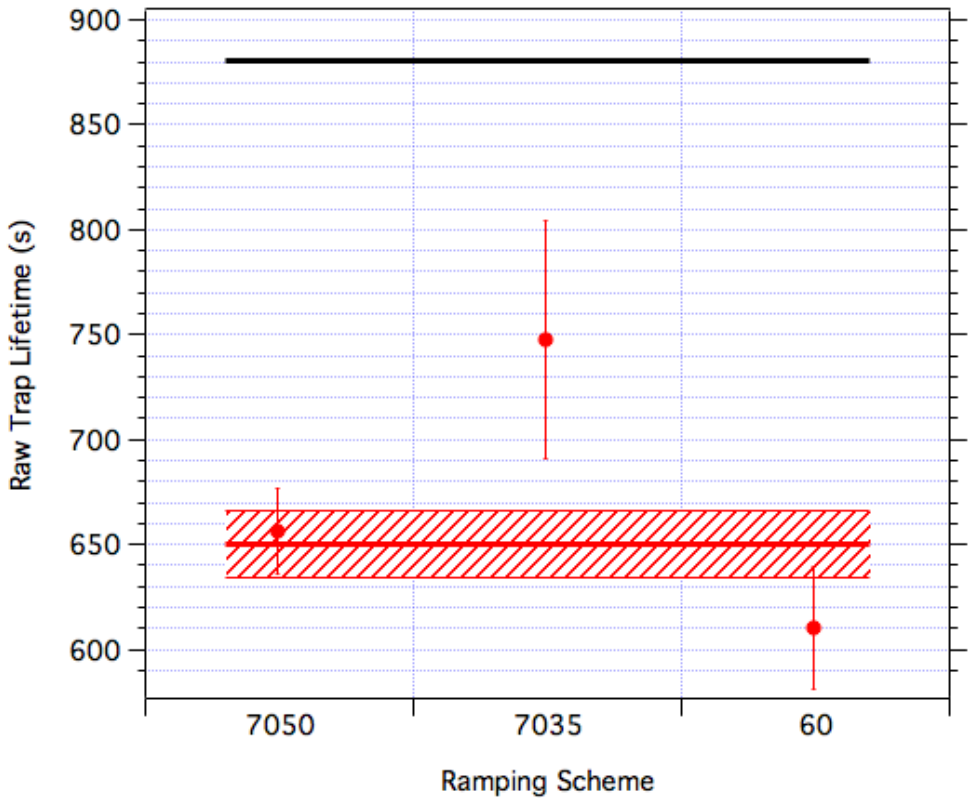
In addition to the production data that was described above two additional data sets are worth mentioning. Warm data in Series 18.2, was used to test the background subtraction. By running the cell at a higher temperature, ≈ 1 K, the effect of the neutron β-decays could be substantially reduced without affecting the backgrounds in the experiment. The results were consistent with no time-dependent background, which suggests that the background subtraction does not introduce a systematic effect in our data.

Finally, the last data was Series 21. It was data that was designed to enhance the systematic effect associated with MTNs. To do this, during the data collection phase the trap was operated at a lower magnetic field state than the filling stage to augment the loss rate due to wall interactions. Also, the data acquisition phase was started while the magnets were ramping. This was done because the population of neutrons was expected to be lost very quickly from the trap due to wall interaction, and therefore, in order to be able to detect the neutron decays, the data acquisition phase had to be started as early as possible. These effects introduced a variety of complications in the analysis process that would require substantial modifications to the analysis code to processes this data. In the end, we were unable to motivate the additional work to develop analysis specifically for this data set, although we did run Series 21 through the traditional analysis. We suspect that there are large systematic effects in the Series 21 result that arise from using analysis that was not designed to deal with its abnormal running configuration. However, as expected, the results were consistent with the effect of MTNs being greatly increased in this data set.

3.5 Pooled Trap Lifetimes

In order to present the final trap lifetime results for the various ramping configurations, it is more useful to pool the similar data sets to make the results less sensitive to fitting biases and to improve the statistical sensitivity. This has been done in Figure 3.18. The results after the systematic corrections can be found in the chapter on systematics, see Section 5, which starts on page 132.

Figure 3.18: Trap lifetime from the pooled data for the main data sets. The weighted trap lifetime is 650 ± 16 s with a reduced χ^2 of 2.42 and a p-value of 0.089. The data points that are being averaged have varying systematic effects that have not yet been accounted for. Therefore, these quality of fit metrics are expected to suffer.



Chapter 4

Monte Carlo Simulation

A detailed MC simulation was developed for this experiment. It was written, run, and analyzed by Kevin Coakley at NIST, Boulder. A number of papers have been written on the topic[23, 45, 46]. My primary role in the MC was to help develop physical models for use in the simulation that were as accurate as possible and to constrain the models either with measurements or our experimental data. I also actively participated in efforts to interpret the results from the MC.

The primary goal of the simulation is to model losses due to wall interactions by the marginally trapped neutrons (MTN)s. This is done by producing a survival probability curve for wall interactions, which is the probability that a MTN will survive until a given time due to wall interactions. Once the survival probability curves were developed, quantities like the systematic effect to the trap lifetime were determined after the fact in post processing. In the post processing, the creation time of UCN was simulated, the different loss mechanisms including neutron β-decay were convolved, and a simulated detection rate was obtained. The systematic effect was extracted using the simulated detection rate. Breaking up the simulation into a calculation of the wall loss survival probability curve and the calculation of the systematic effect in post processing results in a substantial improvement in performance by reducing the amount of time spent simulating neutron trajectories.

Throughout this chapter, I will discuss the Monte Carlo simulation in the simplest case where the magnets are not flushed. There are some technical differences in the simulation when ramping is introduced. I will briefly discuss a few of these differences at the end of the chapter.

As a starting point for discussing the simulation, let us walk through the process of simulating an UCN. When an UCN is born, its initial location and kinetic energy are sampled from a stochastic model built on our understanding of the apparatus. The magnetic field at the location of the UCN is evaluated at each time step allowing the equations of motion to be integrated using symplectic integration. UCNs with sufficient energy may eventually strike the cell wall; When this occurs, the simulation detects a collision and simulates a wall interaction.

The first task in simulating a wall interaction is to calculate a loss probability by solving a one-dimensional Schrödinger equation with a potential that is composed of a set of step potentials corresponding to the average nuclear potentials of each of the materials that make up the cell walls. The exit angle of the UCN after the wall interaction must also be modeled. A diffuse reflection model, specifically a Lambertian model, is used due to the rough nature of the TPB coated ePTFE inserts that line the cell. In the simulation, each time that an UCN interacts with the wall, its survival probability decreases. Once an UCN's survival probability falls below a predetermined threshold, it is considered lost, and the simulation of that neutron ends. At that point, the process begins again for the next UCN.

This method allows a large number of neutrons to be simulated in parallel and for the quantities of interest, like the systematic effect, to be calculated after the fact in post processing. In post processing, the random creation time of the UCN during the loading phase is convolved with the wall loss curve. In this experiment, different populations of neutrons are subject to different combinations of loss mechanisms. These loss mechanisms may or may not be detectable. Therefore for each UCN in post processing, the simulation determines the loss mechanisms that removes this particular UCN, when the loss occurs, and if the loss mechanism is detectable. If all of these conditions are true, a count is added to the detection rate at the corresponding time. The process of determining the loss time is much faster than simulating a realistic neutron trajectory, which is what makes this division of the simulation tasks more computationally efficient. I have mentioned that different UCNs are subject to different loss mechanisms; As an example, MTNs are subject to wall losses, ^{3}He absorption, and neutron β-decay whereas fully trapped UCNs are only subject to the latter two effects. As a final step in the post processing, the decay curve is used to estimate the systematic effect introduced if the MTN systematic effect is ignored. This process is performed many times, which allows both the mean systematic effect and its standard deviation to be evaluated as the systematic correction and an estimate of the corresponding uncertainty.

A potential, ultimate goal for the MC would be to determine a running configuration in which the systematic error due to MTNs is sufficiently small that no correction is necessary. However, the process of limiting the MTN systematic effect, ramping the magnetic trap, also has the effect of decreasing the signal to noise in the experiment. Because of this, it is possible for the ramping scheme to be too conservative. Once the MTN systematic effect is smaller than the other systematic effects, running in a more aggressive ramping scheme will further reduce the signal to noise of the experiment, but will result in an increasingly small reduction in the total systematic effect. This motivates a carefully chosen ramping scheme that limits the MTN effect to an appropriate level for the data being taken but does not throw out additional neutrons if it is not necessary. This corresponds to our $70-35-70\%$ data. In some of the data that has already been taken, the $70-50-70\%$ data, a less conservative flushing scheme was

used, which will increase the systematic effect due to MTNs. Although this is undesired for extracting a neutron lifetime, it is helpful when attempting to validate the MC by making the systematic effect large enough that fewer data runs or simulated neutrons are required to see the systematic effect.

This chapter will systematically go through each of the models used in the MC in detail. Studies of the potential systematic effects associated with each of the models have been published elsewhere[23], and they will be included here as well.

4.1 Initial UCN Spatial Distributions

An initial step in the MC is to determine the distribution from which the initial locations of UCN will be sampled. In the early stages, the MC used a cylindrical beam with a constant cross section. Since its inception, the MC has been updated to include a realistic beam profile that was extracted from beam images that were taken during the commissioning of the experiment. Sensitivity studies were performed to understand the effect of beam divergence. Both of these features can be turned on and off in the simulation to help estimate the sensitivity of the MC to these effects.

4.1.1 The Beam Profile

Measurements of the spatial neutron flux were taken during the commissioning of the experiment. An example can be seen in Figure 2.1. I have included another copy of the same figure in Figure 4.2 for convenience. This image was taken at the approximate location of the collimator, and it is included here to show the extent to which the flux is non-uniform. The effect of the cadmium crosshairs, which are used to align the beam image with the apparatus, can be seen in the beam image. Before they can be used in the simulation, the cross hairs must be removed. Also, smoothing is applied to reduce the sample variability. The beam image is designed to be larger than the collimator. Therefore, collimation is applied artificially to the beam image before it is used in the MC.

The first task was to clean the beam image of various effects that are not representative of the neutron fluence during the experiment. These effects include the cadmium cross hairs and the support structure, which can be seen on the outside of the beam image as a ring. These features tend to be characterized by a sharply defined increase in intensity and therefore a large derivative in the beam map. To make the metric for isolating these features more sensitive to the fractional change in intensity, the absolute value of the percent difference is used. By calculating this parameter and comparing it to a threshold, these portions of the beam image can be isolated and removed. The percent difference was calculated between pairs of adjacent

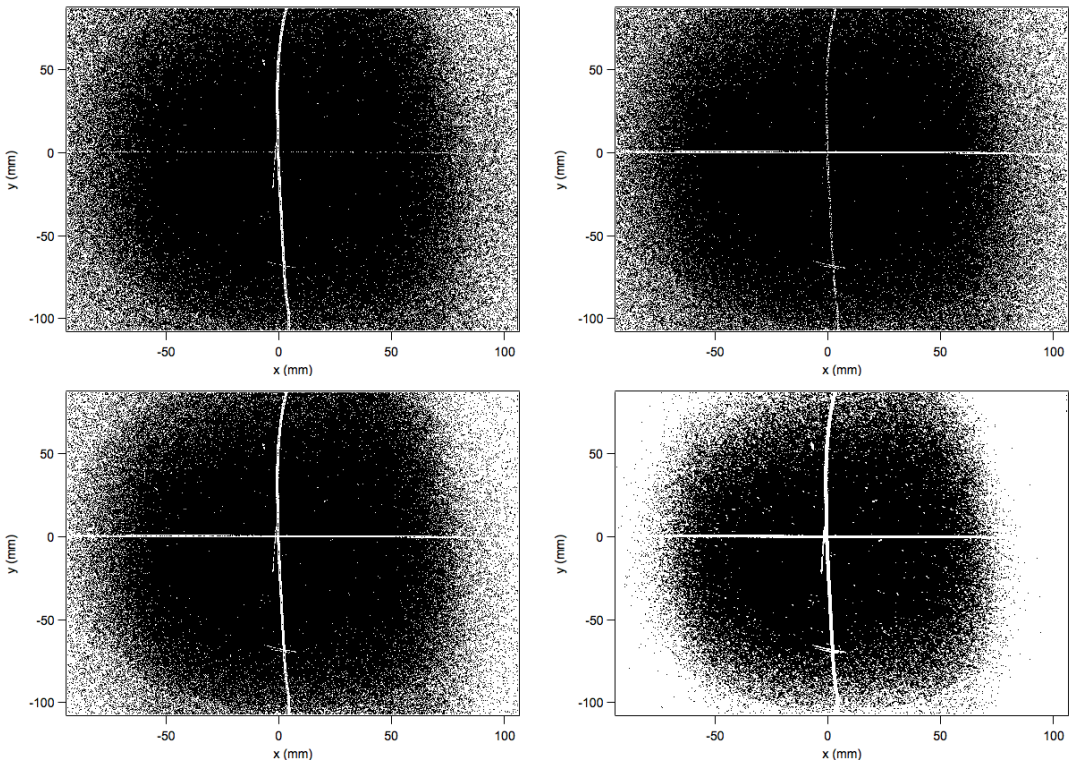
points in both the x and y directions. The threshold for removing these portions of the beam map was varied to be as low as possible while removing the dominant features in the beam image. An absolute threshold of 0.07 in the percent difference was used to produce the figures in this section. A different beam image was used in the MC, which used a threshold of 0.13. This is mentioned to illustrate that these thresholds must be adjusted for the individual beam map in order to be effective at isolating the imaging artifacts while preserving as much of the rest of the beam image as possible.

Using this threshold, a mask was made to remove the beam image defects. The mask was given either a value of 1 or Nan for parts of the beam image that should be included or excluded respectively. To be more explicit, if a particular pixel of the beam image was above the threshold, the mask for that pixel was set to a value of Nan. All of the other pixels were set to a value of 1. This allowed the mask to be applied to a beam image or for multiple beam masks to be combined by multiplying them together. Taking advantage of this, the masks for the x and y scans were combined. The resulting mask contained some points inside the crosshairs where the percent difference plateaued, which allowed these points to fall under the cut threshold. To remove these points, any portions of the mask that were adjacent to a Nan value were removed. This process was repeated until the plateaus in the beam defects were completely removed. The results of each of the steps of this process can be seen in Figure 4.1.

Once the mask is created, it is applied to the beam image, which results in the defects in the beam image being replaced with Nan values. To estimate the beam image in these regions, a matrix filter was applied that replaced any Nan values with the median in a 3x3 box around the Nan value. This filter had to be applied repeatedly to slowly fill in the regions of the beam image that were removed by the mask. Finally, the collimation is applied with a simple circular mask that is centered on the cadmium crosshairs with a radius defined by the collimator. The result at each step in this process can be seen in Figure 4.2. Kevin Coakley then applied an adaptive weighting smoothing algorithm to diminish the sampling error and decrease the effect of any defects that remained in the image, which results in the final beam image used in the simulation. The final beam image before (left) and after (right) the smoothing can be seen in Figure 4.3.

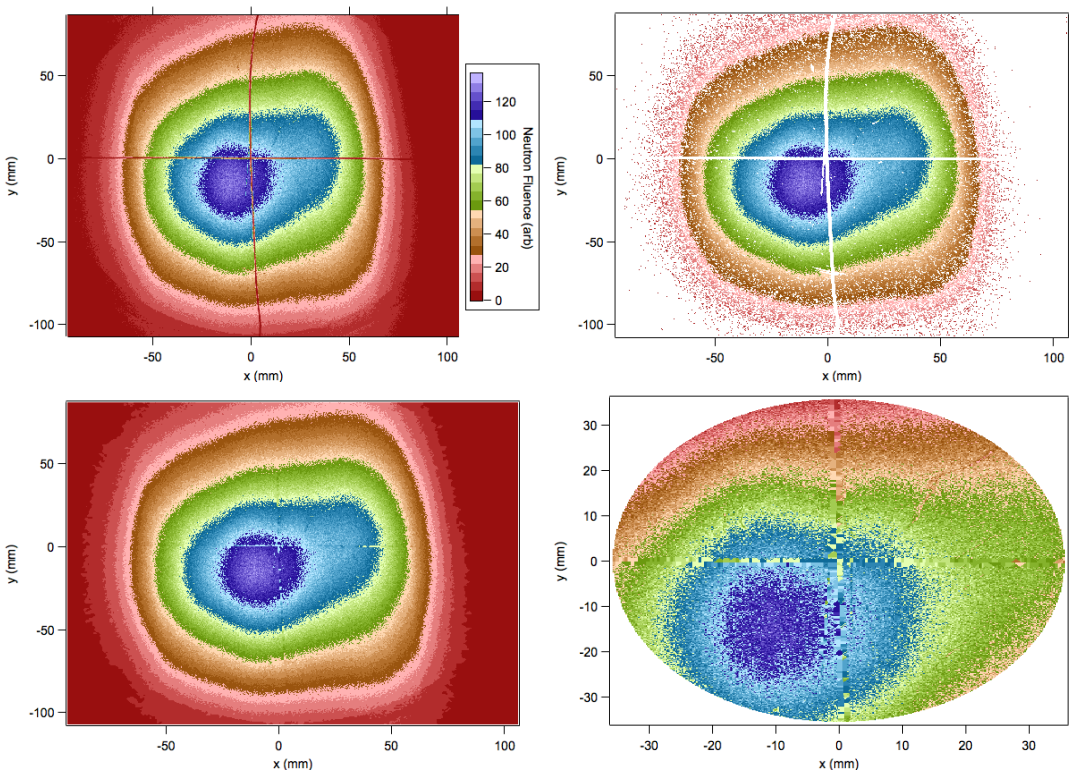
This beam image was taken years before the start of data collection. As a result, there is some question about how representative the beam image is of the beam fluence during the experiment. For example, the beam image could be sensitive to changes in alignment of various components of our apparatus or changes to the reactor infrastructure that could have occurred between data collection and when the beam images were taken. I do not know of a reason to disbelieve this beam image, and even in the case where it differs from the beam fluence during the operation of the experiment, it still shows the effect on the MC of a non-uniform, realistic beam image. Therefore, it is worthwhile to include this more realistic beam image in the MC. It

Figure 4.1: Images showing the process of creating a mask to remove imaging artifacts from the beam image. [Top Left (Right)] The mask from applying a percent difference threshold of 0.07 in the horizontal (vertical) directions. [Bottom Left] The combination of the horizontal and vertical masks. [Bottom Right] The combined mask after some final cleaning to remove features associated with plateaus inside the image defects. This cleaning also resulted in the removal of a substantial amount of the outside of the mask, which is expected to have a minimal effect on the final beam image because these portions of the beam image will be discarded by the collimation.



has been estimated, using the MC, that the realistic beam image causes a 7.5(1.6) s systematic effect[23] in the case of a static trapping configuration when compared to the uniform beam profile. This suggests that a realistic beam image is important for accurately estimating the systematic effect associated with MTN in this experiment. However, the cause of this systematic effect is a change in the ratio of MTN to fully trapped neutrons, and therefore, the flushing technique is expected to substantially reduce this systematic effect in the production data. Also, the beam image is expected to be much more representative of the beam fluence during the operation of the experiment than a uniform beam profile. Therefore, this estimate is expected to be a conservative limit on the systematic effect.

Figure 4.2: Beam images after removing the imaging artifacts and applying collimation. (Top Left) An unprocessed version of the beam image. (Top Right) The beam image after the mask is applied. (Bottom Left) The beam image after each NaN value is replaced with the median value of the 3x3 box centered on the NaN value. (Bottom Right) The rebuilt beam image after collimation is applied.

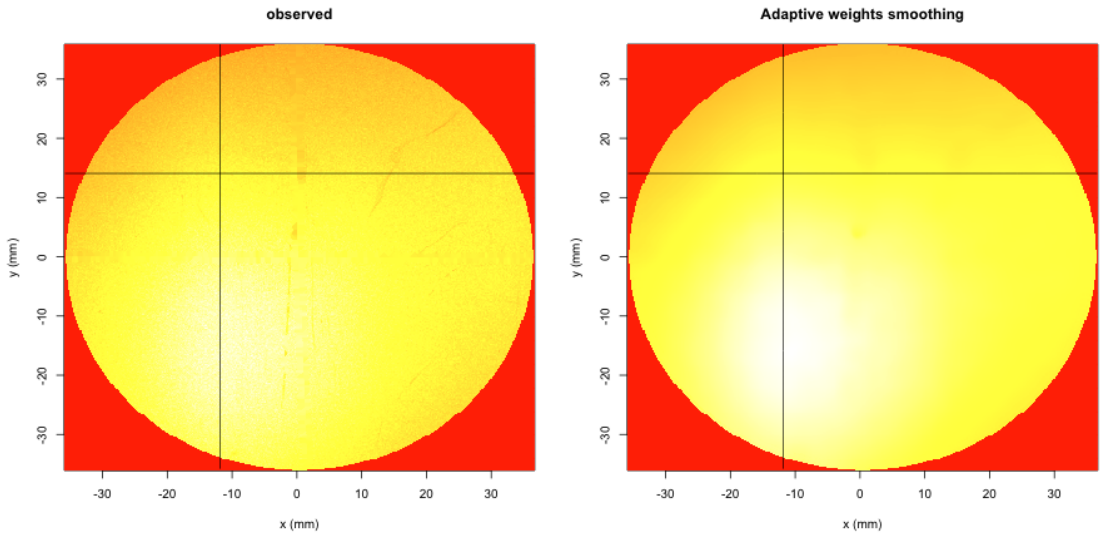


4.1.2 Modeling Divergence and the Effect of an Artificial Production Boundary

In the simulation, an artificial boundary is included in the cell that prohibits any UCN from being created above the maximum in the magnetic potential on the upstream side of the cell. Because of the way that this is implemented, it was convenient to estimate the effect of this boundary at the same time as the effect of the beam divergence. Therefore, these somewhat unrelated effects are combined below.

The divergence of the cold neutron beam will affect the initial distribution of UCN locations in the trap. This part of the simulation is described in more detail elsewhere[23]. To summarize, by assuming that the estimates of the angular distribution in the mosaic crystals of the monochromator correspond directly to the angular divergence of the cold neutrons, a maximum beam divergence of $\approx 2°$[11] was assigned to the beam. To add this to the simulation, the loca-

Figure 4.3: The final beam image used in the Monte Carlo simulation[23] before (Left) and after (Right) an adaptive weights smoothing algorithm.



tion of the cold neutron is sampled from the beam profile to determine its location as it passes through the collimator. Then, two random variables are sampled to determine the orientation of the cold neutron. In our simulation, the angles that determine the divergence were sampled from a uniform distribution in the range $0 \leq \theta \leq 2\pi$ and $0 \leq \phi \leq \phi_{max}$ although a truncated Gaussian distribution may be a more reasonable model than the uniform distribution[23]. Once the orientation of the cold neutron is determined, the final parameter to constrain creation of the UCN is the distance from the collimator to where the UCN is created. This parameter is sampled from a uniform distribution between the -37.5 cm to 37.5 cm. Let us consider the geometry of the apparatus to give these values some context. This range corresponds, at least approximately, to the location of the maximum of magnetic potential along the axis of the trap. On the downstream end of the cell, this is also the location of the start of the light guide. There is, however, no corresponding physical boundary at the upstream side of the cell. Therefore, UCN could be produced below the lower limit. However, they would be repelled from the trap. Some of these neutrons could have an initial velocity allowing them to make it over this potential barrier and into the trap, but they will tend to be very high-energy UCN and so will be lost very quickly from the trap. It was found that an upper boundary was needed for the simulation to converge. Therefore, this boundary was implemented although it could affect the results of the simulation. To estimate the systematic effect associated with the choice of the upper boundary, a sensitivity study was performed where the location of the upstream boundary was varied about the indicated value. The sensitivity of the MC to the location of

the upstream boundary was found to be $-0.55(0.40)$ s/cm[23]. Therefore, we conclude that we are not particularly sensitive to the location of this boundary.

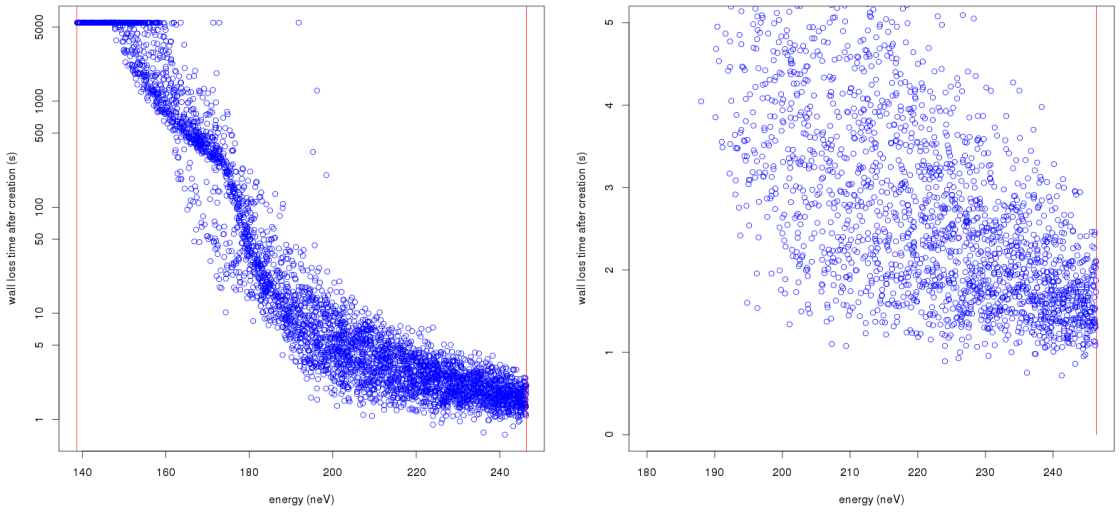
Similar to the beam profile, the effect of the beam divergence on the simulation is to change the initial energy distribution of the MTN in the trap and the ratio of MTN to fully trapped neutrons. The systematic effect associated with the beam divergence is estimated by varying the maximum angle of the divergence in the simulations and was found to be $-0.97(0.16)$ s/°[23] in the static ramping case. Once again, like in the case of the beam profile, this systematic effect is expected to be greatly reduced in the case where the magnetic trap is flushed to preferentially purge MTNs from the trap. It is important to note that the sensitivity studies are the only simulations in which beam divergence was included. Therefore, an additional correction to account for the beam divergence must be applied to all of the systematic corrections calculated from the MC. The size of this correction is expected to scale roughly with the size of the total systematic correction. Therefore, this additional correction will be scaled by multiplying by the ratio of the systematic correction in the static data over the systematic correction in the ramping data. For the static data, the correction is -2.9 s[23] for a 45 s systematic correction. Therefore, the additional systematic correction in a ramping data set with an estimate systematic correction ignoring beam divergence of Δ_{MTN} should be $-2.9 \times \frac{\Delta_{MTN}}{45s}$.

4.2 Initial UCN Velocity Distribution

The kinetic energy of the UCNs is sampled from $f(v) \propto v^2$. Qualitatively this is a statement that the UCN uniformly fill the momentum phase space. This distribution is not bounded, i.e. the number of UCNs will continue to increase at higher energies. Without a cutoff mechanism of some kind, this would suggest that an infinite number of UCNs are produced with infinite energy. This is obviously non-physical, however from the standpoint of both the simulation and this experiment there is a cutoff to this distribution. Any neutrons with total energy much above the minimum potential energy of the cell wall will very quickly strike the wall (< 1 s) and will have a high absorption probability with each wall interaction. Simulations were performed to test how quickly UCNs of different energies were lost from the trap by wall losses. The results can be seen in Figure 4.4. Above $\approx$ 200 neV, the UCNs are lost in less than a few tens of seconds. In the experiment, there is a delay between closing the beam, the end of the neutron filling stage, and the start of data collection that is ≥ 142 s. Any given UCNs must survive this duration in order to be detected. Because of their short lifetime in the trap, a negligible number of UCNs with energies greater than 200 neV are expected to make it into the data collection phase. An upper threshold of 245 neV was chosen as a conservative limit that allows some of the UCN above the less conservative threshold of 200 neV to be monitored in the simulation as a precaution. The choice of any cutoff threshold above 200 neV is expected to have a negligible

effect on the simulated survival probability curve and hence the trap lifetime.

Figure 4.4: Simulated energy-dependant survival duration for UCN in the trap due to wall losses[23] for UCN. All of the UCNs in this simulation have energies higher than the trap depth of 139 neV. Shows that above $\approx$ 200 neV no simulated neutrons extend past a few tens of seconds after their creation. This neglects β-decay, therefore the actual decay rate will be slightly faster. We also see that all of the neutrons near the lower energy threshold of the simulation are not lost due to wall interactions during the simulation.



As mentioned above, neutrons that do not strongly interact with the wall require more processor time to simulate. UCNs with sufficiently little total energy are unable to interact with the wall will never be lost by wall interaction and therefore do not need to be simulated to calculate the effect of MTNs. The magnetic field maps were used to determine the field minimum on the wall, and it was found to correspond to a potential energy of 139 neV[23]. Therefore, this kinetic energy was used as the lower limit of the velocity sampling distribution.

As an aside, Figure 4.4 also shows that there are many neutrons near the lower energy threshold, the minimum potential energy on the cell wall, that are not lost due to wall interactions during the simulation. This suggests that there are classes of MTNs that have lifetimes due to wall interactions that are much longer than the neutron β-decay lifetime. In other words, there are classes of MTNs with stable orbits that can go extremely long times without interacting with the cell walls. These UCNs are extremely computationally demanding because they are not lost before the end of the simulation, which requires that they be simulated for the full duration of the simulation. Comparatively, higher energy MTNs are lost very quickly allowing

the tracking of the UCN to be stopped early.

To summarize, the velocity of the MTN is sampled from a v^2 distribution between 139 neV and 245 neV. Once the magnitude of the velocity is selected, the orientation of the velocity vector is sampled by sampling two random variables from uniform distributions, one for each angle. This fully constrains the initial location and velocity of each MTN. The systematic effect introduced by both the lower and upper UCN energy thresholds is expected to be negligible.

4.3 Modeling Realistic UCN Trajectories

Once the initial velocity and location for an UCN are determined, it is tracked through realistic trajectories inside the apparatus. This involves modeling the potential that the UCNs experience and integrating the equations of motion to get the trajectories of the UCNs as they move through the helium. It also involves simulating a wall interaction if a MTN strikes the cell wall.

4.3.1 Evaluating the Potential

Assuming adiabatic spin transport, the magnetic potential simplifies to $V_B(x) = \mu|B(\vec{x})|$. Adding in the effect of gravity, we get a total potential for the UCN of

$$V(\vec{x}) = \mu|B(\vec{x})| + m_n g y.$$

The equations of motion are integrated to determine realistic neutron trajectories, using symplectic integration. The details of our methodology have been published elsewhere[23]. To evaluate the total potential, the magnetic field must be evaluated at arbitrary locations in the cell.

A realistic map of the magnetic field was calculated by Chris O'Shaughnessy using the Radia add-on in Mathematica. By modeling realistic wire wrappings for the magnet and numerically solving the Biot-Savart law, the magnetic field can be calculated at any point in space. For the simulation, the fields were calculated on a grid with 5 mm spacing along the axis of the cell and 1 mm spacing in the plane perpendicular to the cell axis. The field is then interpolated from this grid using a fourth order tensor spline method to evaluate the field and gradient at the location of the simulated UCN for every time step.

The same potential grid was used to create the magnetic field maps used throughout this work and to evaluate the minimum potential on the cell wall, which is the energy used to distinguish fully trapped neutrons from MTNs. More information on the magnetic field map and the interpolation method have been published previously[23, 17].

4.3.2 Wall Interactions

When a simulated UCN crosses the boundary of the cell wall, a wall interaction is simulated, which consists of two parts. First, the probability that the MTN was lost in the wall interaction is calculated. Then if the MTN survived the wall interaction, the exit angle of the UCN is determined. The models used to simulate the wall interaction will be discussed in the following sections.

Wall Loss Curve

The wall loss curve is a lookup table that lists the probability that a neutron will be lost when it interacts with the wall as a function of its perpendicular kinetic energy. It is calculated using the one-dimensional Schrödinger equation and the optical model. The optical model operates under the assumption that the neutron wavelength is sufficiently large, in comparison to the nuclear spacing in the material, that it experiences the average potential due to a large number of nuclei instead of interacting with the individual nuclei. These average potentials are calculated with[47]

$$V = \frac{2\pi\hbar^2}{m_n} \sum_i N_i a_i$$

and

$$W = \frac{\hbar v}{2} \sum_i N_i \left(\sigma^i_{abs} + \sigma^i_{loss}\right),$$

where i is the nuclear isotopes of the material, N_i is the nuclei number density, a_i is the scattering length, σ^i_{abs} is the absorption cross section, and σ^i_{loss} is the sum of the cross sections of any additional loss mechanisms. In this experiment, any neutrons that are not reflected are lost from the experiment. Incoherent scattering causes neutrons to be emitted isotropically. Therefore, half of the neutrons travel inward and the other half travel outwards. The neutrons traveling outward are lost from the experiment. To account for this, we include half of the incoherent cross section in the imaginary potential as an additional loss mechanism. An up to date set of scattering lengths and absorption cross sections can be found on the NIST website at `http://www.ncnr.nist.gov/resources/n-lengths/`. The real and imaginary potentials calculated with these tabulated values are included in Table 4.1 along with the material thicknesses.

The contribution to the imaginary potential from the inclusion of half of the incoherent scattering cross section is quite large, in particular for hydrogen. This converts the TPB and acrylic light guide from neutron reflectors to relatively strong neutron absorbers. As a result, the inclusion or exclusion of the incoherent scattering cross section in the imaginary potential has a large effect on the output of the MC simulation. We are treating this as an uncertainty in the wall loss curve in that we are accounting for it in our uncertainty budget. If the assumption

that the incoherent scattering can be modeled as an additional loss mechanism is incorrect, a correction will be required. At that point, the uncertainty in the simulation from the wall loss model will need to be reassessed, which should result in a smaller uncertainty. There is also a contribution to the uncertainty in the wall loss model that comes from the uncertainty in the incoherent scattering cross section, but we neglect that source of uncertainty here.

Table 4.1: The scattering parameters, potentials, densities, and materials thicknesses for the wall materials[48] that are needed to calculate the wall loss curves. The potentials are relative to the average nuclear potential of ^{4}He. * The indicated values were borrowed from Chris O'Shaughnessy's thesis[17].

Material	Thickness	V	W	ρ
	μm	neV	neV	g/cm^3
^{4}He (He)	∞	0	0	0.14
TPB ($C_{28}H_{22}$)	1.5	34.6469	-0.125235	1.079
Gore-Tex (C_2F_4)	2000	15.899	-1.25556E-5	0.6
Graphite (C)	1000	140.04	2.64285E-5	1.82
Boron Nitride (BN)	2000	176.641	-2.84091	2.1
Acrylic PMMA ($C_5O_2H_8$)*	∞	27.8	1.39E-3	1.19
Teflon (C_2F_4)*	2000	121.1	4.26E-5	2.15

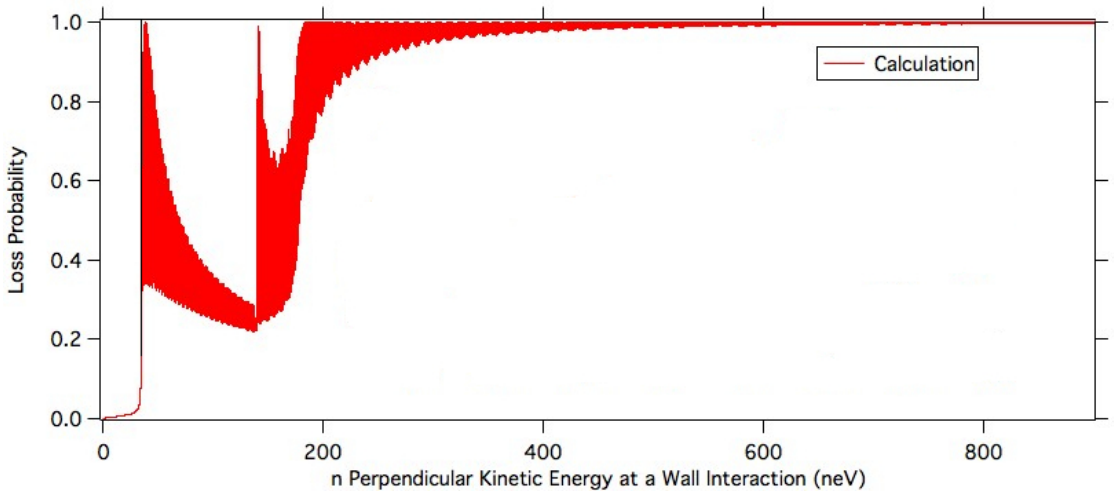
Nucleus	m	a_i	σ_{inc}	σ_{abs}
	AMU	fm	b	b
He	4.002602	3.26	0	0
C	12.0107	6.6460	0.001	0.0035
H	1.00794	-3.7390	80.26	0.3326
F	18.9984	5.654	0.0008	0.0096
B	10.811	5.30	1.7	767
N	14.006	9.36	0.5	1.9

The solution to the boundary value problem is quite simple for a single boundary and can be found in any introductory quantum physics textbook or in *Ultra Cold Neutrons*[47] in the case of UCNs. As the number of materials increases, the complexity of the analytic solution quickly becomes unwieldy. In our case with four boundaries, we decided to solve the boundary value problem numerically in *Mathematica* using code developed by Chris O'Shaughnessy. His code determines the reflection probability as a function of the perpendicular kinetic energy. The lookup table was produced by solving the boundary value problem on a grid from 0 neV to 900 neV with a step size of 0.005 neV. This wall loss curve is an input for the MC and contains all of the information about the scattering probability that is used by the MC simulation.

The wall loss curve that is used in the main MC simulation can be seen in Figure 4.5. The low-frequency step-like features are located at the difference between the individual material potentials and the potential of the LHe. Therefore, the location of these steps along the x-axis is strongly dependent on the scattering parameters. The high-frequency signal is strongly dependent on the widths of the materials. It is worth noting that the loss probability is calculated in terms of the perpendicular kinetic energy, where it is the total energy of the neutrons that are well defined. The perpendicular kinetic energy varies between bounces for the same neutron depending on location and incident angle of the neutron when it strikes the wall. Therefore, it is reasonable to assume that the high-frequency components will wash out (destructively interfere) even for a neutron of a particular energy. Instead, we expect the wall loss model and ultimately the MC simulation to be much more sensitive to the lower frequency components of the wall loss curve.

To estimate the effect of a particular wall loss curve, we compare the case of a wall loss model that includes half of the incoherent scattering cross section in the imaginary potential to the traditional wall loss model described in [47]. The difference in these two simulations corresponds to an 8 s systematic difference in the static ramping configuration. Comparing this to the effect of neglecting wall losses entirely, 45 s, we find that although this is an important parameter in the MC simulation, it is a small fraction of the total correction. That said, as with many of the other systematic effects associated with the MC, this is the systematic correction in the static configuration, and it is expected to be greatly reduced in the data sets where the magnetic trap is flushed.

Figure 4.5: The loss probability of an UNC interacting with the wall as a function of its perpendicular kinetic energy.



Uncertainty in the Wall Loss Curve

The wall loss curve depends on the scattering parameters, thicknesses of the materials, and the density of the materials. The potentials depend both on properties of the nuclei, their masses, the scattering lengths, the absorption cross sections, and incoherent scattering cross sections as well as Planck's constant and the mass of the neutron. Of these parameters, Planck's constant and the mass of the neutron are extremely well known and therefore can be ignored for purposes of estimating the uncertainty. The densities of many of these materials have a wide range of values, which are presumably due to variations in their preparation; the densities used in the calculation were supplied by the manufacturer, and are expected to be accurate. The potentials are proportional to the density. Therefore, the uncertainty in the potentials is estimated to be smaller than 10%. As we have discussed previously, the high-frequency component of the wall loss curve is sensitive to each of the material thicknesses, but because of the random nature of the perpendicular kinetic energy of the neutrons when they strike the wall, we expect this to have a minimal effect on the MC simulation. Finally, this leaves the masses and scattering parameters of the materials to be considered. Both of these quantities are tabulated for the individual isotopes and the natural abundance. By inspecting the nuclear scattering parameters, it is clear that the wall loss curve will strongly depend on the isotopic composition of the wall materials.

In order to quantify the uncertainty due to the isotopic composition, we can calculate the scattering parameters of the natural abundances using the pure isotope scattering parameters and compare the result to the tabulated values. We have found that the disagreement is as much as 50%. In particular, the incoherent scattering cross section has the largest disagreement for carbon and boron. Most of the other calculations agree at better than the percent level. The disagreement, in the case of carbon, is accounted for in the size of the error bars. However, the values for boron remain discrepant even after error analysis with a measured value of 1.7 ± 0.12 and a calculated value of 0.8 ± 0.1. It is not clear what is causing this discrepancy.

Calculating the percent difference between the scattering parameters of the dominant stable isotopes gives an indication of how sensitive the wall loss model is to the isotopic ratio for each particular element. The percent difference for the materials in the cell wall ranges almost evenly from -710% to 155%. Both ^{2}H and ^{10}B have negative coherent scattering lengths resulting in large negative percent differences of -210% and -710%, respectively. This suggests, as someone in the field would probably suspect, that the potentials are quite sensitive to the isotopic abundance. In the case of the incoherent scattering cross section and the absorption cross section, one of the isotopes usually has a much larger cross sections than the other. As a result, the percent difference in these parameters is 200% for almost all of the elements of interest. In conclusion, we find that we are relatively sensitive to the isotopic composition of our cell materials. The optical potentials are linear in the isotopic concentrations, and we have

found that the slope is large for many of scattering parameters.

We have not determined the isotopic abundances of our cell materials. The numbers used in these calculations have been included in Table 4.2. We recognize this as a source of uncertainty in the simulation, and it is a topic where additional work would be motivated for future generations of this experiment.

4.3.3 Angular Dependence of Wall Bounces

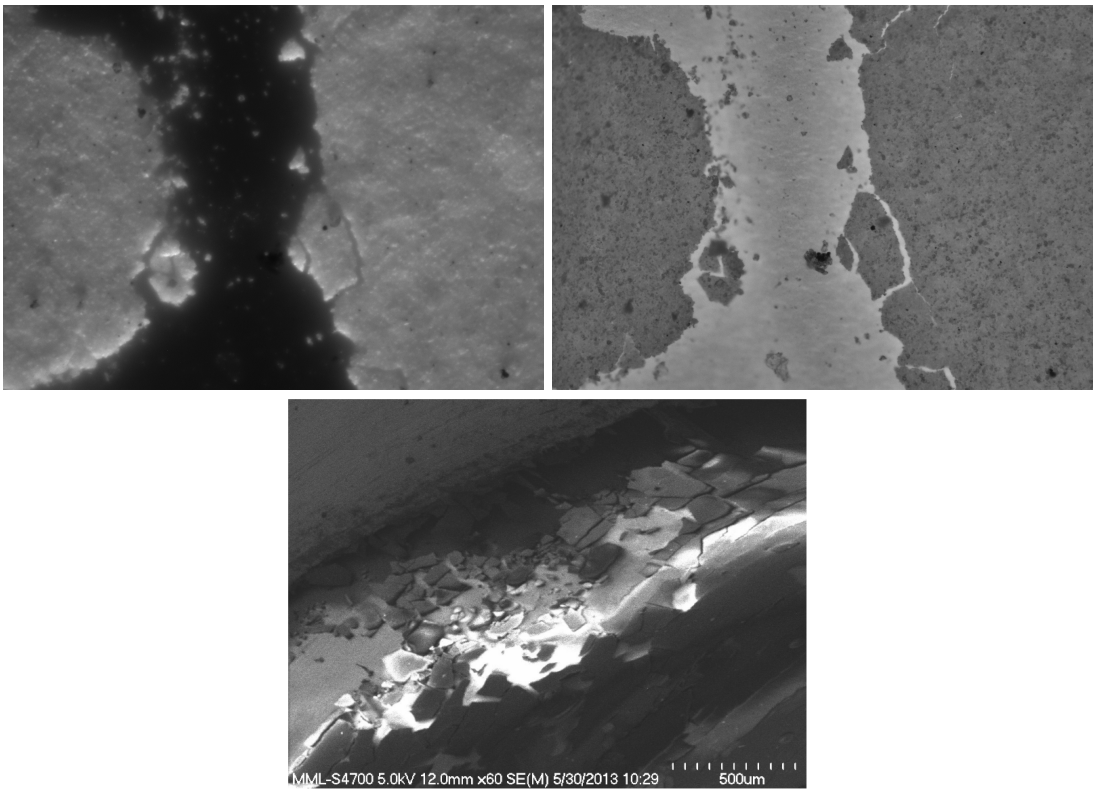
The exit angle of the UCN after a wall interaction needs to be calculated. The surface of the cell walls is expected to be rough because of the TPB coated ePTFE. Therefore, the wall scattering is expected to be diffuse. This is partially motivated by the fact that ePTFE acts as a diffuse scatterer of light, which has a similar wavelength to the UCN in this experiment. The surface roughness has been evaluated previously[17], and additional work has been done recently to image it with both scanning electron microscopy (SEM), atomic force microscopy (AFM), and visually with a microscope, see Figure 4.6 and Figure 4.7.

