Recent Measurement of the Flavor Asymmetry of the Light Proton Sea

by

Daniel Morton

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Physics)
in The University of Michigan
2019

Doctoral Committee:

Professor Wolfgang Lorenzon, Chair
Professor Christine Aidala
Professor Ratindranath Akhoury
Associate Professor Alexander Thomas

Daniel Morton

dhmorton@umich.edu

ORCID iD:0000-0003-3813-1375

Dedication

This dissertation is dedicated to my supportive family and my loving wife in particular. She encouraged and tolerated me throughout my entire graduate program, and I wouldn't be where I am today without her.

Table of Contents

List of Tables

Tables

List of Figures

Figures

List of Appendices

Abstract

Despite the success of perturbative Quantum Chromodynamics in particle physics, it is unable to predict the underlying structure of the nucleon. The observed violation of the Gottfried Sum Rule underscores the failings of pQCD in the low energy regime. The quantity $\bar{d}-\bar{u}$ was measured to be anomalously large with a strong dependence on Bjorken-x. Subsequent studies lead to the development of theoretical non-perturbative models to describe proton structure. The E866 experiment provided Drell-Yan measurements of the flavor asymmetry ratio $\bar{d}/\bar{u}$ as a function of x, but also observed an unexpected drop below unity at large x that suggests $\bar{u} > \bar{d}$.

E906 SeaQuest was proposed to extend the $\bar{d}/\bar{u}$ measurements in x above 0.35 for the first time. The E906 spectrometer was upgraded from the E866 spectrometer to improve sensitivity beyond E866's in the $x = 0.25$ to 0.35 range. Towards this end, SeaQuest utilized Fermilab's 120 GeV Main Injector proton beam, which changes the four momentum-transfer squared from E866's $Q^2 = 54\ GeV^2$ to $Q^2 = 29\ GeV^2$. The beam impinges on targets consisting of liquid hydrogen, deuterium, and solid nuclear targets. Dimuon pairs produced by the beam proton collisions are measured by four detector stations. The Drell-Yan cross-section ratio of the hydrogen and deuterium targets (σ_{pd}/σ_{pp}) was derived by comparing the relative yields. Background species counts far outnumbered the Drell-Yan counts. Boosted decision tree techniques improved the quality of the purity of the yields through the classification of events. Measuring the dimuon kinematics and extrapolating to zero intensity allowed σ_{pd}/σ_{pp} to be derived as a function of x. The $\bar{d}/\bar{u}(x)$ ratio was then extracted from the cross-section ratio $\sigma_{pd}/\sigma_{pp}(x)$.

The measured $\bar{d}/\bar{u}(x)$ ratio agree with that from $E866$ until about $x = 0.23$, after which the two data sets diverge. The ratio never approached unity and instead appeared to stabilize around 1.6, in agreement with some non-perturbative models. These results will likely motivate new PDF global fits and can be used in conjunction with other data to narrow down the origin of the asymmetry in the nucleon sea.

Chapter 1

Introduction and Background

Our human world is made up of atoms, a word meaning ”uncuttable” in ancient Greek. Greek philosophers surmised the existence of a discrete unit making up all things and giving matter its properties. This concept matured into the modern atom. Yet, these atoms are not truly fundamental. They instead consist of a nucleus orbited by electrons. Electrons are fundamental particles, as they are both indivisible and identical. Ergo, they possess discrete mass, charge, and spin. After proving the existence of a concentrated, massive nucleus, Rutherford proposed the proton as a particle that had an equal but opposite charge to the electron, but with much greater mass[1]. These protons were confined in extreme proximity in the nucleus, despite their electromagnetic repulsion. Overcoming their mutual repulsion necessitates the existence of a much stronger force to bind them, governed by its own distinct functional form. This force was pragmatically labeled the strong force. The nature of the strong force and nucleons remained inaccessible for some time, leading to the fallacious assumption that the proton was a fundamental particle like the electron[1].

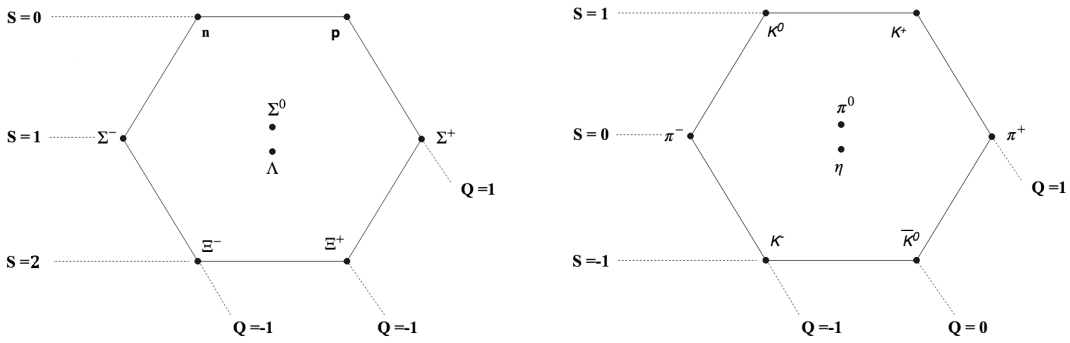
Once scientists were capable of probing the strong force particles, some paradoxically heavier than the parent protons. It was not until the first Deep Inelastic Scattering (DIS) experiments, physics could be disabused of the fundamental ”particle zoo” notion and evolve into the familiar quarks, leptons, and forces that we know today. Furthermore, these nucleons would be discovered to contain an underlying structure. A particular anomaly, the seaquark flavor asymmetry, serves as an instrument by which to explore nucleon structure at low energies. Hopefully, new insight into nucleon structure can help to shape the under-

standing of this fundamental force responsible for most of the mass in our atomic world.