These images were taken with of a sample of TPB coated ePTFE that was created in the prototyping process and had been used for a few years as a demonstration piece. Therefore the poor TPB coverage and the damage to the TPB coverage is a result of the samples hard life and not characteristic of the inserts used in the experimental cell. Additionally, the inserts that were used in the cell should be substantially cleaner (the dark particles on top of the TPB coating in the top right image in Figure 4.6 are thought to be dirt or debris of some kind.)

These images span a variety of length scales and are intended to give a feeling for the characteristic length scales of the surface structure. The two visual microscope images show a 20x magnification of the ridge between two plateaus of TPB. Macroscopic structures of TPB can be seen extending from the surface. In higher quality versions of these images, the ePTFE substrate shows small imperfections, which are believed to be the porous structure that is much more easily see in the AFM images. Using a UV lamp causes the TPB to fluoresce, which very clearly shows the portions of the ePTFE that are coated in TPB. In addition, a scanning electron microscope (SEM) is another tool that has been used to evaluate the structure of the surface. It produces images that are quite objectively prettier than the other methods although I am not sure they provide any additional useful information.

On shorter length scales, AFM was used to measure the profile of the material. These figures show a variety of imaging artifacts that are not thought to be representative of the features of the material. They manifest as horizontal stripes if only one row is affected or as a gradient effect if a large number of rows are affected. They are due to the limited rate at which the AFM tip is allowed to drop between one measurement and the next. The result is that if the tip passes over a tall feature, it may continue to fall for many subsequent data points until it finally

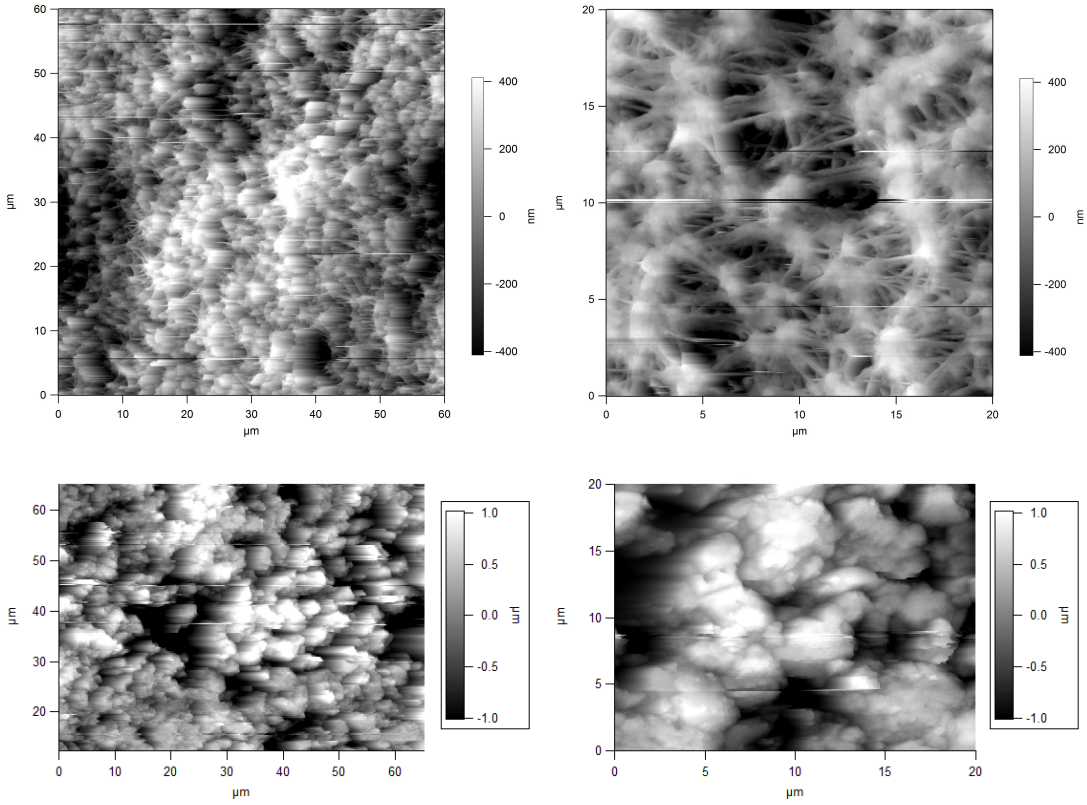
Figure 4.6: Optical and SEM images of TPB coated ePTFE inserts. The optical images are of an old, dirty sample of TPB coated ePTFE front illuminated with white (Top Left) and UV (Top Right) light, which clearly shows the portions or the substrate that are coated with TPB. In these images, small imperfections in the ePTFE substrate can be seen, which are thought to be manifestations of the more complicated features seen in the AFM images, which are presented below. (Bottom) An SEM image of a different TPB coated ePTFE sample showing similar features.



strikes a surface again. This results in a linear drop off in the AFM image for some number of points after the high features. In the affected portion of the image, the reported value is an upper limit on the actual height of the sample at that location. This effect could be minimized by adjusting the operating parameters of the AFM and to some extent with post processing. However, the images presented here are more than sufficient to make the points I want to make.

Recall that ePTFE (Gore-Tex) is a material created from PTFE (Teflon). ePTFE is manufactured by repeatedly stretching PTFE until its properties change. This is due to the formation of the very structures that are evident in the AFM images, globs of Teflon connected by a set of strands. The globs appear to be a few μm in width with strands of varying lengths of ≈ 10 μm. This complicated structure explains why light scatters diffusely from TPB; there are surface

Figure 4.7: AFM images of ePTFE with and without a TPB covering. (Top) Two images of the ePTFE substrate without a TPB coating show the porous nature of ePTFE. (Bottom) Two corresponding images of TPB coated ePTFE show a rough, globular nature to the coated substrate. These AFM images were taken by Terry McAfee, a fellow graduate student at NC State, and were carried out in non-contact mode under ambient conditions using a commercial instrument (Asylum Research MFP-3D). The AFM tips (Budget Sensors, Tap300AL-G) had a nominal radius of 10 nm and a nominal resonant frequency of 300 kHz.



normals pointing in every direction. Therefore even if the light scatters specularly from the surface, the emitted light will be approximately diffuse.

When ePTFE is coated with TPB, the result is large globular structures with characteristic widths between ≈ 2 μm to 10 μm. These images do not probe how the location of the globs of TPB fall with respect to the structure of the underlying ePTFE, and we will not discuss it here. Instead, it is enough to note that the surface is very rough on 1 μm to 50 μm length scales. We believe that this roughness is a reasonable supplemental justification for assuming that UCNs, with wavelengths that are > 40 nm, that scatter from this surface will not do so specularly. In fact, these images seem to suggest that a purely diffuse model might be more reasonable than the Lambertian model that was used in the MC.

The MC uses a Lambertian wall model, which samples the inclination angle from a $\frac{1-\cos\theta}{2}$ distribution and the azimuthal angle from a uniform distribution of 0 to 2π. To estimate the sensitivity of the MTNs to the angular distribution of the wall interactions, a simulation using the Lambertian wall model was compared to a purely diffuse model, where both the inclination and azimuthal angles were sampled from a uniform distribution. The difference was found to be 1.3(1.5) s[23]. This is considered to be a rough estimate of the systematic uncertainty in the MC from imperfect knowledge of the actual angular distribution of UCNs leaving a wall interaction.

4.4 Simulating Ramping Data

When simulating the ramping, there are a few changes that must be made to the simulation. These effects arise because some of the clever techniques that were used in the MC simulation for the static ramping profile are not valid in the case of ramping. This topic is not related to the models that are used in the MC that I participated on. Therefore, this is a little beyond the scope of this work. However, for completeness sake, I will include a brief description of a few of these effects.

In the static ramping case, the total energy of the UCN is fixed. Therefore, there is no need to simulate wall interaction for neutrons below the minimum potential on the wall. When the magnet is flushed, lower energy neutrons can interact with the wall. The location of the neutrons in momentum, position phase space is expected to evolve after the trap is filled with UCN. Therefore, to understand this population of lower energy neutrons during the ramp, they must be simulated during the entire filling and flushing phases. This causes a substantial increase in the computational difficulty of the simulation.

Additionally, in the static case, an average survival probability curve can be calculated using all of the simulated neutrons, i.e. in the simulation, you do not have to account for when the neutron was created in the filling stage because the process is invariant in time. This allows the effect of the random creation times to be accounted for in the post processing by convolving the survival probability with a randomly selected offset to account for the uniform distribution of creation times. In the case of ramping, this invariance is broken. As a result, the random creation time can not be included with a simple convolution and instead the effect must be simulated directly. This also has the effect of making the MC simulation much more computationally intensive.

These additional hits to the computational efficiency were a strong motivation for doing the sensitivity studies for the MC on the static data, which allowed a large number of sensitivity studies to be run. Scaling the systematic uncertainty in the systematic correction by the size of the systematic correction will allow these estimates to be applied to the more computationally

challenging case of the ramping data.

4.5 Conclusions

In the previous sections, I have discussed the various models that are required to simulate UCNs in the UCN Lifetime Experiment at NIST. Throughout, I have attempted to list the systematic uncertainties corresponding to each of these models as they have been presented in Kevin Coakley's paper[23]. The systematic table presented there is included for convenience, see Table 4.3. All of the systematic studies that quantified the systematic uncertainties in the MC are the work of Kevin Coakley. As I stated previously, my contribution was primarily focused on constraining the models that were used in the simulation and estimating reasonable bounds for each of the parameters of the models to allow sensitivity studies to be performed.

From this work, we see that the simulation, and hence the experiment, appears to be most sensitive to our imperfect knowledge of the loss probability curves for wall interactions. However, the majority of the other systematic effects contribute at a similar, if somewhat smaller, level. In the static ramping case, the sensitivity studies suggest that the percent systematic uncertainty in the systematic correction is $\approx 20\%$. In the case of the ramping data, the systematic uncertainties from the model assumptions are expected to be smaller because the MTN will be a smaller fraction of the total UCN population. If we assume a linear scaling between the size of the systematic correction and the systematic uncertainty in the systematic correction, this factor of 1/5 will be the relative size of the uncertainty even in the more aggressive ramping schemes. This suggests that the size of the uncertainty in the systematic correction in the $70-35-70\%$ ramping data will be sufficient for a 1 s measurement of the neutron β-decay lifetime. For a more complete discussion of this topic, see the chapter on systematic effects.

Table 4.2: Detailed scattering parameters for isotopes that compose the wall loss materials that were used to assess the sensitivity to the isotopic abundances of the cell wall materials. These values are from `http://www.ncnr.nist.gov/resources/n-lengths/`.

Isotope	conc	b_{coh}	σ_{inc}	σ_{abs}
		fm	b	b
B	NaN	5.3000	1.7000	767.0000
^{1}0B	0.20	-0.1000	3.0000	3835.0000
^{1}1B	0.80	6.6500	0.2100	0.0055
% Dif	NaN	-2.0611	1.7383	2.0000
B Calc	NaN	5.3000	0.7680	767.0044
H	NaN	-3.7390	80.2600	0.3326
^{1}H	1.00	-3.7406	80.2700	0.3326
^{2}H	0.00	6.6710	2.0500	0.0005
% Dif	NaN	-7.1059	1.9004	1.9938
H Calc	NaN	-3.7390	80.2583	0.3326
He	NaN	3.2600	0.0000	0.0075
^{3}He	0.00	5.7400	1.6000	5333.0000
^{4}He	1.00	3.2600	0.0000	0.0000
% Dif	NaN	0.5511	2.0000	2.0000
He Calc	NaN	3.2600	0.0000	0.0000
C	NaN	6.6460	0.0010	0.0035
^{1}2C	0.99	6.6511	0.0000	0.0035
^{1}3C	0.00	6.1900	0.0340	0.0014
% Dif	NaN	0.0718	-2.0000	0.8816
C Calc	NaN	6.5847	0.0000	0.0035
N	NaN	9.3600	0.5000	1.9000
^{1}4N	1.00	9.3700	0.5000	1.9100
^{1}5N	0.00	6.4400	0.0001	0.0000
% Dif	NaN	0.3707	1.9996	1.9999
N Calc	NaN	9.3592	0.4982	1.9029
O	NaN	5.8030	0.0008	0.0002
16O	1.00	5.8030	0.0000	0.0001
17O	0.00	5.7800	0.0040	0.2360
18O	0.00	5.8400	0.0000	0.0002
O Calc	NaN	5.8031	0.0000	0.0002

Table 4.3: Sensitivity studies for the MC simulation[23].

Effect	Correction	Uncertainty
	s	s
Wall loss model	none	8
Angular reflection model	none	1.3(1.5)
Beam divergence	-2.9	2.9
Beam profile	none	NA
Time step	none	1.9
Choice of z_{min}	none	1
total	-2.9	8.9

Chapter 5

Systematics and Results

The first general class of systematic effects is neutron loss mechanisms other than neutron β-decay. This can be further broken into two types that are treated in different ways, exponential and non-exponential loss mechanisms. The simpler of these is exponential loss mechanism, of which ^{3}He capture and thermal upscattering are examples. In these cases, the lifetime of each of the processes needs to be estimated, and then they can be combined to estimate the β-decay lifetime by assuming that the neutron population decays according to a series of exponentials multiplied together where the lifetimes of the exponentials correspond to the lifetimes of the individual mechanisms. This results in the commonly used expression $1/\tau_{trap} = 1/\tau_\beta + \Sigma_{i=1}^{N} 1/\tau_i$, where the sum is over i the exponential loss mechanisms. The inherent assumption in using the exponential model is that the decay rate follows the first order differential equation, $\dot{N(t)} = -N(t)/\tau_{trap}$, i.e. the decay rate only depends on the lifetime and the number of particles that remain in the trap. These types of loss mechanisms, therefore, can not depend on any information about the neutron, for example, its location, its energy, the magnetic field it is experiencing, its proximity to other things, etc.

If a particular loss mechanism affects neutrons in a way that depends on the neutron's properties or that only interacts with a fraction of the neutron population, the loss mechanism will not follow this simple exponential form. These non-exponential loss mechanisms can not be combined in the same way. In this case, a different method[23] is used to estimate the systematic error and uncertainty. This method was developed primarily for understanding the effect of MTNs, but could also be used for any other non-exponential loss mechanism, for example depolarization through Majorana spin flips.

An alternative type of systematic effect can arise if the experiment detects a varying fraction of the decay events as a function of time, i.e. has a time-dependent detection efficiency. There are a variety of mechanisms that can cause this type of systematic effect including the locations of the neutron decays changing in time coupling with the spatially varying light detection efficiency

and either the hardware pulse height discriminator or the software pulse shape thresholds coupling with a time-dependent gain drift.

Accidental coincidence between background events and a neutron events can have a similar effect on the data by pushing events across a cut threshold. If the background rates vary in time, the fraction of neutrons that are pushed across the cut threshold will also vary in time resulting in a similar systematic effect. Possible mechanisms include aluminum activation from the cryostat, fluorine activation from the Teflon vacuum windows, and neutron-induced luminescence. This is expected to be a smaller effect than the gain drift of the main detection PMTs because it requires accidental coincidence. The rate of accidental coincidence events of a few types are estimated.

The remaining systematic effects are sufficiently unique, and the estimates of their systematic corrections are sufficiently different that there is no benefit in grouping them with other systematic effects. Instead, they are listed individually below.

Error in the calibration of the DAQ clock will cause a corresponding systematic error in the extracted β-decay lifetime. This effect is assessed by looking at events with a well-understood timing distribution and comparing the expected timing to that reported by the DAQ.

The analysis uses a background subtraction method to decrease the background rate in the experiment in a model-independent way. Specific background data is collected that mirrors the background rates in the trapping data as accurately as possible while minimizing the population of UCN. This background data, or non-trapping data, is then subtracted from the trapping data to remove the contribution of these background events. Any difference between backgrounds for these two data types will result in an imperfect background subtraction and a corresponding change in the decay rate histogram, which will lead to a systematic effect. Also, any changes in the detection efficiency for these two types of data could also result in an over or under correction of the backgrounds and will similarly create a systematic effect. The size of this effect is estimated using data taken where the population of UCN was greatly reduced by operating the helium bath around 1 K, where the superthermal upscattering is a substantial loss mechanism that depopulates UCN from the trap. This allows the background subtraction to be performed on data without UCN and thereby allows a clean test of the background subtraction. By comparing the background subtracted decay rate histogram in this warm data to a constant, a limit can be placed on the size of systematic effects associated with an imperfect background subtraction.

The fact that the neutron β-decay takes place inside matter and in the presence of magnetic fields could also modify the trap lifetime. Spectator particles can allow higher order processes to affect the decay rate. Alternatively, the presence of the helium nuclei in the helium bath can change the available phase space for the decay products. Both of these effects can affect the measured lifetime. They have been calculated previously by this collaboration, and those

calculations will be referenced.

Fitting the decay rate histograms to an exponential could cause fitting bias that introduce a systematic error in the extracted neutron β-decay lifetime. Measuring the decay rate for too short of a duration can introduce a systematic effect. Additionaly, low count rates can introduce a systematic error if not proprly accounted for. These effects are estimated using simple MC simulations that use data and background rates that are estimated directly from the data.

In the following sections, each of these systematic effects is described and estimated in detail. The contributions from the various systematic effects are combined in Section 5.13, which starts on page 170. Finally, a discussion of the disagreement between the experimental result after accounting for the systematic effects and the current Particle Data Group world average is discussed.

5.1 Thermal Upscattering

In a process that is very similar to the superthermal production that populates the trap with UCN, UCN can be upscattered to energies that are sufficiently high that they are no longer bound in the trap by interacting with the helium bath. These interactions between the UCN and the helium bath are through interaction with the phonon states, the quanta of excitation in the helium. The UCN do not have enough energy to create a phonon, i.e. the single phonon process is kinematically forbidden. However, an UCN can scatter from an already present phonon, which is a second order process. The theory and measurements of the thermal upscattering rate have not progressed since the last generation of this experiment. The theory predicts a T^{-7} dependence to the lifetime due to the two phonon processes[49]. Measurements of the upscattering lifetime are in rough agreement with the theory[50, 16].

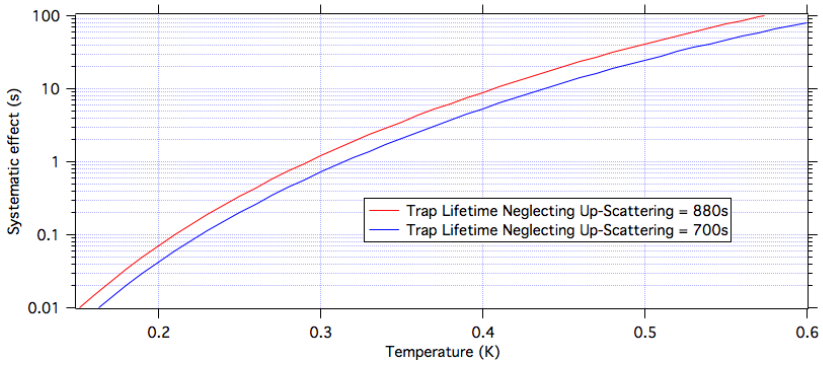
The second generation of this experiment measured the upscattering rate in their apparatus and found it to be consistent with the previous measurements[16]. The majority of the measurements by both of these groups showed a smaller than expected loss rate where it was the coldest data points that showed the best agreement. The author suggested that this could be an indication that at higher temperatures the rotons began to contribute to the physics, pushing any ^{3}He contamination from the measurement cell into the dilution refrigerator where it would not affect the trap lifetime[16].

The quality of the data below 1 K is suspect and, while the measurement and theory disagree by a factor of 2-3, the different theories disagree by an order of magnitude[49]. Despite this disagreement, the clearest path forward is to use the T^{-7} dependence with a coefficient determined from the data in [50]. Using Data Thief, the upscattering data is extracted and fit to the distribution $\lambda = B \times T^7$, which resulted in the estimate $B = 0.78(0.02)\ \times 100/\text{s}$. Using a MC simulation, this is converted to the coefficient in the lifetime $\tau_{up} = A \times T^{-7}$ with a value of

$A = 129(3)$ s. Also, there are contributions to the uncertainty that come from the uncertainty in the temperature measurement, which includes statistical uncertainties as well as systematic uncertainties.

After inflating the uncertainties, I estimate the temperature of the trap to be 280(20) mK. The central value is an estimate of the median temperature during the production data files, and the uncertainty is an estimate of the combined uncertainty of the size of the calibration error, any other systematic uncertainties, and the statistical uncertainty. The statistical uncertainty in the temperature measurements has been estimated by calculating the standard deviations of the temperature log during a data file; the resulting uncertainty was 0.006 K. In light of this measurement, the vast majority of the estimated uncertainty comes from the systematic effects with the primary contribution being uncertainty in the thermometry calibration. Using these values and the method described in Section 5.13, the systematic effect associated with thermal upscattering is estimated to be 0.7(0.4) s after scaling the effect to the Particle Data Group mean lifetime.

Figure 5.1: A sensitivity estimate for the thermal upscattering of UCN via the two phonon scattering process with the ^{4}He in the cell. The lifetime of the process has a T^{-7} temperature dependence. Estimates of the systematic effect were calculated with both a 880.2 s and 700 s lifetime to give a feeling of the systematic effect in the trap lifetime range of this experiment.



To reduce this systematic and its uncertainty further, upscattering data should be taken at least to the operating range of the experiment and the operating temperature of the experiment should be reduced further if possible. Additionally, the calibration of the thermometry should be tested.

5.2 ^{3}He Absorption

^{3}He has a large capture cross section for neutrons, and therefore a small amount of ^{3}He contamination could result in a substantial loss mechanism and corresponding systematic effect in this experiment. The lifetime associated with neutrons inside a medium with an absorption cross section[47] can be extended to the case of ^{3}He via

$$1/\tau_{abs} = N_3 \sigma_3 v,$$

where τ_{abs} is the lifetime due to ^{3}He absorption, v is the velocity of the neutrons, N_3 is the ^{3}He number density, and σ_3 is the ^{3}He absorption cross section. The absorption cross section is inversely proportional to the amount of time that the neutron spends in the vicinity of the ^{3}He nucleus and therefore on the velocity. This cancels out the velocity dependence in the lifetime causing the lifetime to be independent of the neutron velocity. Therefore, the absorption cross section can be tabulated for thermal neutrons and then scaled to the appropriate energy according to $\sigma_3 \equiv \sigma_{th} v_{th}/v$, where $v_{th} = 2200$ m/s and $\sigma_{th} = 5333 \pm 7$ b[51]. Finally, the ^{3}He number density can be rewritten in terms of the density of the liquid and the ratio of ^{3}He to the total number density, $N = N_3 + N_4 \approx N_4$, via $N_3 = \frac{N_3}{N_3+N_4}\rho_4(T) \approx R_{34}\rho_4(T)$, where R_{34} is the ratio of the number density of ^{3}He to ^{4}He and $\rho_4(T)$ is the ^{4}He density, which is constant throughout our temperature range at 0.145 g/cm^3[52]. This leads to the relation

$$\frac{1}{\tau_{abs}} = R_{34}\rho(T)\sigma_{th}v_{th} = 2.6 \times 10^7 R_{34}[\mathrm{s}^{-1}].$$

Plugging in our values we get

$$\tau_{abs} = 3.9 \times 10^{-8}/R_{34}[\mathrm{s}].$$

Using the properties of exponentials, we can relate the lifetimes of interest with

$$1/\tau_{trap} = 1/\tau_{n+\mathrm{other}} + 1/\tau_{abs},$$

and we define the systematic effect due to ^{3}He absorption using the relation

$$\tau_{n+\mathrm{other}} = \tau_{trap} + \Delta\tau_{abs}.$$

In the above, τ_{trap} is the trap lifetime, $\Delta\tau_{abs}$ is the systematic effect due to ^{3}He capture, and $\tau_{n+\mathrm{other}}$ is the lifetime including all systematic effects other than ^{3}He capture. Using these relations, the systematic effect can be extracted as a function of R_{34} and either the trap lifetime

or the true lifetime once all other systematic effects have been accounted for with

$$\Delta\tau_{abs} = \tau_{trap}\left(\frac{\tau_{abs}}{\tau_{abs}-\tau_{trap}} - 1\right)$$

or

$$\Delta\tau_{abs} = \tau_{n+\text{other}}\left(1 - \frac{\tau_{abs}}{\tau_{abs}+\tau_{n+\text{other}}}\right).$$

Before presenting the size of the systematic correction, let us discuss the value of R_{34}. At present, no experiments have performed a direct measurement of a sample with the purity below $R_{34} < 1\times10^{-15}$, which we attribute to the difficulty of a measurement at this level. Historically, the only way to measure purity below the $R_{34} = 1\times10^{-12}$ level was indirectly by taking the purified sample, processing it a second time through the heat flush to concentrate any contamination into a smaller sample, and then perform mass spectroscopy on the concentrated sample at the $R_{34} = 1\times10^{-12}$ level. Using this method, McClintock reports a limit of $R_{34} = 5\times10^{-16}$[53] for a batch processing purifier and $R_{34} < 5\times10^{-13}$[54] for a continuous flow purifier. Although the purities measured previously, at least for the batch purifier, seem to be sufficient for use in this experiment, it was desired to do a direct measurement of the isotopically pure ^{4}He used in this experiment.

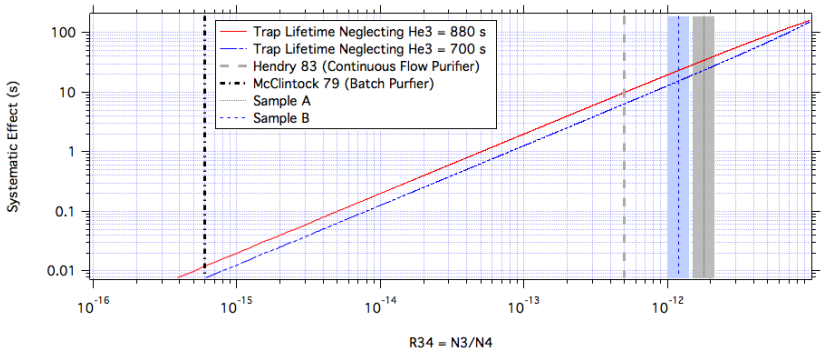
An effort to measure our isotopically pure ^{4}He samples directly continues in collaboration between the ATLAS staff and the collaboration of the UCN Lifetime Experiment at NIST[55]. A sample of isotopically pure ^{4}He was taken directly from the apparatus (A), which may have been contaminated due to unfortunate circumstances that occurred during the extraction, was measured to have a purity of $R_{34} = 1.8(0.6)\times10^{-12}$. A secondary sample (B), which was never admitted to the apparatus and therefore is the least likely to have been contaminated, was measured to have a purity of $R_{34} = 1.2(0.4)\times10^{-12}$. The plot of these measurements and the corresponding systematic uncertainties are shown in Figure 5.2, and the systematic effects are presented in Table 5.1.

The measurement of sample B is expected to be a best case scenario for our experiment. Although it is possible that sample B was contaminated by events that did not contaminate sample A, we consider it unlikely. We are less confident that sample A is the worst case scenario. Because of the circumstances surrounding the extraction of sample A, it is believed that the helium in the sample may not have been representative of the helium in the cell during operation. The sample was the first helium extracted from the cell, and it was extracted when the apparatus was being warmed up more quickly than the apparatus was designed to be warmed. A consequence of this is large thermal gradients that we believe would be in the direction to cause a heat flush to purify the sample that was extracted. The thermometry was not designed to accurately constrain the thermal gradients in this atypical operating configuration, which

Table 5.1: Systematic effect corresponding to measured helium purities from our apparatus and other sources. The column labeled Effect (τ_{trap} = 700 s) shows the systematic effect in the case of our data, which has a trap lifetime of approximately 700 s. In contrast, the column Effect ($\tau_{n+\text{other}}$ = 880.2 s) indicates the size of the systematic effect in the case where helium absorption is the dominant systematic effect in the experiment.

Sample	Effect (τ_{trap} = 700 s)	Effect ($\tau_{n+\text{other}}$ = 880.2 s)
	s	s
A	23.	34.
B	15.	23.
Hendry87	6.3	9.8
McClintock78	0.008	0.012

Figure 5.2: A sensitivity estimate for the ^{3}He absorption of UCN. $R_{34} = N_3/N_4$, the ratio of the number densities of ^{3}He and ^{4}He. Various purity measurements are included to show the state of the measurement techniques.



prevents meaningful quantitative estimates from being calculated. Although we can not prove that a heat flush caused the sample to have less contamination than the helium in the cell during operation, it is plausible. Without a sample that is thought to be characteristic of the actual helium in the apparatus during operation, the only obvious choice is to list the systematic effect associated with sample A as a reasonable assumption with the caveat that it is possible that the actual systematic effect was much larger than is calculated from this sample. Our collaboration proposes that this effect is the explanation for the large disagreement between the Particle Data Group world average and our measured value. The purity measured in sample A of $R_{34} = 1.8(0.6) \times 10^{-12}$ corresponds to a systematic correction of $\Delta_{abs} = 29(10)$ s when scaled to the Particle Data Group β-decay lifetime.

If we take the trap lifetime, correct for the other dominant systematic effects, and then use the resulting lifetime to calculate the concentration of ^{3}He required to account for the systematic difference between the resulting lifetime and the PDG world average, the result is

$R_{34} = 3.9 \times 10^{-8}/\tau_{abs} = 1.14 \times 10^{-11}$. This value is higher than what was measured from the previous apparatus of $R_{34} = 4.2 \pm 1.5 \times 10^{-12}$, which would require that our helium was further contaminated in the decade since those measurement samples were taken. This is plausible. Therefore, our conclusion is that not only is it conceivable that the ^{3}He absorption systematic is substantially larger than what was measured, but it could also account of the disagreement between our measured value and the Particle Data Group world average. Furthermore, it seems plausible that a substantial portion of the systematic effect that was observed in the previous experiment could have been caused by ^{3}He contamination because it seems unlikely that the entire systematic difference was caused by MTNs as they suggested[16]. This reevaluation of their work is motivated by the results of the Monte Carlo simulation, which as the Monte Carlo has been refined has continued to report systematic corrections that are too small to account for the discrepancy between their trap lifetime and the PDG mean lifetime. Their apparatus differs somewhat from ours, and we have not rerun the simulation with their geometry. However, from the results of our sensitivity studies, it seems unlikely that the differences between the two apparatuses could change the systematic effect of MTNs enough to account for the discrepancy that they saw in their data. This situation provides strong motivation for producing a new batch of isotopically pure ^{4}He that is experimentally verified to have a purity of less than $R_{34} \approx 1 \times 10^{-15}$ before taking additional data is warranted. This has motivated the creation of a new purifier to produce a new batch of isotopically pure ^{4}He for operating the UCN Lifetime Experiment at NIST, which is discussed in Chapter 6.

5.3 Marginally Trapped Neutrons

Marginally trapped neutrons (MTNs) are a population of neutrons that have more energy, but only slightly more, than the magnetic potential energy of the trap. Having more energy than the potential energy of the trap allows the MTNs to interact with the cell wall. In these wall interactions, the MTNs can be lost from the trap, which introduces a loss mechanism and systematic effect into our experiment. If the MTNs were quickly lost from the trap, they might not survive into the data acquisition phase, in which case they would not be measured and would not introduce a systematic effect. Having slightly more energy than the potential energy of the trap, the MTN's interact infrequently and weakly with the walls. This allows these MTNs to survive into the data acquisition phase, which introduces a systematic effect in our experiment if they are lost during the data collection phase.

This systematic effect is strongly energy dependent and therefore does not have an exponential time dependence. A method for limiting the population of MTNs has been developed, which reduces the corresponding systematic effect. After the cell is filled with UCN and the beam is closed, the quadrupole magnet is ramped down. This increases the size of the UCN

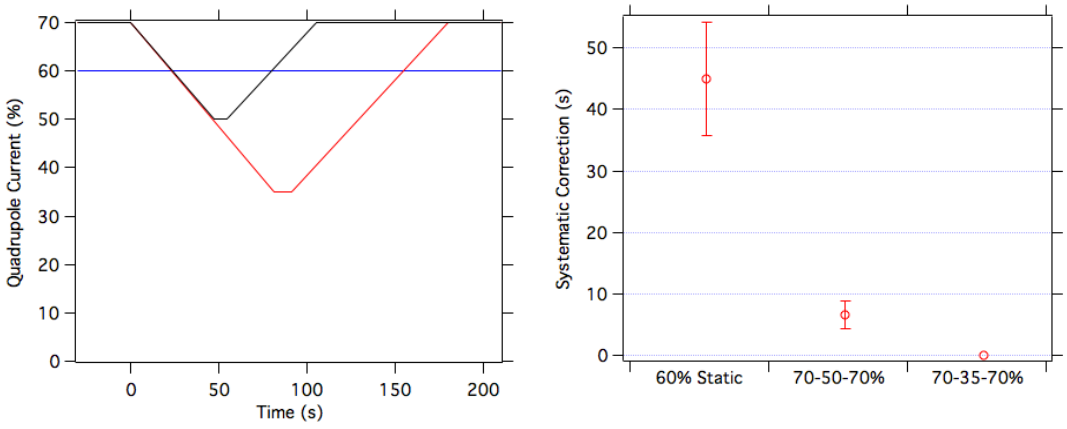
orbits causing them to interact more frequently and strongly with the cell walls, which increases the rate that UCN are purged from the trap. The magnetic field is left at a lower strength for ≈ 10 s, and then the quadrupole is ramped back up to operational strength. The size of the orbits is reduced, and the wall losses decrease. This ramping or purging of the trap preferentially removes MTNs. By carefully selecting the ramping schedule of the trap, the population of MTNs and the corresponding systematic effect can be greatly reduced. By ramping the magnetic field to approximately half of the starting field, the systematic effect can be reduced well below the level required for a 1 s measurement of the neutron β-decay lifetime.

To estimate this systematic effect, a detailed Monte Carlo simulation was developed by Kevin Coakley at NIST, Boulder. The Monte Carlo uses detailed models of the beam, cell materials, and magnetic fields to track simulated neutrons through realistic trajectories inside the apparatus. A survival probability curve due to wall losses is tabulated, and in post processing, it is convolved with the neutron β-decay to estimate the systematic effect associated with the MTNs. Additionally, the Monte Carlo allows new ramping profiles to be developed and for the associated systematic effect to be understood before using the precious beam.

Simulations were run, and data was taken in three primary ramping configurations, see Figure 5.3. The specific ramping configurations that were chosen in this work were motivated by a large number of constraints, and alternative choices may have resulted in a clearer understanding of how the systematic correction scales with the ramping profile. For example, if the 60% static ramping profile had been 70% static instead, it would have made a simpler comparison between the cases where the magnets were and were not ramped. This compromise primarily stemmed from a desire to more quickly acquire statistics with a deeper trap potential and the magnets being less stable early in operation. The 60% static data was the first data taken with this apparatus, and at the time, the magnet could not be operated more aggressively because of the threat of quenching. When the magnets became more stable and ramping was implemented, it was decided that having the greater statistics of a deeper trap was more important than having a more straightforward comparison to the static data that had already been gathered. Therefore, $70-35-70\%$ and $70-50-70\%$ ramping data was taken instead of $60-30-60\%$ data. Actually, some $60-30-60\%$ data was collected. However, not enough data was taken with this ramping scheme to get a statistically meaningful result, and therefore, there was little utility in simulating this ramping scheme.

If the magnets were not ramped, the systematic effect was found to be large. For the static 60% data, the systematic effect was 45(2.2)(8.9) s, where the first uncertainty is the statistical uncertainty in the simulation and the second is an estimate of the systematic uncertainty in the simulation due to inaccuracies in the modeling assumptions. The systematic effect in the most aggressive ramping scheme included here, the $70-35-70\%$ ramping data, is reduced to $\leq 0.03(0.01)$ s. The systematic correction, in this case, is estimated by scaling the systematic

Figure 5.3: Sensitivity estimate for the MTN systematic effect. (Left) Three ramping schemes in which data was collected and simulations were performed. (Right) The corresponding systematic corrections for the ramping schemes shown. A similar study has been published elsewhere[23].



correction in the $70-50-70\%$ ramping case by the relative size of variations in the survival probability curves that are simulated in the Monte Carlo simulation. Also, the statistical uncertainty has not been estimated, and the systematic uncertainty has been estimated by scaling the value for the $70-50-70\%$ by the ratio of the systematic corrections. Below, I will discuss the computational challenges in estimating the systematic correction in data sets with greatly reduced MTN populations, which is what motivates the use of these scaling arguments instead of a brute force simulation.

As the ramping schemes become more aggressive and the systematic effect associated with the MTNs decrease, the computational difficulty of the Monte Carlo increases rapidly. This is because in these ramping schemes, few MTN survive until the data acquisition phase and very few survive until the end of the data acquisition phase. In order to obtain a statistically well-determined result, the survival probability curve must contain the contribution from a large number of MTNs that have survived until the end of the data acquisition phase. In order to assess the statistical uncertainty in the systematic correction, the number of simulated MTNs must be increased by a large factor in order to use a boot-strapping resampling method to assess the statistical variations by observing how the results vary under resamplings of the simulated data. This use of computational resources was deemed inappropriate, and therefore the scaling estimates were used instead.

The scaling arguments are thought to be reasonable and are expected to be accurate to within a factor of five. The systematic effect in the $70-50-70\%$ ramping data already does not dominate our systematic corrections, and we believe that the systematic effect in the $70-$

$35 - 70\%$ data will be greatly reduced. We consider this to be sufficient evidence that the systematic effect due to the MTNs is already sufficient for a 1 s measurement of the neutron β-decay lifetime. If in the future, this is found not to be the case, an even more aggressive ramping scheme can be implemented to reduce this systematic effect further.

There is one other effect that has to be accounted for with the MTN systematic correction. The simulations that are used to estimate the systematic effect of MTNs do not include beam divergence. A sensitivity study was run with beam divergence in the static case[23]. However, as opposed to the other models that were studied, in the case of the beam divergence the most accurate model, which corresponded to a 2° beam divergence, was not introduced into the simulation that was used to calculate our systematic effects. Therefore, this effect has to be accounted for after the fact by applying a correction to the MTN systematic correction. In the static case, the 2° beam divergence resulted in -2.9 s systematic correction. The size of the correction is expected to be smaller in the ramping configurations that have smaller systematic corrections for the MTNs. Therefore, the size of the systematic effect is estimated as the correction in the static case scaled by the ratio of the systematic correction in the ramping case and the systematic effect in the static case. This is similar to what was done with the errors from the model assumptions that were calculated in the static ramping configuration and scaled in the ramping cases. For the static case, the correction due to the beam divergence results in about a 5% reduction in the total systematic correction. Because of this linear scaling, the relative reduction in the systematic correction is the same in the ramping cases.

In each of the ramping schemes, the estimates for the systematic effects have been for the uncorrected trap lifetime. In order for the size of these corrections to be more comparable to the other systematic effect, I will scale their values to the PDG world average. The results can be seen in Table 5.2.

Table 5.2: Systematic corrections due to MTNs for the different ramping schemes. The values are tabulated for the trap lifetimes, τ, which are included for convenience, and after scaling the values to the PDG mean lifetime of 880.2 s. The parentheses are the statistical uncertainty and the square brackets are systematic uncertainties due to model assumptions in the simulation.

Ramping	τ	$\Delta_\tau(\tau)$	$\Delta_\tau(880.2)$
	s	s	s
60%	671(45)	45.0 (2.2)[8.9]	59.0(2.9)[11.7]
$70 - 50 - 70\%$	656(21)	6.6(1.9)[1.3]	8.9(2.5)[1.8]
$70 - 35 - 70\%$	748(57)	0.03[0.01]	0.04[0.01]

5.4 Imperfect Gain Correction

The system used to measure the gain is described in Section 2.11, which starts on page 52. The method of calculating the gain and correcting the pulse area and pulse height is discussed in Section 3.2.4, which starts on page 90. Here the systematic effect due to imperfections in the gain correction is estimated. This is done by comparing the extracted lifetime for two methods of calculating the gain in carefully chosen data sets. The methods of calculating the gain will be discussed, then the differences in the two methods will be used to estimate the sensitivity of the trap lifetime to our gain calculation.

The first method, which was developed by Karl Schelhammer and is described in his thesis[18], fits the gain to a double exponential with a constant offset. The gains do not fit well to a single exponential. However, the double exponential fit appears to capture the trend in the gain drift. The second method uses a Loess time averaged method, which was discussed in more detail in Section 3.2.4, which starts on page 90.

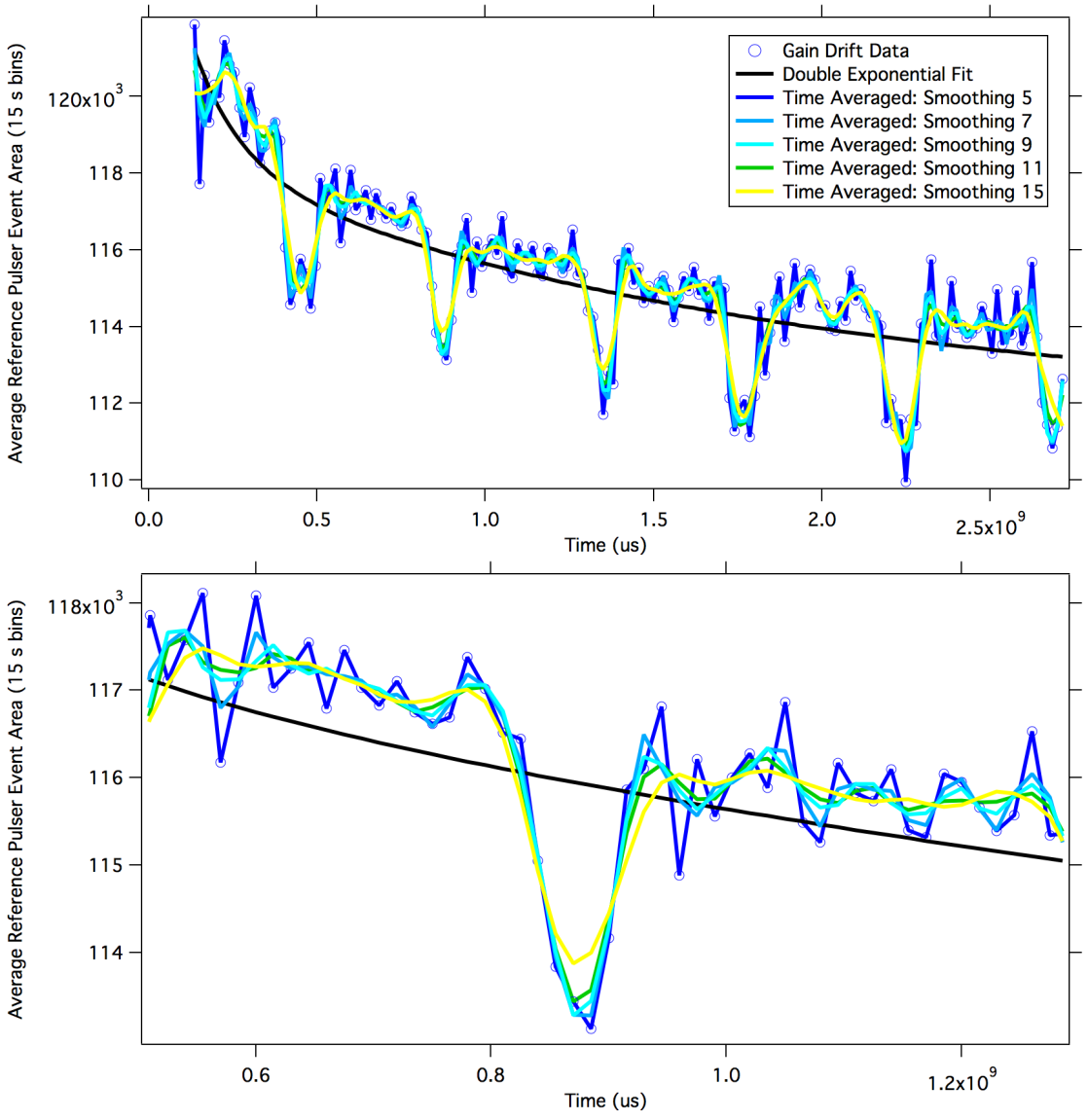
These methods have different strengths and weaknesses, which can be taken advantage of to estimate the sensitivity of the trap lifetime to the gain calculation method. If the gain drift follows a double exponential, which is true for much of the data, the fitting method will be less sensitive to statistical fluctuations than the time averaged method. However, there is a class of data files where, in addition to the gain drift, the gain also alternates between two gain values. In this data, the solenoid power supply that was used to compensate the magnetic field near the main detection PMTs was set to a higher current draw than it was designed to provide. This resulted in an oscillation in the current, the corresponding magnetic field, and as a result, the gain of the main detection PMTs. The exact shape of these gain jumps are not well known, but they have a period that is roughly $\approx 4 \times 10^2$ s and a fractional amplitude, relative to the constant offset in the gain drift fit of $\approx 4\%$, which is $\approx 50\%$ the amplitude of the total gain drift in the file. The fitting method is unable to capture these periodic jumps in the gain, therefore in these files, the time averaged method is much more successful. The smoothing factor of the time averaged method was selected as a compromise between a large smoothing window to decrease statistical fluctuations and a smaller window that could more quickly adjust to the gain jumps. An example of the time averaged gain calculation for a variety of smoothing factors can be seen in Figure 5.4.

There are two cases to consider, data with and without gain jumps. Examples of each of these are shown in Figure 5.5. When there are no gain jumps, we will assume that a double exponential with an offset is the underlying functional form of the gain drifts, that the difference in the two methods is due to the uncertainty of the time-averaged method, and that this difference is an estimate of the size of the systematic effect. We determine the fractional systematic effect when scaled to a percentage to be $-0.0075 \pm 0.0397\%$. In the case of the PDG mean lifetime, this

corresponds to a systematic effect of -0.07 ± 0.35 s.

In the case of data with gain jumps, the curve fitting method is unable to capture the functional form of the underlying data. In this case, the difference in the trap lifetimes for the two methods is an indication of the sensitivity of the trap lifetime to an easily identifiable miscalculation of the gain drifts. This difference provides an upper limit on the size of the systematic effect and comes out to $-0.8 \pm 1.8\%$. This corresponds to a systematic correction of -7 ± 16 s when scaled to the PDG mean lifetime. We estimate that the uncertainty in the

Figure 5.4: Figure of the time averaged gain drift calculation as a function of the smoothing factor for the entire duration of s13r10 (Top) and zoomed in on a gain jump (Bottom).



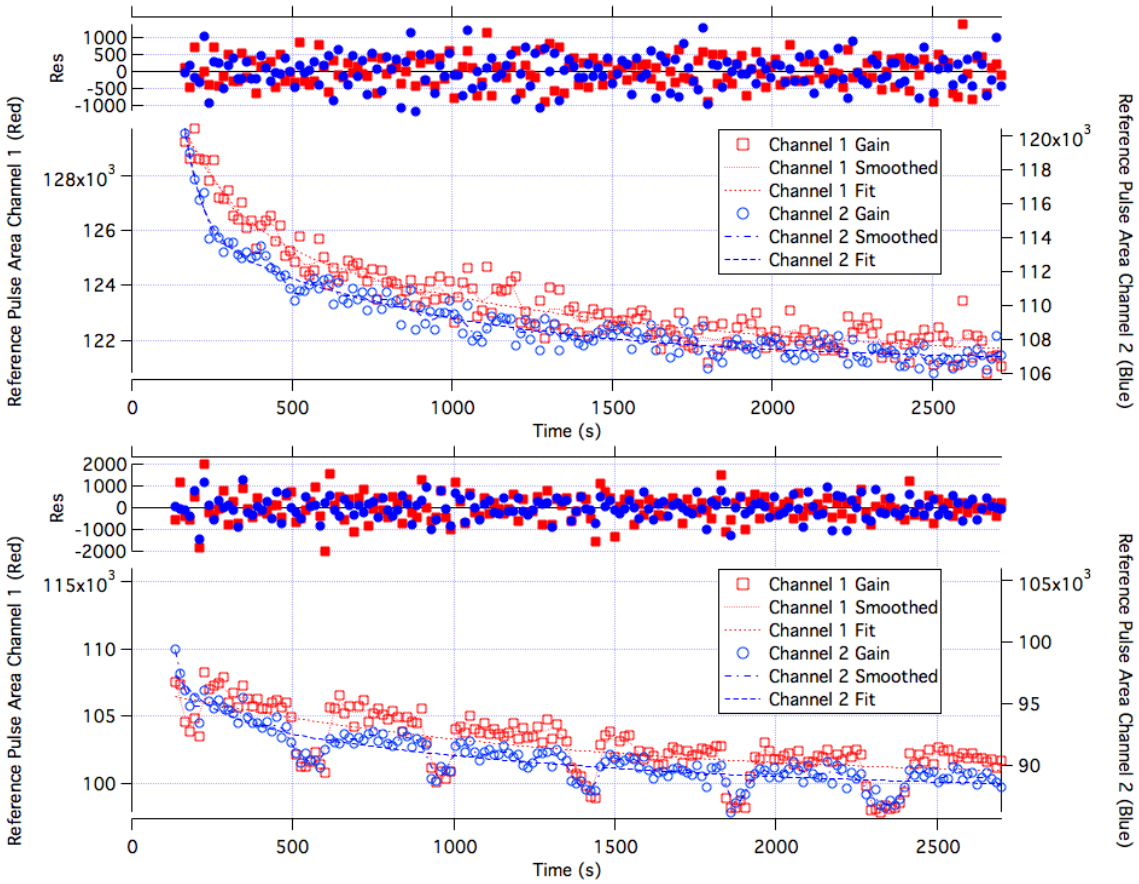
trap lifetime after the gain correction is much smaller than this value. This verifies that our experiment is sensitive to gain drifts and that the gain monitor is essential for the operation of this experiment.

The following section on time-dependent cut efficiencies also relates to gain drifts but is a distinct effect. The estimate described above was specifically designed to isolate the systematic effect associated with incorrectly calculating the gain drifts.

5.5 Time-Dependent Cut Efficiencies

One of the primary goals, when developing the detection electronics, was to make them as fast as possible to help leverage the use of pulse shape discrimination to reduce the backgrounds. As a result of this effort, the detection electronics were able to resolve the timing distribution of the neutron events in the data. For many of the small events, the response of the detection

Figure 5.5: Representative gain drifts in files with and without gain drifts. The file without gain jumps (Top) is s16r7. The file with gain jumps (Bottom) is s15r44.



electronics to each individual PE that composed the event could be resolved in time, which caused a weakening in the correlation between the pulse area and pulse height. Additionally, a pulse height discriminator was included in the detection electronics to reduce the low-energy backgrounds like neutron-induced luminescence. When the effect of the pulse height discriminator and the weak correlation between pulse height and pulse area for neutron events are taken into account together, some of the even relatively large neutron events fall below the pulse height discriminator threshold. This prevents these events from being detected. Because the events are not digitized, the gain drifts can not be corrected for in the analysis, which allows these events to circumvent the gain correction and consequently cause the type of systematic effect that the gain monitor was designed to remove.

This systematic effect in the data that we have taken was found to be much larger than is acceptable. However, it can be drastically reduced in future data by either reducing the threshold of the pulse height discriminator to a level where the vast majority of the single PE events pass the discriminator threshold or by adding an integrator into the detection electronics to help restore the correlation between the pulse area and the pulse height for these slow events. Reducing the pulse height discriminator will increase the data rate in our experiment by a factor of ≈ 3. Current technology has improved to the point where this increase in the data rate or even one a substantially larger increase could be allowed without introducing insurmountable systematic effects due to deadtime effects. Introducing an integrator into the detection electronics has its own set of complications. Obviously, from what has been found here, it is very important to consider if a different choice of detection electronics could introduce an additional systematic effect. That said, with some care, it should not be hard to reduce this systematic effect by more than an order of magnitude.

A detailed Monte Carlo simulation was designed to estimate the size of this effect. The simulation works by creating realistic voltage traces and then comparing those voltage traces to a simulated pulse height threshold. A gain drift is extracted from the data and applied to the simulated pulse shape parameters. Each voltage trace is assigned a decay time that is sampled from an exponential distribution. After simulating a large number of events, the decay times are histogrammed for events that passed the pulse height discriminator and for all events. The histogram including all of the events returns fit coefficients that agree with the inputted lifetime of 880.2 s, while the events that pass the pulse height discriminator threshold show a systematic shift in the lifetime. This difference is applied to the trap lifetime as a systematic correction. However, this systematic effect can be greatly decreased by redesigning the trigger logic before new data is taken. For an 880.2 s lifetime, the systematic shift is estimated to be 15.7(2.5) s.

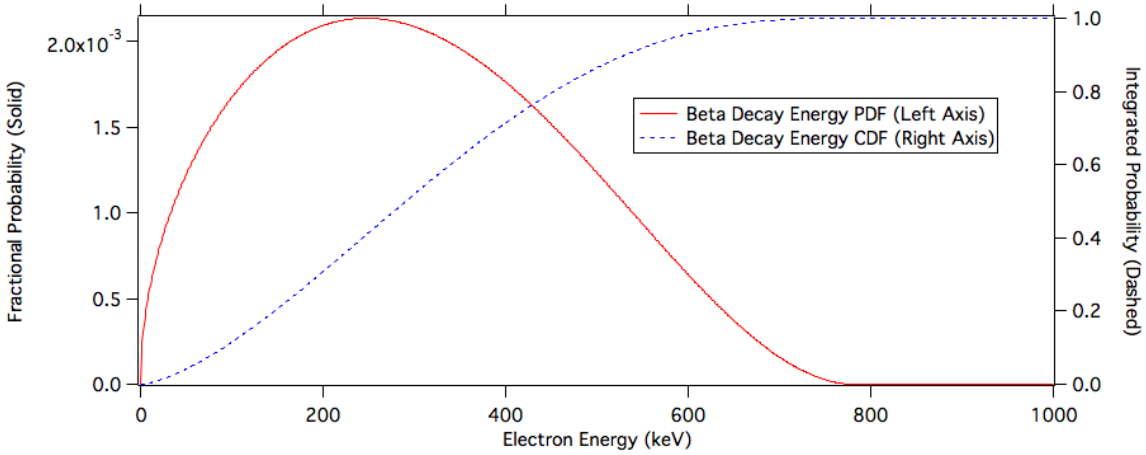
In order to simulate a realistic neutron population in this way, a variety of models needed to be developed. In the following, the models that were used in this simulation are discussed. The timing distribution for the neutron events was sampled from a pure exponential model.

The energy of the electron from the neutron β-decay was calculated using kinematics to be

$$PDF = \sqrt{E^2 + 2 * E * m_e}(Q - E)^2(E + m_e),$$

where $m_e = 511$ keV is the electron mass and $Q = 782.3$ keV is the q-value of neutron β-decay. The resulting probability density function (PDF) and the corresponding cumulative density function (CDF) can be seen in Figure 5.6.

Figure 5.6: PDF and CDF of the electron energy from neutron decays.



To determine the number of photons that strike the photocathode and the resulting signal, spatial variations in the detection efficiency must be accounted for. First, the location of the neutron β-decay must be modeled. This was done with a uniform model, an approximate model using a truncated z^4 distribution, and model where locations along the axis were determined from Kevin Coakley's Monte Carlo simulation for MTNs, which is discussed in Chapter 4. These models can be seen in Figure 5.7.

The next step is to determine the light collection efficiency of photons as a function of their location in the cell. This has been simulated by Chris O'Shaugnessy[17] and is described in the chapter on the apparatus. In this work, a few alternative detection efficiency curves were used to assess the sensitivity of the simulation to variations in the model. These models can be seen in Figure 5.8.

Using the models described above, the location and energy of the neutron β-decay can be converted into a number of photons striking the photocathode and a corresponding time for the event can be simulated. The number of photons that strike the PMTs must be split between the two main detection channels, which is modeled using the binomial distribution. The next

step is to make a realistic voltage trace for that number of photons. This also involves a few steps, and additional models must be developed. The goal is to produce a template for a single PE event, determine the timing distribution of photons for neutron β-decay events, and then using the single PE template as the response of the detection system to a single photon and to populate a realistic voltage trace using the timing distribution of photons from neutron events.

The single PE template was calculated by averaging events with very small pulse area and pulse height in which the maximum of the trace was in the range 200 ns to 360 ns after the start of the trace. This requires not only that the event that triggered the DAQ was small, but that it was also the dominant event in the trace, i.e. it removes events where there is a large peak

Figure 5.7: Three models for the spatial distribution of UCN inside the trap.

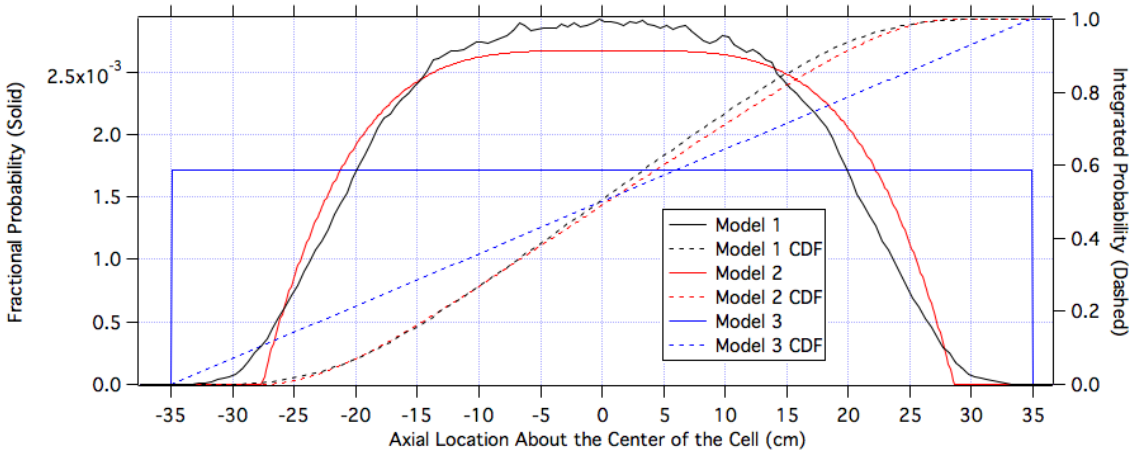
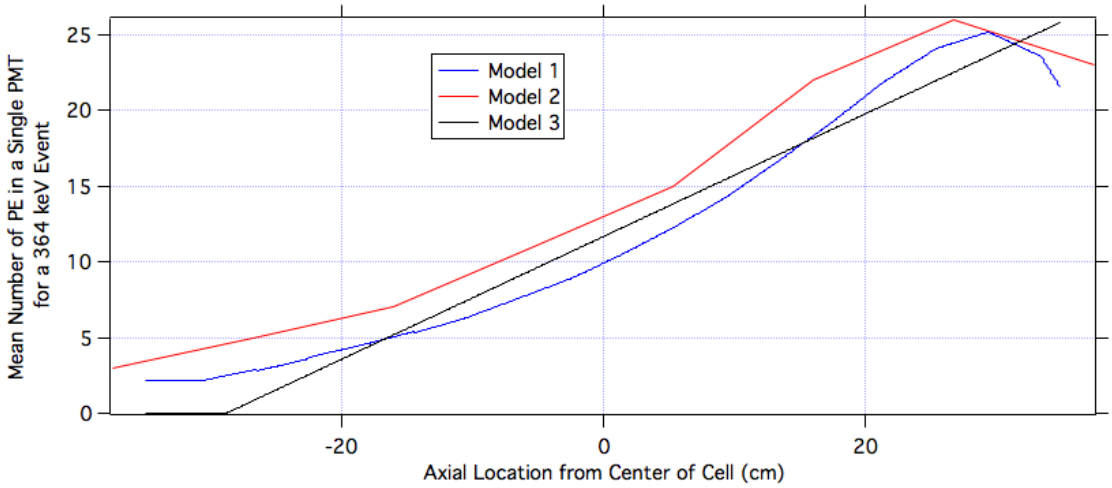
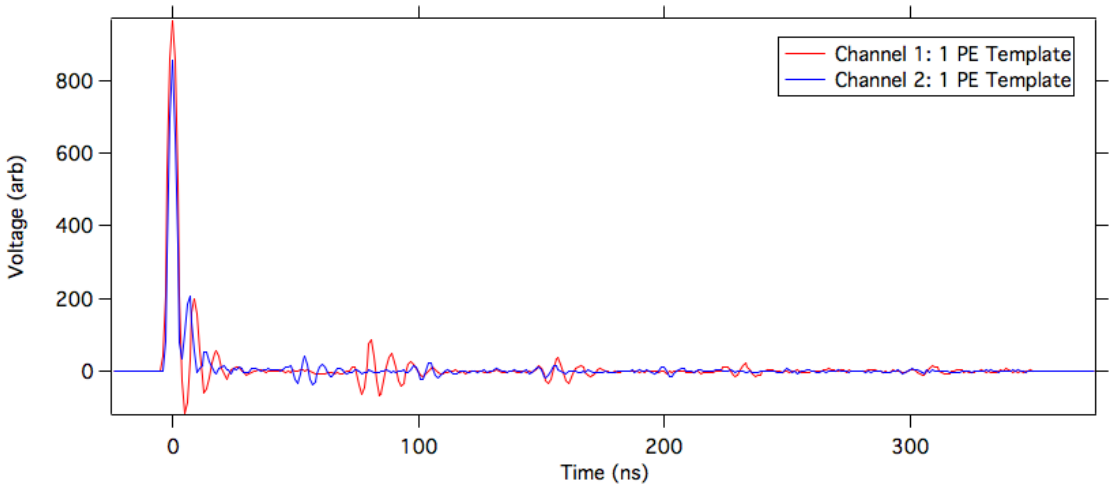


Figure 5.8: The light detection efficiency models used in the simulation to estimate the effect of gain drift and the pulse height discriminator.



outside of the region of interest. Some of the traces could have single PE elsewhere in the trace, but they would have to be smaller than the peak inside the trigger region. Before the averaging, the traces were aligned by the maximum of the trace. Finally, the ends of the template were truncated. The resulting single PE template can be seen in Figure 5.9. It is worth noting that a more accurate single PE template might be able to be obtained by looking for single PE events that fall outside of the main detection region. The main pulse could cause systematic effects in extracting the single PE template. However, since the single PE event is not subject to the discriminator threshold, the discriminator can not bias the template. The difference between these two methods is expected to have a small effect on this work, and this alternative method was not pursued here.

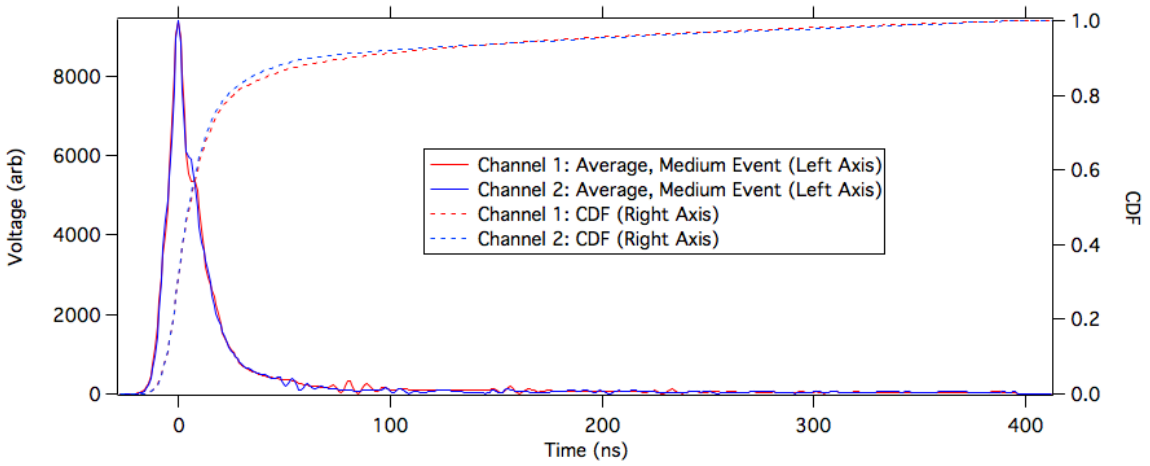
Figure 5.9: Template of the single PE events constructed from small events that triggered the detection system. In the simulation, it is used as the impulse or response of the system to a single photon striking the photocathode of the main detection PMTs.



The neutron β-decay timing distribution was estimated using a similar method on medium, slow events. Large events were excluded to prevent non-linearity in the detection electronics from affecting the results. In small events, the kurtosis becomes less effective at selecting on the pulse timing. Using relatively large events allows a clean separation between the fast and slow events. These events were aligned, averaged, and truncated. With no additional processing, the regions around the center of the averaged peak were negative. This seems to indicate that the baseline subtraction on the voltage traces is being slightly overestimated and when the effect is combined by averaging many events it becomes noticeable. I suspect that the effect of this on the analysis is negligible, but for this simulation, the PDF of the medium events must

be positive definite so that the locations of the photons can be sampled uniformly from the CDF. Therefore, the constant offset was subtracted off by hand, and the averaged trace was further truncated to remove regions outside of the main pulse that had heights that were less than zero. The resulting averaged trace can be seen in Figure 5.10. As an aside, we suspect that this template should be representative of any events that produce scintillation light in the helium bath including not only neutron β-decays but also muons, ^{3}He absorption, and any other mechanisms that ionize the helium in the helium bath.

Figure 5.10: Template of a helium scintillation event, which has a slow timing distribution. This timing distribution is expected to be representative of neutron β-decays, ^{3}He absorption events, and any backgrounds that result in scintillation in the helium bath. In the simulation, it is used as the timing distribution of photons striking the photocathode for these slow events.



Using the models developed here, simulated voltage traces for neutron events can be simulated. However, these simulated traces lack some of the random variation that one would expect in the real voltage traces. To be more representative of the actual traces, the size of the individual PE impulses should be able to vary, and there should be background noise on the voltage trace. Ultimately, it was found that the effect on the simulation from these noise models was quite small in comparison to the effects that dominated the estimates. However, until the noise was modeled, it was not clear how large the effect might be.

Variations in the size of the individual PE impulses comes about primarily from variations in the number of secondary electrons emitted from the first dynode in the PMT. These fluctuations are therefore expected to approximately follow a Poisson distribution. To incorporate this into the simulation, a random scaling for each impulse was sampled from a Gaussian distribution with a characteristic standard deviation of 0.2 and a mean of 1. This is only a rough estimate

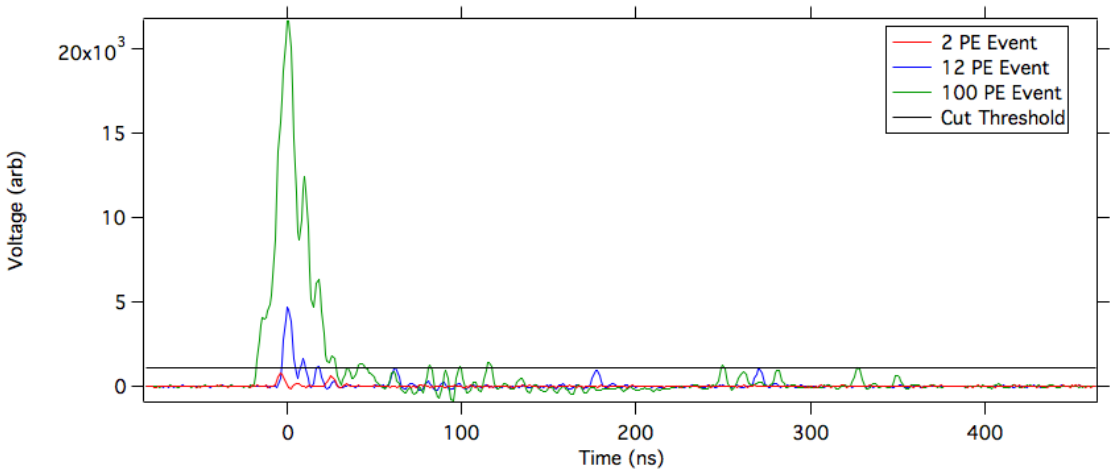
of the variation in the actual pulses. In sensitivity studies, the result only depended weakly on the size of the scaling. This is an indication that the simulation is not sensitive to noise of this type, and therefore, an effort to refine the model was not necessary.

The random noise in the anode current of the PMTs was assessed independently for the two main detection channels by calculating the standard deviation of regions of the voltage trace in which there were no peaks. The standard deviation of these zero regions were 26.8(22.3) in channel 1(2). This effect was accounted for by adding a random value to each bin in the simulated voltage trace that was sampled from a Gaussian with a mean of zero and a standard deviation that was appropriate for that channel.

After including all of these effects, a realistic voltage trace can be calculated in each channel. An example of a simulated voltage trace for a 2 PE, 12 PE, and 100 PE event can be seen in Figure 5.11 along with an estimate of the pulse height discriminator threshold.

^{3}He capture events could have a large impact on the results of the simulation. They were added to the simulation using the same methodology described above with the exception of the energy spectrum, which for ^{3}He capture events is monoenergetic at 765 keV.

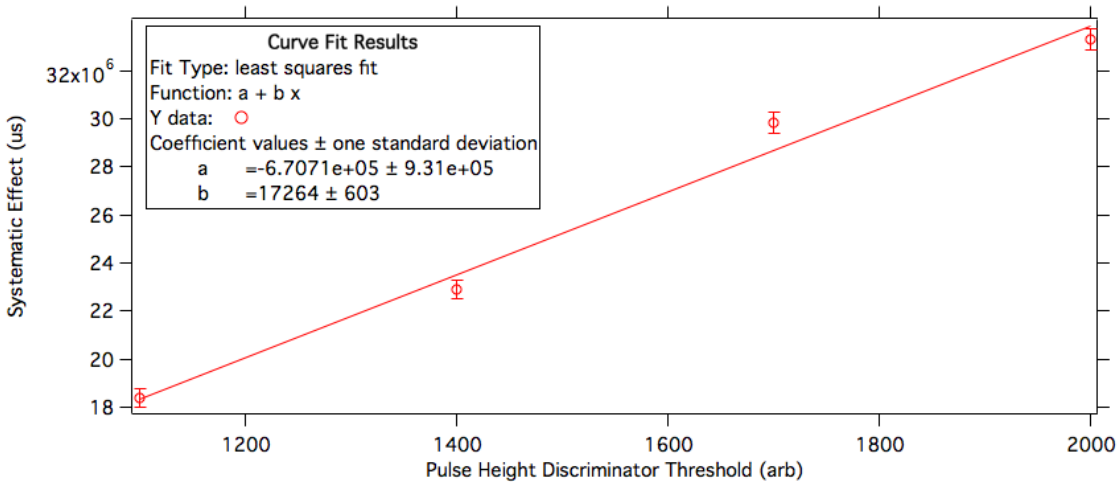
Figure 5.11: Realistic, simulated voltage traces obtained by convolving the light detection efficiency, the neutron spacial distribution, the electron energy distribution, the single PE response, and the photon timing distribution. The simulated pulse height discriminator is shown for reference before it is scaled in time to simulate the effect of the gain drifts. The broad timing distribution of these simulated events results in weak correlation between the pulse area and pulse height for small numbers of PE, which allows some of the low-energy neutrons to fall below the discriminator threshold.



A gain drift was extracted from the data and applied to the cut threshold to simulate the

effect of a gain drift on the pulse height discriminator. First, the threshold of the discriminator must be estimated. The exact location of the pulse height discriminator with respect to the height of an average single PE event has not been accurately determined but is expected to be between 1.0 and 1.8 times the average height of a single PE event. The most accurate determination of the threshold was obtained by fitting the low-energy portion of the gain corrected pulse area histogram to a combination of Gaussians corresponding to the response of the individual PE peaks. The amplitudes of the Gaussians were allowed to vary independently, but the x-offsets were all constrained by a single scaling that corresponds to the location of the single PE peak. The discriminator threshold was modeled in the fit as a Hill function with a fixed base and max at zero and one, respectively. Using this method the threshold was estimated to be approximately 1000 in pulse height. Sensitivity studies were run on the pulse height threshold, and the systematic effect was found to be very sensitive to the location of the pulse height discriminator. Between 1100 and 2000 the systematic effect was found to be roughly linear, see Figure 5.12.

Figure 5.12: Sensitivity study of the simulated discriminator threshold on the time-dependent cut efficiency systematic effect.

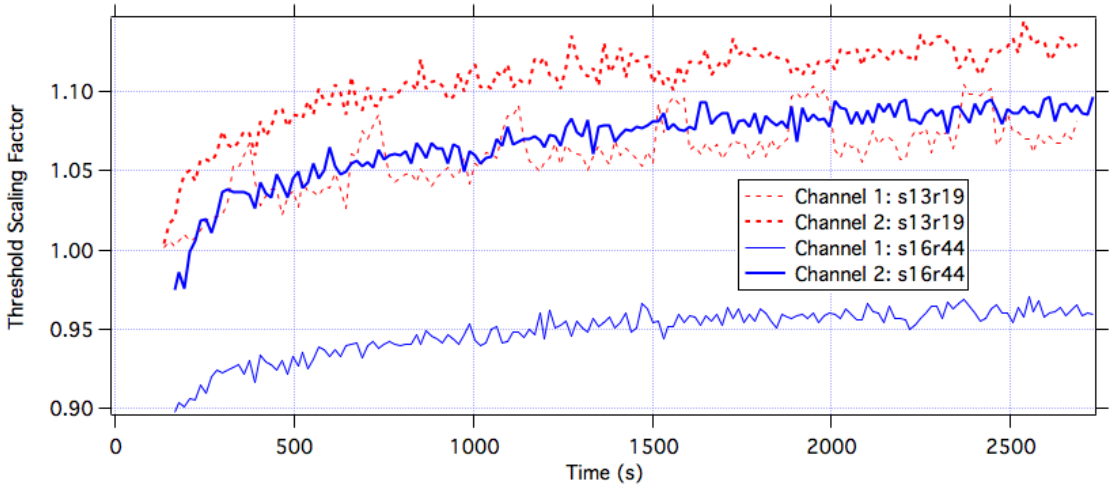


To estimate the uncertainty in the systematic correction due to the uncertainty in the pulse height discriminator threshold, a threshold value and its uncertainty had to be selected. We are confident that the threshold is slightly above the single PE peak height, but we do not know its exact location. We chose a plausible value of 1000 ± 50. Using this value and the sensitivity to the cut threshold shown in Figure 5.12, the contribution to the uncertainty in the systematic correction from the threshold value was calculated to be 1.4 s, which was estimated in a simple

Monte Carlo simulation.

Once the height of the discriminator threshold is determined, the gain drift array from the analysis is converted into a time-dependent scaling of the threshold by inverting the gain and multiplying by a constant. Using this constant scaling causes an offset between the gains of two different files to influence the calculated systematic effect. This motivates this method over a scaling calculated from the gain drift in each file separately. This is a consistent way of calculating the threshold so that different gain drift arrays can easily be used to test the sensitivity of the simulation to the exact shape of the gain drift. The constants were chosen based on the steady-state gain at the end of s8r37 to be 118240(117295) in units of non-gain adjusted pulse area for channel 1(2). A few examples of the scaling curves can be seen in Figure 5.13.

Figure 5.13: Examples of the time-dependent multiplicative scaling arrays used to simulate the effect of gain drifts on the hardware pulse height discriminator. The scaling factors were usually within about 10% of unity and the gain typically drifted by about 10%.



In conclusion, neutron events can be simulated that match our expectation. Realistic voltage traces are simulated, pulse shape metrics are calculated, and timestamps are simulated. The effect of the hardware, pulse height discriminator is simulated by comparing the voltage traces to a threshold that varies in time according to gain drifts that have been extracted from the data. This allows separate timestamp histograms to be simulated for events that pass the pulse height discriminator and for all events. The difference in the extracted lifetime for these two populations is an estimate of the systematic effect. By repeating the simulation many times and calculating the mean and standard deviation of the ensemble of simulations, a statisti-

cally meaningful systematic correction, and its uncertainty can be estimated. As mentioned previously, the systematic correction with our most realistic models comes out to a 15.7(2.5) s systematic correction. The uncertainty is estimated in sensitivity studies where the sensitivity of the result to the particular models that were used is evaluated.

The simulation used experimental data to constrain the majority of the processes that could contribute to this effect. However, there were a few topics that were not taken into account. First, the pulse height discriminator is not a step function; there is some squishiness to the threshold. This effect was not included in this simulation. Second, this simulation has modeled how neutrons interact with the combined effect of the pulse height discriminator and the gain drifts. However, there will be a similar interaction for the background events. Without a compelling energy spectrum, spatial distribution, and light collection efficiency for the background events the corresponding simulation for the background events could not be performed. However, if these background events create a similar systematic effect, it can be estimated by testing the background subtraction, see Section 5.7, which starts on page 155 for a discussion that limits the size of this effect. Also, the pulse area spectra in the simulation do not agree with the pulse area spectra in the data, see Figure 5.14. In this figure, there are three different thickness of lines thin, medium, and thick. The thick lines highlight the primary point of the figure, while the thin and medium lines add supplementary information. The thick solid line is the background subtracted, gain adjusted pulse area spectrum from s16 and the thick dash dotted line is the simulated PE spectrum that includes both neutron β-decay and ^{3}He absorption events. These two traces show characteristic shapes that are inconsistent and a linear scaling does not resolve the discrepancy. The data does not show any indication of a rounding off in the spectrum until it is very quickly removed by the lower pulse area cut whereas the simulation shows substantial rounding on the low PE side of the distribution. Additionally, the data extends to much higher energy than the simulation. We are confident in the PE to pulse area calibration at approximately the 20% to 30% level, however this seems to suggest that, if our calibration is accurate the data and simulation do not agree. The medium (thin) lines show the simulated spectrum when the pulse height discriminator is (is not) included in the simulation. The dashed data (red) is the simulated ^{3}He absorption events, the solid (blue) data shows the simulated n β-decay events, and the dashed dotted (black) data shows the simulated, combined spectrum including both the n β-decay and ^{3}He absorption events.

The size of the systematic effect reported in the previous generation of the experiment was 8 s[16]. In this work, the inclusion of the gain monitor and gain correction should have greatly reduced the systematic effect associated with gain drifts. Instead, there is a larger systematic effect in this experiment than what was estimated in their work. This can be accounted for by comparing the size of the gain drifts in the two experiments. They reported a 2% gain drift while this experiment's gain drifts were about 8% to 10%. Therefore, it is reasonable that the

systematic effect in our experiment could be larger than what they reported.

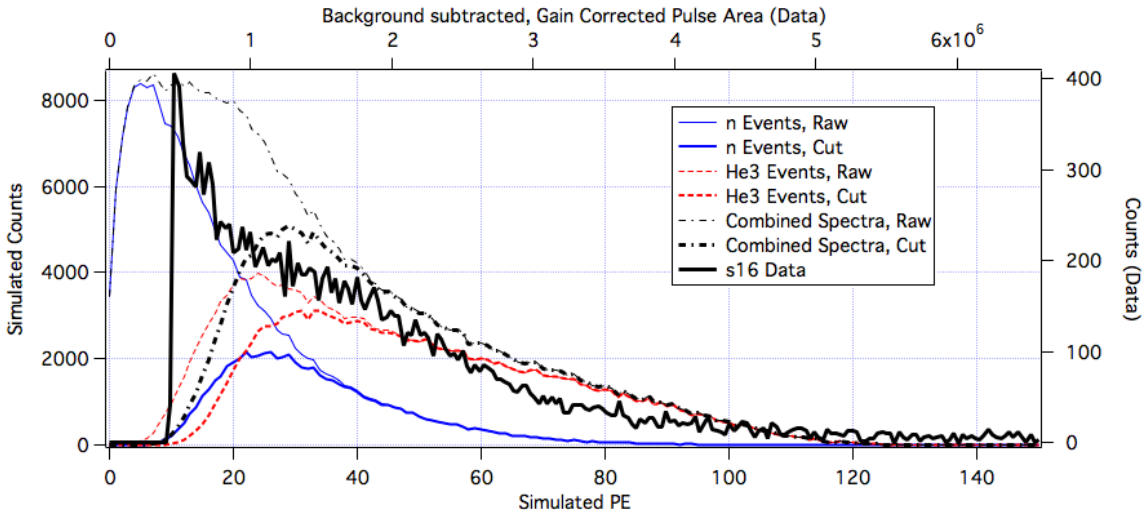
5.6 Drift in the Clock

Any inaccuracies in the DAQ clock will result in a systematic effect. An estimate of the systematic effect was calculated in the section on the performance of the apparatus and can be found here Section 2.13.3, which starts on page 62. There, the most conservative estimate of the systematic effect was calculated to be $\sigma_\tau/\tau = 3 \times 10^{-5}$. Assuming the PDG's world average lifetime, this results in a systematic effect of 0.03 s, which is more than a few orders of magnitude below the statistical uncertainty of the experiment. Therefore, it will be neglected in this work.

5.7 Imperfect Background Subtraction

This experiment relies on our ability to remove the contribution to our timestamp histograms that comes from background events in a model independent way by directly measuring our background rates. At some level, these measured background rates are expected to differ in the trapping and non-trapping (background) data due to differences in the background rates, detection efficiencies, or through other mechanisms. This will introduce a time dependence that survives the background subtraction and a systematic effect in the trap lifetime. In the second

Figure 5.14: The energy spectra for the neutron like events in the time-dependent pulse height discriminator systematic effect simulation with or without the effect of the pulse height threshold with a comparsion to the data.



generation experiment, a clever method for estimating the size of this systematic effect was developed. After they had finished taking production data, the isotopically ultrapure ^{4}He in the cell was intentionally contaminated with ^{3}He, and additional diagnostics data was taken. The addition of the ^{3}He to the cell was not expected to affect any non-UCN induced backgrounds. This resulted in the backgrounds being almost identical to the production data except that no neutrons survive into the data acquisition phase. This ^{3}He data showed a time dependence that was consistent with a constant and very successfully limited how large this systematic effect could be[16].

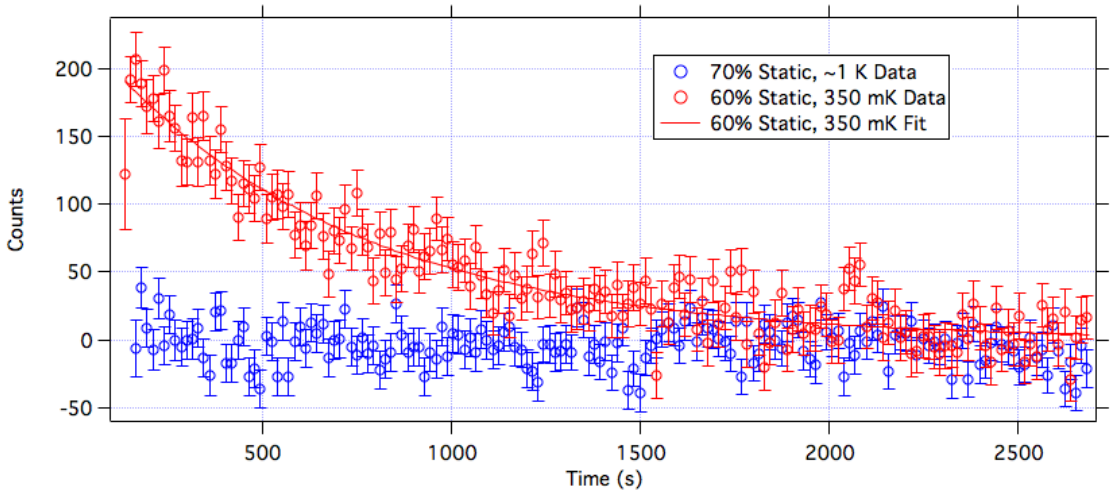
The process of admitting ^{3}He into the cell ruins the cell. If the cell was ever used after exposing it to large quantities of ^{3}He, it would be very challenging to prove that no ^{3}He still existed in the cell materials that could contaminate the new ultrapure ^{4}He. Therefore this is the last data that should be taken with a particular generation of the experiment.

For this generation of the experiment, we were hoping to be able to take additional data in the future. Therefore, when we stopped running for the long shutdown in 2012, we did not admit ^{3}He into the cell. As a result, we do not have ^{3}He systematics data to test the background subtraction. Without this data, I will do my best to present some mechanisms that could conceivably introduce systematic effects associated with the background subtraction as well as arguments for limiting the size of these systematic effects. First, I will use warm data to apply a similar constraint to the size of systematic effects associated with an imperfect background subtraction.

Our best data for constraining the quality of the background subtraction is s18p2, which is warm, $\approx$ 1 K, 70% static data. At 1 K the trap lifetime including an 880 s lifetime for β-decay and a 129 s lifetime for upscattering comes out to 112 s. Therefore, this data will also greatly reduce the number of trapped neutrons in the data with minimal effects on the backgrounds. The results can be seen in Figure 5.15. If the ratio of the amplitude in the warm data and the cold data could be determined, that ratio could be used as an indication of the fraction of background events that made it through the background subtraction. To estimate the systematic effect one could add the contribution from that ratio multiplied by the background rates to the background subtracted data. By fitting to the resulting histogram and comparing the result to the fit to just the background subtracted data, the systematic correction could be calculated. Unfortunately, we have not come up with a convincing method for extracting this ratio of the amplitudes. With a 112 s expected lifetime in the warm data, our delay time is too large, and the majority of the neutrons will have decayed before the start of data collection. This means that the data is not able to constrain the amplitude if we use the expected lifetime. It also, however, suggests that there is no evidence that there are background events making it through the background subtraction. In order to obtain a quantitative estimate, we fit the cold data to an exponential with an offset. Then we fit the warm data where we held the lifetime

to the value determined with the cold data. Note that the lifetime in these two data sets is not expected to be the same, this is an invalid assumption that is being used to help extract a reduction in the amplitude. The result was a fractional reduction in the amplitude of 0.03. Taking this at face value, and using the method described above, the systematic correction is 0.03 s on a lifetime of 675 s. When this is scaled to the PDG mean lifetime the systematic effect is 0.04 s. Therefore, we find that the systematic effect can be neglected for this data. The fact that the warm data shows little time dependence indicates that neutron-induced luminescence and magnetic focusing can not be large effects. Despite this, they will be considered below.

Figure 5.15: Timestamp histograms of the warm data demonstrating the effectiveness of the background subtraction. When the cell is warm and consequently the number of UCNs in the trap is greatly reduced there is no evidence of UCNs in the background subtracted data. Cold data is also shown for comparison.



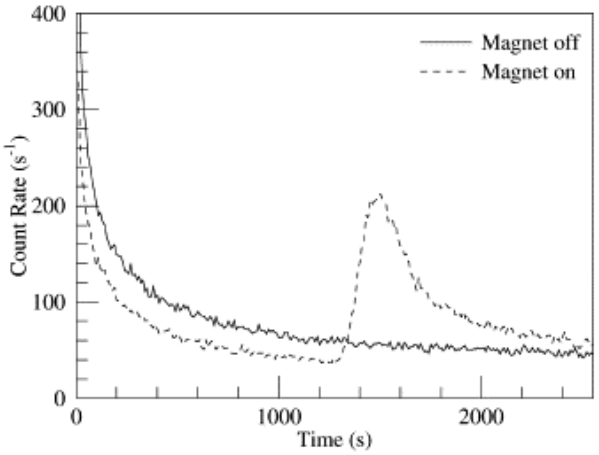
5.7.1 Neutron-Induced Luminescence

Neutron-induced luminescence is the name of a process in which BN produces uncorrelated photons after it has been exposed to neutrons. Because this experiment measures light, neutron-induced luminescence could result in background events. The experiment was carefully designed to limit the number of photons from neutron-induced luminescence that make it into the detection system using opaque graphite shields, for more information see Section 2.3, which starts on page 17.

It is reasonable to expect differences in the amount of neutron-induced luminescence in the trapping and non-trapping data because most of the UCN that are trapped in the trapping

data would have been absorbed on the BN in the non-trapping data. One might think that this would be a very small effect because only a small fraction of the cold neutrons are converted into UCN using the following argument. Of the UCN that are produced, about half, i.e. the wrong spin state, will be absorbed on the BN in the trapping data and all of the UCN will be absorbed in the non-trapping data. Therefore, the ratio of interest is the half of the UCN population over the cold neutron population. The integrated neutron fluence has been measured to be about $4.7 \times 10^6 \ \frac{1}{\mathrm{cm}^2\mathrm{s}}$[8]. The cross-sectional area of the beam is 163 cm^2 we see that the number of cold neutrons every second is orders of magnitude greater than the total number of UCNs trapped in a trapping data file. Therefore the effect of the wrong spin state is expected to be much smaller than the effect of fluctuations in the cold neutron beam, which suggests that this would be a very small effect. However, the neutron-induced luminescence has been shown to be stabilized by magnetic fields[19]. This is demonstrated very convincingly in Figure 5.16. The magnetic field states are very different in the trapping and non-trapping data, and this could cause a substantial difference in the lifetime associated with neutron-induced luminescence in the trapping and non-trapping data, and consequently cause a systematic effect. The measurements of this magnetic stabilization of the neutron-induced luminescence, which has been termed magnetic pinning of the luminescence centers, are not sufficiently accurate to strongly constrain this effect. Therefore, the only way to place a sufficiently stringent constraint on this effect is using the warm data, which suggests that the effect is negligible.

Figure 5.16: Evidence of magnetic pinning of neutron-induced luminescence from boron nitride. "The time dependence of the luminescence signal with no magnetic field (—) and when the magnetic field is energized during the irradiation and de-energized 1275 s after the irradiation ends. (- - - -)" [19].



5.7.2 Magnetic Focusing Effects

The magnetic trap will interact with both the cold neutrons and UCN through their dipole moment. There are a few ways that the trap could influence the cold neutrons in such a way that the background rate might differ in the trapping and non-trapping data. In both of the cases, the effect comes about from the bending of the cold neutron beam, which as one might expect is a very small effect. In the first case, the cold neutrons are bent differently in the trapping and non-trapping data because of the difference in the magnetic state for these two magnetic ramping schemes. This results in the cold neutrons striking slightly different portions of the cell. If this results in exposing the beam to parts of the apparatus that are more prone to activation or if it couples with geometric variations in the detection efficiency, it could result in a difference in the measured background rate in the trapping and non-trapping data and consequently a systematic effect. In the second case, the cold neutrons are bent away from the center of the trap. This changes where the UCN are created in the trap and therefore the initial UCN energy distribution. In this case, only the trapping files need to be considered. First, let us estimate the size of the displacement of a cold neutron due to the presence of the magnetic field because this number will give a feeling of the length scales that we need to be sensitive to. We will find that this length scale is sufficiently small that both of these types of systematic effects are expected to be negligible in our data.

The cold neutrons, like the UCN, will interact with their magnetic dipole moment through the potential $V = -\vec{\mu} \cdot \vec{B}$ or equivalently through the force, $F \propto \mu \nabla B$, where the magnetic field gradient can be estimated via $\nabla B \approx \frac{\Delta B}{\Delta r}$. Figure 5.17 shows the results of a simulation to calculate the magnetic field strength as a function of the location inside the experiment using realistic coil wrappings, which we can use to estimate the magnetic field gradient. To simplify the calculation, consider a cold neutron that is traveling parallel to the axis of the cell and consider a gradient that is perpendicular to the axis of the cell. From the figure, the magnetic field is shown to decrease by $\Delta B \approx 3$ T from the cell wall to the center of the trap, a radial distance of $\Delta r \approx 6$ cm. Assuming no initial radial velocity or displacement, integrating the force twice results in the following equation for the displacement due to the magnetic bending $x = \frac{Ft^2}{2m} = \frac{\mu \Delta B t^2}{2m\Delta r}$. Finally using the velocity of the cold neutrons $v = 444$ m/s and an approximate path length of cold neutrons in the cell $L \approx 1$ m, the duration can be estimated to be $t = v/L = 2$ ms. Plugging these values in the deflection of the beam is estimated to be 4 mm. Keep this length scale in mind as we consider the two mechanisms below.

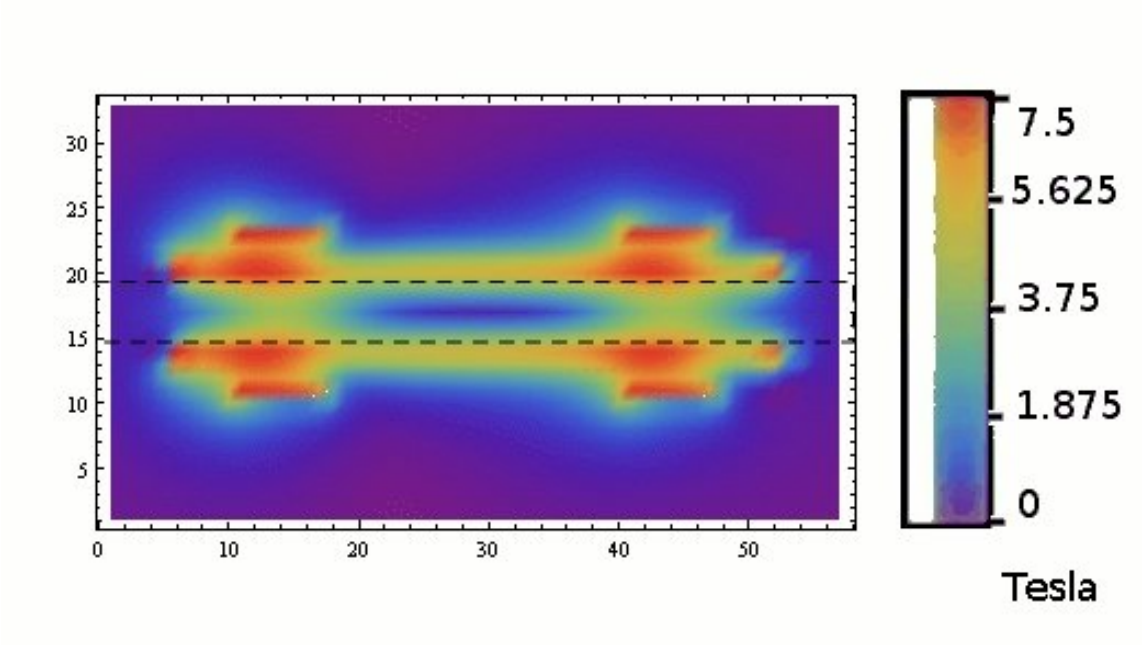
If you recall, the first effect was the deflection of the cold beam causing the beam to strike portions of the cell that are either prone to activation or that couple with spatial variations in the detection efficiency. The broad nature of the beam in the apparatus and the small length scale that we are considering makes it difficult to fathom how the bending of the cold neutron

beam could cause a non-negligible difference in the activation of the apparatus between the trapping and non-trapping data. There may be small effects if the strongest part of the beam is deflected toward or away from gaps in the BN or graphite shielding, but it is hard to imagine that such a small deflection would cause a sufficiently large systematic effect to be important to this experiment. Therefore, it is expected to be negligible.

A second contribution to this effect could result from spatial variations in activation coupling with the detection efficiency, which varies throughout the cell[17]. The detection efficiency changes little on the length scales that we are concerned with here and therefore this effect is also expected to be negligible.

The final effect has to do with the change in the initial energy distribution of the UCN in the cell and how a change in the fraction of MTNs might introduce a change in the trap lifetime. This is a complicated effect, but it could be simulated in the Monte Carlo. Instead, we will point out that the effect on the initial energy distribution will be very similar to the Monte Carlo simulations that were already run to test the effect of beam divergence on the trap lifetime. In those simulations, it was found that the trap lifetime was changed by about 1 s/degree of divergence[23]. The divergence follows a linear dispersion whereas this focusing effect will be curved, but the general effect is expected to be very similar. A 1 degree divergence

Figure 5.17: A cross sectional view of the simulated magnetic field strength from the magnetic trap from Chris O'Shaugnessy's thesis[17]. The dotted lines are intended to indicate the approximate location of the cell wall. The numbers on the x and y axis are in centimeters.



over the same 1 m path length results in a 1.7 cm displacement, which is quite a bit larger than the displacement from magnetic focusing. Also, the beam divergence simulation used a static ramping profile, i.e. the magnets were not flushed to preferentially remove MTN, and the effect is expected to be smaller for flushing data. Therefore, the possible systematic effect is expected to be < 1 s in static data even smaller in the flushing data.

In summary, we have considered a few mechanisms through which focusing of the cold neutrons by the magnetic trap could introduce a systematic effect in this experiment, and we have found that for each of them, the systematic effect is expected to be negligible in the current data.

5.8 Neutron β-decay in Matter

In the Appendix H of his thesis[56], Clinton Reed Brome discussed the effect on the trap lifetime of the neutron decay occurring in matter. The presence of the helium changes the available phase space for each of the decay products. By Fermi's Golden Rule, one expects this will affect the decay rate. He estimates this effect to be ≤ 0.01 s. He also considers the possible effect on the matrix element due to the presence of additional nuclei, which is found to have an even smaller effect. Both of these effects are negligible for this generation of the experiment.

5.9 Fitting Bias

The fitting algorithm that is used to extract the trap lifetime can introduce a systematic effect. The size of the systematic effect is most strongly correlated with two parameters. First, the number of counts in the bins, which is a manifestation of Poisson statistics. The second parameter is the duration of the observation stage, which is due to the fitting algorithms inability to distinguish a constant offset from a shorter lifetime if the observation stage is too short for the exponential to flatten out. A set of recommendations for the UCN Lifetime Experiment at NIST were published by Kevin Coakley[46] before data collection began and either the number of trapped neutrons or the background rates were measured.

To estimate this effect using parameters that are more representative of the data, a simple Monte Carlo simulation was developed. An exponential process and a constant process were simulated and were fit to an exponential with a constant offset. The range of the fit was adjusted to simulate different observation phases and the difference in the extracted lifetime from the simulated lifetime was taken as an estimate of the systematic effect. The simulation was performed many times with each set of simulation parameters to allow the mean and standard deviation of the systematic effect to be quantified. The mean is the size of the systematic effect, and the standard deviation is its uncertainty.

The size of the systematic effect is expected to depend on both the limits of the fit and on the data rates, which in the case of the aggregated data is related to the number of file pairs of the given data type. To make the simulations as representative as possible, the neutron rate, background rate, and the number of data pairs were based on the values from the data. Finally, we want to do sensitivity studies to determine how the systematic effect varies with these parameters.

In these sensitivity studies, the results of the simulations were shown to be insensitive to variations in the input parameters at the operating parameters that correspond to our experimental data. When the fit ranges were substantially shorter than in the data, the result was a large negative systematic effect that decreased substantially with increasing data rate but was still very large at the highest data rates that were simulated. At the largest fitting ranges and low data rates, the simulation showed a positive systematic effect, which is attributed to low count rates at later times. However, once 8 to 10 data pairs were included in the data series the systematic effect stabilized with a value that was consistent with zero. For the simulations with the lowest data rates that were simulated, the dependence on the number of files fit well to a single exponential. The simulations with the second lowest data rate already fit well to a constant. However, the data also appeared to be consistent with the exponential with the same lifetime as the simulations with the lowest data rate. Since a constant dependence would not result in an additional systematic effect at low numbers of data pairs and a fitting bias is expected in this limit, we assumed that the simulation followed an exponential dependence with the same lifetime as the lowest data rate. With either a constant fit or an exponential fit, the systematic effect from the fitting bias was quite small once it was extrapolated to the parameters from the data. Because the exponential method had a more realistic form for small numbers of data files, it was used.

The simulations were found to be computationally demanding, for a few second estimate of the systematic effect for a single simulation takes a few days of simulation time. Therefore, to make a more efficient use of our computational time, I decided to focus on simulating the data set with the fewest number of data files as it was the data set that was most likely to have a non-negligible systematic effect. The data set was s13, in which only 4 data pairs were used in the analysis. In s13, which has a trap lifetime of 619 s, the systematic effect was found to be 0.06 ± 0.19 s. The systematic effect is assumed to be smaller in all of the other data series and as a result negligible at our statistical sensitivity. Therefore, it is reasonable to list the systematic effect as being $\ll 1$ s at the PDG mean lifetime.

5.10 Random Coincidence

In this section, we discuss random coincidence in general. At the end of the section, a few physical mechanisms that involve random coincidence are discussed in more detail including PMT afterpulsing and neutron-induced luminescence in boron nitride.

In this experiment, there are a few mechanisms that create uncorrelated photons, which are a type of background that will result in a single PE in just one of the main detection PMTs. Because of the coincidence requirement in the detection electronics, the majority of these events will not trigger the DAQ. The small fraction of events that make it past the coincidence requirement do so by random coincidence, which is when two random events occur at the same time, one in each channel, and as a result, can pass the coincidence requirement. Alternatively, if a single event is able to trigger the coincidence requirement, a uncorrelated photon that is coincident with this main event can add to the voltage trace. In this case, the random coincidence does not create a new event, but it may change the pulse shape parameters of the primary event.

As the data rate increases, the fraction of time where one of the main detection channels has triggered its discriminator threshold also increases, and therefore, there are more opportunities for an uncorrelated event to trigger the second channel and consequently the DAQ. This also means that, since our data rate is time-dependent, the rate of random coincidence events will vary in time, which can introduce a systematic effect when extracting the trap lifetime.

Any random coincidence with the backgrounds will be removed to a high order by the background subtraction due to the similarity in the background rates between the trapping and non-trapping data files. We also do not expect to be very sensitive to backgrounds that are coincident with neutrons. In this case, the neutron is expected to trigger both channels, and the background event is added to one of the voltage traces. For the event to affect the pulse area or pulse kurtosis, it has to fall inside the region of interest. Even if it does, the majority of these background events are expected to be small events. Therefore, if there is a background signal in one of the traces, it will only effect the pulse area slightly, pushing it to a higher value. For a few events, the pulse area may be pushed above the lower or upper pulse are cut thresholds by random coincidence, but this is expected to be a tiny effect. Coincident events will also affect the kurtosis metric. However, this effect is expected to be of a similar size and as a result quite small. Therefore, although this is a systematic effect in the experiment, it is expected to be negligible at our current statistical sensitivity.

5.10.1 PMT Afterpulsing

PMT afterpulsing is a particularly malicious background mechanism in this experiment. It is a large part of the motivation for using two main detection PMTs. PMT afterpulsing occurs

when gas, typically helium, contaminates the vacuum inside the PMT. The electron cascades can ionize the gas in the PMT, which is then accelerated through the electric potential. When it strikes a surface, the ions can liberate electrons starting a secondary event that trails the primary event by about 1 μs. In the previous generation of this experiment, some of the after pulsing events were found to be quite large[30]. Because these events are large, they can pass a lower pulse area cut. Therefore the detection system was designed to include two main detection PMTs, which takes advantage of the fact that this mechanism occurs inside the PMT. Therefore, it will not affect the other PMT. Since coincidence is required between the two main detection PMTs, the only way for the DAQ to be triggered is through random coincidence. Since the majority of the background events, i.e. potential random coincidence candidates, are small requiring that each event passes a 3 PE pulse area cut in both main detection PMTs limits this background source substantially. Our experiment takes advantage of this methodology. However, it also uses the kurtosis cut, which preferentially selects events that are much slower than the response of the PMTs. The afterpulsing will have a very fast timing distribution, in fact, it should have the fastest timing distribution of any multi-photon event in our experiment. Therefore, even if these events make it past our pulse area cut, it is very reasonable to assume that the vast majority of any remaining afterpulsing events will be removed by the kurtosis cut.

It is possible to limit the rate of PMT afterpulsing using the production data to evaluate the performance of the apparatus. In particular, it is convenient to use the reference pulser events, which make up a large fraction of the data, are uniform in size, and should not be strongly correlated with the other events. The background rate can be constrained by calculating the pulse area in regions outside of the main pulse of the event. Specifically, the regions from 0 s to 200 s, 800 s to 1000 s, 1000 s to 1200 s, and from 1200 s to 1400 s do not typically contain portions of the main peak for a reference event. Figure 5.18 illustrates an afterpulsing event and the timing of these windows. The results have been histogrammed in Figure 5.19. As expected, the majority of the reference events have an area that is less than the signal from a 1PE event in these regions. The shape and height of the corresponding pulse area histograms depended strongly on the particular window. In particular, the fraction of events with a > 1 PE signal increased as the window got further from the start of the voltage trace. I would expect the first window, which is before the trigger, would have a lower rate if the events were correlated with the main pulse of the event, but it is very interesting that the number of events continues to increase for windows that are later and later relative to the start of the trace. Perhaps this says something interesting about the afterpulsing mechanism and the internal structure of the PMT.

We can use the rate of > 1 PE pulse area in these alternative windows to constrain the rate of PMT afterpulsing. Because the rate is highest in the last window and the delay between the initial pulse and the afterpulse is expected to be about the length of our digitization window,

we will use the rates in the last window as our estimate. We find that about 11% of reference pulser events have a >= 1 PE signal in the window from 1200 ns to 1400 ns after the start of the digitization window. A sizable fraction of these events extend to much higher pulse area with only a few events with more than 30 PE in a single channel. By taking into account the average number of PE in a reference event, the probability of each primary electron causing a

Figure 5.18: An example of a trace that is a possible candidate for PMT afterpulsing in channel 1. The alternative windows that are used to find afterpulsing and random coincidence are identified with diagonal hash shading. In this case of the traces shown here, the window from 1000 ns to 1200 ns has a signal that corresponds to ≈ 5 PE and the window from 1200 ns to 1400 ns has a signal of ≈ 13 PE in channel 1 and there is no corresponding signal in channel 2.

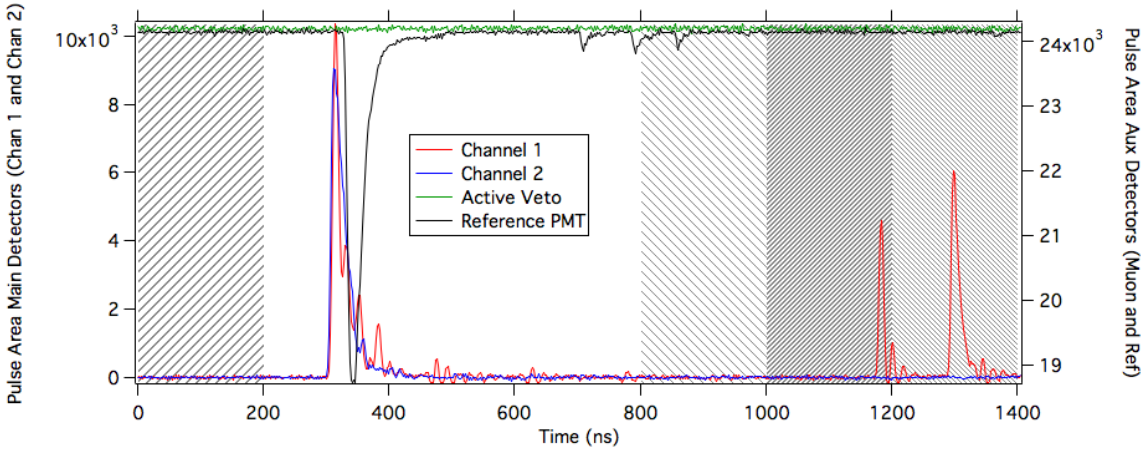
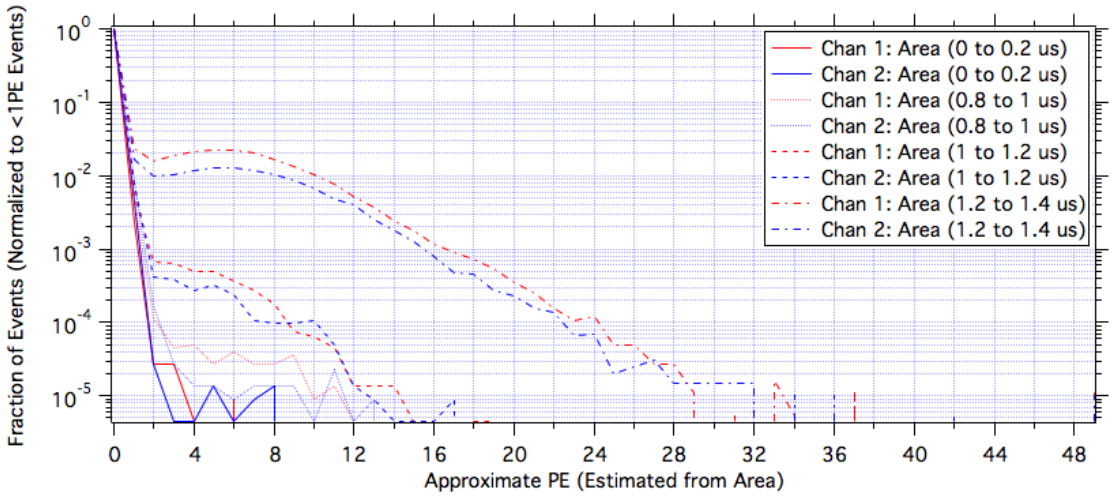


Figure 5.19: Histogram of the pulse area for reference events that are in windows outside of the region of interest to constrain the rate of PMT afterpulsing. The number of counts is strongly dependent on the window that is selected.



PMT afterpulsing peak can be estimated to be $\approx 0.5\%$.

Recall that there is also a veto window of 4.6 μs after each event during which the trigger logic is turned off. This will help prevent the DAQ from triggering on PMT afterpulsing events if they stray beyond the range of our digitization window. Also, recall that for PMT afterpulsing events to trigger their own event, they require random coincidence with an event that passes the discriminator threshold in the other channel, which drastically suppresses this background. When we combine these effects, it seems unlikely that PMT afterpulsing, despite being correlated with the data rate, will have a non-negligible systematic effect. This can be further motivated by looking for evidence of PMT afterpulsing in the data. One place that this could arise is in timing histograms that calculate the amount of time between an event of a given type and the next event. I did this for three types of initial events reference events, large slow event, and large fast events, where I define large to correspond to about 11 PE and the fast threshold is 10 in the local kurtosis, for this work I only looked at channel 1. For the reference events, the timing distribution seems to be well described by a single exponential, i.e. it is linear on semi-logarithmic plot, see Figure 5.20. Both the fast and slow large events show additional components at shorter times. The PMT afterpulsing should occur equally for all events.However, the reference events do not show these additional components. Therefore, this additional time dependence must be the result of other physical mechanisms and not the effect of PMT afterpulsing. Any deviation from a single time dependence in the reference events could be an indication of afterpulsing making it through the coincidence requirement and the pulse shape cut thresholds. I consider the fact that there is no evidence of an additional time dependence in the reference timing histogram as evidence that the effect of PMT afterpulsing is negligible in our data.

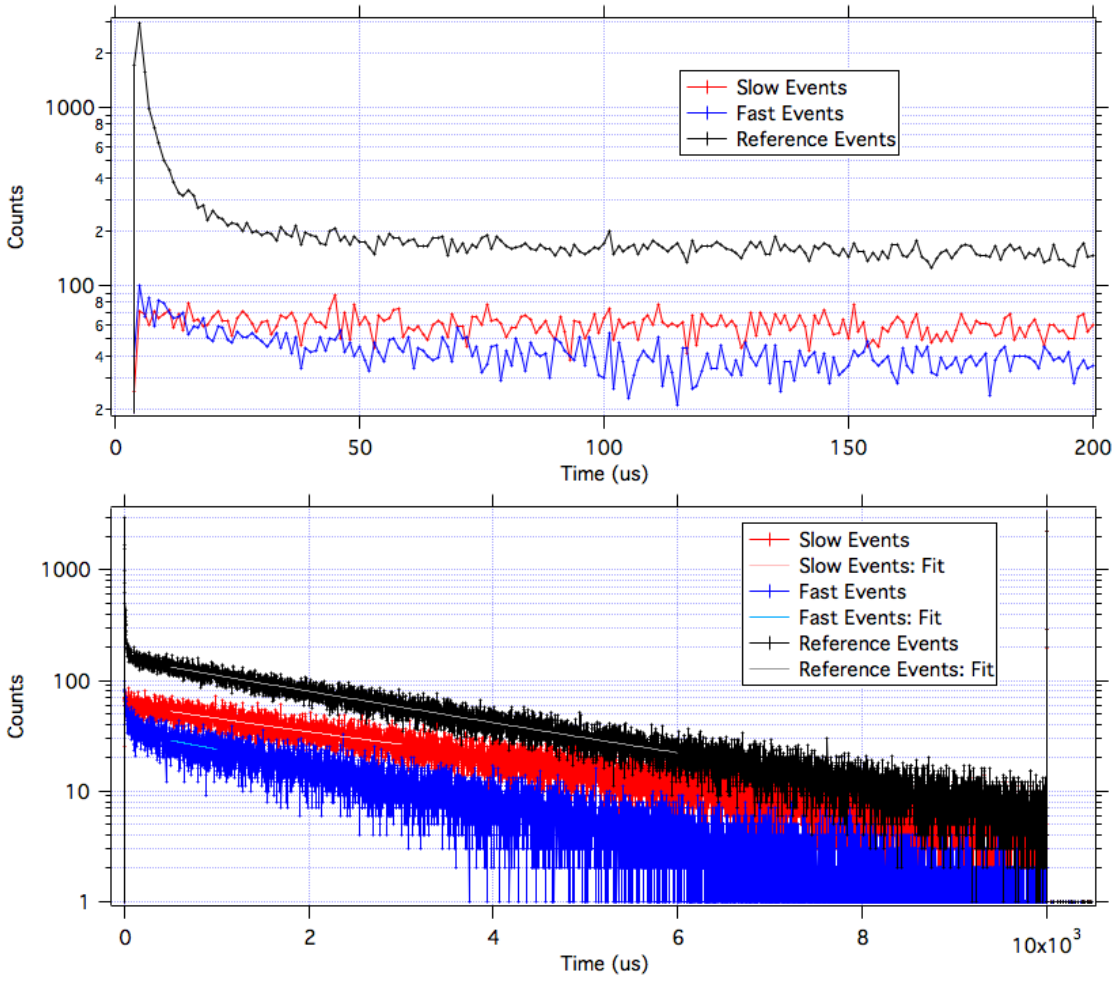
5.10.2 Neutron-Induced Luminescence

Neutron-induced luminescence of Boron Nitride is a process in which Boron Nitride spontaneously emits uncorrelated single photons after being exposed to neutrons. Because these events are uncorrelated, they will require random coincidence to trigger the DAQ. Once again the coincidence requirement and an individual, lower pulse area cut in each channel can drastically reduce this effect. In comparison to the PMT afterpulsing, the pulse area cut is expected to be more effective at removing these background events because they will primarily be single PE events. In contrast to the PMT afterpulsing, if there are multiple uncorrelated photons that allow the event to pass the lower pulse area cut, they are not expected to be tightly correlated in time. This broadness to the timing distribution is expected to cause the kurtosis cut to be much less effective at removing these background events.

Because the pulse height discriminator is above the height of a typical single PE peak, the

majority of these events are expected to require random coincidence in both the main detection channels to pass above the discriminator threshold. This means that the vast majority of these

Figure 5.20: Duration-between-events histograms indicating that there is minimal PMT after-pulsing in the data. Graphs of the time between an event of a given type and the next event in the data. There are three types of data indicated reference events, large slow events, and large fast events. In this case, slow and fast are identified by a local kurtosis threshold of 10 and large events are ones where the pulse area in channel 1 is > 11 PE. (Top) The three types of events show very different timing distributions at short times. (Bottom) By fitting to the highest statistics data to a single exponential and an offset over the range where the number of counts in each bin $\approx \geq 20$ a lifetime for the long term component is determined. This lifetime is then held fixed in the other data sets over fit ranges where the count rate is similarly large, and the fit shows good agreement, which indicates that the long-term component is the same in the different data types. This suggests that although the timing distribution is different at short times, at long times it is consistent in the different data types.



events are not digitized. It is expected that the majority of the luminescent photons that make it into the data are coincident with a single event that by itself was able to pass the trigger threshold in both channels. Below, we will discuss how we placed a limit on the rate of luminescent photons in the experiment. We found that it is very unlikely that more than a few PE could be introduced into an event from luminescence. Therefore, these events are not expected to have a large impact on any of the pulse shape metrics.

Using an analysis similar to that described above for the PMT afterpulsing, we can estimate the rate of uncorrelated single PE in our experiment and thereby limit the rate of luminescent photons. This time we chose the first window, i.e. the window before the trigger, because it has the lowest rate. Recall that Figure 5.18 shows the location of the windows with an example set of traces. Presumably, all of the windows include the uncorrelated single PE events, the rate being lowest in the first window seems to suggest that it includes the least amount of other background pulses. We find that about 0.2% (0.4%) of the reference events have an area equivalent to a single PE event in this early window in channel 1 (2). We can estimate the rate of these uncorrelated photons by dividing by the duration of the window, which comes out to 1×10^4 s^{-1} (2×10^4 s^{-1}) for channel 1 (2). This is larger than the data rate of all of the digitized events in this experiment, which is what motivated the use of a pulse height discriminator before the digitization to limit deadtime and the amount of data that needed to be stored.

For uncorrelated photons to trigger the coincidence, we need the timing of the two events to be within the coincidence window, which is 258 ns. Also, the pulse must pass above the hardware pulse height discriminator, which is set above the height of a 1 PE pulse. The height of 1 PE pulses will vary, but to simplify the calculation let us assume, that they do not. This would require that two 1 PE pulses overlap with each other, which requires a duration between them of $\lessapprox$ 10 ns. Assuming that there is a 0.2% (0.4%) chance of a single PE pulse inside a 200 ns window, the chance of it falling within a 10 ns window is 0.01% (0.02%.) The probability of two uncorrelated photons falling within the same 10 ns window to trigger the discriminator is this value squared, approximately 0.0001% (0.0004%) in channel 1 (2). To calculate the probability that this happens in channel 1 within 258 ns of it occurring in channel 2, we multiply the probabilities in the two channels getting $\approx 4 \times 10^{-8}$%. This percentage is small enough that we do not expect more than a few events composed entirely of uncorrelated photons in each data file and we conclude that they will introduce a negligible systematic effect.

Taking into account the variable size of single PE events, the rate will be somewhat higher than this, but this demonstrates that requiring coincidence strongly suppresses uncorrelated photons and we believe that any systematic effect will be negligible.

Alternatively, one of the channels could be triggered by luminescent photons while the other channel is triggered by some other background. This will still be suppressed by the coincidence requirements and the lower pulse area cut. Therefore, we conclude that although there are a

variety of ways that luminescent events could be detected with our system, these mechanism do not contribute a non-negligible systematic effect.

5.11 Adiabatic Condition and Spin Flips

A neutron that satisfies the adiabatic condition has sufficiently little angular inertia that it can reorient itself inside the field gradients supplied by the magnetic trap. As the adiabatic condition begins to fail, the UCN become frustrated, and there is a probability of a spin flip, which results in the neutron being lost from the trap. The question of how well the adiabatic condition is satisfied in our experiment has been raised repeatedly, which justifies a calculation of how well the neutrons in our trap satisfy the adiabatic condition.

The standard test for adiabaticity is

$$\frac{\delta B}{\delta t} \ll \frac{2\mu_n B}{\hbar} \approx 180 \times 10^6 \frac{T}{s}.$$

Time derivatives in the magnetic field can arise from the neutron moving in a field gradient, the magnetic fields being ramped, or a combination of these two effects.

The field gradient can be estimated as a change in the magnetic field over a distance divided by the time to cross that distance or as the change in the magnetic field times the velocity of the neutron. The trap depth is about 3 T. An upper bound on the highest field gradient in the cell can be achieved by taking the trap strength and dividing by half the cell width, ≈ 6 cm. This results in a field gradient of roughly

$$\nabla B \approx \frac{3T}{0.06m} = 52\frac{T}{m}.$$

Neutrons in our trap are limited to velocities $v_n \leq 10\frac{m}{s}$, which is the velocity of a 500 neV neutron. Combining these two values we get

$$\frac{\delta B}{\delta t} \approx \frac{\Delta B}{\Delta x}\frac{\Delta x}{\Delta t} = \nabla B v_n \approx 500\frac{T}{s} \ll \frac{2\mu_n B}{\hbar}.$$

Obviously, UCN traversing the trap easily satisfy the adiabatic condition.

When the field is ramped, a UCN may experience a slight increase or decrease in the time-dependent magnetic field. In the production data, the magnetic fields have never been ramped faster than $\frac{\Delta B}{\Delta t} = \frac{4T}{30s} = 0.13\frac{T}{s}$. This results in a fractional increase of in $\delta B/\delta t$ of 0.3% and therefore is a perturbation to the previous estimate. Therefore, it is reasonable to conclude that spin flips due to non-wall interaction, non-Majorana spin flips have a negligible effect on the trap lifetime.

The ramp rate that was used here is the average ramp rate. However, when the ramp

command is given, the computer software breaks the ramp command up into small steps every few seconds to achieve the desired average ramp rate. This results in an average ramp rate that is consistent with what was presented previously. However, if the inductance is low enough, the magnetic field ramp rate could be much higher than the average ramp rate directly proceeding a ramp command. Using the inductance of the quadrupole magnet, 58 mH, and the maximum voltage that the quadrupole power supplies can supply, 5 V, we can calculate the maximum ramp rate of the magnet to be 86 A/s. This corresponds to a shortest duration to ramp from 0% to 70% of 28 s. Therefore, we find that the ramp rate that was calculated previously is a very reasonable estimate of the maximum ramp rate possible with the combination of these power supplies and the quadrupole magnet.

Majorana spin flips occur when a neutron is exposed to a region with a weak magnetic field. They can be limited by developing a trap with no zero field regions. This is true for our trap, and we expect Majorana spin flips to be a negligible loss mechanism in this experiment.

5.12 Neutrons in the Background Data

Our collaboration has discussed neutrons surviving in the background data as a systematic effect[18]. It is true that at some, very small, level UCN will be trapped in the background data. However, this would not introduce a systematic effect. Once the magnetic trap is ramped back up in the background data file, any UCN that have not left the trap will be trapped and will experience the same loss mechanisms as the UCN in the trapping data. Therefore, the effect of these neutrons will be a decrease in the statistical sensitivity of the experiment by removing some fraction of the UCN signal during the background subtraction, but they would not cause the trap lifetime to change and therefore would not introduce a systematic effect.

The fraction of UCN that survive into the data acquisition phase of the background data could be estimated with the MC simulation, however, since they do not introduce a systematic effect and the fraction is expected to be very small, an effort to estimate this has not been warranted.

5.13 Combining the Systematic Effects

To obtain a fully corrected result, the systematic corrections must be combined, which requires a little bit of care. A representative number for each of the systematic corrections should be presented so that their relative sizes can be compared.

To calculate the final trap lifetime, the MTN systematic effect is accounted for first. The post processing of the MC provides a lifetime and uncertainty after taking into account the effect of MTNs. This lifetime and its uncertainty are used as the starting point for the calculation

of the final result. Then the scalar systematic corrections are applied. In this case, the only non-negligible, scalar systematic correction is the pulse height discriminator threshold. For this correction, a linear scaling is applied by multiplying the systematic correction by the ratio of the trap lifetime and the lifetime that was used to calculate the systematic effect. Similarly, the uncertainty in the systematic correction is scaled by the same ratio. The trap lifetime after linear corrections and its uncertainty will be referred to as $\tau_{lincorr}$ and $\sigma_{lincorr}$.

Then the effect of the exponential loss mechanisms is simulated in a simple MC simulation. To obtain a realization of a realistic fully corrected lifetime, lifetimes for each of the loss mechanisms and the partially corrected trap lifetime are sampled from realistic distributions. The partially corrected lifetime is sampled from a Gaussian distribution with a mean of $\tau_{lincorr}$ and a standard deviation of $\sigma_{lincorr}$. The exponential loss mechanisms depend on variables that have uncertainty. To account for how that uncertainty will result in uncertainty in the loss lifetime, each of the variables is also sampled from a Gaussian distribution with a mean and standard deviation corresponding to their value and uncertainty. These sampled variables are used to evaluate a lifetime for the loss mechanism. Once values are sampled for each of the lifetimes, they are combined according to the equation

$$\frac{1}{\tau_{lincorr}} = \frac{1}{\tau_\beta} + \Sigma_i \frac{1}{\tau_i},$$

where τ_β is the fully corrected lifetime, and τ_i is the lifetime of the ith exponential loss mechanisms. How the exponential lifetimes are sampled is shown in the two cases with exponential loss mechanisms.

The lifetime for the thermal up-scattering depends on the temperature, T, the coefficient A, and the corresponding uncertainties. Therefore, the lifetimes were sampled according to

$$\tau_{up} = (A_{avg} + gnoise(A_{sig})) \times (T_{avg} + gnoise(T_{sig}))^{-7},$$

where for a variable, X, X_{avg} is the mean of the variable, and X_{sig} is the uncertainty in the variable. Here $gnoise(arg)$ is a function that randomly samples from a Gaussian distribution with a mean of zero and a standard deviation determined by the argument arg. The mean and standard deviation of A and T are determined as described in the section on thermal up-scattering.

For the ^{3}He lifetime, the only uncertainty that is included is the uncertainty in the measurement of $R34$, which is expected to dominate. The value and its corresponding uncertainty are described in the corresponding section. The lifetime associated with ^{3}He absorption is sampled according to

$$\tau_3 = 3.9 \times 10^{-8} / \left(R34_{sig} + gnoise(R34_{sig})\right).$$

A corrected lifetime, τ_β, is calculated for each simulation and the standard deviation of the ensemble of corrected lifetimes is the uncertainty in the final lifetime. The final trap lifetime is calculated using the best estimates for each of the loss mechanisms. The difference between using the mean of the simulated lifetimes and the best estimate is a few tenths of a second.

This method makes it easy to account for asymmetric error bars in the systematic effects. Although the variables on which the exponential lifetimes depend are expected to follow Gaussian distributions, after accounting for how the lifetime depends on the variables, the resulting distribution may not be Gaussian. The histograms of the simulations are checked by hand to verify that the histogrammed values are accurately described by a Gaussian distribution. This ensures that reporting the uncertainty as the standard deviation is reasonable. A statistical test was not performed to verify this, but we found that the histograms fit well to a Gaussian and the center of the Gaussian agreed with the mean. The agreement between the results from the fit and the standard statistical parameters was taken as evidence that the statistical parameters accurately determined the average systematic shift and its uncertainty.

We also want to be able to compare the size of the different systematic effects. Additionally, it would be beneficial to get a feeling of how large each systematic effect would be if it were the dominant systematic effect. To achieve both of these goals, the systematic corrections are scaled to their expected size on the PDG mean lifetime value. This inherently assumes that the size of the systematic effects scales linearly with the trap lifetime, which is expected to be true for most of the systematic effects.

In Table 5.3, the quantitative estimates for our systematic effects are tabulated. Due to variations in the calculation methods, data with different lifetimes were used to calculate many of these systematic effects. In column one, the lifetimes at which the systematic effect was calculated is indicated, τ, and the corresponding systematic correction, Δ_τ. The systematic effects is also presented after being scaled by the ratio of the PDG mean lifetime and τ to correspond to the size of the systematic effect if it were the dominant systematic effect, $\Delta_\tau(880.2)$.

5.14 Final Results and Discussion

Using the methodology developed in the previous section, the fully corrected lifetimes can be calculated for the three main data sets. The values are included in Figure 5.21 and also Table 5.4.

After applying all of the systematic effects that can be accounted for, a large discrepancy remains between our measured value and the PDG mean lifetime of 880.2 ± 1 s. As was discussed in the section on ^{3}He absorption, we think that this is likely due to additional contamination that was present in apparatus during operation that was not properly represented in the sample that was measured. Before we discuss this and the implications in more detail, let us first discuss the systematic effects that have been determined, see Table 5.3.

Table 5.3: A summary of the systematic corrections for the experiment. The values are tabulated for the trap lifetimes, τ, and after scaling the values to the PDG mean lifetime of 880.2 s. The parentheses are the statistical uncertainty, and the square brackets are systematic uncertainties due to model assumptions in the simulation. $*$ This indicates that the value was calculated as a percent change in the lifetime instead of being calculated at a particular lifetime value.

Systematic	τ	$\Delta_\tau(\tau)$	$\Delta_\tau(880.2)$
	s	s	s
Upscattering	880.2	0.7(0.4)	0.7(0.4)
He3 Absorp.	880.2	29(10)	29(10)
MTN 60%	671(45)	45.0 (2.2)[8.9]	59.0(2.9)[11.7]
MTN $70-50-70\%$	656(21)	6.6(1.9)[1.3]	8.9(2.5)[1.8]
MTN $70-35-70\%$	748(57)	0.03[0.01]	0.04[0.01]
Gain Corr.	*	–	-0.07(0.35)
Detection Eff.	880.2	15.7(2.5)	15.7(2.5)
Clock Drift	*	–	0.03
Background Sub	675	0.03	0.04
Fitting Bias	619	$\leq$0.06(0.19)	$\leq$ 0.09(0.27)

It is clear that most of the systematic effects are controlled at or below the level required for a 1 s measurement of the neutron β-decay lifetime. There are a few systematic effects that appear to be exceptions to this, which should be discussed. First, recall that the systematic effect due to MTNs can be controlled at the level required for a 1 s measurement. In this case, the data sets that experience larger systematic corrections were taken intentionally in order to benchmark the Monte Carlo simulation. Time-dependent drifts in the detection efficiency also resulted in a large systematic correction in the data that is discussed in this work, but in the corresponding section, a method for reducing this systematic effect to the level required is presented. Finally, ^{3}He absorption has a larger than expected systematic effect, which requires additional discussion.

Recall that the systematic correction for the ^{3}He absorption that is tabulated here is from a direct measurement of helium that was extracted from the apparatus. The size of this systematic effect demonstrates that the contamination in the sample was larger than was expected. Our understanding was that the helium that we used in the experiment had a level of contamination of less than a few parts in 10^{15}. However, as the initial data came in and a large, unexpected systematic effect started to become more likely, an effort to eliminate possible candidates for this systematic effect commenced. As more and more of the potential systematic effects were vetted, ^{3}He contamination became an increasingly likely candidate, but it is very challenging

Table 5.4: The final, fully corrected lifetimes for the main data sets. For reference, the PDG mean lifetime and the quoted uncertainty are included, which is inflated artificially to account for the disagreement in the reported values[3].

Data Type	τ	σ_τ
	s	s
60%	702	49
$70-50-70\%$	696	23
$70-35-70\%$	790	62
Weighted Avg.	707	20
PDG mean	880.2	1

to measure. This provided additional motivation for work that was already underway in collaboration with Argonne's ATLAS facility to directly measure the helium contamination in a sample from our apparatus. In the course of that work, our method was refined to the point where we believed a direct measurement of the purity required for this experiment could be performed.

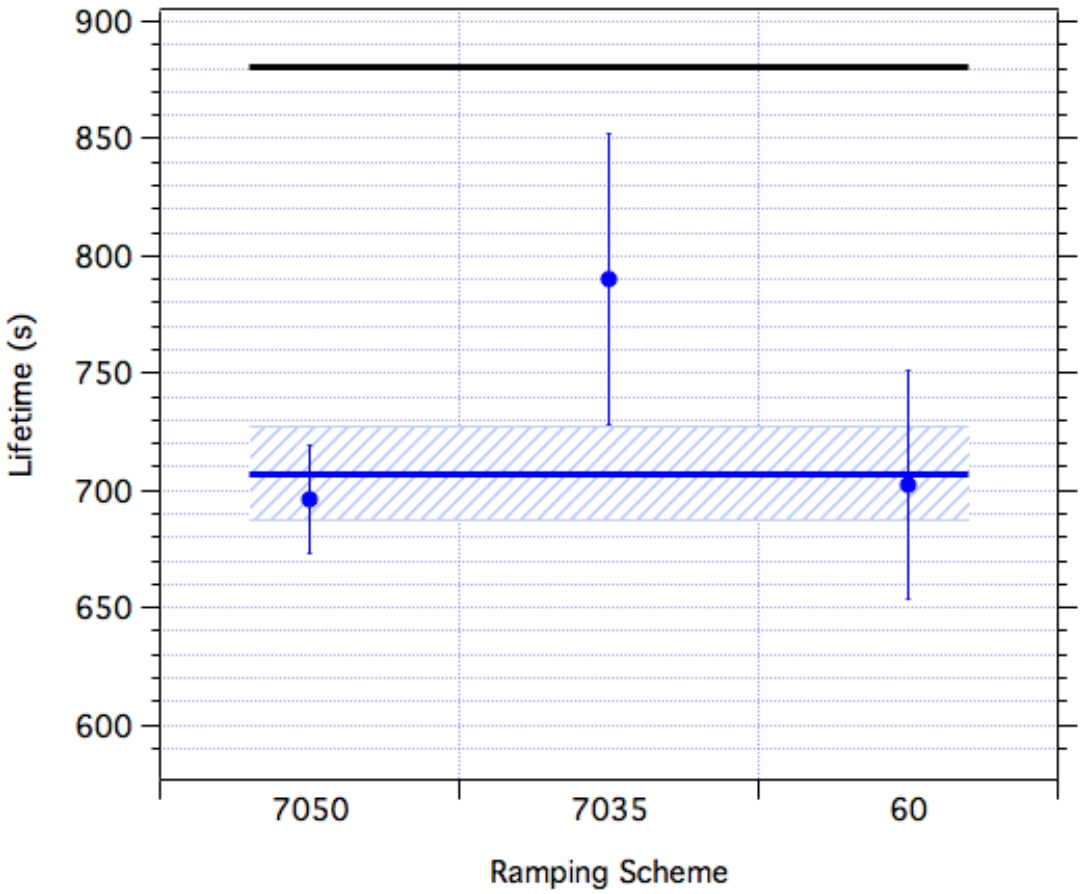
At that point, the sample that was measured at Argonne National Lab was the only remaining sample of helium from our apparatus. We also knew that because of the rupturing of the vacuum window on the cell that the sample was not an ideal sample. If the vacuum window had not ruptured, all of the helium would have been extracted into high-pressure storage volumes at room temperature, and the sample from our experiment would have come from the entire volume of helium, which would have been characteristic of the purity of the helium in the cell during operation. Instead, because of the vacuum window rupture, only the first helium that was extracted from the cell was not lost. When the purity measurement of the sample was measured, it was clear that the contamination in that sample was not sufficient to explain the difference between our measured value and the PDG mean value. This motivated a discussion about whether the first helium that was extracted from the cell would be characteristic of the helium in the cell during operation. As mentioned in the section on the ^{3}He absorption, we realized that the thermal gradients in the cell would be in the appropriate direction to purify the helium as it was extracted from the cell using the same mechanism that the helium purifier uses, the heat flush. For the heat flush to be established, very specific operating conditions must be met that include a small mass flow, a relatively stable system, and temperatures around the 1 K to 2 K range. Unfortunately, the thermometry in the apparatus was not designed to determine if the apparatus was suitable for a heat flush when the experiment was being warmed. Therefore, we are not able to give conclusive evidence that there was or was not heat flush purification that caused the sample that was measured to be purer than the helium in the

apparatus during operation.

All of the systematic effects, with the exception of ^{3}He, have been shown to be controlled at the level required for a 1 s measurement. Although we are not able to prove whether or not additional ^{3}He contamination is causing the remaining systematic effect, we believe that it is likely. Without verifying that ^{3}He can be controlled at the level required for a 1 s measurement, it would not be motivated to take additional data with this method. Therefore, a clear path forward is to produce a new batch of isotopically pure ^{4}He and to verify that its purity is at or below the level required for a 1 s measurement of the neutron lifetime.

To achieve this goal, a new heat flush purifier has been commissioned and operated at NIST. This purifier is the topic of Chapter 6. This work was being completed at the same time as

Figure 5.21: The final lifetimes after correcting for all of the significant systematic effects. The weighted fully corrected lifetime is 707 ± 20 s with a reduced χ^2 of 1.001 and a p-value of 0.37.



the analysis of the lifetime data and the development of the method for directly measuring the purity of the helium samples at ATLAS. Unfortunately, the first sample that was produced with the new purifier was created after the most recent set of purity measurements that are described in [55]. Our collaboration is currently in the process of requesting beam time to measure the new sample. Once that sample is measured, the purifier can be run again to produce more samples or a new batch of isotopically pure ^{4}He for the UCN Lifetime Experiment at NIST.

Chapter 6

^{4}He Purifier

In the UCN Lifetime Experiment at NIST, LHe is used both as a moderator for the UCN as well as a scintillator for detecting the neutron β-decay events. The experimental cell is at the center of the cryostat and houses the volume of LHe. It is also the region of the apparatus where the UCN are produced and stored throughout the measurement. ^{3}He has a large absorption cross section for neutrons. As a result, any ^{3}He in the cell will provide an additional loss mechanism and therefore will decrease the trap lifetime. The isotopic ratio must be less than $R_{34} = \frac{n_3}{n_4} \leq 1 \times 10^{-15}$ to limit the systematic effect to $1 \times 10^{-3}\tau_n$, see Section 5.2, which starts on page 136, where n_3 and n_4 are the number density of ^{3}He and ^{4}He atoms in the sample and τ_n is the neutron β-decay lifetime. The isotopic purity of natural sources varies between $R_{34} \approx (1 \times 10^{-6}$ to $1 \times 10^{-7})$[57]. Consequently, a substantial level of purification is required for this experiment.

A new batch of isotopically pure ^{4}He is needed for the UCN Lifetime Experiment at NIST. To produce this isotopically pure ^{4}He, a purifier has been built at NC State that is based on the heat flush technique and closely mirrors the design by the McClintock group, which provided the original batch of isotopically pure ^{4}He used in the UCN Lifetime Experiment at NIST.

The following sections describe methods of purifying ^{4}He with a more detailed discussion on the method used in our purifier, the heat flush method. A brief theoretical description of the heat flush method is followed by experimental verification. The remaining sections describe some of the components of the purifier, its performance, and suggested upgrades to the purifier that I had a leading role in designing, constructing, and operating. Procedures for operating the purifier can be found in Appendix E, which starts on page 238

6.1 Introduction

A variety of methods could conceivably be used to purify ^{4}He by taking advantage of how the properties of the two isotopes differ as a function of the temperature, pressure, or any other relevant operating parameters. For example, one could leverage the relative pumping speed of the two isotopes to preferentially remove ^{3}He from a sample. To the best of my knowledge, only three methods of isotopic purification have ever been published. They are distillation, superleaks, and the heat flush. The following paragraphs describe the two more frequently published methods, superleaks and the heat flush.

The use of superleaks is the older of the two techniques. Superfluid can move through passageways that are too small for normal-state LHe to pass through. A device that takes advantage of this property is called a superleak. Since ^{3}He is part of the normal fluid, it can not pass through the superleak. Therefore any helium that passes through the superleak will be purified ^{4}He. This allows for a high level of purification and a sufficiently large output flow rate to be feasible for producing liquid liter quantities of purified ^{4}He. As an example, Ref [58] uses a superleak in a continuous flow purifier.

The second method is known as the heat flush method. The McClintock group has published quite a few papers about their use of heat flush purifiers to produce isotopically pure ^{4}He[54]. In the heat flush method, a roton flux is established that sweeps ^{3}He from a heater to a cold source. Similar to the method described above, helium mixture is siphoned through the purifier and, in this case, the heat flush produces a potential barrier that the ^{3}He can not overcome.

The McClintock group has used heat flush purifiers exclusively. This is motivated by concerns of contamination because, in practice, a small amount of ^{3}He will always pass through a superleak. Additionally, helium is notorious for absorbing into materials and desorbing out over long time scales. Therefore, additional contamination could come from ^{3}He that has been absorbed into a superleak and may desorb out during purification contaminating the ^{4}He. The large effective surface area of superleaks will exacerbate this effect. The heat flush method, in comparison, allows the use of a purification region that has a much smaller effective surface area, which substantially limits the effect of desorption of ^{3}He from the surface. It seems prudent to follow in the footsteps of McClintock and to build a purifier that uses the heat flush method. This warrants a more detailed discussion of this purification technique.

The heat flush is a method of purification of ^{4}He that takes advantage of the quantum mechanical nature of LHe to create an isotopic purity gradient. LHe changes state into a superfluid at 2.17 K. This novel state has non-Newtonian properties that do not exist in classical fluids. For example, it has extremely high heat conductance for a fluid and extremely low viscosity. In the superfluid state, the liquid is a quantum liquid meaning that many of the properties of the fluid can no longer be described as an ensemble of molecules with ballistic trajectories bouncing

off of each other. Instead, it begins to act like a quantum state where the energy of the system can be thought of as excitations above some ground state. In this case, the ground state is the liquid at 0 K. The properties of the system at higher temperatures are brought about by adding excitation quanta to the system thereby exciting the liquid into a higher energy state.

The excitations can be broken into distinct types based on the ratio of their energy to momentum. In this model, it is the properties of the ground state in addition to the properties and densities of these excitation quanta that determine observable properties of the system as a function of the temperature. As an example, the density of the liquid can be written as $\rho(T) = \rho_s(T) + \rho_n(T)$. The first term is the contribution of the ground state. It contains the superfluid properties of the liquid and therefore is labeled with a subscript "s". The second term is the combined contribution from all of the excitation types. The excitations are labeled with a subscript "n" to signify that they make up the normal fluid contributions to the fluid properties. The contribution from the excitations is temperature dependent because the number density of each of the types of excitations changes as a function of the temperature. Empirically modeling the observable properties of LHe as consisting of a superfluid and normal fluid contribution provides a reasonably effective model for predicting the properties of the superfluid.

The heat flush can be described as the bulk transport of ^{3}He due to scattering from a roton wind. Rotons are one of the elementary excitations in helium and have a large cross section for scattering ^{3}He of 1.6×10^{-14} cm^2 at 1.27 K[59]. Therefore, a roton wind with sufficient flux directed along the length of a tube purifies one end of the tube as it sweeps all of the ^{3}He to the other end.

In the 1 K to 2 K range, rotons are the largest contributor to the properties of the superfluid, and therefore the dominant excitation. Since superfluid helium is a quantum fluid, any heat that is applied to the fluid can be thought of as the creation of excitations. A heater, therefore, is a source of rotons. A refrigerator will correspondingly act as a sink of rotons. Hence a roton wind and a resulting isotopic purity gradient can be realized by placing a heater on one end of a flushing tube and an evaporative helium refrigerator on the other end. The roton wind sweeps all of the ^{3}He contamination from the end with the heater toward the cold end, thereby purifying the helium near the heater.

To use the isotopically pure ^{4}He for the UCN Lifetime Experiment at NIST, it must be extracted from the warm end of the flushing tube. This is performed by applying suction to the warm side of the flushing tube to induce a small mass flow. The mass flow effects both the ^{3}He and ^{4}He, however for sufficiently small mass flows the ^{3}He velocity due to the heat flush can overcome the mass flow. This allows the ^{4}He to be extracted from the warm side of the flushing tube while flushing the ^{3}He contamination to the cold side of the flushing tube. In principle, this is how the heat flush method works. The rest of the complication of the purifier results from the logistics of setting up the roton wind, maintaining stable operation, and extracting

the isotopically pure ^{4}He.

6.1.1 ^{3}He Heat Flush

Following the derivation presented in Hendry's thesis[60], the heat flush is modeled by the one-dimensional equation of gas attack. The solution is an exponential where the length scale of the purification is a struggle between how effectively the heat flush forces the ^{3}He through the flushing tube, which manifests itself in the normal fluid velocity, and the rate at which the ^{3}He diffuses against this current, the ^{3}He diffusion constant,

$$u(x) = u_0 \exp\left(-\frac{v_n x}{D}\right) = \exp\left\{\left(\frac{-x}{5\times10^{-4}\,(T+5.8)}\frac{\dot{Q}}{A}\right)\right\}.$$

Here $u(x)$ is the ^{3}He concentration along the flushing tube; u_0 is the purity of the helium on the cold end of the flushing tube, for typically natural purity $u_0 \approx 2\times10^{-7}$; v_n is the normal fluid velocity, of which ^{3}He is a component; D is the diffusion coefficient of ^{3}He in ^{4}He; $\dot{Q}$ is the heater power in W; A is the flushing tube cross-sectional area in m^2; x is the length of the flushing region in m; and T is the temperature in K. This functional form indicates that the purification is much more sensitive to the heater power and the flushing tube length than the operating temperature because of the large offset to the temperature in the denominator. The temperature must, however, remain in the 1.2 K to 1.8 K range where the rotons are still the dominant excitation.

The location in the tube where the purity reaches the desired level can be estimated using conservative values for the operational parameters of the purifier and solving the equation of gas attack for the distance. In our geometry, the inner diameter of the flushing tube is 0.013 m. During typical operation, the temperature was approximately 1.6 K, and the heater was operated at 0.5 W. Purification by a factor of roughly 1×10^{-8} is needed to reach the desired purity of 1×10^{-15}. Using these parameters, this level of purification occurs 0.06 m from the location of the heater. Adding a second factor of 1×10^{-8} allows a substantial safety factor and would take $\approx$ 0.12 m. The flushing tube consists of a multiple, straight, vertical sections that are connected by elbows. Each of the vertical sections is roughly 0.3 m in length, for a combined flushing tube length of over a meter. This gives us confidence that the purifier remains insensitive to any breakdown in the one-dimensional solution to the gas attack equation due to the bends in the flushing tube and suggests that we have an ample safety margin ensuring that the purifier will perform as expected to the extent that the theory is valid.

One mechanism that could lead to the breakdown of the one-dimensional equation of gas attack is convection in the flushing tube. This will be discussed in the list of possible upgrades to the system, see Section 6.4, which starts on page 213.

Operating Temperature Range

We want to establish an operating temperature range in which the heat flush performs satisfactorily. The effectiveness of the heat flush method is expected to vary substantially with the temperature. The operating temperature range of the purifier is estimated to be between 1.2 K to 1.8 K.

Below 1.2 K, the roton density starts to drop substantially. At 1.1 K, the rotons are no longer the dominant excitation[61]. Correspondingly, the mean free path of ^{3}He increases, which allows diffusion of the ^{3}He to start competing with the heat flush. As the temperature rises, the normal fluid constitutes a larger fraction of the mass of the fluid. Therefore, a smaller normal fluid velocity is required to maintain the concentration gradient in the flushing tube. At the high-temperature limit of 1.8 K, the normal fluid velocity starts to fall dangerously close to the mass flow velocity of the LHe through the purifier. If the normal fluid velocity does not overwhelm the induced flow velocity through the flushing tube, ^{3}He may be able to diffuse against the roton wind. To estimate this effect, let's calculate the induced flow velocity through the purifier and the normal fluid velocity.

The linear flow velocity inside the flushing tube is estimated to be 1.6×10^{-4} m/s by scaling the extraction rate of isotopically pure ^{4}He from the purifier to the cross-sectional area of the flushing tube. With this linear flow rate, the average transit time for ^{4}He passing the flushing tube is ≈ 6000 s.

The normal fluid velocity is a simple function of measured properties of the liquid and operating parameters of the purifier[52]. My notation differs somewhat from what is presented there;

$$v_n = \left(\dot{Q}/A\right) / \left(\rho S T\right),$$

where $\dot{Q}$ is power flux down the flushing tube, A is the cross-sectional area of the flushing tube, ρ is the total density of the fluid, S is the entropy per unit mass, and T is the temperature. The density and entropy are tabulated in Ref [62]. Using the geometry of the apparatus and our telemetry, we estimate $A = 5.3 \times 10^{-4}$ m^2 and $T = 1.5$ K. $\dot{Q}$ is more challenging to estimate. The power output of the flushing heater is a reasonable estimate of this quantity, but the actual power flux through the liquid will be somewhat lower due to whatever portion of the heater power is carried away by mechanisms other than roton transport down the flushing tube. Other possible loss mechanisms include thermal conduction down the copper tube, thermal conduction through the extraction line, radiative loss, convection in the vacuum can, and the helium "reflux". We can break this value into an unknown constant representing the fraction of the heater power transmitted through the flushing tube as rotons, ϵ, and the operating power

of the heater, $\dot{Q}_h$,

$$\dot{Q} = \epsilon \dot{Q}_h.$$

We do not have data to constrain ϵ. However, it seems unlikely that $\epsilon \leq 0.5$, which is the situation where half of the heater power output is carried away by mechanisms other than the roton flux down the flushing tube. This is because of the large thermal conductivity of superfluid helium, the diameter of the flushing tube being much larger than the diameter of the extraction tube, and the expectation that other heat loss mechanisms have a much smaller effect. Therefore, we expect that ϵ alone will introduce an error into our calculation that is somewhat less than a factor of 2. Using $\epsilon = 0.75$, we get a normal fluid velocity of $v_n = 0.016$ m/s.

We find that the estimated normal fluid velocity is a couple of orders of magnitude larger than the induced mass flow through the flushing tube. To the extent that the velocity of ^{3}He is well approximated by the normal fluid velocity, this suggests that the mass flow velocity is not large enough to overcome the roton wind. As a caveat, the normal fluid velocity is inherently a statistical quantity; therefore even if the average ^{3}He velocity can be estimated by the normal fluid velocity, the individual ^{3}He atoms will vary about this average due to random fluctuations caused by the random walk motion associated with the diffusion. Therefore, although the difference of a few orders of magnitude is suggestive that the heat flush method will work, without a proper calculation of the variation in individual velocities, it is not sufficient to show that no ^{3}He atoms will be able to overcome the heat flush.

Experimental Verification

A large set of complementary experiments forms the foundation of the heat flush method. In the following section, I list a few notable experiments that demonstrate the heat flush method in a straightforward way. These experiments conclusively demonstrate that the heat flush can be used to purify ^{4}He.

The simplest experiment used a pair of helium volumes connected by a flushing tube[59]. A thermal gradient was set up in the flushing tube, and a mass spectroscopy measurement was performed on the helium vapor from the cold bath. The signal in the mass spectrometer showed a marked change in the relative purifications once the thermal gradient was established. The measurement was performed on the ^{3}He rich side to take advantage of the two relative rates being more similar on that side. The relative pumping speeds of ^{3}He and ^{4}He result in a substantial systematic effect; despite this, their work still demonstrates the effectiveness of the heat flush method.

Fukuda and Hirai took it a step further by performing NMR on the ^{3}He in the flushing tube[63]. This provides a separate method for observing the purification and allows direct observation in the region of the heat flush. It is clear, from these works and a variety of others,

that the heat flush method is quite effective at these relatively high concentrations of ^{3}He where the change in concentration can be measured with well-established measurement techniques. Using these methods to estimate the concentration on the warm side of the flushing tube still requires measuring the ratio of ^{3}He to ^{4}He at the 1×10^{-15} level to obtain the limit required for this experiment, which has not previously been technically feasible.

To avoid measuring purity at the 1×10^{-15} level, the McClintock group designed a reverse concentration cryostat. It processes a sample of isotopically pure ^{4}He through a flushing gradient in such a way that it concentrates all of the ^{3}He into a small volume. At the end of processing, the concentrated sample is isolated and extracted for measurement using traditional mass spectroscopy. Neither the details of the sample processing nor the data are accessible, and the author gives the impression of doubts in the reliability of the reverse concentration cryostat. However, they report an upper threshold of $R_{34} < 5 \times 10^{-13}$ and say that they suspect that the actual value is much better.

At present, no samples have ever been measured at the desired level $R_{34} = 1 \times 10^{-15}$ that came from a continuous flow purifier even with the use of the reverse concentration cryostat. This could either be a failure of the purification process, the measurement process, or a breakdown of the theory at high purity. We believe it is likely that the theory is valid and that the previous results are due to the difficulty in eliminating ^{3}He backgrounds in the measurement process. Building a new purifier, taking new samples, and measuring these samples at Argonne National Lab (ANL) is a logical step towards understanding the disagreement between the experimental results and expectation.

6.2 Apparatus

A new purifier has been constructed and commissioned to produce a new batch of isotopically pure ^{4}He for the UCN Lifetime Experiment at NIST. A cartoon of the purifier is presented in Figure 6.1. The purification occurs in the flushing tube. Natural purity LHe enters the purifier through NV1. It precools as it passes through a heat exchanger. Next, it continues to cool as it passes through the small tube of the flushing tube, which is wrapped around the large tube of the flushing tube and soft soldered in place providing thermal contact. It passes into the cold end of the flushing tube just under the copper cone of the 1K Pot. Here it is exposed to the roton wind from the flushing heater, which is located on the far end of the flushing tube. The ^{3}He is overwhelmed by the roton wind and flushed to the cold end of the flushing tube where it is slowly pulled into the 1K Pot via NV2. The isotopically pure ^{4}He is pulled through the flushing tube, NV3, and the other side of the heat exchanger before exiting the cryogenic volume via a vacuum jacketed extraction line to room temperature for storage. The mass flow that pulls the ^{4}He through the purifier is supplied by a diaphragm pump on the gas handling

system.

The cooling power of the purifier is provided by evaporative cooling of ^{4}He in the 1K Pot. It can be supplied with helium from the cold end of the flushing tube via NV2, directly from the main helium bath through NV4, or a combination of the two.

Each of these components will be described in more detail below. Additionally, many subsystems that were not described here will also be introduced. The following sections list the operational parameters of many of the purifier systems. Information is included that influenced the design of the purifier, may help others select components when designing a similar apparatus, or is needed for calculations of either the purity or performance characteristics of the purifier. Additional information is given to identify and estimate systematic uncertainties of the sensor measurements that give diagnostic information during operation.

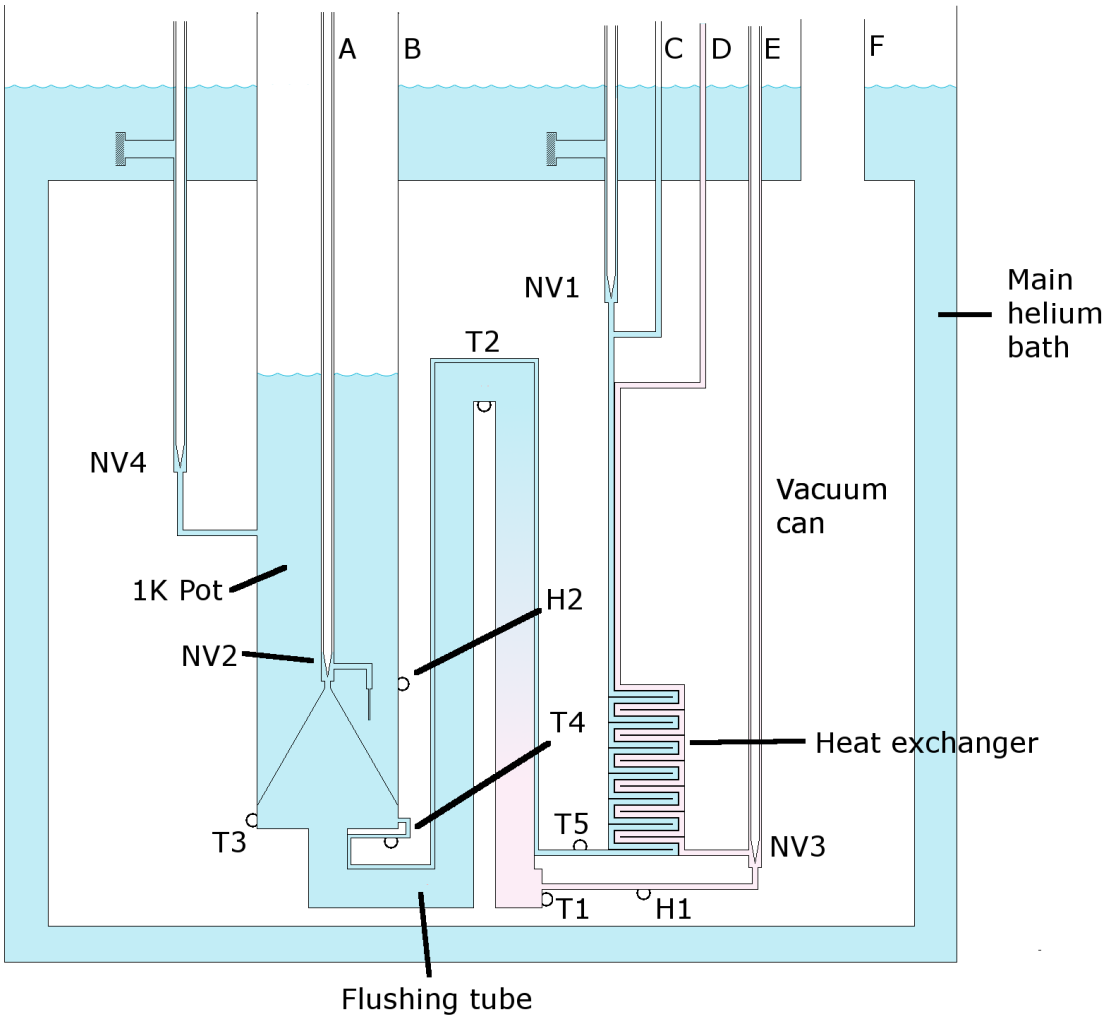
6.2.1 Pureloop

The primary components that influence the performance of the purifier are inside the vacuum can. They house and transport the superfluid helium during the operation of the purifier. The material and relative size of various components determine key information including the heat loads, the cooling power, and the liquid flow rates, all of which are important parameters for understanding the operation of the purifier. The pureloop includes the flushing tube, the 1K Pot, and the heat exchanger. The heaters are also instrumental in characterizing the heat flush. Therefore, a brief description of their design, mounting method, and operating powers are included in this section.

Flushing Tube

The flushing tube is the region of the purifier where the concentration gradient is established. Helium throughout the flushing tube is exposed to the roton wind from the flushing heater, which is attached a few inches past the warm end of the flushing tube on a 1/8 in copper tube. As mentioned previously, the concentration gradient takes place very close to the cold source, therefore in typical operation, only the first 0.12 m of the flushing tube contain a sizeable fraction of ^{3}He. This suggests that during typical operation, the majority of the flushing tube is filled with isotopically pure ^{4}He, which causes it to act like a buffer volume and makes the purifier less sensitive to short term fluctuations. To fit the flushing tube in the vacuum can, it is comprised of a set of 4x ≈ 0.3 m long vertical sections that are connected by pairs of elbows acting as U junctions at the tops and bottoms. The flushing tube has an outer diameter of ≈ 0.016 m and an inner diameter of ≈ 0.013 m. The helium inlet takes natural purity helium from the outlet of the heat exchanger. It coils around and soft soldered to the flushing tube to further cool the incoming helium and to provide a radial component to the heat flush. The

Figure 6.1: A diagram of the purifier. Natural purity helium is represented by light blue. isotopically pure ^{4}He is pink. The flushing heater is designated H1 and is located between the heat exchanger and the flushing tube. The 1K Pot provides evaporative cooling. The vacuum can isolates the pure loop from the main helium bath. The locations of the thermometers and heaters are indicated, where the thermometers are labeled T# and the heaters H#. "A" and "E" are the needle valve assemblies for NV2 and NV3 respectively. At the purifier top plate they both have KF-16 pumpout ports to assist in evacuating the system. "B" is the 1K Pot pumpout line. "C" is the helium inlet pumpout line. "D" is the ultrapure extraction line. "F" is the vacuum can pumpout line.



radial heat flush guides ^{3}He from the walls toward the center of the tube where the flushing velocity is highest[60]. The pitch of the wrap is such that the inlet tube completes a revolution around the large tube every few centimeters.

1K Pot

The 1K Pot provides the cooling power for the purifier. It takes advantage of evaporative cooling of LHe to maintain temperatures as low as 1.4 K. The 1K Pot is effectively a pumping line, and increasing the pumping speed is one of the key design parameters for minimizing the base temperature of the purifier. Therefore, the 1K Pot was designed to have the largest inner diameter allowed by the other spatial constraints.

To achieve this goal, the 1K Pot has two junctions that allow transitions to larger diameter tubes as additional space becomes available further up the purifier. The vacuum can has the tightest spatial constraints; inside the vacuum can the 1K Pot uses a tube with a 3.6 cm ID and 0.5 m length. About 0.1 m above the vacuum can top plate, the 1K Pot transitions to a 4.8 cm ID tube that extends to the top plate of the purifier. Finally, the 1K Pot extends above the purifier top plate with a 7.3 cm ID tube of length 0.33 m. This allows sufficient room for a KF-50 fitting to be welded to the side of the tube for connecting to the GSX via a 2 m KF-50 pumping line. It also provides sufficient space on the 1K Pot top plate to accommodate the handle for NV2, an electronics feedthrough, and a one-way safety relief valve. All of the structural tubes of the 1K Pot are made out of stainless steel.

The 1K Pot houses the NV2 assembly, which runs the entire length of the pot to the top of the flushing cone. There are also three small support tubes that provide additional structural support. The flushing cone is the bottom of the 1K Pot and acts as the boundary separating it from the flushing tube. The conical geometry results in an axial normal fluid velocity that is independent of the location along the cross section. This promotes effective transport of the ^{3}He into the 1K Pot through the flushing cone[60]. The cone is 0.2 m tall.

Finally, the 1K Pot contains two level meters for determining the liquid level, which are described in their own section. A helium fill port allows helium from the main bath to fill the 1K Pot through NV4. Helium can also enter the 1K Pot from the flushing tube, through a capillary and then NV2. The capillary is 0.029 m long and has an ID of 3×10^{-4} m.

Without radiation baffles, the 1K Pot top plate, which is at 300 K, would have a direct line of sight to the coldest part of the purifier. Radiation baffles were included to limit this heat load. The 1K Pot has 3 copper baffles in the region of the flushing cone, the bottom 0.2 m of the flushing tube. Above the copper baffles are an additional 5 brass baffles before the transition to the 0.048 m OD tube. The five lowest baffle were used to hold the level meters and therefore, at least some of them are expected to be below the liquid level when operating the purifier. An additional set of ≈ 6 stainless steel baffles are positioned above the first extension in the 1K Pot. All of the baffles are spaced approximately uniformly in their corresponding section of the 1K Pot pumping line. The stainless steel baffles, in particular, were designed to prevent direct line of sight from each baffle to the baffle that is two lower. They are sized to fit snuggly to

prevent direct line of sight down the wall of the tube, and vent holes with an OD of ≈ 1 cm are included on alternating sides of the baffle to allow the system to pump more effectively. The wiring for the level meters is woven through these pumpout vents.

Heat Exchanger

A heat exchanger is used to precool the helium entering the pure loop. The original design consisted of a pair of concentric tubes wrapped into a helix around a mandrel and then annealed to maintain their shape. During the heat treatment, the weld joint on the stainless steel tubes failed. This design was expensive and had a long lead time because it used non-traditional tube sizes to balance the impedance of the inner tube and the passage between the two tubes.

A new heat exchanger with a simpler design was constructed from two parallel 1/8 in copper tubes of total linear length 0.91 m. The 1/8 in copper tubes were tied together with stainless steel wire, wound into a spiral shape, and then soft soldered together. The soft solder provides thermal contact between the tubes allowing the cold liquid leaving the system to cool the 4.2 K liquid entering the pure loop from the main helium bath. This design has a much lower surface area contact between the two tubes, which limits the effectiveness of the heat exchanger; however, it is cheap, easily assembled, and can be easily extended if additional cooling is required by adding length to the heat exchanger.

Thermometer T5 was included to assess the performance of the heat exchanger to determine if the design needed to be modified. Unfortunately, T5 gave sporadic readings during operation that prevented a clear assessment of the performance of the heat exchanger. T5 is attached to the helium inlet 1/8 in copper tube, leading away from the heat exchanger. The flushing heater is on the helium extraction 1/8 in copper tube on the same side of the heat exchanger. With this geometry, T5 is about 0.08 m of 1/8 in copper tube away from the flushing heater. This is not a direct heat path because the two components are on separate tubes, but they do have relatively good thermal conductivity because of their contact in the heat exchanger. When the flushing heater is operated, the temperature of T5 is found to increase substantially. During typical operation, it reached values as high as 4.2 K, and the temperature jumped around sporadically. It is not clear if the helium leaving the heat exchanger is really at 4.2 K, i.e. the heat exchanger failed to properly cool the incoming helium. Alternatively, the reading on the thermometer could be unrepresentative of the helium temperature.

It seems likely from this behavior that, at the least, the heat exchanger was not operating at full efficiency because the flushing heater was warming the copper tubing too much. It is unclear if the heat exchanger was cooling the incoming liquid with the thermometer placement that was used while operating the purifier. To address this, thermometer T5 should be moved as far from the heater as possible to limit this effect. Instead of measuring the temperature of the

inlet helium after the heat exchanger, an alternative is to measure the temperature difference of the extracted helium across the heat exchanger. Moving T5 to the far side of the heat exchanger on the extraction line would allow this measurement to be performed by comparing T5 and T1.

Heaters

Two heaters were installed in the system. The locations of these heaters can be seen in Figure 6.1. The heat flush heater, designated H1, provides the heating power that creates the roton wind. It is attached using an electronics mount, see Section 6.2.2, which starts on page 196, to a 1/8 in copper tube a few centimeters beyond the warm end of the flushing tube. The flushing heater does not use a Faraday cage like the thermometers; instead, it is directly glued to the electronics mount with Stycast 2850-FT epoxy, which acts both as an electrical insulator and thermal conductor. It is a 220 Ω thin film resistor, with a typical operating power of 0.45 W.

The second heater, H2, is part of an absorption pump. Before describing what an absorption pump is, let us describe the problem that they are used to solve. When the system is below 4.2 K, any residual helium will begin to preferentially condense on cold surfaces in the vacuum can. This reduces the pressure in the vacuum can and consequently decreases the heat load transferred from the warm surfaces to cold surfaces by the exchange gas. A larger cold-surface area to vacuum volume results in a reduction in the base pressure and a smaller heat load from the exchange gas. An additional effect occurs if the cold surface is below the lambda transition. A superfluid film can form on the surfaces in the vacuum can, and using the substantial heat conduction of the superfluid helium film itself or through a helium "reflux", can thermally tie the cold surfaces to warmer surfaces. This can drastically increase the heat load and modify the thermal gradients in the vacuum can. The size of both of these effects is correlated with the thickness of the superfluid film. They can be reduced with an absorption pump, which is a device with a very large effective surface area and an embedded heater that is attached to a cold surface, preferably the coldest surface in the vacuum can. This large surface area on the coldest part of the vacuum can allows a large amount of the residual gas to be absorbed and it results in a thinner film throughout the vacuum can. This reduces the ability of the superfluid film to couple to warmer components.

To deactivate an absorption pump, the heater is continuously operated. It can be cleaned by powering the heater when the system is not actively being cooled and is sitting at 4.2 K. This evaporates the helium that is condensed on the surface of the absorption pump and allowing it to be pumped out of the vacuum can.

The absorption pump used in this experiment consists of an exposed, activated charcoal surface with an embedded heater. The structure of the absorption pump comes from a thin copper plate that is approximately 4 cm long, 2 cm wide, and 0.8 mm in thickness. First, the

copper plate is bent to match the curvature of the 1K Pot, where it will be mounted. Both the center 2 cm of the copper plate and the heater are coated in Stycast 2850-FT epoxy, and then the resistor is embedded in the epoxy on the plate. Activated charcoal, which provides the large effective surface area, is pressed into the epoxy and then the epoxy is allowed to cure. Vacuum grease is applied to the back of the copper plate to improve thermal contact and then it is mounted to the 1K Pot by fastening quick ties around the uncoated parts of the copper plate.

Two thin film resistors were included in the absorption pump, a 1 kΩ backup resistor and the primary resistor, which is 120 Ω. The absorption pump is not used during normal operation, but as a reference point, a typical power output to warm the 1K Pot or prevent helium from cryopumping on the absorption pump is 7 mW.

The flushing heater and absorption pump are powered with a Phillips / Fluke PM 2813 programmable power supply. It has three channels with maximum output parameters of 30 V, 10 A, and 60 W, which is more than sufficient to power the heaters for the purifier.

6.2.2 Sensors

The thermometry, level meters, and pressure gauges provide diagnostic information for understanding the state of the purifier. In the following sections, I will describe these systems. The mounting schemes and their motivations are also discussed.

Thermometry

The thermometry was the primary data for understanding the state of the purifier. As a result, it was important to carefully select the locations of each of the thermometers. Our choices followed those of Hendry[60]. The thermometers on the two ends of the flushing tube are the most important thermometers in the system. Therefore custom feedthroughs were designed to allow the thermometers to rest inside the LHe. In addition, duplicate thermometers were placed on each end of the flushing tube, which were mounted externally, as were all of the other thermometers in the purifier. The custom feedthroughs and the external mounts are described in Section 6.2.2, which starts on page 198.

The locations of the thermometers can be seen in Figure 6.1. The warm side of the flushing tube has the thermometers T1a (internal) and T1b (external). The cold side of the flushing tube has T3a (internal) and T3b (external). Thermometer T2 is located roughly half way along the flushing tube. Thermometer T5 measures the temperature of the helium entering the small tube fo the flushing tube after it passes through the heat exchanger. Finally, thermometer T4 measures the temperature of the helium entering the cold side of the flushing tube just under the flushing cone. Recall that between T5 and T4 the inlet helium passes through a 1/8 in

copper tube that was wrapped around and soft soldered to the flushing tube. Therefore the difference between T5 and T4 is an indication of how effectively heat is transferred between the small and large tubes of the flushing tube.

The resistance bridge that was used, a Picowatt AVS 147, has eight channels, which constrained the number of thermometers that could be used. The thermometers were uncalibrated and therefore needed to be calibrated before operating the purifier. This calibration was performed inside the purifier. In one calibration run, seven thermometers can be calibrated, leaving the last channel for a calibrated thermometer. Because seven thermometers are sufficient to operate the purifier, a single calibration run was performed. This set of thermometry was chosen because it gives the best diagnostic information about the operation of the purifier with the limited number of thermometers available.

The thermometers used were 10 kΩ RuO thermometers, which are typically used in cryogenic systems with large magnetic fields. Although the purifier does not have strong magnetic fields, these resistors were used because they were readily available and could be borrowed from the UCN Lifetime Experiment at NIST. As mentioned previously, these thermometers were purchased without a calibration curve. In the following sections, I will discuss the calibration process, the mounting scheme for the thermometers, the control software, and the resistance bridge.

Calibration

The calibration was performed inside the purifier using a custom Stycast feedthrough that allowed the calibrated thermometer and 7 uncalibrated thermometers to be suspended at the same approximate location in the helium bath. The feedthrough was placed at the cold end of the flushing tube just under the flushing cone. This location was chosen because it is the coldest part of the pure loop, which ensures that the calibration will extend below the expected temperatures of the thermometers while operating the purifier. During the production run, the base temperature will be slightly higher due to the additional heat load incurred by the mass flow that is required to extract the isotopically pure ^{4}He. The LHe environment provides ample thermal conduction to minimize any temperature differences between the thermometers during the calibration process. The calibration consisted of cooling the purifier into the 1 K to 2 K range and then taking 4-lead resistance measurements at a variety of temperatures in that range. The temperatures from the calibrated thermometer were transferred to the uncalibrated thermometers to obtain curves of the resistance as a function of the temperature. These curves were then fit to a power law to calibrate the individual thermometers. Secondary calibration data was taken at 4.2 K and 77 K by submerging the thermometers into a LHe transfer dewar and a small LN test dewar.

Table 6.1: The thermometry designations with the Picowatt AVS-47 channel for both the calibration and purification configurations. The % Dif column lists the percent difference of the resistance between each thermometer and the average of all the 10K RuO thermometers during the calibration run. This gives an indication of the variation between the individual resistors from the same batch and shows a variation consistent with expectation of less than a few percent.

Designation	Description	Purification Channel	Calibration Channel	% Dif
T1a	Flushing Tube; Warm End; Internal	AVS0	AVS3	0.38 ± 0.03
T1b	Flushing Tube; Warm End; External	AVS1	AVS1	1.30 ± 0.06
T2	Flushing Tube: Middle: External	AVS2	AVS2	-0.93 ± 0.04
T3a	Flushing Tube; Cold End; Internal	AVS4	AVS4	-0.41 ± 0.02
T3b	Flushing Tube; Cold End; External	AVS5	AVS5	-0.97 ± 0.03
T4	Helium Inlet to Flushing Tube; External	AVS6	AVS6	0.18 ± 0.05
T5	Helium Inlet Post Heat Exchanger; External	AVS7	AVS7	0.45 ± 0.07

The calibrated thermometer was a 1 kΩ RuO thermometer that was calibrated by David Kendellen in his thesis work[64]. In his thesis, the thermometer was designated T10 and was the thermometer at the bottom of the heating element assembly. It was calibrated by comparison to a factory calibrated Lakeshore GR-200A-30 Ge thermometer. The calibration was transferred to the 1 kΩ RuO thermometer and then fit to the functional form

$$R = a \exp(b/T^c),$$

where the fit coefficients were $a = 1111.8\ \Omega$, $b = 0.7597\ \mathrm{K^c}$, and $c = 0.7514$. The fit was performed in the range of 0.3 K to 5 K.

A Picowatt AVS-147 resistance bridge was used to measure the resistances. It can only measure one resistor at a time and requires a few seconds to stabilize before the resistance measurement becomes accurate as mentioned in Section 6.2.2, which starts on page 194. The calibration process requires that either the temperatures are stable during the measurement cycle or that temperature changes are accounted for by interpolating the calibrated thermometer temperature to the value at the time of the measurement of the resistance for the secondary thermometer. For our calibration, we decided only to include data where the temperature was stable in time to prevent any unnecessary uncertainty from the interpolation process. To achieve a stable temperature, the valve on the 1K Pot, V40, was throttled. When V40 was opened slightly, the temperature would begin to drop and would stabilize at a lower temperature over the course of 5 min to 10 min. Once the temperature was sufficiently stable, the valve was opened slightly, and the system began to approach the next equilibrium temperature. Calibration data was only taken at sequentially lower temperatures because it took much longer for the system to equilibrate after V40 was closed than when it was opened. The dedicated calibration data was taken between 11:15 AM and 9:15 PM on July 11th, 2014, see Figure 6.2. Specifically, points 1694 to 2176 from the resistance logs on that day were used.

To remove the events where the temperature is changing quickly, the derivative of the resistance arrays for each channel were individually required to be less than a threshold of 26 Ω/min. Figure 6.3 shows the portions of the calibration data selected by this cut. The threshold was chosen to qualitatively preserve as much of the equilibrium temperature data as possible without including any of the quick temperature changes directly proceeding a change in the valve position. This cut removes 44% of the data from AVS channel 1 during the mentioned time period. Selecting the threshold was a tradeoff between the number of points included in the fit and the inclusion of data points where the temperature was not stable. The parameters of the fit curve were insensitive to variations in the threshold value. The threshold of 26 Ω/min was taken as a conservative value. Figure 6.4 shows how the residual varies as the threshold is changed.

Figure 6.2: Resistance data for each of the 10 kΩ RuO thermometers from the calibration run.

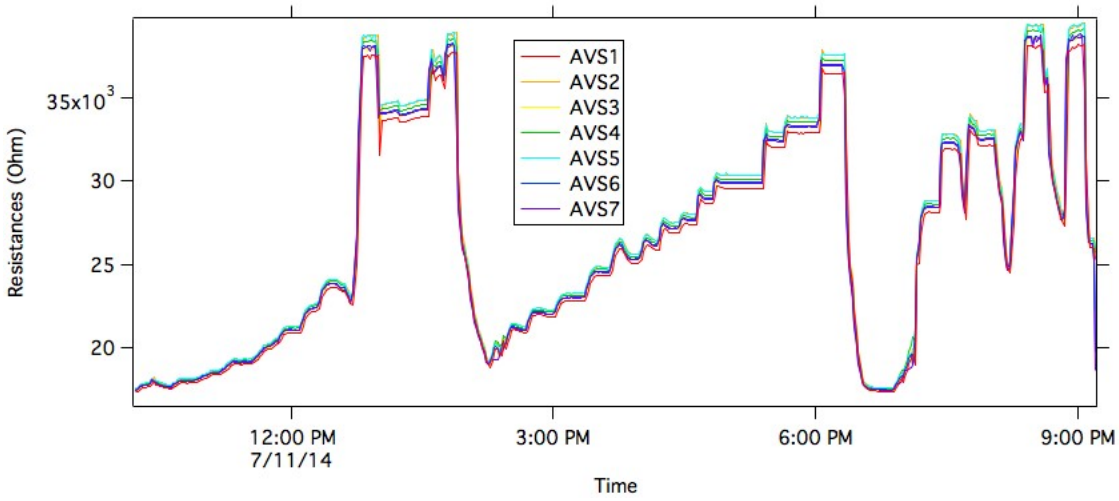
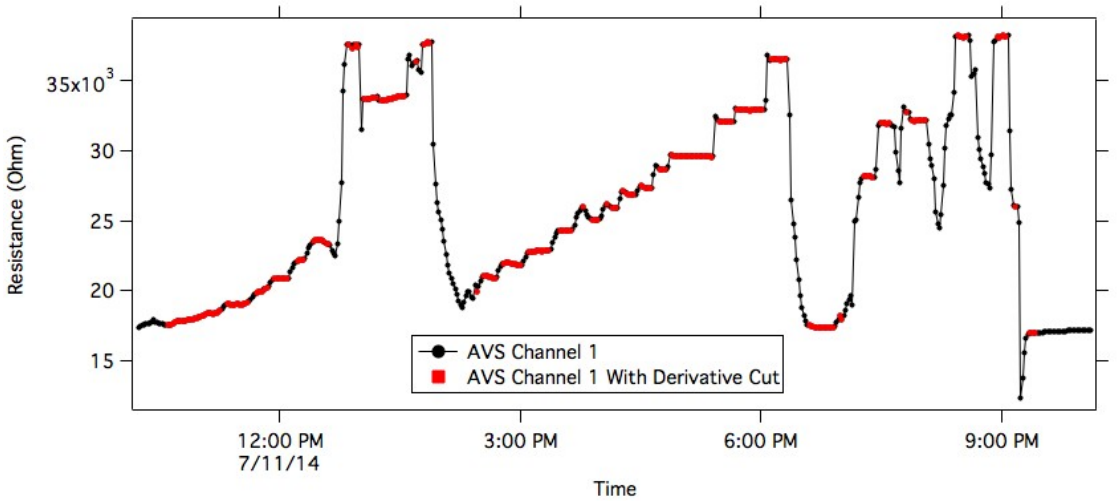
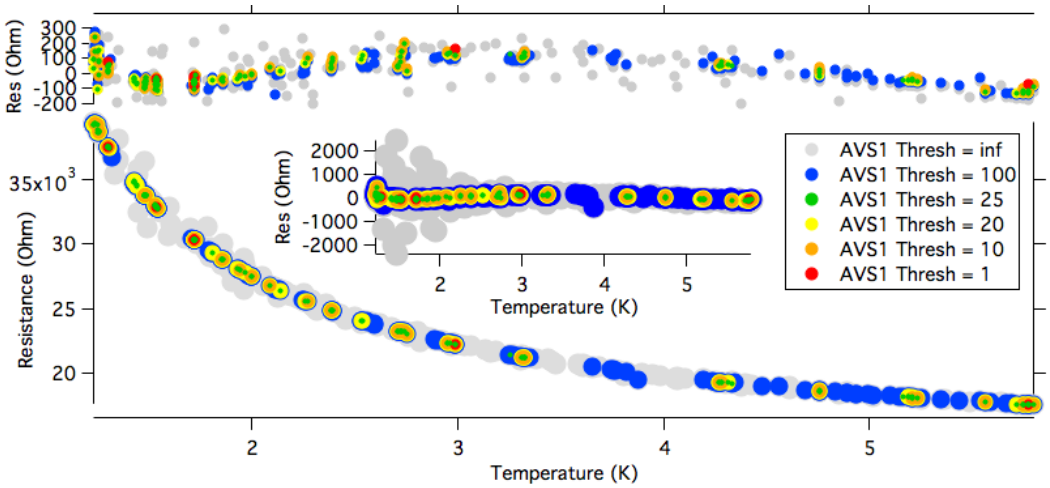


Figure 6.3: Resistance data for AVS1 with (red) and without (black) a derivative cut to remove data taken when the temperature was changing.



The calibration curve for the calibrated thermometer was given an offset in the resistance as a normalization to account for the resistance of the leads and solder joints in the portions of the circuit that only have two leads. The offset was chosen to put the calibration in agreement with the LHe boiling temperature, which was measured by submerging the thermometers into a LHe transfer dewar. The boiling point of LHe is expected to be 1458.51 Ω according to the calibration curve supplied by Kendellen; the resistance of the 1 K RuO thermometer was measured to be 1378.5 Ω while submerged in a LHe transfer dewar at atmospheric pressure. The difference, 80.022 Ω, is added to the resistance before applying the calibration curve.

Figure 6.4: Sensitivity study of the derivative cut threshold on the fits and residuals of thermometry calibration data.



After the calibration offset is applied, the temperature calibration is transferred from the calibrated thermometer to each of the other channels for every data point that passes the derivative cut. For each channel, the temperature is plotted as a function of the measured resistance, and the result is fit to a power law or the form $R = A \times T^{pow} + y0$, where the resistance is in Ω and the temperature is in K. The power law seems to capture the general shape of the data. However, the residual shows a definite trend about zero that suggests a small systematic effect. The magnitude of the systematic effect is less than a few percent, see Figure 6.5, which is sufficient for the needs of this experiment for an absolute calibration. The relative calibration is expected to be substantially better because the calibration transfer was performed while the thermometers were in good thermal contact in the LHe bath.

The curve fit parameters from the calibration are presented in Table 6.2. The channel that was used in the calibration measurement differed slightly from the channel during the operation of the purifier. Table 6.1 indicates the channel used for each thermometer separately for the calibration and operation and presents the disagreement in the temperature calibrations for the different thermometers. We expect any systematic effects in the temperature measurements from using different channels in the calibration and production to be negligible.

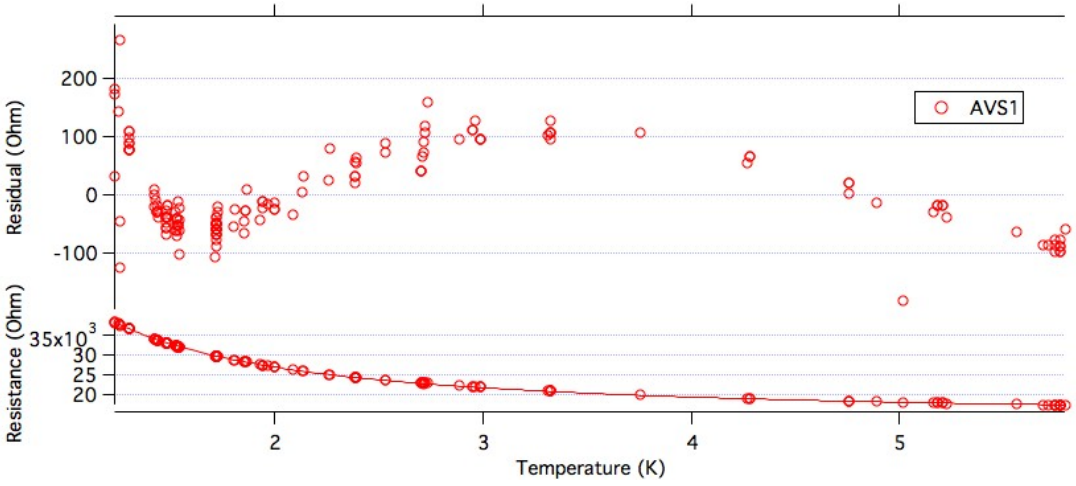
Resistance Bridge

The thermometry is measured using a Picowatt AVS-47 Resistance Bridge that was borrowed from the UCN Lifetime Experiment at NIST's Oxford dilution refrigerator. The AVS-47 is an eight channel bridge with a variety of resistance ranges. It is interfaced to LabView via an

Table 6.2: Calibration fit parameters for the thermometry. The channel of the AVS that each thermometer was attached to for the calibration run and both the calibration fit parameters and their uncertainties for each channel.

AVS Channel	y0	σ_{y0}	A	σ_A	pow	σ_{pow}
	Ω	Ω	Ω	Ω		
AVS 1	14331.5	41.3	31029.3	29.8	-1.30588	0.00522
AVS 2	14554.8	46.5	32522.9	34.2	-1.32792	0.00570
AVS 3	14468.1	48.1	31650.1	35.4	-1.32193	0.00603
AVS 4	14533.8	54.4	32106.2	40.1	-1.32719	0.00686
AVS 5	14550.9	42.8	32524.3	31.5	-1.32715	0.00532
AVS 6	14419.1	46.6	31743.2	33.3	-1.31421	0.00583
AVS 7	14354.4	60.1	31622.8	41.6	-1.30735	0.00747

Figure 6.5: A calibration curve for AVS1 with the final fit curve and residual.



AVS-47 Computer Interface, which allows the settings on the bridge to be computer controlled and allows temperature logging to be automated. The resistance of the thermometry varied between about 38 kΩ at 1.25 K and 19 kΩ at 4.2 K as can be determined from the calibration fit parameters in Table 6.2. The two closest readout scale ranges were 0 kΩ to 20 kΩ and 0 kΩ to 200 kΩ. To prevent the measurements from going off scale during typical operation of the purifier, the 0 kΩ to 200 kΩ resistance range was used. The resistance values lie at one end of the resistance range where the response of the AVS bridge is less linear. This results in a small systematic effect in our temperature calibration. It is expected to be a very small effect on the purifier operation because of the insensitivity of the purifier to the absolute temperature as long as it remains in the range of 1.2 K to 1.8 K.

As mentioned previously, the resistance bridge takes time to stabilize at the measurement value. This amount of time is strongly correlated with how far the previously measured resistance value was from the current resistance value. To account for this, the LabView code measures the resistance for an adjustable duration before it records the value. In the case where there is a large change in the resistance between two subsequent thermometer measurements, it is necessary to increase this delay time so that the bridge has sufficient time to stabilize. This increase in the delay time will then limit how quickly a measurement cycle can be completed and thereby the timing resolution of temperature measurements. If the delay time is not properly adjusted and is too short to allow the resistance value to stabilize, it will result in a large systematic effect in the resistance measurement. Therefore, any thermometry that is added to the system should be matched to the resistance of the thermometers that exist to minimize the delay needed between temperature measurements.

Electronic Mounts

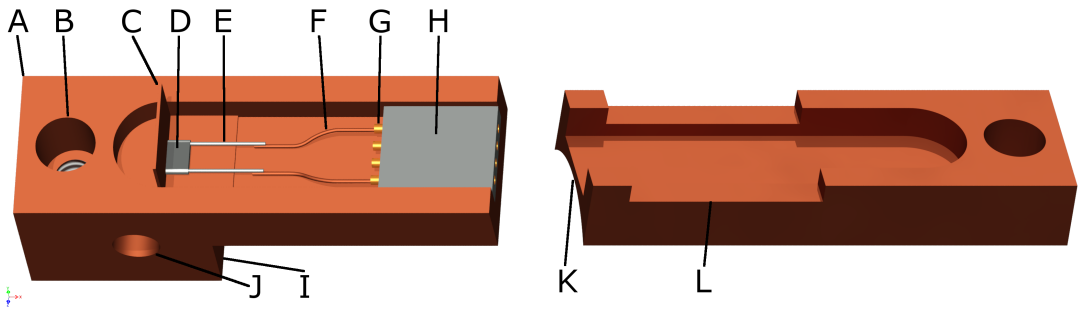
The thermometry requires a mounting scheme that allows the thermometers to be attached to a variety of tube sizes with strict spatial constraints. The thermometer should be as strongly coupled to the LHe as possible, securely attached to help limit vibrations, and finally, it must be protected from direct exposure to RF interference. To tackle this set of constraints, a mounting scheme was based on a typical design used at NC State[65]. A simplified view of two of the types of electronics mounts can be seen in Figure 6.6.

The design uses a copper bar with a groove milled out that extends almost the full length of the mount. This leaves a swath of removed material that has the shape of an arched doorway. The length of the channel is sufficient to store the thermometer, its Faraday cage, and a 4-pin MicroTech connector. The width of the channel in the electronics mount was chosen to snuggly fit the MicroTech connector. The mounts used a female connector, which is positioned intop of the copper flap is folded over the thermometer and crimped in place to securely wrap the thermometer in a copper shield.

If a small amount of epoxy is used, when the connector is removed, any remaining epoxy can be broken off with pliers without damaging the connector. If, however, the epoxy coats all the way around the corner of the connector, because too much epoxy was used, it is very easy to damage the connector when attempting to remove the epoxy.

The RF shielding is provided by a Faraday cage that is built into the electronics mount. It is constructed from an approximately 2.5 mm wide, 5 mm long, and 0.8 mm thick piece of soft copper sheeting. The copper sheeting is folded in half width wise, and then one side is soldered to the electronics mount at the top of the arch with the opening facing the MicroTech connector. A piece of Kapton tape is applied to the inside of the copper flap to provide electrical

Figure 6.6: Image of the computer aided design models of the electronics mounts. The lengths of the wires are not to scale. (Left) An electronics mount with the Faraday cage before it is crimped over the thermometer. This electronics mount is for 1/8 in copper tube. Key components of the electronics mount are labeled, "A" is the body of the electronics mount; "B" is a through hole and shows the bottom piece and its threaded hole for securing the electronics mount by clamping it to the tube; "C" is the copper Faraday cage; "D" is the RuO thermometer; "E" is the permanent leads; "F" is the secondary leads; "G" is the golden solder cups of the MicroTech connector; "H" is the MicroTech connector itself; "I" is the bottom piece of the electronics mount; and "J" is the hole that fits the 1/8 in copper tube. To simplify the figure, the Kapton tape and vacuum grease on the Faraday cage are not shown. Also, neither the solder joints nor the electrical insulation on the solder joints are shown. (Right) An electronics mount, for mounting to a 1.6 cm outer diameter tube, shown without its thermometer. "K" indicates the curvature of the mount, which matches that of the tube. "L" indicates a cutout in the top surface of the mount, which acts as a guide to keep the cable ties from slipping off of the mount.



isolation, and a thin dab of vacuum grease is applied to the Kapton tape to promote thermal conduction. The thermometer is placed in the vacuum grease on the bottom of the copper flap, and then the top of the copper flap is folded over the thermometer and crimped in place to securely wrap the thermometer in a copper shield.

To limit the number of times that the thermometers must be heated, they are soldered to a set of permanent leads. The permanent leads connect to a pair of secondary leads that attach to the MicroTech connector. Alternatively, the thermometers could have been directly soldered to the leads in the calibration feedthrough, but when the thermometers were moved from the calibration feedthrough to their location for the operation of the purifier, they would have needed to be re-soldered. Each time the thermometers are soldered their calibration could be invalidated, or the thermometer could be destroyed. This motivated the use of the permanent leads.

Both the permanent and secondary leads consist of a pair of 2 cm long, coated cryogenic wires. They were made from the same stock of Quad-Lead 32 AWG wire used throughout the purifier, which for this application was separated into individual wires. Because both the

permanent leads and secondary leads consist of two wires, the four lead measurement starts on the far side of the male MicroTech connector. This is expected to result in a small constant offset to the measured resistance values.

The design was modified to allow the mounts to be attached to tubes with three different outer diameters, 0.32 cm; 1.6 cm; and 5.1 cm. The 0.32 cm tubing is the aforementioned 1/8 in copper tube. For this tube, a semicircular channel was removed from the bottom of the mount such that the 1/8 in copper tube fits halfway into the channel. A second piece with a matching groove is used to clamp the mount to the tube. The clamping force was applied with a 6-32 bolt, a through hole in the electronics mount, and a corresponding tapped hole in the clamp. The electronics mounts for the 1.6 cm and 5.1 cm OD tubes were mounted by including curvature in the bottom surface of the mount that allowed it to sit flush against the corresponding tube. This second mounting scheme was secured in place using standard nylon cable ties. As mentioned previously, images of the CAD model of the electronics mounts can be found in Figure 6.6 for the 1/8 in copper tube (left) and 1.6 cm OD tube (right).

The cable ties were tested through many cryogenic cycles without failure. They are expected to contract more than the metal components at cryogenic temperature ensuring that the connection remains tight. They are easy to place; remove; and, if installed with a cable tie gun, can be attached with a repeatable tension. Consequently, cable ties have proven to be very convenient in cryogenic applications. The electronic mounts were attached with cable ties of tensile strength #10 - #50. Longer cable ties, which happened to be larger, were used on the larger structural tubes by the necessity of the larger circumference.

Custom Electrical Feedthroughs

To allow a more accurate measurement of the temperature of the helium, three custom feedthroughs were designed that allowed the thermometers to rest in the LHe inside the flushing tube. They were designed to fit snuggly in holes drilled through the large tube of the flushing tube. The first hole is located just below the 1K Pot, and the other is on the warm end of the flushing tube.

The feedthroughs consist of a brass tube with a copper alignment jig that constrains its orientation with respect to the flushing tube and provides the mating surface for the epoxy joint between the feedthrough and the flushing tube. Cryogenic wires were threaded through the brass tube and epoxied in place. The entire assembly was then epoxied to the flushing tube.

The brass tube has a 6 mm OD and 1 mm ID. The copper alignment jig has a square cross section of width 1 cm with a through hole for the 6 mm brass tube. The bottom surface of the jig was milled to match the radius of curvature of the flushing tube. Once these pieces were soft soldered together, they both constrain the rotational degrees of freedom and creates

the structural integrity of the feedthrough. The joint was leak checked using modified rubber stoppers after the alignment jig was soft soldered to the brass tube. Heating the feedthrough will ruin the epoxy joints. Therefore, the soft solder joint must be leak free before continuing.

Quad-lead wires were separated, and sleeves of Teflon shrink tubing were prepared to slide over both ends of the wire. The centers of the wires were coated in Stycast 2850-FT epoxy and then inserted into the brass tube. The wires were worked around to pull epoxy into the tube. Strain relief was provided by encasing the end of the tube in epoxy and using Teflon shrink tubing to provide support for the wires. To do this, a plastic cup was placed over the end of the tube and filled with epoxy. The Teflon tubes were threaded over the wires, from both ends, until one end of the Teflon tubes was embedded in the epoxy. The feedthroughs were allowed to cure in this state, and then the epoxy joint was leak checked.

The permanent leads of the thermometers were soldered to the feedthrough wires and cleaned with both alcohol and water. The exposed part of each wire and the solder joint was coated in Kapton tape, and then the wires were taped together. Additionally, Kapton tape was folded over the thermometers to prevent any exposed metal surfaces from shorting to the walls of the flushing tube.

In order to epoxy the feedthrough to the flushing tube, Stycast 2850-FT epoxy was applied to the flushing tube, and then the feedthrough was placed into the epoxy. The epoxy was manipulated with wooden applicators to ensure good contact between the epoxy and the copper and then the epoxy was allowed to cure. After the epoxy had cured, the system was leak checked with the leak detector on the flushing tube while spraying helium on the feedthrough. If the joint was not found to leak with this test, the vacuum can was put on; evacuated; and then back filled with helium and the system was leak checked again. Finally, the purifier was thermally cycled to 77 K, and then the system was leak checked at room temperature one final time.

There was a high failure rate in the production of the custom feedthroughs due to the feedthroughs leaking. The largest contribution to the failure appears to have been the result of using a tube with an inner diameter that was too large in the original design, which resulted in too much strain during thermal cycling. This appears to have been alleviated when the design switched from 3/8 in copper tube, with a relatively large inner diameter, to the custom 1 mm inner diameter brass tube. Once a feedthrough was shown to be leak tight, no leaks seemed to form unless the feedthrough was exposed to extreme heat.

These custom feedthroughs were essential in understanding the temperature in our system, which was obvious from the large discrepancy between the temperature readings of the thermometers that were mounted internally when compared to their external counterparts. This verifies that inferring the temperature of the helium from thermometers attached to external mounts introduces a large systematic effect.

When operating the purifier for the production of isotopically pure ^{4}He, the feedthroughs

only require two wires, however, for the calibration, it was necessary to have eight sensors in the same location. This was done with the same general design, but with sixteen wires. It was not necessary to increase the inner diameter of the brass tube to fit the additional wires.

Wiring

All of the wiring used in the vacuum can for both the thermometry and the heaters was Lakeshore product WGL-23-100; a Quad-Lead, 32 AWG wire. This is a lower gauge wire than is used in many low-temperature experiments like the UCN Lifetime Experiment at NIST. The larger wire was chosen to make the wiring less prone to breaking. It is also easier to handle and solder. Using the lower gauge wire limited both the frequency of breaks and the turnaround time for fixing them. The heat load due to the flushing heater is much larger than the additional heat load of conduction down the larger wire. Therefore, this is expected to have a negligible effect on the operation of the purifier or its base temperature.

Quad-lead wiring was used instead of the typical quad-twisted pair wiring. This means that all four wires ran parallel to each other for the entire run as opposed to twisted pair wiring, which, as the name suggests, twists the individual wire pairs around each other. The twisted pair wiring reduces the inductance of the circuit. This can reduce the noise or increase the response time of a circuit. As a result, it is a common practice in many high-precision experiments. Non-twisted wiring was used to allow a larger diameter wire than is available in twisted pair. With the limited sensitivity of the purifier to the temperature, it is not necessary to limit this additional uncertainty in the resistance measurements at the expense of smaller wiring.

Pressure Gauges

The primary task, outside of general diagnostics, that requires pressure gauges is calculating the dilution factor when cleaning portions of the purifier before storing a sample of isotopically pure ^{4}He. In the dilution process, the system is evacuated, refilled with isotopically pure ^{4}He, and then evacuated again. The ratio of the initial base pressure over the refill pressure is the dilution factor. The dilution factor is the fractional reduction of contamination in the system. This is under the assumption that the pumping speed is the same for both ^{3}He and ^{4}He and that they mix effectively during the refilling process.

The pressures at these two stages differ by a few orders of magnitude. The pressure when the system is evacuated is a few mTorr, whereas at the height of the refill the pressure is a large fraction of atmospheric pressure. Because of the large difference between these two pressures, two distinct types of pressure gauges were used in this experiment, both of which will be discussed below.

The first type of gauges was model 275 Convectron Pirani Vacuum Gauges from Granville-

Phillips. They extend down to a pressure of a few mTorr. They were calibrated for use with nitrogen, not helium gas, but a separate calibration curve for helium is available from the manufacturer, allowing the pressure measurements to be corrected. The pressure in a helium environment is approximately 150% of the reported value. These gauges were typically used to monitor an approximate vacuum pressure in different portions of the system, so a high degree of accuracy was not needed. The second type of pressure gauges was compound, mechanical gauges that the UCN Lifetime Experiment at NIST had on hand.

The compound gauges were used at higher pressures where the convectron gauges are known to overestimate the pressure. For example, the compound gauges would still read -30 inHg when the convectron went off scale, with a reading of 1000 Torr. From this, it is clear that there is substantial uncertainty in the pressure measurements. To limit the uncertainty, both types of gauges were included so that the appropriate gauge could be used depending on the current pressure in the system. To account for the uncertainty in the pressure measurements, the dilution was performed to a much higher level than should be required, to provide a safety margin.

Convectron gauges were included on the extraction line of the PGHS, on the UGHS, and on the SGHS. Each of these portions of the gas handling system also had a compound, mechanical gauge. The pressure gauge symbol on the gas handling system schematic for these three gauges represents the approximate location of both the convectron and compound gauges. There was also a compound gauge on the 1K Pot, which was included primarily as a general diagnostic. The sensitivity of the gauge was not high enough to provide an alternative method for estimating the temperature of the 1K Pot by measuring the vapor pressure. The vacuum can had a convectron gauge without a compound gauge. Finally, both of the pressure gauges on the DGHS were compound, mechanical gauges.

6.2.3 Other

The following components are not instrumental in calculating the properties of the heat flush. Instead, they are included because they provide some context for the performance calculations, e.g. the internal dimensions of the dewar are useful for calculating the helium consumption, or because I thought a description of the system might be useful.

LHe Level Meters

Two resistive level meters were used in the purifier, one in the main helium bath and the other in the 1K Pot. They were read with AMI 135 level monitors. A third, capacitive level meter was designed and fabricated at NC State and was also located in the 1K Pot. The capacitive level meter was a backup in case the resistive level meter could not handle the large helium vapor

flow in the 1K Pot during typical operation. The resistive level meters were quite robust and therefore the capacitive level meter, which has a smaller response, was not particularly useful.

The capacitive level meter consisted of a stainless steel rod with an outer diameter of 3 mm and a stainless steel outer sleeve with an inner diameter of 5.8 mm. The capacitance gauge was designed to match the capacitance of the gauge used in the UCN Lifetime Experiment at NIST for a total capacitance of $25 \times 10^{-12} F$ when empty, which required a length of 0.137 m. The threaded rod had tapped holes on each end to allow a Delrin jig to be attached that positioned the outer tube around the threaded rod. Holes were drilled in both the outer tube and the Delrin jigs to allow helium to fill and drain from the region between the rod and sleeve. One lead was soldered to the sleeve, and the other was clamped down by the screw used to attached one of the Delrin jigs. The dielectric constant of LHe is about 1.05, therefore, as the level of LHe rises from the bottom of the level meter to its top, the capacitance is expected to change by about 5%. This is neglecting the dielectric constant of gaseous helium which is a factor of 50 smaller at this temperature.

The 1K Pot level meter was 0.27 m long, and the bottom of the sensor was located approximately 0.07 m above the bottom of the cone. During operation, the 1K Pot was adjusted to maintain a level of 50%. When the level was below the bottom of the level meter, it was not possible to tell how close the 1K Pot was to being empty, which could result in instability in the temperature. When the level rose in the 1K Pot, the base temperature of the purifier increased. Operating the 1K Pot when it was half way full gave time to adjust for fluctuations in the level while keeping the temperature of the purifier in an acceptable range.

The top of the level meter in the main helium bath was attached to the bottom radiation baffle. The height of the radiation baffle was chosen so that the bottom of the level meter was very close to the bottom of the vacuum can. The meter that was used was 0.75 m long. The higher of the purifier fill lines in the main helium bath is about 0.25 m above the bottom of the vacuum can. In good agreement with this value, the behavior of the purifier was observed to change when the level read $\approx 28\%$ indicating that the first of the fill line was no longer submerged in LHe.

Gas Handling System

The gas handling system contains all of the pumps, gauges, valves, and tubing necessary to both prepare the purifier for running and to operate it. See Figure 6.7 for a detailed schematic of the entire gas handling system. The gas handling system has been broken up into separate sections geographically to make the procedures for operating the purifier more clear. Each section will be discussed in the following paragraphs.

Before describing the system, there are a few general design concepts that were incorporated

throughout the gas handling system. Due to the high sensitivity to ^{3}He contamination, exposure of the system to the atmosphere could contaminate the apparatus. In particular, ^{3}He diffusion through non-metallic gasket materials and small leaks in the gas handling system are potential sources of contamination. To combat the diffusion, the system was designed to expose the isotopically pure ^{4}He product to only all-metal components. This included the use of valves with all-metal seals, all-metal pressure gauges, and both a diaphragm pump and compressor that were designed with this in mind. To address small leaks in the gas handling system, it is leak checked first with H_2 gas, and once it is found to be below the detection limit with hydrogen, it is leak checked with helium. The use of the H_2 allows large leaks to be discovered without exposing the system to unnecessarily large quantities of helium. Following up with helium leak checking allows a higher sensitivity to be obtained. This methodology constrains the leak rate in the system, but small leaks that are below the detection limit can still slowly contaminate the system over time, in particular, if the system is left under partial vacuum for long periods of time. To prevent this type of long-term contamination, the gas handling system is backfilled with isotopically pure ^{4}He from the purifier and then evacuated a few times at the start of production. This dilutes and removes a large fraction of any long-term contamination. The rest of the production run is performed within a period of a few days to limit the amount of contamination that can make it into the system from slow sources.

As was mentioned previously, gas in the isotopically pure ^{4}He samples was only exposed to all metal surfaces. The components it is exposed to use solder, weld, or VCR connections of 3/8 in stainless steel tubing. These tubes are too small to be effective as pump lines. Therefore on pump lines like V12, the valve is an all metal seal valve and just after the valve, the system transitions to KF pumpout lines to increase the effective pumping speed of the system. The KF standard uses non-metallic gaskets. Therefore, the gas for the isotopically pure ^{4}He samples can not be exposed to these portions of the gas handling system. There are four such VCR to KF transitions in the gas handling system. They are V3, V12, V51, and a set of pumpout ports on the dumps that are not shown in the gas handling system schematic. In each of these cases, the clean portions of the gas handling system were only exposed to non-metal gaskets when the system was actively being pumped. on and the KF components were part of the pumping line. Therefore, the amount of ^{3}He diffusion into the system is expected to be negligible. The exception to this is the clean part of the PGHS. This section of the gas handling system was exposed to contaminated helium at least daily. If the purifier was cold; the flushing gradient was not operating; and the system was not actively being watched, the extraction line was opened to the main vacuum line. This connected it to the overpressure valve as a safety precaution to prevent trapped gas and an explosion hazard in the event of a catastrophic failure of the vacuum jacket of the dewar. Therefore, the clean part of the PGHS, as opposed to the rest of the clean gas handling system, has to be cleaned every time the flushing gradient is started. In

summary, the gas handling system has been carefully designed to maximize the pumpout speed and safety of the system while limiting the possibility of ^{3}He contamination.

The first two parts of the gas handling system are borrowed from the UCN Lifetime Experiment at NIST and take advantage of its infrastructure for manipulating and storing isotopically pure ^{4}He in a system that was designed to limit ^{3}He contamination. In this work, this system is broken up into two different sections geographically to make the procedures for operating the purifier more clear. The first section, the UGHS, contains low-pressure storage volumes, multiple vacuum ports, and a diaphragm pump for manipulating the isotopically pure ^{4}He product. It connects to the second part, the DGHS, which has a compressor and a bank of high-pressure gas cylinders for long term storage of the isotopically pure ^{4}He. In addition to these, there are two other components to the gas handling system. The third part is the SGHS, which is a small addition to an auxiliary connection on the UGHS that allows samples of isotopically pure ^{4}He to be taken from the system. The final portion is the PGHS, which also connects to the UGHS. It is a very flexible system that allows the isolation, evacuation, and leak checking of the purifier through a large number of possible configurations. These systems will be described in more detail below.

The UGHS is the central portion of the gas handling system. It connects to the extraction line of the PGHS at V10, the SGHS at V50, and the DGHS at V20. It can be evacuated through V12. The pressure is monitored with a compound mechanical pressure gauge and a convectron. The diaphragm pump, which is described in more detail in its own subsection, pumps the isotopically pure ^{4}He out of V10 and into temporary storage in the dumps, which are a set of 4x 800 L, stainless steel, low-pressure storage volumes. The direction of the action of the diaphragm pump can be switched using a set of tubes and valves that are not shown in the schematic. It can be bypassed entirely via V11. The dumps are equipped with a set of pumpout ports with custom, all-metal seal, KF-40 valves, which are not shown in the schematic and allow the dumps to pumped out over time scales of approximately 8 h instead of multiple days through the 1/4 in stainless steel tubing. They are used during the initial pump out of the system as well as to assist in pumping out the dumps when they are flushed with the isotopically pure ^{4}He product before producing samples.

The DGHS gas handling system contains a custom built compressor with all metal diaphragms for compressing the isotopically pure ^{4}He from the dumps into a bank of eight high-pressure gas cylinders. V23 can be used to bypass the compressor. V21 is connected in series with a one-way regulator allowing isotopically pure ^{4}He back into the storage bank if the compressor is not on or if it is unable to keep up with the pressure buildup in the dumps. The DGHS has two compound pressure gauges for measuring the pressure on both sides of the compressor. V22 and V24 can be used to isolate the compressor from the rest of the system.

The SGHS uses a VCR flex line to connect from V50 to V52, which is directly welded to the

sample bottle. The flex line allows the sample bottle to be submerged in LN when the sample is being taken to increase the amount of gas that can be stored in the sample bottle. A VCR tee before the flex line allows the flex line to be evacuated after switching out sample bottles through V51 and a roughing pump. A convectron gauge is included on SGHS to constrain the pressure in the sample bottle for the process of flushing out contamination before filling the sample bottle.

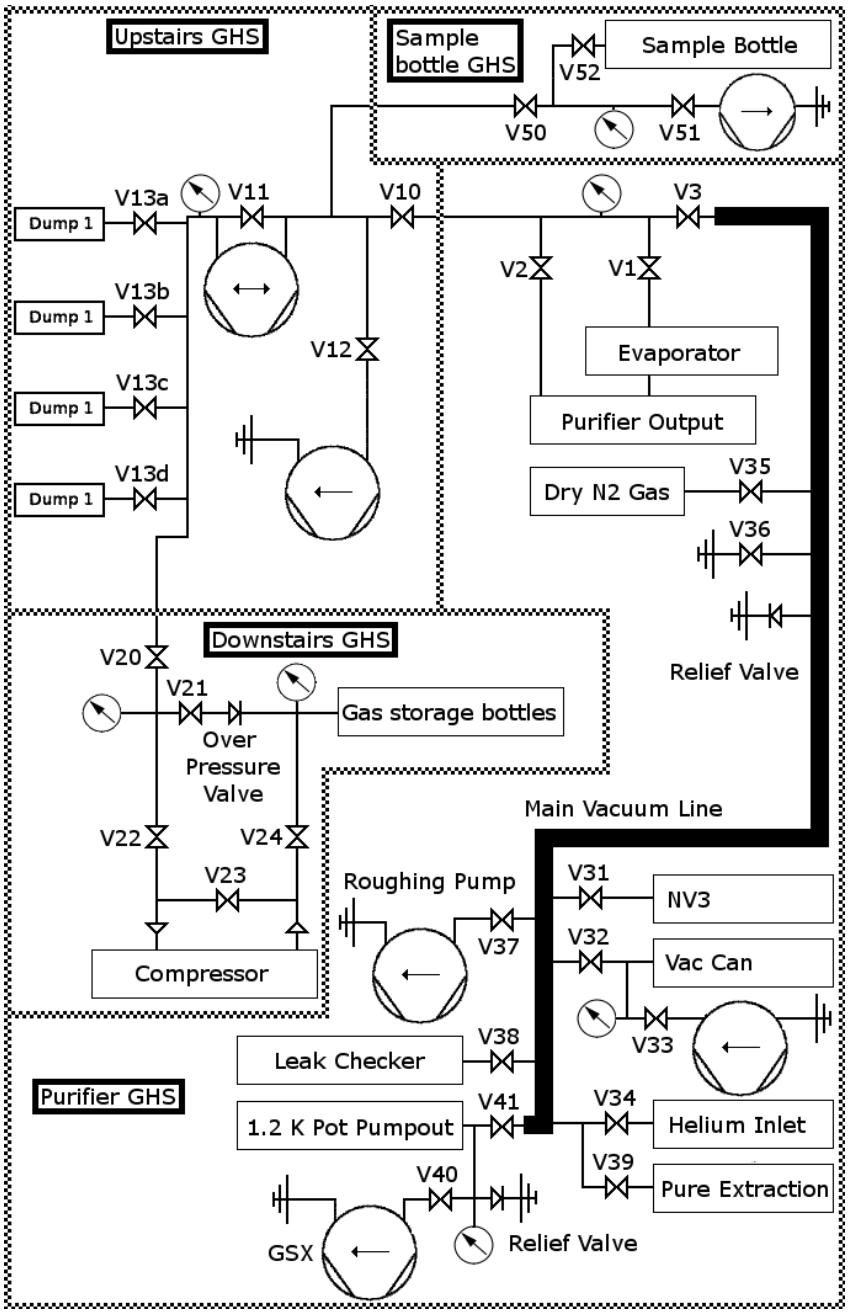
Finally, the PGHS is the addition to the gas handling system that provides the functionality needed for preparing, cooling, and operating the purifier. It contains the isotopically pure ^{4}He extraction line (purifier output), which then proceeds through the evaporator and into the UGHS through V10. V2 allows the evaporator to be bypassed. Both a compound gauge and convectron monitor the pressure of the extraction line. V3 connects the extraction line to the main vacuum line, which connects to a variety of pumpout ports all over the purifier including V31 to NV3, V32 to the vacuum can pumpout line, V34 to the helium inlet pumpout line, and V41 to the 1K Pot pumpout line. The main vacuum line is evacuated with the roughing pump via V37 or the leak detector via V38 depending on the current pressure in the system. The ability to connect all of the pumpout ports allows the system to be pumped down more quickly and results in a faster response in the leak detector when preparing the purifier for operation. Two, one-way relief valves are included, one on the 1K Pot top plate and the other on the main vacuum line. V35 can be used to backfill the main vacuum line and any attached components to dry nitrogen or helium gas. This port is equipped with a flow meter, which allowed a positive pressure of helium gas to fill the purifier while the system is being cooled with LN to prevent nitrogen from making it into the purifier and freezing during the subsequent cooldown to LHe temperatures. V36 can be used to expose the main vacuum line to atmosphere.

The 1K Pot and the vacuum can both have separate pumpout ports, V40 and V30 respectively, and pressure gauges. The pump that evaporatively cools the 1K Pot is a GSX 160/1750 dry pumping system and is described in a subsection that follows. During typical operation, the vacuum can is pumped on with the leak detector, which allows the amount of helium coming out of the vacuum can to be measured. Once the purifier is cold, the base pressure in the vacuum can is typically better if the vacuum can is allowed to cryopump by closing both V32 and V33.

Evaporator

The evaporator was designed to warm the isotopically pure ^{4}He that is extracted from the purifier before it enters the UGHS. The evaporator consisted of 1 cm OD, soft annealed copper refrigeration tubing of length 15 m. It was coiled into a helix and soldered in place. This helical shape minimized the footprint of the evaporator allowing it to be submerged in a bucket of water to act as a thermal reservoir. The evaporator was not needed in practice because of the

Figure 6.7: The schematic of the gas handling system.



small extraction rate. During typical operation, the extraction line barely cools below room temperature. As a result, the bucket of water was not used.

Needle Valve Design

The needle valves were a slight modification of the design by the McClintock group. The primary modifications include moving the threads of the needle valve to the needle instead of the top plate of the purifier and including a pump out port at the top of the needle valve assembly to allow the needle valve tube to be evacuated more quickly before operation. All of the dimensions were modified slightly to bring them from metric to standard units. During commissioning of the apparatus, the portion of the needle valve assemblies located in the vacuum can were modified further by moving the solder joints further from the threads to reduce the likelihood of solder getting into threads when the joints were being soldered.

Electrical Connectors

Microtech GM-4 and GF-4 electrical connectors were used in the vacuum can. They are small, 4-pin connectors with a resin body and solder cups and pins that are both gold. The body of the connectors are 0.007 m long, 0.006 m wide, and 0.002 m thick. These connectors are used in cryogenic experiments at both NIST and NC State. They are easy to use, handle cryogenic cycling well, are quite small, and are robust.

Dewar

The dewar that was used for the purifier consisted of a metallic cup of inner diameter 0.254 m and length of 1.12 m. The metallic cup was supported from the top flange of the dewar by a G10 cylinder of length 0.46 m resulting in a total internal height of 1.52 m. The top flange outer diameter was 0.32 m. The center of the o-ring groove had a diameter of 0.27 m with a width of 0.006 m. The hole pattern consisted of 12 x 10-32 holes on a diameter of 0.2985 m. The dewar had a vacuum jacket with multiple layers of superinsulation to provide the insulation and did not have a LN volume.

Vacuum Pumps and the Leak Detector

The McClintock purifier used a Roots Blower to pump on the 1K Pot to provide cooling for the purifier. In our system, this was replaced with an Edwards GSX 160/1750 Dry Pumping System. This pump consisted of a mechanical booster backing pump as well as an industrial dry screw pump. This pumping system has substantial pumping speed at high pressures and being a dry pump it limits the possibility of contaminating the purifier. It is a very high-tech pump including features like a web server to monitor the status of the sensors inside the pump as well as to display running information. A digital interface allows substantial modification of the operation including independently modifying the maximum pumping speed of each of the

pumps, which was used extensively while running the purifier as is described in the purifier operating procedures that are included in the appendix.

The diaphragm pump was a model PU 462-N035.0.4.91 rotary pump that was produced by KNF Neuberager Inc. It was borrowed from the UCN Lifetime Experiment at NIST where it was used to back the compressor to help compress helium into the high-pressure bottles. In the purifier, this pump was used to help pump helium out of the extraction tube and into the dumps. It is designed to be used on pressures less than or equal to atmospheric pressure, and it was able to maintain a backing pressure on the extraction line of -12 inHg, which was enough to maintain flow through the purifier even as the pressure in the dumps approached atmospheric pressure. Even with the diaphragm pump, it was not possible to achieve atmospheric pressure in the dumps.

The leak detector that was used was a Leybold model Inficon UL 200. Two oil-free roughing pumps were also used, an Edwards XDS10 and an Edwards XDS35i. The XDS35i was primarily used to evacuate the main vacuum line via V37. The XDS10 and the leak detector were both moved between various pump out ports as needed.

LHe Fill Line for the Pure Loop

The fill lines were constructed from 3/8 in stainless steel tubing and were soldered together with copper water tube fittings. They connected to the needle valve assemblies with tees just above the vacuum can top plate. At the edge of the vacuum can top plate, an elbow redirects the tube downward, which ends with an off the shelf, sintered copper plug that acts as a filter. The sintered copper plug is designed to be a sound muffler and is part number 4450K3 from McMaster Carr. The fill lines extend about half way to the bottom of the of the vacuum can.

The Main Helium Bath and the Purifier Top Plate

The main helium bath has two fill lines and a vent. The vent is connected to a KF tube that directs the exhaust from the cryogen fills away from the purifier top plate. Heating tape is wrapped around all of the other tubes on the purifier top plate to prevent them from freezing during the cryogen fills. Before the addition of the heater tape, there were a few issues with the feedthroughs freezing during the cryogen fill. During one of the cryogen fills where the electrical feedthroughs froze, a wire on the inside of the feedthrough broke, which motivated the addition of the heater tape.

Four 1.5 mm stainless steel radiation baffles were used to limit the radiative heat load. They were suspended from 4x 10-32 stainless steel threaded rods and were roughly equally spaced so that the bottom baffle was 0.75 m from the bottom of the vacuum can. The level meter was connected to the bottom radiation baffle, and this height was chosen so that the bottom of the

level meter was at the bottom of the vacuum can. A brass cone was included on both of the fill ports that caught the stinger for the LHe fills and provided a sealing surface. A 3/8 in stainless steel tube that is soldered to the bottom of the brass cone extended the fill line to the bottom of the helium volume.

Off-the-Shelf Electrical Feedthroughs

Military spec feedthroughs were used to transmit the wires from atmosphere into the vacuum can and the 1K Pot. The feedthroughs performed as expected with little problems once heater tape was used to prevent them from freezing during helium transfers. The 1K Pot feedthrough required 6 wires for the capacitance and resistive level meters. A 10 pin connector was purchased from Detoronics with part number DT02H-12-10PN and the corresponding connector MS3116F-12-10S. The vacuum can has two, 19 pin feedthroughs with part number DT02H-14-19PN and MS3116F-14-19S. One feedthrough has four, 4 wire connections for the RuO thermometers and one, 2 wire connection for one of the heaters. The other feedthrough has three, 4 wire connections for the RuO thermometers and one, 2 wire connection for the other heater.

6.3 Performance

The purifier was successfully operated at NIST from 1/16/2016 to 1/20/2016. During this time a batch of isotopically pure ^{4}He was created and stored in the dumps. Sufficient isotopically pure ^{4}He was produced to make multiple samples for measurement at ANL. In this section, we present metrics that can be used to quantify the performance of the purifier. This, along with the design information given previously, is hoped to be useful for constraining the design of future apparatus and for estimating reasonably attainable performance.

6.3.1 Isotopically Pure ^{4}He Production Rate

The isotopically pure ^{4}He extraction rate is an important parameter for determining the cost of producing isotopically pure ^{4}He. A faster extraction rate allows more freedom in the operation by limiting the amount of helium and time consumed on necessary, but ultimately secondary tasks like diluting any ^{3}He contamination in the system before collecting samples. Diluting contamination from the system is done using isotopically pure ^{4}He extracted from the purifier to backfill a given volume to pressure near atmospheric pressure and then evacuating it again. The extraction rate determines how long it takes to fill the volume being purged to any given pressure. Therefore a higher extraction rate can allow a larger dilution factor for each purge by filling the system to a higher pressure in the same amount of time. This allows the system to

be cleaned to a higher level resulting in a larger safety margin with regards to the amount of residual ^{3}He contamination in the system.

The extraction rate of the purifier has been estimated by measuring the rate at which the pressure increases while extracting isotopically pure ^{4}He into a single 800 L dump. The extraction rate was estimated to be about 290 L/h of gas at atmospheric pressure. At this rate, it would take approximately 76 h to produce the 21 L of liquid required to operate the experiment. These estimates are very rough due to the lack of sensitivity of the pressure gauge on the UGHS. A gauge with higher sensitivity would allow a more accurate measurement to be made.

The purifier was expected to have a larger extraction rate. This is probably due to a flow impedance in the extraction line. A likely explanation is the presence of a small amount of solder that is partially blocking the passage of one of the solder joints between a 1/8 in copper tube and either the fittings on the heat exchanger, flushing tube, NV3, or the vacuum can top plate. In theory, this would not be particularly challenging to fix. However, it would involve soldering in the vacuum can.

Soldering in the vacuum can is risky for a couple of reasons. It is very common for new leaks to form while attempting to fix a joint in the vacuum can. This is primarily due to the limited space in the vacuum can resulting in many tubes and solder joints being in close proximity to the joint being reflowed. Soldering in the vacuum can could also result in damage to the wiring, thermometry, and the glue joints on the custom flushing tube electronics feedthroughs. The wiring and thermometry can be removed to prevent damage, however cryogenic thermometry and wiring is quite fragile; removing and rewiring this amount of thermometry will likely result in some damage to the wiring. The glue joints on the flushing heaters can not be removed, which puts the two calibrated thermometers inside the flushing tube at risk when anything is soldered in the vacuum can. Obviously, the amount of risk is strongly related to the proximity between the thermometer and the solder joint being reflowed. This is particularly important because these thermometers have been calibrated. If they are destroyed or damaged to the extent that the calibration curve is no longer accurate, a second calibration run will be required, which is both expensive and time-consuming. Ultimately, it was determined that the extraction rate was high enough to produce isotopically pure ^{4}He both to verify the purity of the extracted helium and to provide a new batch for the UCN Lifetime Experiment at NIST.

6.3.2 Helium Consumption

Helium consumption is another parameter that is directly related to the cost of operating the purifier. An ideal system could be designed to use an entire transfer dewar to minimize the losses during the LHe transfers; allow a sufficiently large bath of helium that the purifier could

be run continually, at full power, for a long shift without requiring additional fills; and would leave enough helium in the dewar to keep it cold until the next time the system was to be run. For this purifier, these design constraints were not taken into account during the design, but despite this proved reasonable.

Helium consumption can be broken into two components that affect the operation procedures and cost of operating the purifier differently. The first is how quickly the level in the main helium bath decreases. This determines how long the purifier can be run continuously without a transfer, and therefore affects the fraction of the time that isotopically pure ^{4}He is produced. The second factor is the amount of helium that is required to cool the apparatus to 4.2 K and to fill the apparatus with LHe.

The level in the main helium bath decreases by about 4 %/h while operating the purifier and 1 %/h if the 1K Pot is not being pumped on. As a reference, the purifier can be operated while the level in the main helium bath is in the range 30% to 100%. Below about 30% the level drops below one of the helium fill lines and reliable purification is no longer possible.

During a typical initial helium transfer, which involves cooling the system from 77 K to 4.2 K and filling the main helium bath, about 60 L of liquid is consumed. This leaves sufficient helium in the transfer dewar for a second fill at the end of the following day. Hence, this is enough helium for approximately two days of operation. It takes approximately 3 h to get the purifier operating in a stable state, to start extracting isotopically pure ^{4}He, and to clean the purifier. Each of these tasks requires pumping on the 1K Pot and therefore results in the higher helium consumption rate. Assuming that the purifier is producing isotopically pure ^{4}He for about 8 h a day, it will take approximately nine days and about 450 L of LHe to produce enough isotopically pure ^{4}He for the operation of the UCN Lifetime Experiment at NIST. Realistically, a sixth dewar will be required to account for difficulties during purification and any delays during operation. This is operating under the assumption that the system is allowed to warm back up to temperatures substantially higher than 4.2 K every other day. Warming the purifier between fills will help prevent any issues from superfluid films forming in the vacuum can. It is unclear if this is necessary due to our inability to determine if there is superfluid film flow in the vacuum can, which will be discussed in the following section.

6.3.3 Periodic Oscillations in the Load on the Pump.

A periodic oscillation in the temperatures of the pure loop was discovered when V41 is completely open, and the GSX is allowed to pump at its highest speed. The system cools down normally to about 1.4 K, and then an almost sinusoidal oscillation in the temperatures occurs with an amplitude of ≈ 0.2 K and a period of ≈ 10 min.

Along with the oscillation in temperatures, there was a corresponding oscillation in the load

on the pump, which could be observed auditorily or by monitoring the pumping speed. This led to the conclusion that the effect could be due to the GSX overheating and a protection circuit limiting the maximum speed until the temperature recovers. The manufacturer did not support this theory, and by monitoring the temperatures inside the pump with its built-in thermometers, it was ruled out.

Another alternative is that there is an oscillatory load on the 1K Pot in this running configuration. The primary bogeyman in helium experiments below the λ transition is superfluid films. I am aware of no experiments with superfluid films that discuss periodic phenomenon. However, presumably an effect that is this large would have to be due to a helium "reflux"[66]. A helium "reflux" is an extremely efficient heat conduction mechanism in superfluid films where the superfluid film transports a steady flow of liquid to a heat source where it evaporates and preferentially recondenses on the coldest surfaces. This forms a continuous heat exchange process, that is similar to convection but has a much larger effective thermal conductivity because each atom that completes the cycle transmits its entire heat of vaporization in addition to some heat capacity between the cold bath and the heat source. Thermal conduction through the helium "reflux" is typically limited by the mass transport through the superfluid film creep, which increases with pressure and decreases with temperature.

Superfluid flow does not start in a superfluid film until a critical thickness of the film is reached[67], which varies with both the pressure and the temperature. Therefore, one requirement for the helium "reflux" is that there is enough helium in the vacuum can to coat the surfaces with at least this thickness of film. A precise value is not required here. However, to give a feeling, at 1.3 K macroscopic heat transfer beings when the film reaches a thickness of about two superfluid layers. At 2.0 K about ten superfluid layers are required before heat transport begins. As more layers are added to the superfluid film, the heat conduction of the helium "reflux" increases. There are two prominent candidates for mechanisms that could have allowed a sufficient amount of helium into the vacuum can to reach the critical thickness, not properly evacuating the exchange gas from the system or either a superfluid or normal fluid leak somewhere in the pure loop of the vacuum can.

If the helium in the vacuum can came from a superfluid leak, then it should be possible to remove the majority of the helium in the vacuum can by pumping on the vacuum can with the heaters energized while the purifier is above the λ transition. After pumping on the vacuum can and heating the pure loop with absorption pump overnight, the periodic effect was still present. This either suggests that the superfluid leak is sufficiently large to immediately re-saturate the vacuum can or the film, if it exists, probably results from either a normal leak or the effective pumping speed of the system being too low to properly evacuate the vacuum can. Either way, it is conceivable that enough helium exists in the vacuum can to allow the helium "reflux".

The helium "reflux" could conceivably create a periodic effect if, as the temperature de-

creased, which results in an increase in the effective thermal conductivity of the helium "reflux", the GSX was unable to keep pace with the increased heat load. This would result in an increase in the pressure, a corresponding increase in the temperature, and finally a decrease in the effective thermal conductivity of the helium "reflux". The decrease in the effective thermal conductivity could allow the GSX to catch back up restarting the periodic loop.

The measurements of the pressure in the vacuum can are not of sufficient quality to establish if the pressure was high enough at our operating temperature to allow superfluid creep, much less to adequately constrain the effective thermal conductivity of the hypothetical helium "reflux". Our calculations are inconclusive being both consistent with no helium "reflux" and with the majority of the flushing heater power being carried away by the "reflux" instead of by roton production. This makes improvements in the pressure measurement of the vacuum can an important upgrade to the purifier, which is one of the suggested improvements in the following section. Without further measurements, it is not possible to say with confidence if the helium "reflux" is truly the cause of this periodic phenomenon. However, it does seem plausible.

The periodic effect can be circumvented by limiting the maximum pumping speed of the mechanical booster (MB) on the GSX. This method is effective and results in a sufficiently low base temperature for the operation of the purifier. It allowed our effort to be put into the operation of the purifier instead of an investigation to understand this periodic effect.

6.4 Future Work, Issues, and Suggested Improvements

During the process of running the experiment, possible improvements were discovered. I have compiled the following list of suggested upgrades and issues with the purifier to assist in both future efforts to operate this purifier and the design of future purifiers.

1. Breaking up the flushing tube into a series of vertical sections results in gravitational instabilities and convection in half of the vertical sections. This is because LHe has a maximum in its density at the λ transition. Therefore a vertical tube of superfluid LHe will be unstable if warmer liquid is above colder liquid; in the case of a heat flush purifier, if the heater is above the cold bath. In our case where the flushing tube is broken up into a hand full of vertical sections, some of the vertical sections will have gravitational instabilities, and therefore, convective mixing is expected. This was also true for the purifier by the McClintock group to the best of my knowledge. Because each vertical section is longer than what is required to obtain the desired purity, this is not expected to prevent the purifier from functioning. That said, it may be a good idea to redesign the flushing tube to take into account the gravitational instabilities.

   It is hard to imagine an inexpensive method for producing a flushing tube that is a meter

long, fits inside a reasonably sized vacuum can and is not susceptible to gravitational instabilities. A vertical helix with the heater at the bottom of the helix and the cooling bath at the top would have no gravitational instabilities. I do not know if it is possible to have a copper tube bent into a helix that would fit in the vacuum can. The system would also have to be completely redesigned to allow the 1K Pot to connect to the flushing tube at the top of the vacuum can. In conclusion, I do not see a clear method for solving this problem. However, it is a part of the design that has room for improvement.

2. It would be beneficial to redesign the flushing tube to consist of a pair of concentric tubes. The small tube of the flushing tube is meant to create a radial component to the flush by warming the wall of the larger tube. It would be better for this heat source to be uniform throughout the flushing tube; in our case, the small tube is wrapped around the larger tube in a helix. By switching to a concentric tube geometry, the radial component of the heat flush should be much more uniform.

3. The helium "reflux" is a process with a circular evaporation-recondensation loop of a superfluid LHe film that has a very high heat conductivity due to the large flow rate of superfluid film creep. In the 1 K to 2 K range, the heat conductivity of this effect can be orders of magnitude larger than the conductivity of copper[68, 69]. It is conceivable that this mechanism could short out the flushing heater allowing the power generated by the heater to be carried away by the helium "reflux" instead of by roton creation in the flushing tube. With the telemetry that was installed, it was not possible to constrain the pressure in the vacuum can accurately enough to eliminate the possibility of superfluid film creep. Therefore, it is conceivable that the effectiveness of this purifier could be greatly diminished. If the purifier were to be run again, it would be of paramount importance to ensure that the pressure in the vacuum can is a few orders of magnitude below the vapor pressure at the temperature that the purifier is operating. This will prevent superfluid film creep and therefore the helium "reflux". It is possible that the pressure was low enough in the vacuum can already and that there was no helium "reflux". In that case, improving the pressure measurement in the vacuum can would be sufficient. Otherwise, increasing the effective pumping speed of the system will result in a lower base pressure.

   Using a shorter KF vacuum tube on the vacuum can pump out port would substantially improve the effective pumping speed of the system. The tube that was used was a 2 m long KF-25 tube. It was chosen because it was what we had on hand, but was much longer than needed. If instead the pump was placed on a raised platform near the purifier top plate and a much shorter pump out line was used, it would increase the effective pumping speed of the system by a factor of $\approx 2 - 3$. This would improve the performance of the purifier when extracting the exchange gas from the vacuum can after the purifier is filled

with LHe. It would reduce the thermal conductivity of the helium "reflux" by limiting the thickness of any superfluid films that exist in the vacuum. It would also make burning off the helium films in the vacuum can more effective, which is done by pumping on the vacuum can while warming the pure loop and 1K Pot above 4.2 K using the heaters.

4. Adding a second all metal seal valve at V52 would allow the flex line to remain under vacuum when a new sample bottle was attached to the system. This would drastically decrease the amount of atmosphere allowed into the gas handling system and would decrease the time it takes to pump the system back down after changing sample bottles.

5. I suggest reworking the purifier procedures for cleaning the PGHS from the extraction line to V3 and V10. To measure the amount of gas that was being purged so that the dilution factor could be calculated, the extraction line was blocked resulting in a rise in the pressure in the pure loop and a corresponding increase in temperature. I do not have evidence that this caused an issue with the flushing gradient. However, I think it would be better to throw away ultrapure product for some amount of time with V1 open instead of pumping the system back out with V1 closed. The difficulty is in determining the proper amount of time to throw out the ultrapure product. I suggest using the procedures as they are written first and measure how long it takes to clean the purifier. I would then use that duration, possibly modified by a safety factor if the duration is not very long, as the amount of time to throw out the ultrapure. This should result in a much more stable temperature and therefore may result in a more stable flushing gradient. Ultimately, at least the same amount of helium should be flushed through the system with this method. The main concern that I have with this modification is that the product that is being used to flush the system may not mix as thoroughly with any contamination in the PGHS, for example ^{3}He that is close to V10 might be less likely to be removed with this new method if the ^{3}He and ^{4}He do not mix properly.

   Alternatively, the extraction line could be fitted with an overpressure valve that leads to a second vacuum pump. Setting the pressure of the overpressure valve to -5 inHg would result in a much more stable pressure in the extraction line. With this modification, instead of the pressure building up in the extraction line when the system was being cleaned, it would be vented through the overpressure valve. When the cleaning was finished, the overpressure valve would not throw out any product because -5 inHg is a bit higher than the pressure on the extraction line when the purifier is running.

6. The needle valves leaked substantially and should be remade. There were three primary issues associated with this. First, it severely affected the method for leak checking. To leak check the purifier, we were required to use rubber stoppers on the fill ports so that

the pressure would get low enough in the purifier for the leak detector to function. To use the rubber stoppers, the sintered copper filters were broken off. After leak checking is completed, the rubber stoppers are removed, and the sintered copper filters are glued back in place. If future leak checking was required, the filters would have to be broken off and re-glued once leak checking was finished. The cure time on the Stycast epoxy that we used was 8 hr resulting in a substantial delay if leak checking was required. The second manifestation of the leaky needle valves was our inability to control the level in the 1K Pot. The needle valves leaked sufficiently that, even when NV2 was closed, the GSX would continue to pull additional liquid into the 1K Pot, increasing the level. This resulted in an increased heat load on the purifier as the liquid level rose higher in the 1K Pot to portions of the 1K Pot that were at higher temperatures. It also resulted in an increase in the base temperature of the purifier. Luckily, the temperature was still in the desired range, but it would be much better to be able to control the level in the 1K Pot. Finally, the third effect of the leaky needle valves was that nitrogen could get into the pure loop and the 1K Pot during the LN fill. Any nitrogen in the purifier will freeze during the LHe fill and could cause blockages or further degrade the quality of the seal through the needle valves. This last issue was fixed by using a high-pressure helium gas bottle to maintain a positive flow through the PGHS and into the main helium bath through both the 1K Pot and the pure loop. This was done during both the LN fill and the start of the LHe fill to limit the amount of nitrogen that made it into the system.

7. The ID of the helium fill lines on the 1K Pot and the pure loop inlet should be decreased. The current fill lines have a 1.6 cm OD, which is much larger than needed. In addition, the fill lines should be extended to within an inch of the bottom of the vacuum can. This could allow the purifier to be operated for longer because in the current design the operation has to be ceased when the first fill line falls below the liquid level. If the fill lines are extended, operation could continue until the purifier started warming up. We will not know when this will happen until the fill ports are extended and the purifier is operated again. During operation, the level in the main helium bath decreases by a rate of about 0.1 %/min. Therefore, if the purifier could be run until the level was 20% instead of 28%, it would extend purification by over an hour.

8. There were a sufficient number of open connectors on the feedthroughs to put another calibrated thermometer in the vacuum can. During the thermometer calibration, the calibrated thermometer and the 10 kΩ RuO thermometers were cooled together to transfer the calibration to the 10 K RuO thermometers. Only seven thermometers could be calibrated at a time. It was decided that it was not worth the time and monetary investment to calibrate an additional thermometer by performing another calibration run. Therefore,

in the future, it would be possible to add an additional calibrated thermometer to the system. I recommend putting the additional thermometer on the extraction tube after the heat exchanger. The difference between T5 and 4.2 K was meant to be used to assess the performance of the heat exchanger; however, T5 did not produce meaningful readings during the operation of the purifier because it was too close to the flushing heater. Including this additional heater and comparing it to T1 will allow an independent assessment of the performance of the heat exchanger. T5 should also be moved as far away from the flushing heater as possible.

9. A small thermometer cell should be designed that embeds the thermometers inside the helium for use with the 1/8 in copper tubes. To move the thermometer inside the tube, a larger ID will be required. A 1.6 cm OD tube should be sufficient, with couplers for the 1/8 in copper tube on each end. The 1.6 cm tube would have a custom feedthrough with the two wires for the RuO thermometer. The two couplers would have to be brazed to the tube first and leak checked with rubber stoppers. The part would then have to be soft soldered to the 1/8 in copper tube tubes and leak checked, again with rubber stoppers. Finally, the feedthrough would be glued in place, and the part would be leak checked with the rest of the pure loop. If this worked, it would give us more confidence that the reported temperature was representative of the helium.

10. During the cryogen fills, the cold gas from the exhaust port would freeze components of the purifier top plate. On one occasion, one of the solder connections on the electronics feedthrough in the vacuum can broke during the cryogen fill, which was attributed to the thermal cycling. Therefore, heater tape was wrapped around the most strongly affected parts of the purifier top plate, and a tube was connected to the exhaust port to help direct the exhaust away from the other components on the purifier top plate. The heater tape was operated at higher currents during the cryogen transfers and was also used when the 1K Pot was being pumped on because the helium vapor would also cause the purifier top plate to freeze. After these modifications, there were no additional issues with wires breaking.

Chapter 7

Conclusions

This thesis presents work that has been completed as part of a continued effort to measure the neutron β-decay lifetime at NIST using magnetically trapped ultracold neutrons. The apparatus has been successfully operated, and neutron trapping has been demonstrated. The experimental apparatus and the initial neutron trapping data are described. The performance of a variety of subsystems used in the experiment been evaluated and additional tests have been recommended to improve the performance evaluations.

Analysis software has been developed to distil and extract the useful information from the digitized voltage traces into convenient pulse shape parameters that are used to tag and remove background events. The analysis also corrects for a variety of effects including gain drifts, detector deadtime, and a firmware issue in the DAQ cards. Background rates are measured in an alternative operating scheme and directly subtracted from the data to further improve the signal to noise. Finally, similar data files are pooled to limit systematic effects associated with low counting statistics. Using these methods, a trap lifetime is extracted for three primary data sets.

Quantitative estimates of the systematic effects have been presented, which includes an additional contribution to the time-dependent detection efficiency systematic effect that has been discovered in this work. Substantial improvements in the models of the Monte Carlo simulation for marginally trapped neutrons have been realized, and sensitivity studies have been performed to estimate both a systematic correction and the uncertainty in the correction from the model assumptions. Additionally, the Monte Carlo has been benchmarked showing good agreement that is statistically limited between the predicted systematic corrections and the data. Recent improvements in the AMS techniques that have been developed in collaboration with Argonne National Laboratory's ATLAS facility will allow direct measurement of ratios of ^{3}He to ^{4}He at the level required for a 1 s level of the neutron β-decay.

After applying the systematic corrections, a final trap lifetime of 707 ± 20 s is obtained,

which suggests that there are additional systematic effects that have not been accounted for properly. Additional ^{3}He contamination is the most likely candidate for the remaining discrepancy between the PDG mean lifetime the measured result.

To address the possibility of additional ^{3}He contamination, new samples of ultrapure ^{4}He must be produced and measured using the AMS methodology that has been developed. A continuous flow, ^{4}He purifier has been commissioned and operated to produce new samples of ultrapure ^{4}He. The purifier is designed to limit potential ^{3}He contamination sources by exposing the ultrapure product to only all-metal surfaces. Slight modifications to the UCN lifetime's gas handling system at NIST have allowed the purifier to be operated on site and for the ultrapure product to be produced directly into the gas handling system there that is designed for the long term storage of ultrapure ^{4}He. Operating procedures for the purifier are presented. AMS measurements of the new samples are pending. If the purity of these samples is found to be at the level of a few $\times 10^{-15}$, it will provide motivation for taking additional data to determine if the unaccounted for systematic effect has been addressed.

REFERENCES

[1] F. E. Wietfeldt and G. L. Greene. Colloquium: The neutron lifetime. *Rev. Mod. Phys.*, 83:1173–1192, Nov 2011.

[2] D. Griffiths. *Introduction to Elementary Particles.* Wiley-VCH, 2008.

[3] C. Patrignani and Particle Data Group. Review of particle physics. *Chinese Phys. C*, 40(10):100001, 2016.

[4] G. J. Mathews, T. Kajino, and T. Shima. Big bang nucleosynthesis with a new neutron lifetime. *Physical Review D*, 71(2):021302, 2005.

[5] M. Kobayashi and T. Maskawa. CP-violation in the renormalizable theory of weak interaction. *Progress of Theoretical Physics*, 49(2):652–657, 1973.

[6] I. S. Towner and J. C. Hardy. The evaluation of Vud and its impact on the unitarity of the Cabibbo–Kobayashi–Maskawa quark-mixing matrix. *Reports on Progress in Physics*, 73(4):046301, 2010.

[7] J. C. Hardy and I. S. Towner. Superallowed $0^+ \rightarrow 0^+$ nuclear β decays: 2014 critical survey, with precise results for Vud and CKM unitarity. *Physical Review C*, 91(2):025501, 2015.

[8] J. S. Nico, M. Arif, M. S. Dewey, T. R. Gentile, D. M. Gilliam, P. R. Huffman, D. L. Jacobson, and A. K. Thompson. The fundamental neutron physics facilities at NIST. *Journal of research of the National Institute of Standards and Technology*, 110(3):137, 2005.

[9] A. T. Yue, M. S. Dewey, D. M. Gilliam, G. L. Greene, A. B. Laptev, J. S. Nico, W. M. Snow, and F. E. Wietfeldt. Improved determination of the neutron lifetime. *Physical review letters*, 111(22):222501, 2013.

[10] D J Salvat, E R Adamek, D Barlow, J D Bowman, L J Broussard, N B Callahan, S M Clayton, S Currie, E B Dees, W Fox, P Geltenbort, K P Hickerson, A T Holley, C Liu, M Makela, J Medina, D J Morley, C L Morris, S I Penttil, J Ramsey, A Saunders, S J Seestrom, E I Sharapov, and S K L Sjue. Storage of ultracold neutrons in the magneto-gravitational trap of the UCNτ experiment. *PHYSICAL REVIEW C*, 052501:1–6, 2014.

[11] C. E. H. Mattoni, C. P. Adams, K. J. Alvine, J. M. Doyle, S. N. Dzhosyuk, R. Golub, E. Korobkina, D. N. McKinsey, A. K. Thompson, L. Yang, et al. A long wavelength neutron monochromator for superthermal production of ultracold neutrons. *Physica B: Condensed Matter*, 344(1):343–357, 2004.

[12] S. N. Dzhosyuk, A. Copete, J. M. Doyle, L. Yang, K. J. Coakley, R. Golub, E. Korobkina, T. Kreft, S. K. Lamoreaux, A. K. Thompson, et al. Determination of the neutron lifetime using magnetically trapped neutrons. *Journal of research of the National Institute of Standards and Technology*, 110(4):339, 2005.

[13] P-N. Seo, K. J. Coakley, J. M. Doyle, F. H. DuBose, R. Golub, E. Korobkina, S. K. Lamoreaux, H. P. Mumm, C. M. OShaughnessy, G. R. Palmquist, et al. Progress towards a precision measurement of the neutron lifetime using magnetically trapped ultracold neutrons. In *AIP Conference Proceedings*, volume 842, pages 811–813. AIP, 2006.

[14] C. M. OShaughnessy, R. Golub, K. W. Schelhammer, C. M. Swank, P-N. Seo, P. R. Huffman, S. N. Dzhosyuk, C. E. H. Mattoni, L. Yang, J. M. Doyle, et al. Measuring the neutron lifetime using magnetically trapped neutrons. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 611(2):171–175, 2009.

[15] P. R. Huffman, K. J. Coakley, J. M. Doyle, C. R. Huffer, H. P. Mumm, C. M. OShaughnessy, K. W. Schelhammer, P-N. Seo, and L. Yang. Design and performance of a cryogenic apparatus for magnetically trapping ultracold neutrons. *Cryogenics*, 64:40–50, 2014.

[16] L. Yang. *Towards precision measurement of the neutron lifetime using magnetically trapped neutrons*. PhD thesis, Harvard University Cambridge, Massachusetts, 2006.

[17] C. M. O'Shaughnessy. *Development of a Precision Neutron Lifetime Measurement: Magnetic Trapping of Ultracold Neutrons*. PhD thesis, 2010.

[18] K. W. Schelhammer. *Measurement of the Beta-Decay Lifetime of Magnetically Trapped Ultracold Neutrons*. PhD thesis, North Carolina State University, 2013.

[19] P. R. Huffman, C. R. Brome, J. S. Butterworth, S. N. Dzhosyuk, R. Golub, S. K. Lamoreaux, C. E. H. Mattoni, D. N. McKinsey, and J. M. Doyle. Magnetically stabilized luminescent excitations in hexagonal boron nitride. *Journal of luminescence*, 92(4):291–296, 2001.

[20] J. S. Butterworth, C. R. Brome, P. R. Huffman, C. E. H. Mattoni, D. N. McKinsey, and J. M. Doyle. A removable cryogenic window for transmission of light and neutrons. *Review of scientific instruments*, 69(11):3998–3999, 1998.

[21] C. E. H. Mattoni. *Magnetic Trapping of Ultracold Neutrons Produced Using a Monochromatic Cold Neutron Beam*. PhD thesis, Harvard University, 2002.

[22] S. N. Dzhosyuk, C. E. H. Mattoni, D. N. McKinsey, A. K. Thompson, L. Yang, J. M. Doyle, and P. R. Huffman. Neutron-induced luminescence and activation in neutron shielding and scintillation detection materials at cryogenic temperatures. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 217(3):457–470, 2004.

[23] K. J. Coakley, M. S. Dewey, M. G. Huber, C. R. Huffer, P. R. Huffman, D. E. Marley, H. P. Mumm, K. W. Schelhammer, A. K. Thompson, A. T. Yue, et al. Survival analysis approach to account for non-exponential decay rate effects in lifetime experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 813:84–95, 2016.

[24] K. Tsuchiya, K. Egawa, K. Endo, Y. Morita, N. Ohuchi, and K. Asano. Performance of the eight superconducting quadrupole magnets for the TRISTAN low-beta insertions. *Magnetics, IEEE Transactions on*, 27(2):1940–1943, Mar 1991.

[25] S. N. Dzhosyuk. *Magnetic Trapping of Neutrons for Measurement of the Neutron Lifetime*. PhD thesis, Harvard University Cambridge, Massachusetts, 2004.

[26] G. Citver, S. Feher, P. J. Limon, D. Orris, T. Peterson, C. Sylvester, J. C. Tompkins, et al. HTS power lead test results. In *Proceedings of the 1999 Particle Accelerator Conference*, volume 2, pages 1420–1422. IEEE, 1999.

[27] T. G. Miller. Determination of α/β ratio for liquid helium. *Nuclear Instruments and Methods*, 32(2):239–241, Feb 1965.

[28] M. Stockton, J. W. Keto, and W. A. Fitzsimmons. Ultraviolet emission spectrum of electron-bombarded superfluid helium. *Phys. Rev. A*, 5:372–380, Jan 1972.

[29] D. N. McKinsey, C. R. Brome, S. N. Dzhosyuk, R. Golub, K. Habicht, P. R. Huffman, E. Korobkina, S. K. Lamoreaux, C. E. H. Mattoni, A. K. Thompson, L. Yang, and J. M. Doyle. Time dependence of liquid-helium fluorescence. *Phys. Rev. A*, 67:062716, Jun 2003.

[30] D. N. McKinsey. *Detection of Magnetically Trapped Neutrons: Liquid Helium as a Scintillator*. PhD thesis, Harvard University, 2002.

[31] J. R. Kane, R. T. Siegel, and A. Suzuki. Scintillations in normal and superfluid helium. *Physics Letters*, 6(3):256–257, Sep 1963.

[32] J. S. Adams, Y. H. Kim, R. E. Lanou, H. J. Maris, and G. M. Seidel. Scintillation and quantum evaporation generated by single monoenergetic electrons stopped in superfluid helium. *Journal of Low Temperature Physics*, 113(5-6):1121–1128, 1998.

[33] J. M. Flournoy, I. B. Berlman, B. Rickborn, and R. Harrison. Substituted tetraphenylbutadienes as fast scintillator solutes. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 351(2):349–358, 1994.

[34] C. H. Lally, G. J. Davies, W. G. Jones, and N. J. T. Smith. UV quantum efficiencies of organic fluors. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 117(4):421–427, Oct 1996.

[35] V. M. Gehman, S. R. Seibert, K. Rielage, A. Hime, Y. Sun, D. M. Mei, J. Maassen, and D. Moore. Fluorescence efficiency and visible re-emission spectrum of tetraphenyl butadiene films at extreme ultraviolet wavelengths. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 654(1):116–121, Oct 2011.

[36] J. D. Jackson, S. B. Treiman, and H. W. Wyld. Possible tests of time reversal invariance in beta decay. *Phys. Rev.*, 106:517–521, May 1957.

[37] P. Boulet, J. Gérardin, Z. Acem, G. Parent, A. Collin, Y. Pizzo, and B. Porterie. Optical and radiative properties of clear PMMA samples exposed to a radiant heat flux. *International Journal of Thermal Sciences*, 82:1–8, Aug 2014.

[38] H. M. Zidan and M. Abu-Elnader. Structural and optical properties of pure PMMA and metal chloride-doped PMMA films. *Physica B: Condensed Matter*, 355(1-4):308–317, Jan 2005.

[39] Hamamatsu Photonics K. K. Photomultiplier handbook: Third edition.

[40] J. Uribe, Hongdi Li, H. Baghaei, M. Aykac, Yu Wang, Yaqiang Liu, Tao Xing, and Wai-Hoi Wong. Effect of photomultiplier gain-drift and radiation exposure on 2D-map decoding of detector arrays used in positron emission tomography. In *2001 IEEE Nuclear Science Symposium Conference Record (Cat. No.01CH37310)*, volume 4, pages 1960–1964, 2001.

[41] P. B. Coates and J. W. Andrews. Measurement of gain changes in photomultipliers. *Journal of Physics E: Scientific Instruments*, 14(10):1164, 1981.

[42] M. Yamashita. Time dependence of ratedependent photomultiplier gain and its implications. *Review of Scientific Instruments*, 51(6):768–775, Jun 1980.

[43] Burle 8854 data sheet.

[44] J. Beringer and Particle Data Group. Review of particle physics. 86, 2012.

[45] K. J. Coakley. Stochastic modeling and simulation of marginally trapped neutrons. In S. J. Seestrom, editor, *Next Generation Experiments to Measure the Neutron Lifetime: Proceedings of the 2012 Workshop*, Santa Fe, NM, 2014. World Scientific.

[46] K. J. Coakley. Statistical planning for a neutron lifetime experiment using magnetically trapped neutrons. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 406(3):451–463, 1998.

[47] R. Golub, D. Richardson, and S. K. Lamoreaux. *Ultra-cold neutrons*. CRC Press, 1991.

[48] Certain materials are identified in this paper to foster understanding. such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials identified are necessarily the best available for the purpose.

[49] R. Golub. On the storage of neutrons in superfluid 4He. *Physics Letters A*, 72(4-5):387–390, Jul 1979.

[50] R. Golub, C. Jewell, P. Ageron, W. Mampe, B. Heckel, and I. Kilvington. Operation of a superthermal ultra-cold neutron source and the storage of ultra-cold neutrons in superfluid Helium4. *Zeitschrift für Physik B Condensed Matter*, 51(3):187–193, Sep 1983.

[51] V. F. Sears. Neutron scattering lengths and cross sections. *Neutron News*, 3(3):26–37, Jan 1992.

[52] J. Wilks. *The properties of liquid and solid helium.* Clarendon Press, 1967.

[53] P. V. E. McClintock. An apparatus for preparing isotopically pure He4. *Cryogenics*, 18(4):201–208, Apr 1978.

[54] P. C. Hendry and P. V. E. McClintock. Continuous flow apparatus for preparing isotopically pure 4He. *Cryogenics*, 27(3):131–138, Mar 1987.

[55] H. P. Mumm, M. G. Huber, W. Bauder, N. Abrams, C. M. Deibel, C. R. Huffer, P. R. Huffman, K. W. Schelhammer, R. Janssens, C. L. Jiang, R. H. Scott, R. C. Pardo, K. E. Rehm, R. Vondrasek, C. M. Swank, C. M. O'Shaughnessy, M. Paul, and L. Yang. High-sensitivity measurement of ^{3}He $-^4$ He isotopic ratios for ultracold neutron experiments. *Physical Review C*, 93(6):065502, Jun 2016.

[56] C. R. Brome. *Magnetic Trapping of Ultracold Neutrons.* PhD thesis, Harvard Univeristy, 2000.

[57] J. H. Coon. He^3 isotopic abundance. *Phys. Rev.*, 75:1355–1357, May 1949.

[58] H. Yoshiki, H. Nakai, and E. Gutsmiedl. A new superleak to remove He3 for UCN experiments. *Cryogenics*, 45(6):399–403, Jun 2005.

[59] G. A. Herzlinger and J. G. King. Measurement of diffusion in 3He-4He solutions and determination of the 3He-roton cross section. *Physics Letters A*, 40(1):65–66, Jun 1972.

[60] P. C. Hendry. *The Isotopic Purification of ^{4}He.* PhD thesis, University of Lancaster, 1985.

[61] T. P. Ptukha. Thermal conductivity and diffusion in weak He3-He4 solutions in the temperature range from the 0.6-degrees-K. *Soviet Physics JETP-USSR*, 13(6):1112–1119, 1961.

[62] J. S. Brooks and R. J. Donnelly. The calculated thermodynamic properties of superfluid helium4. *Journal of Physical and Chemical Reference Data*, 6(1):51–104, Jan 1977.

[63] K. Fukuda and A. Hirai. An NMR experiment on the flow in 3He-4He II mixtures. *Physica B+C*, 82(2):343–346, Apr 1976.

[64] D. P. Kendellen. *Cryogenic Design for the Neutron Electric Dipole Moment Experiment.* PhD thesis, North Carolina State University, 2012.

[65] Professor David Haase (North Carolina State University). personal communication.

[66] B. V. Rollin and F. Simon. On the "film" phenomenon of liquid helium II. *Physica*, 6(2):219–230, Feb 1939.

[67] E. Long and L. Meyer. Superfluidity and heat transport in the unsaturated helium-II film. *Phys. Rev.*, 98(6):1616–1622, Jun 1955.

[68] D. P. Kendellen and D. G. Haase. Measurement and modeling of thermal flow in an enclosed tube containing superfluid helium film. *Cryogenics*, 57:134–139, Oct 2013.

[69] P. J. Nacher, M. Cornut, and M. E. Hayden. Compression of 3He by refluxing 4He: A model for computing HEVAC effects in 3He-4He mixtures. *Journal of Low Temperature Physics*, 97(5-6):417–443, Dec 1994.

APPENDICES

Appendix A

Energy Calibration

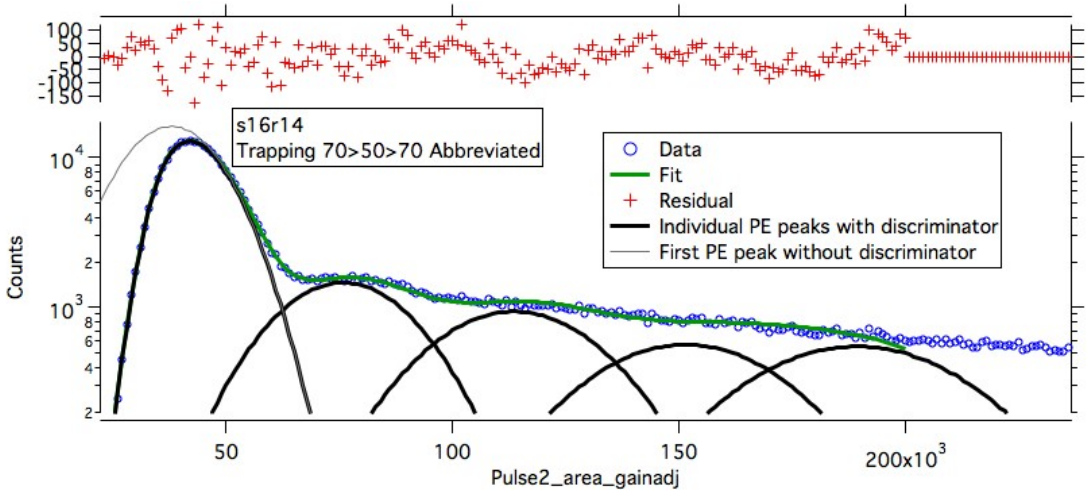
Throughout this work, I have used units of the gain adjusted pulse area and the number of PE almost interchangeably. Both the lower and upper pulse area cuts were made on the effective number of PE instead of the actual pulse area metric. From the point of view of understanding the backgrounds and cuts, the number of PE is typically the most straightforward parameter to consider, but the pulse area is the parameter that is calculated. The interchangeability of these two parameters is allowed for a few reasons. First, a calibration between the gain adjusted pulse area and the number of PE has been determined, which is the topic of this section. Secondly, the experiment is insensitive to small inaccuracies in the absolute calibration. By performing the pulse area and pulse height cuts on the gain adjusted pulse area, these cuts are effectively being performed with respect to the location of the reference pulser. The reference pulser has been shown to be very stable in time, which suggests that the mean number of photons and hence the number of PE in the reference events is also stable in time. By performing the cuts with respect to the reference pulser, the cuts are effectively with respect to the number of PE. A small inaccuracy in the calibration will cause a variation in the exact location of the cut threshold in the number of PE, but the cut threshold will be stable in time. In this section, I will present the method used to determine the calibration between the gain adjusted pulse area and the number of PE.

An example of the gain adjusted pulse area histogram can be seen in Figure A.1. At low pulse area, the contribution from the individual PE can be seen. If the mean of these PE peaks can be determined, a calibration between the gain adjusted pulse area and the number of PE can be obtained. To do this, a reasonable model for the shape of the PE spectra can be fit to the data. For a first pass model, the individual PE peaks can be assumed to be Gaussian, and the hardware pulse height discriminator can be modeled as a hill function. The resulting model is

$$y\left(x\right)=\frac{1}{1+\left(\frac{xhalf}{x}\right)^{rate}}\sum_{i=1}^{n}A_i\exp\left(-\left(\frac{x-x_i}{w_i}\right)^2\right),$$

where x is the gain adjusted pulse area, $xhalf$ is the location of the midpoint of the hill function, $rate$ determines the squishiness of the hill function, i is the individual PE peaks where A_i is the amplitude of the Gaussian, x_i is the mean of the Gaussian, and w_i is its width. Depending on the range of the fit that is performed only a limited number of Gaussians have to be included in the fit, and as a larger fit range is used and the contributions from each of the PE peaks becomes less obvious, the fitting algorithm is more likely to fail due to a singular matrix error.

Figure A.1: Fit to the gain corrected pulse area spectra showing the contribution from the individual PE peaks. The first PE peak is shown without the inclusion of the hill equation in gray as an indication of the effect of the discriminator threshold. This graph demonstrates that the discriminator threshold interacts with the second PE peak to a much lesser extent and the interaction with the higher PE peaks is negligible.



Using physical intuition, the number of free parameters can be reduced by introducing a single variable that defines the locations of all of the PE peaks and another variable to describe the widths of the PE peaks. Unfortunately, a method for relating the amplitudes for the PE peaks was not developed, so every peak included in the fit requires one free variable to describe its height. This allows the relatively complicated pulse area spectra to be modeled with a greatly reduced number of free parameters.

The variations in the response for the individual PE is what causes the width of the individual PE peaks. If the response of the individual PE that make up the larger peaks is assumed

to be uncorrelated, which corresponds to assuming that the response of the PMTs is linear, they will add like the square root of the sum of the squares of the individual uncertainties. This results in a simple scaling argument between the widths of the individual PE peaks. The width of the ith peak will be given by $w_i = \sqrt{i} \times w_1$. Under the assumption that the response of the PMTs in linear, the mean of the PE peaks can be assumed to be linear in the number of PE, i.e. $x_i = i \times x_1$. Combining these two assumptions, the model of the PE spectra becomes

$$y\left(x\right) = \frac{1}{1 + \left(\frac{xhalf}{x}\right)^{rate}} \sum_{i=1}^{n} A_i \exp\left(-\left(\frac{x - i \times x_1}{\sqrt{i} \times w_1}\right)^2\right),$$

where x_1 is both the location of the single PE peak and the conversion between the gain adjusted pulse area and the number of PE and w_1 relates to the variation in the size of the individual PE peaks. Figure A.1 shows a fit to the gain adjusted pulse area spectra using this model. In this case, the fit included the effect of five Gaussians, and the fit range extends to approximately the mean of the fifth Gaussian. At high gain adjusted pulse area the fit falls below the data, which is the effect of ignoring the contribution from the higher PE peaks. Because the contribution from the higher order peaks is small, the fit function becomes unreliable if a larger number of PE peaks are included with the same fit range because of singular matrix failures. The location of the reference pulser was determined in terms of the PE conversion calculated with this method and was found to be 27.25(25) PE in channel 1(2).

Appendix B

Gain Monitor: Additional Information

There are a few additional tests that were performed on the reference pulser system, which might prove useful and are discussed in the following. The section ends with an unsuccessful attempt to separate the effects of the various mechanisms that contribute to the gain drifts.

B.1 Missing Reference Pulser Events

We have found that there are missing reference pulser events. Since the reference pulser timing is known to high precision, the location of the next reference pulser event can be estimated with high accuracy. When this is done we are very successful in predicting the location of the next reference events, but a small fraction of the reference pulser events are missing, see Figure B.1 for an example of two consecutive missing reference pulser events. In a typical production data file, s13r10b1m0.dat, 22% of the 9.9×10^5 events were tagged as reference events. To calculate the number of missing reference events, the reference events that were affected by bad timestamps or the software veto are removed. After removing these events, 219196 good reference events remain. In addition, there are 1436 missing reference events in this time period, or equivalently 0.66% of the reference events are missing.

There are a couple of possible explanations for the missing reference events that can be easily thrown out. It has been verified that there is not an event at the time of the missing reference event. Therefore, it is not an issue of the reference threshold being set too high or of light getting to the main detection PMTs but not to the reference PMT by whatever mechanism. It has also been verified that there was not an event very close to, but slightly before the reference event, which suggests that this is not a deadtime issue.

Another possible explanation, which is much harder to test, is that the reference pulser

events are being lost by the hardware pulse height discriminator. The exact shape of the discriminator threshold is difficult to assess. The reference pulser is not located particularly far from the pulse height threshold, in terms of the width of the reference pulser peak. This can be seen by looking at Figure B.2. Although this is a pulse area histogram the reason that the distribution does not extend to zero in pulse area is that the pulse height discriminator has removed the small events. The lowest peak in the pulse area histogram is substantially modified by the hardware pulse height discriminator. Therefore, the lowest peak in pulse area histogram is an indication of the location of the hardware discriminator threshold in pulse area, which in this plot can be compared to the reference pulser location. Without knowledge of the width and location of the discriminator threshold, our ability to estimate the fraction of reference events below this threshold is limited. The pulse height discriminator is set a little above the height of a single PE peak. We can get a conservative (overestimate) of the number of reference events below the pulse height discriminator by assuming the cut threshold is at zero pulse area.

For file s13r1b1m0.dat, in channel 1 (channel 2), the mean is 4.8 σ (4.1 σ) from the zero value. Assuming a perfect discriminator with a threshold set to zero, when in reality the threshold is somewhat higher, suggests that $2 \times 10^{-8}\%$ ($4 \times 10^{-7}\%$) of reference events fall below a zero pulse area threshold. Because of the coincidence requirement, a reference event will be not be detected if the signal in either main detection PMT falls below the discriminator threshold. Therefore, assuming the response in the two channels is uncorrelated, the probability of a missing reference event is given by the sum of the two probabilities minus their product in order to account for double counting. The probability of a missing reference event using this conservative estimate comes out to $4.2 \times 10^{-7}\%$. In the opposite limit, where the signal in the two main detection PMTs for a reference event are perfectly correlated the probability is given by the smaller of the two probabilities, resulting in an even small fraction of missing events. Despite the fact that this is a conservative estimate, it is hard to imagine how this effect could be large enough to account for all of the missing reference pulser events.

In conclusion, we find that a small fraction of the reference events is missing. The mechanisms that come to mind that could cause this do not seem to be large enough to account for the number of reference events we are missing. Although it would be nice to understand this effect, we do not believe it would introduce a systematic effect and therefore we do not think it is worth pursuing further at this time.

B.2 Reference Event Timing

As mentioned previously, the reference pulser is operated at 100 Hz. The timing of the reference pulser can be cross checked by looking at the timestamp of the reference pulser events in the data. A timing spectrum of the reference events is plotted in order to show the difference in time

between consecutive reference pulser events in the top graph of Figure 2.9. The width of the distribution is roughly 3 μs. Some width to the distribution is expected due to the random path that the light takes to the detectors, however much less than what is observed here. Therefore, it is suspected that the structure in the timing spectrum must come from uncertainty in the DAQ timing, the signal generator, or the mechanism for producing the light and not from variations in the time of flight of the light itself. It seems likely that this is due to the timing of the signal generator, not the DAQ cards, and therefore it is not expected to cause a systematic effect.

Figure B.1: Plot of an atypical timestamp array with two consecutive missing reference events, where reference events have been distinguished. The reference events are observed at 100 Hz to high precision, as expected. The time that these events were expected is shown in green and black for the first and second missing reference events respectively.

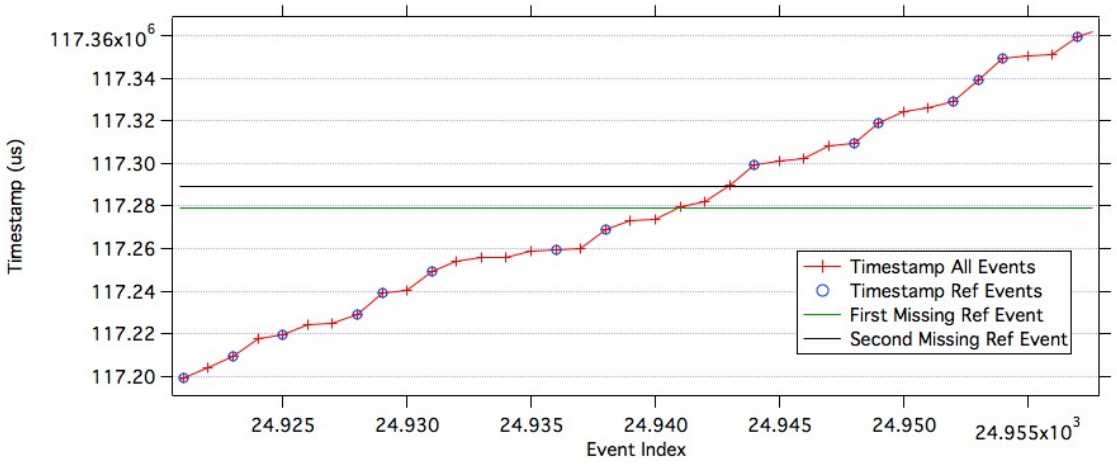
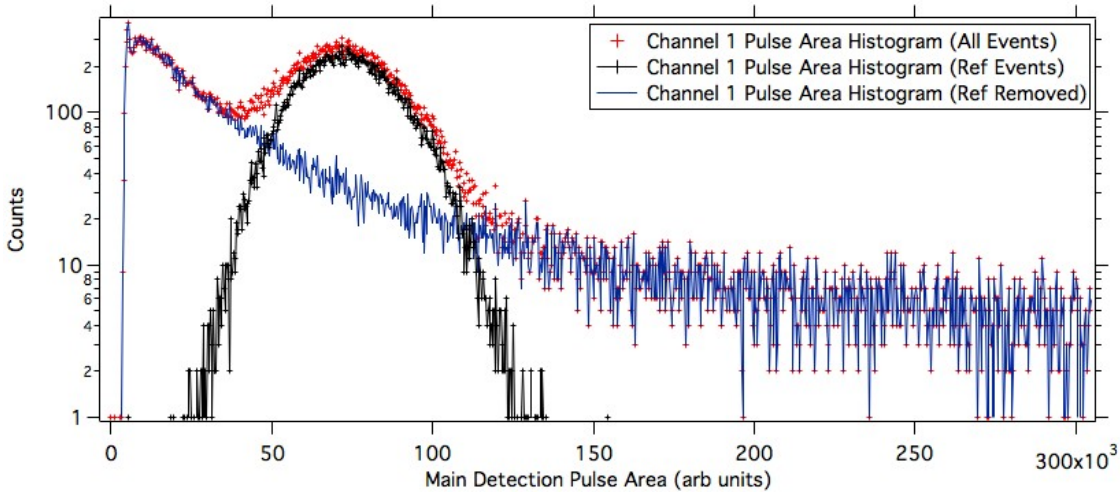


Figure B.2: Main detection pulse area spectrum including and excluding reference events.



This topic is discussed in more detail in Section 2.13.3, which starts on page 62 as a general performance issue instead of something solely related to the gain monitor.

B.3 Evaluating Mechanisms that Contribute to the Gain Drifts

Any work to improve upon the design of the gain monitor system should be informed by an understanding of the mechanisms that dominated the current design. Along these lines, a list of possible gain drift mechanisms is presented, and to the extent possible with the data previously gathered, a limit on the size of the gain drift due to each of these effects is estimated. The mechanisms that will be considered here are PMT blinding, heating of the PMT electronics, and relaxation of the magnetic shielding. Each of these mechanisms is expected to result in a gain drift that approaches a steady state value. As a result, it may be possible to model the individual gain drifts with an exponential and a constant offset. The following paragraphs investigate each of these mechanisms to determine which mechanisms dominate the gain drifts in this experiment.

The warming of the PMT electronics can be separated from the other effects by taking diagnostic data where the voltage of the main detection PMTs was changed periodically and the location of the reference pulser in the main detection channels was monitored. This diagnostic data was taken with the magnetic fields off so that time-dependent magnetic effects could not influence the gains of the main detection PMTs. As mentioned in Section 2.5.3, which starts on page 33, one of the bases of the two main-detection PMTs was upgraded with resistors with a low sensitivity to the temperature. The gain drift in the original base, which corresponds to channel 2 of the main detection system, experiences a gain reduction of $8.5 \pm 0.3\%$ that approaches the steady state value with a time constant of 56 ± 4 s. The temperature induced gain drift in channel 1 was sufficiently small that the fitting algorithm suffered a singular matrix error due to its inability to separate out the effect of the lifetime and amplitude. Therefore, in order to quantify the reduction in the gain drift in the temperature stable sockets, it was assumed that the time constant was the same in both of the main detection channels. This allowed the fit to converge and resulted in a relative gain drift of $0.4 \pm 0.3\%$ in channel 1, which corresponds to a fraction reduction in the temperature dependent gain drifts of 0.05 ± 0.04 when compared to the PMT base using standard resistors. Using this information, we estimate that the use of temperature stabilized PMT bases causes a reduction in the temperature induced gain drift by at least a factor of 0.13 with a 95% confidence level.

In our data, the shortest duration between ramping up the PMTs and the start of data collection is 57 s. This results in an additional reduction of the fractional gain drift by a factor of $>= \approx 1/e$. Because of the short lifetime, this effect will become more dominant in cases when the PMTs are ramped closer to the start of data collection; and therefore, it can be minimized

effectively by ramping up the PMTs as far before data collection as possible.

Next, the gain drifts due to ramping the magnets are considered. Even after the magnets finish ramping, the gain will continue to drift as the magnetic shielding saturates and the apparatus approaches its steady-state magnetization. Because the solenoid and quadrupole magnets were not ramped at the same time, the magnetic effects can be further broken down by separating out the contributions of the two magnets. The solenoid magnets were typically ramped up at the beginning of a 20 h period of data collection, while the quadrupole magnets were ramped for each production data file.

The solenoid-ramp-induced gain drift can be investigated using the diagnostic runs that follow the cryogen fills, which use the magnet overshoot script. It is descried in Appendix D, which starts on page 237. During these scripts, the data collection system is operating while the solenoid magnets are ramping up. The location of the reference pulser peak in $pulse1\_area$ fits well to an exponential in the region after the magnet finishes ramping. Therefore by monitoring the location of the reference pulser peak in the main detection PMTs and fitting to an exponential with an offset it is possible to extract the amplitude, lifetime, and y-offset associated with the gain drift. The gain fits to a 1500 ± 100 s lifetime with a fractional gain drift amplitude of $\approx -40\%$ from the moment that the solenoid is set to 70%, which corresponds to a $\approx -2\%$ gain drift amplitude at the start of data collection of the first file after the cryogen fill. This effect is much smaller than the total gain drift in a typical file, however, because it primarily affects the first file after a cryogen fill, it could potentially cause an imperfect background subtraction. Despite this, the gain drift is sufficiently small that any systematic effect is expected to be negligible.

In the majority of the data, the gain drifts due to PMT blinding and ramping the quadrupole magnets are difficult to isolate because of the high correlation between the time of the beam close and the ramping of the quadrupole magnets. The best method for isolating these mechanisms is to compare trapping and non-trapping files in static ramping data. In this case, the non-trapping files have a quadrupole ramp just before data collection, and the trapping files do not. Unfortunately, in our data, the vast majority of the static data has gain jumps, which add a layer of difficulty in extracting the desired information. The series 14 data uses a static ramping configuration and does not experience gain jumps. When we explored this data, we found that the functional form of the gain drift after correcting for the solenoid gain drift and temperature effects was better described by a linear model than an exponential model. It is not clear what is causing this functional form, and it suggests that a simple understanding of both the beam induced and the quadrupole ramp gain drifts can not be extracted in this way.

This is not a particularly satisfactory place to leave the gain drift investigation, however, it is what was reasonably achievable with the data that was gathered. Additional diagnostics data should be taken, in the future, to clearly separate out these mechanisms. Suggestions for

designing the diagnostic data are included in the apparatus chapter in the section on additional data types to take, see Section 2.14.5, which starts on page 70.

Appendix C

Neutron Energy Classifications

Table C.1 lists the energy, wavelength, temperature, and velocity for many of the common neutron energy classifications and a couple of specific energies that pertain to this experiment.

Table C.1: Typical neutron energy classifications. * The nuclear optical model potential of ^{58}Ni is a commonly accepted definition of the maximum ultracold neutron (UCN) energy, however it is not particularly useful in this experiment. ** This is the energy of the monochromatic cold neutron beam supplied to the UCN Lifetime Experiment at NIST.

Classification	Energy (neV)	Wavelength (nm)	Temperature (K)	Velocity (m/s)
	k_BT	$h/\sqrt{2m_nk_BT}$	T	$\sqrt{2k_BT/m_n}$
Ultracold	100	90.4	1.16×10^{-3}	4.37
Ultracold*	350	48.3	4.06×10^{-3}	8.18
Ultracold	<500	40.4	5.80×10^{-3}	9.78
Cold	$>1.00\times10^{6}$	0.904	1.16×10^{1}	4.37×10^{2}
Cold**	1.03×10^{6}	0.890	1.20×10^{1}	4.44×10^{2}
Thermal	2.50×10^{6}	0.572	2.90×10^{1}	6.91×10^{2}
Epithermal	3.00×10^{6}	0.522	3.48×10^{1}	7.57×10^{2}
Cadmium	$>4.0\times10^{6}$	0.452	4.64×10^{1}	8.74×10^{2}
Epicadmium	8.00×10^{6}	0.320	9.28×10^{1}	1.24×10^{3}
Slow	5.00×10^{7}	0.128	5.80×10^{2}	3.09×10^{3}

Appendix D

Magnet overshoot script

As mentioned elsewhere, to perform a cryogen fill the magnets must be ramped down to ensure the safety of the operator. This includes both the quadrupole, which is ramped every few hours during the data collection process, and the solenoid, which is only ramped for cryogen fills. The ramping of the solenoid introduces an additional magnetization gain drift into the data when cryogen fills are performed. The gain drifts due to the solenoid ramp are quite large and were observed to extend into the first data file after the cryogen fill. To combat this, a magnet overshoot script was developed to limit the portion of the solenoid gain drift that extended into the data. During the magnet overshoot script, the detection system was engaged so that the gain, the position of the reference pulser peak in the main detection channels, could be monitored during the script. The solenoid was ramped above the operating current for the data to be taken and held for 10 min. In the case of 70% data, the solenoid was ramped to 80%. Then the solenoid was ramped down to the operating current, the current data file was ended after an additional 15 min, the system waited another 5 min, and another test data file was started to observe the gains for an additional 5 min. The final data file was included separately so that the pulse shape histograms in the data acquisition program would refresh allowing the shape of the pulse shape histograms to be inspected without the broadening from periods with atypically high gain drifts during the magnet ramping.

These magnet overshoot scripts reduced the effect of the gains in a couple of ways. Ramping the solenoid above its operating current the rate at which the magnetic shielding saturates and the support infrastructure magnetizes increase allowing the gains to stabilize more quickly. The script also enforced that the magnets were given an additional ~ 25 min to stabilize at or above the solenoid operating current before the start of a production data file. This resulted in a 60% increase in the time that the gain drift could stabilize before the start of data acquisition in a production data file.

Appendix E

Purifier Procedures

This chapter consists of a series of procedures for running the purifier. Each section is a distinct procedure with its own motivation followed by the procedure. The procedures are arranged according to the expected order that they will be used with a couple of additional procedures at the end. They are meant to indicate the methodology used when running the purifier so that it can be evaluated and understood. Many of these procedures, in particular for situations where the purifier was cold, were only used a few times and as such, they are not as refined as they would have been if the purifier had been run more extensively.

E.1 Baking the Sample Bottles

Baking vacuum equipment is a commonly used method to promote the removal of gasses that have been absorbed into the surfaces of the vacuum components. Heating the surfaces while the system is under vacuum results in an accelerated outgassing from the surfaces and therefore can result in a reduced base pressure of the vacuum system. In the case of the purifier, the goal of baking out the chamber is centered around facilitating the removal of any contamination in the system more than a desire to improve the system's base pressure.

No permanent gas handling system existed for baking out sample bottles. Instead, a system was cobbled together, from scrap pieces on hand, when it was needed. The system was built out of VCR components because that is the connection style that was used on the sample bottles. This small conductance of the 1/4 in VCR lines necessitates a very small gas handling system to maximize the effective pumping speed.

1. The system should include a small turbo station, a convectron gauge, and connections for sample bottles.

2. Start evacuating the system.

3. Wrap heater tape around the bottles and tie it in place.

4. Wrap the outside of the bottles and the heater tape with aluminum foil to help contain the generated heat.

5. Energize the heater tape using a variac at a setting of 80.

6. Leave the system for an hour and then measure the temperature with the infrared thermometer.

7. Leave the system pumping overnight with the heater tape on.

8. Once the system is baked out, close the valve to the bottle, turn off the heaters, shut down the pumps, and allow the system to cool off before disassembly.

E.2 Cooling Down to 4 K

The process of cooling the purifier can be broken into three main steps, which take two days to complete. First, the vacuum of both the vacuum jacket on the extraction line and the vacuum can are verified, and the pureloop is backfilled with helium gas. Second, the purifier is cooled to 77 K using LN, and the system is left to equilibrate overnight. The following day, the system is cooled to 4 K using LHe.

E.2.1 Prepare the System for Cooling Down

1. Pump out the vacuum jacket on the dewar with the leak detector. For reference, it takes ≈ 66 min with the leak detector to reach a pressure of 9.5×10^{-3} Torr.

2. Pump out the vacuum jacket on the extraction line with the leak detector. For reference, it takes ≈ 40 min with the leak detector to reach a pressure of 9.5×10^{-3} Torr.

3. Pump out the vacuum can with the leak detector to limit the amount of sublimation in the vacuum can. It takes $\approx 1.5 \times 10^{4}$ s and with the leak detector to reach a pressure of 1.4×10^{-2} Torr.

4. Backfill the pureloop and the 1K Pot with helium gas to prevent any ice from forming in the pureloop during the LN fill. This could cause blockages or impedances in the pureloop or, if ice gets into a needle valve assembly, could prevent it from sealing properly.

   (a) Pump out the pureloop and 1K Pot with the roughing pump. It takes $\approx 2.2 \times 10^{3}$ s with the roughing pump to reach a pressure of 3.8 Torr.

   (b) Backfill the pureloop and 1K Pot with helium gas to atmospheric pressure.

E.2.2 Cool to LN

1. Attach a helium bottle to the pureloop and overpressurize it with helium gas. This is to ensure no nitrogen gets into the purifier during the LN fill. Use a flow meter on the gas bottle to verify that there is a constant flow of helium gas from the pureloop into the main helium bath during the LN fill. When the main helium bath is overpressurized, during the fill, the flow meter will require adjustment to maintain a constant flow.

2. Backfill the vacuum can with 200 mTorr of helium gas, as measured by the convectron, to act as an exchange gas. This will result in a pressure of ≈ 100 mTorr at 77 K; recall that the actual pressure of the helium gas is higher than the value reported by the convectron.

   (a) Fill the vacuum can with more than atmospheric pressure of helium gas.

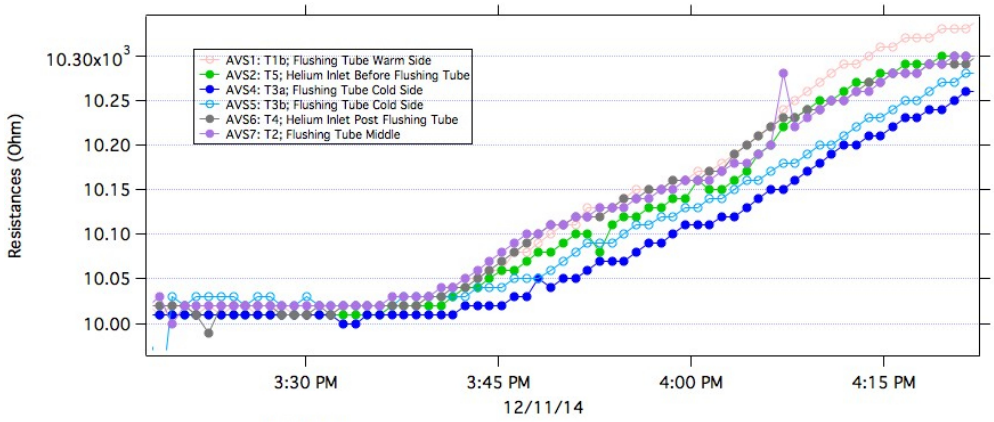
   (b) Briefly, vent the vacuum can to atmosphere to reduce the pressure to atmospheric pressure.

   (c) Evacuate the vacuum can until the convectron reads 200 mTorr when the vacuum can is not actively being pumped on.

3. Start the LN fill. Adjust the flow meter and regulator to maintain a positive flow from the pureloop into the main helium bath. Previously, it has taken $\approx 9 \times 10^2$ s for the temperature in the pureloop to start to respond.

4. Once the thermometers start to respond the fill is very close to being completed, and the level should be measured.

5. Below is an account of one of our LN fills. The conclusions that I come to from this is that the fill should be stopped ≈ 2 min after the temperatures in the pureloop start to respond.



During the cooldown shown above, the fill was started at 3:18 PM. At about 3:45 the resistances read ≈ 10.04 kΩ, the fill was stopped, and the level was measured to be about 1 in. The system was allowed to sit to see if the temperatures would stabilize. When they had not recovered by 4:00, the fill was started again until 4:05 when the level was measured to be about a foot.

These numbers varied over the few fills that were performed. It is thought that the variations were due to changes in the flow rate from the transfer dewar. As such, the numbers here should be taken as rough estimates.

E.2.3 Cool to LHe

1. Siphon any remaining LN out of the main helium bath by inserting a small tube that can be sealed at the purifier top plate into one of the fill ports and then pressurizing the main helium bath with helium gas via the main helium bath exhaust port. This will force the LN out of the main helium bath through the small tube. During the siphoning, attach the same gas bottle to both the main helium bath exhaust and the pureloop to prevent a pressure differential that could force nitrogen gas into the pureloop.

   (a) To prepare for siphoning put the purifier in the following state.

      i. Close V37 so that the main vacuum line is not being pumped on.

      ii. Verify that V32 and V33 are closed to isolate the vacuum can, which should contain ≈ 100 mTorr of helium exchange gas.

      iii. Close NV1 and NV4 to limit flow through the helium inlets between the main helium bath and both the pureloop and the 1K Pot.

      iv. Close V10 to isolate the UGHS from the PGHS.

      v. Open the pureloop and 1K Pot to the main vacuum line.

         A. Close V40 and open V41 to connect the 1K Pot to the main vacuum line.

B. Open V34 to connect the helium inlet pumpout to the main vacuum line.

C. Open NV2 to connect the 1K Pot to the flushing tube.

D. Open NV3 to connect the extraction line to the flushing tube.

E. Open V1, V2, and V3 to connect the extraction line to the main vacuum line.

(b) Put a tee in the rubber hose running from the flow meter on the helium cylinder. Attach one of the lines to the main helium bath exhaust and the other to V36.

(c) Include a second tee with a valve to atmosphere on the helium bath exhaust to allow the pressure to be released in a controlled manner to stop the siphoning process.

(d) Using a small OD SS tube of at least 2 m in length and an assembly of rubber stoppers and quick connect fittings, insert the tube into a fill port in the main helium bath and then seal the tube so that no gas can flow around the outside of it.

(e) Using the settings on the flow meter and helium regulator to pressurize the main helium bath with helium gas until a small steady stream of LN begins to be siphoned out of the top of the small tube.

(f) When LN stops coming out of the tube, the siphoning process is complete. Turn off the gas flow and vent the main helium bath using the valve to atmosphere on the main helium bath exhaust.

(g) Remove the small tube and put the stopper back in the fill line.

(h) Remove any impedances from the main helium bath exhaust.

(i) Install an exhaust tube on the main helium bath exhaust port to redirect the helium vapor away from the other components on the purifier top plate to help prevent freezing of o-ring seals or electronics. Typically a 3 m section of KF25 bellowed tubing was used.

(j) Remove the tee on the helium line and return the system to its original state with a rubber hose going from the flow meter on the helium bottle to the connector on V36. This setup is required to allow the pureloop to be pressurized for the first couple of minutes of the LHe transfer while there is still some gaseous and LN in the system, however a connection from the helium gas cylinder to the main helium bath exhaust is no longer needed.

2. Put the stinger in and prepare for a helium transfer.

3. The state of the purifier should be as follows:

(a) As mentioned previously, during the LHe fill, the pureloop and 1K Pot should be overpressurized with helium gas to prevent nitrogen from getting into the pureloop during the start of the LHe fill.

i. NV3, V1, V2, and V3 are open to connect the pureloop to the main vacuum line.

ii. V40 is closed and V41 is open connecting the 1K Pot to the main vacuum line.

iii. NV2 is open to connect the 1K Pot and the flushing tube.

iv. V31 and V34 are open to connect the auxiliary pumpout lines for the pureloop to the main vacuum line.

v. V37 is closed to prevent the main vacuum line from being pumped on.

vi. V36 is open to allow a small flow of helium gas into the the pureloop and 1K Pot through the main vacuum line.

(b) V32 and V33 are closed isolating the vacuum can, with about 100 mTorr of helium exchange gas.

4. Turn on the sorb heater to 1 V to prevent helium gas from condensing on the absorption pump while there is still exchange gas in the vacuum can. If, during the fill, the temperatures appear to turn over before 4.2 K the heater power should be turned down. In particular, T3 being higher than the other thermometers suggests that the sorb heater is supplying enough power that it is preventing the purifier from cooling to 4.2 K and therefore the heater power should be reduced.

5. Measure and record the level in the helium transfer dewar before starting the LHe fill.

6. Start pumping out the vacuum can. Immediately start the LHe transfer. The hope is that this will result in there being enough exchange gas in the vacuum can for the start of the fill, but it will be mostly pumped out as the temperatures approach 4.2 K.

7. Adjust the helium cylinder regulator to maintain a positive flow from the pureloop into the main helium bath.

8. After a minute or two, turn on the LHe level monitors. If they read 100% make sure that the signal cable is properly plugged into the back of the level monitor.

9. Once the pressure in the transfer dewar falls below 1 PSIG, close V36 and move the line from the helium cylinder from the helium dewar to the back pressure port on the helium transfer dewar. Adjust the helium regulator to maintain 3 PSIG in the transfer dewar. At this point, all but trace amounts of nitrogen should be diluted out of the main helium

bath and any remaining nitrogen should have solidified already. Therefore removing the gas flow through the pureloop should not allow any nitrogen into the purifier.

10. When the main helium bath is full, end the cryogen fill.

11. While the temperatures stabilize, measure and record the level in the transfer dewar and prepare for operation.

12. Continue pumping on the vacuum can with the sorb heater running, for as long as possible, to remove as much of the exchange gas as possible.

13. Start up the GSX pump, which takes about an hour to warm up.

14. At this point, the purifier is full of helium and ready to run.

E.3 Starting the Flush

This procedure is used to start the flushing gradient in the flushing tube. It is designed to reliably put the purifier in a well-behaved state.

1. The current state of the system should be as follows.

   (a) The main helium bath should be full of LHe.

   (b) V33 should be open, V32 should be closed, and the leak detector should be attached to V33 so that the vacuum can is being pumped on by the leak detector.

   (c) V37 should be open and attached to XDS-37i to evacuate the main vacuum line.

   (d) V40 and V41 are closed and the GSX should be running so that the GSX is at full speed but isolated from the rest of the system.

   (e) All heaters should be off.

   (f) Variac that powers the heater tape that is used to prevent freezing of the purifier top plate should be set to 60%

   (g) V10 is closed isolating the PGHS and the UGHS.

   (h) V1 should be open and both V2 and V3 should be closed isolating the pure extraction line from the main vacuum line and setting up the system to use the evaporator.

   (i) V31, V34, and V39 should be closed to isolate the auxiliary pump out ports.

   (j) V35 and V36 should be closed to isolate the auxiliary ports on the main vacuum line.

(k) NV1 and NV4 should be open allowing LHe to enter the pureloop through the fill ports.

(l) NV2 and NV3 are closed to isolate the flushing tube from both the 1K Pot and the extraction line.

2. Cool the pureloop from 4.2 K to 1.4 K.

   (a) Slowly open V40 all the way. The mechanical booster (MB) on the GSX should spin down to 15% to 20%, but it should recover after ≈ 0.5 hr. As it recovers, the temperature should come down to ≈ 1.4 K.

   (b) Close NV4 once the level comes up to about 50% in the 1K Pot. Try to maintain that level, by adjusting NV4. It is possible that even if NV4 is closed all the way the level will continue increasing. In this case, leave the valve closed.

   (c) When the temperature reaches 1.4 K, turn down the MB to 45 Hz using the menus on the handheld control, see Section E.8, which starts on page 251 for a description of how to do this.

   (d) The temperature should come up a little bit to ≈ 1.6 K. Wait for the system to become stable. It should take less than 10 min. As a reference point, on 1/4/2015 the system warmed from 1.4 K to 1.5 K over about 400 s.

   (e) Close V33 to see if the pressure decreases in the vacuum can. If it decreases, leave V33 closed otherwise open it back up. The pressure decreasing is a sign that cryopumping is currently outperforming the vacuum pump and should be allowed to take over. The pressure in the vacuum can should be monitored periodically and V33 should be opened if there is any sign of the pressure coming up in the vacuum can.

3. Turn on the flushing heater to 10 V in order to establish the flushing gradient. All of the temperatures should come up a little bit more, in particular, T5 should come up to 4.2 K and then remain stable. Fluctuations in the temperature of T5 seem to be an indication of instability in the purifier and may be a precursor to the 1K Pot crashing. If T5 is fluctuating, turn down the heater power until T5 becomes stable to prevent a crash. The system is ready to be cleaned once the heater is on and the purifier has been stable for 10 min.

4. Clean ^{3}He from the PGHS as discussed in Section E.4.1, which starts on page 246.

5. The purifier is ready to run.

E.4 Cleaning the System of ^{3}He Contamination

There are a couple of different portions of the gas handling system that could require cleaning. The general process is similar for all of them. The goal is to backfill the contaminated section with ultrapure from the extraction line and then evacuate it to dilute any contamination. The dilution factor of any contamination can be estimated by measuring the maximum and minimum pressure during this process. The cleaning process is complicated slightly because the purifier must remain stable during the entire process or there is a risk of allowing ^{3}He through the purifier.

The stability of the thermometry is the primary method for determining the stability of the flushing gradient. It is assumed that minimizing thermal fluctuations will reduce the chance of ^{3}He making it through the flushing tube. The temperatures throughout the pureloop vary if the pressure on the extraction line is not constant in time. A pressure buildup occurs if V1 is closed isolating the extraction line. Therefore, the procedure has been modified to stop extraction at V10 instead of V1 and once V10 is closed immediately open V3 allowing the ultrapure being produced to be thrown out by the main vacuum line. V3 is then throttled to maintain an approximately constant pressure in the extraction line.

The following section discusses the process of cleaning the PGHS. This is followed by a procedure for cleaning the UGHS, which can be easily extended to any other portion of the gas handling system by connecting the desired portion of the gas handling system to the UGHS.

E.4.1 Cleaning the PGHS

1. The flushing gradient should already be set up, see Section E.3, which starts on page 244.
2. V10 should be closed to isolate the PGHS from the UGHS.
3. V1, V2, and V3 should be closed to isolate the extraction line from the main vacuum line.
4. The main vacuum line is going to be used to throw out ultrapure product from the purifier; therefore, everything must be isolated from the main vacuum line. This is done by closing V31, V32, V34, V35, V36, V38, V39, and V41.
5. Open V37 to pump out the main vacuum line.
6. Open V1 and V3 to start pumping on the extraction line.
7. The pressure on the PGHS should read in the range -30 inHg to -27 inHg, and the temperatures may come down a little bit.

8. To calculate the dilution factor, the amount of gas being thrown out must be measured. Therefore, instead of pumping on the extraction line for an arbitrary amount of time, the PGHS is backfilled with ultrapure product and then evacuated. This process of flushing the PGHS with the ultrapure product is performed twice. By calculating the maximum and minimum pressure during each flush, the dilution factor can be calculated.
   (a) Close V37 and let the pressure come up.
   (b) Once the pressure comes up to -10 inHg close V1. Let the pressure stabilize and then record the pressure.
   (c) Open V37 to evacuate the system.
   (d) Once the pressure reaches $\leq 4 \times 10^1$ mTorr on the convectron close V37.
   (e) Record the pressure once it stabilizes.
   (f) Repeat this process of backfilling and then pumping out the system one more time.
9. With V37 open, open V1 to start pumping on the extraction line.
10. Throttle V3 to maintain a pressure of roughly -11 inHg in the PGHS.
11. Let the system pump like this for 15 min to make sure that the system is stable while slowly throwing out helium.
12. The system is now clean.

E.4.2 Cleaning Sections of the gas handling system

1. The PGHS should have just been cleaned. The system should be in the state at the end of that procedure.
2. Isolate any portions of the gas handling system that do not need to be cleaned and will not be used. For example, close V20 to isolate the downstairs, V13a-d to isolate the dumps, and V50 to isolate the SGHS.
3. Open V12 to pump on the UGHS with a turbo pump.
4. Continue until the pressure reads $\leq 4 \times 10^1$ mTorr on the convectron, then close V12 and record the pressure once it stabilizes.
5. Close V3 and then immediately open V10. The immediacy is important because the volume of the tubing between V3 and V10 is sufficiently small that the pressure comes up pretty fast. Opening V10 connects the extraction line to a much larger volume and prevents a spike in the pressure, which could affect the flushing process.

6. Once the pressure reaches -10 inHg, which should happen quickly if both V20 and V13a-d are closed, close V10 and then immediately throttle V3 to maintain a pressure of -11 inHg in the extraction line.

7. Let the pressure stabilize in the UGHS and then record the pressure to calculate the dilution factor.

8. Open V12 to start pumping the system out.

9. Once the pressure is $\leq 4 \times 10^1$ mTorr on the convectron, close V12, let the system equilibrate, and then record the pressure.

10. Backfill and evacuate the system one more time.

11. The system should be clean.

E.5 Attaching a Sample Bottle to the SGHS.

Connecting sample bottles to the SGHS allows some atmosphere into the system at the location of the connection. This gas will contain ^{3}He and therefore must be pumped out and purged with ultrapure to limit the contamination to the system. The sample bottles are connected by a long section of flexible tubing with a VCR fitting on either end. The flexible tubing allows the sample bottle to de dunked in LN when the sample bottles are being filled to increase the amount of gas stored. With this arrangement, the volume of the flexible tubing is the amount of atmosphere that is allowed into the gas handling system. Below is a set of procedures for purging the flexible tubing and the SGHS after attaching a bottle. This procedure is designed to isolate as much of the gas handling system as possible and to limit the amount of atmosphere admitted to the gas handling system. A substantial improvement to this method is suggested in the suggested improvements at the end of Section 6.4, which starts on page 213.

1. The sample bottles should be baked out and evacuated as described in Section E.1, which starts on page 238.

2. The UGHS should be cleaned and evacuated, see Section E.4.2, which starts on page 247.

3. Isolate the portions of the gas handling system that are not required.

   (a) V13a-d should all be closed isolating the dumps from the UGHS.

   (b) V20 should be closed to isolate DGHS from the UGHS.

4. V10 should be closed to isolate the PGHS from the UGHS.

5. Ensure that the UGHS is completely pumped out.
   (a) Close the valve to the fill line that will be used, V50.
   (b) Open the pneumatic valve.
   (c) Put the small turbo station on the pumpout port attached to V12 and, once it spins up, open V12.
6. Attach the sample bottle to the SGHS and remove the atmosphere that entered the system.
   (a) Attach the sample bottle to the VCR connection on V52. Leave the bottle closed so that it is not exposed to the SGHS, which is currently filled with atmosphere.
   (b) Open V51 to evacuate the SGHS with the small roughing pump.
   (c) Once the pressure is $\leq 4 \times 10^1$ mTorr, close V51 to isolate the roughing pump.
   (d) Open V50 to allow the remaining atmosphere to be pumped out by the small turbo station.
   (e) Proceed to the next step once the pressure on the UGHS convectron is $\leq 4 \times 10^1$ mTorr.
7. Clean the SGHS of any ^{3}He contamination by flushing it twice following the procedure Section E.4.2, which starts on page 247 with V50 open.
8. Clean the sample bottle of any contamination, ^{3}He or otherwise, by flushing it once.
   (a) Note: After the SGHS has been cleaned, the sample bottle should be cleaned as a safety precaution. The sample bottles have not been exposed to concentrated helium before they are used, so cleaning should not be required.
   (b) Pump on the UGHS until the SGHS convectron gauge is $\leq 4 \times 10^1$ mTorr.
   (c) Close V50 to isolate the SGHS. Record the pressure on the SGHS convectron gauge, open V52, and record the pressure again. If there is evidence that the vacuum in the bottle was spoiled, the sample bottle should not be used.
   (d) Open V50 and leave the system pumping until the pressure is $\leq 4 \times 10^1$ mTorr.
   (e) Once the pressure in the bottle is $\leq 4 \times 10^1$ mTorr, close V12, wait for the pressure to stabilize, and then record the value.
   (f) Use the procedure for cleaning sections of the gas handling system, Section E.4.2, which starts on page 247, to purge the UGHS and the SGHS including the sample bottle by leaving V52 open.

(g) Evacuate the bottle until the pressure is $\leq 4 \times 10^1$ mTorr.

(h) The sample bottle is now cleaned and ready for use.

E.6 Collecting Ultrapure Samples

This procedure describes the process of collecting a sample of the ultrapure product from dump 1 and putting it into a sample bottle.

1. Before beginning, any portions of the UGHS and SGHS that could be contaminated should be cleaned. The sample bottle should have been attached and cleaned. The procedures for completing these tasks were included previously.

2. The state of the system should be verified.

   (a) V13a-d are closed to isolate the dumps.

   (b) V10 is closed to isolate the purifier.

   (c) V20 is closed to isolate the downstairs gas handling system.

   (d) V50 is closed to isolate the SGHS.

   (e) V52 is closed to isolate the sample bottle.

3. Open V51 to start evacuating the SGHS.

4. Open V12 to start evacuating the UGHS with a turbo pump.

5. Continue pumping on both volumes until pressure in each volume is $\leq 4 \times 10^1$ mTorr on their corresponding convectron gauges.

6. Once the pressure is $\leq 4 \times 10^1$ mTorr in the SGHS, open V52 to attach the sample bottle to the rest of the SGHS.

7. Once the pressure is $\leq 4 \times 10^1$ mTorr in the SGHS, close V51, and open V50 to switch the SGHS from the roughing pump to the small turbo station on the PGHS.

8. Once the pressure is $\leq 4 \times 10^1$ mTorr in both parts of the gas handling system, close V12 to stop evacuating the gas handling system.

9. Record the pressure.

10. Open V13a to allow the isotopically pure ^{4}He from dump 1 into the sample bottle.

11. Record the pressure.

12. Dunk the sample bottle into LN. It should be submerged all the way to the valve. Let the bottle sit for 10 min then record the pressure again.
13. Close V52 to isolate the sample bottle.
14. Close V50 to isolate the SGHS from the UGHS.
15. Let the bottle warm up, then remove it from the flexible tubing.
16. Put a blank off into the VCR connector on the bottle.
17. Label the bottle.
18. If the purifier is still running, this is a convenient time to attach a new sample bottle.

E.7 Emergency Shutdown

This procedure describes the steps that should be taken in the case of an emergency in which the purifier must be left in the safest possible state as quickly as possible. It is designed to be the shortest number of steps required to leave the purifier in a safe state. It should take less than a minute to perform.

1. Close V10 to isolate the UGHS from the PGHS.
2. Turn off the heaters
3. Verify that V37 is open and that the roughing pump is running.
4. Open V1 and V3 to connect the flushing tube to the over pressure valve on the main vacuum line.
5. Close V40 and open V41 to connect 1K Pot to the main vacuum line.

E.8 Changing the Mechanical Booster Speed on the GSX Pump

When the GSX was allowed to pump on the 1K Pot at full speed with V50 open all the way, a periodic variation in the load on the pump occurred. This effect has been described in the purifier chapter, see Section 6.3.3, which starts on page 211. Decreasing the maximum speed of the mechanical booster (MB) on the GSX pump was found to prevent this periodic effect from occurring. Therefore, I have included instructions on varying the speed of the mechanical booster below. The default password for interacting with the GSX is 202.

1. Note: Change the MB speed on the GSX requires the PDT, at least to the best of my knowledge.
2. Hit the Setup button on the PDT.
3. Navigate through “Set Sequences” > “Speed Control” > “Set 2nd MB Speed” and change the value to 45 Hz for 44% or 101 Hz for 100%. Hit “Enter” to record the value.
4. Hit the “Setup” button on the PDT.
5. Navigate through “Command Menu” > “2nd MB Speed” then select “On” and hit “Enter”.
6. At this point, the MB will attempt to spin to the set 2nd speed value.