1.1 Proton Structure

1.1.1 Quark and Parton Models

Nucleon structure begins with the first characterization of the strong force. Yukawa determined a functional form of nuclear force, mediated by the exchange of light mesons. After the discovery of the aforementioned heavy baryons, questions arose that could not be answered by the current understanding of the strong force. The constituent quarks were effectively hidden by confinement, as they could not be isolated in a measurement. Instead, allowed decays and interactions were described through a series of conservation laws. After the addition of the "strangeness" property, the baryon octet, decuplet, and meson octet were formed as shown in figure 1.1[2]. Gell-Mann augmented the strong force by implementing SU(3) symmetry, pairing off particles and anti-particles correctly. This symmetry successfully predicted the later discovered Ω^- particle[2, 3].



(a) Illustration of the Baryon Octet.

(b) Illustration of the Meson Octet.

Figure 1.1: Eight Fold Way Particle Organization: a) is the baryon octet and b) is the meson octet. The rows are strangeness on the left, charge on the right, and isospin on the x-axis in a), with neutrons and protons at -1/2 and 1/2 respectively.

Gell-Mann's "Eight Fold Way" laid the groundwork for what would become known as the constituent quark model. In 1964, 3 flavors of spin 1/2 quarks with charge $\pm\frac{2}{3}$ and $\pm\frac{1}{3}$ were predicted[4]. Many believed that this formalism was only a mathematical construction

until a series of MIT-SLAC DIS experiments confirmed quarks as a physical property. DIS can resolve distances smaller than the proton size using high energy electrons capable of passing through a proton. By measuring the electron outgoing distributions, individual quarks become observable. DIS experiments can also give information on the momentum fraction of a quark by measuring the change in momentum of the scattering lepton.

In 1967 SLAC found that the DIS pattern was consistent with a proton internal structure consisting of 3 point particles[5]. The conclusion was that protons consisted of at least 3"partons", known later as the valence quarks: up-up-down for the proton and down-down-up for the neutron. Both quark sets appear to naively add to a total nucleon spin of 1/2. This would lead Feynman to construct the parton model to describe scattering observations[1].

The electron had a probability to interact with these partons individually instead of the entire proton. Feynman employed the hadronic property called Bjorken scaling to simplify the parton descriptions, effectively making them primarily depend on relative kinematics. Therefore, the parton model implements the use of a new variable, Bjorken x. Bjorken x is the momentum fraction which the parton carries, given by [5]

$$x = \frac{Q^2}{2p_i\dot{q}}, \tag{1.1}$$

where Q^2 is the interaction energy scale, p_i is initial state momentum and q is the virtual photon momentum, and $p_i\dot{q}$ is equivalent to the final state hadron mass multiplied by the virtual photon energy.

Bjorken constructed this variable after having observed that nucleon scattering cross-sections did not depend significantly on energy scale, in sharp contrast to elastic scattering results. This implies that at high energy, hadrons behave as point-like constituents.

Feynman extended the concept to his partons, effectively incorporating a phenomenological description of asymptotic freedom[1]. At high energies, the individual partons are measurable. The probability distribution resulting from these measurements is the parton

distribution function (PDF) of that parton. The PDFs are key to making predictions of products in particle physics and making experimental determinations. The PDFs represent one of the first insights into the underlying structure of the nucleon.

1.1.2 Quantum Chromo Dynamics

The quark model still had theoretical problems explaining the Δ^{++} and later the Ω^{-} particles due to the Pauli Exclusion principle, both having 3 of the same quarks. Greenberg explained this utilizing a "color" charge of red, blue, and green (r, g, b) and their anti-charges $(\bar{r}, \bar{g}, \bar{b})$ to govern the strong force[6]. The colors act as the strong force analogue to electric charge with $r + g + b$ being colorless. This color force is mediated by massless vector bosons called gluons. Using these elements, a non-Abelian gauge invariant quantum field theory can be constructed. The differences from electromagnetic force give rise to the SU(3) symmetric Lagrangian

$$\mathcal{L}_{\mathrm{QCD}} = \bar{\psi}_i \left(i(\gamma^\mu D_\mu)_{ij} - m\,\delta_{ij} \right) \psi_j - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu}_a. \tag{1.2}$$

The first term represents the fermion interactions $D_\mu = (\partial_\mu - i\sqrt{\pi\alpha_s}\lambda_a A^a_\mu)$ with quark fields ψ, and the second term the gluon force field tensors where $F^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - \sqrt{4\pi\alpha} f^{abc} G^\mu_b G^\nu_c$. The associated SU(3) gauge group generators are the 8 matrices represented by λ_a and are associated with the color charge combinatorics, with $[\lambda_a, \lambda_b] = i f_{a,b,c} \lambda^c$. This formalism became known as Quantum Chromo Dynamics (QCD), the fundamental quantum field theory utilizing SU(3) color charge to describe strong force interactions between quarks and gluons. Deconstructing the first term, it becomes clear that color, associated with the a, b, c indices in Eq. 1.2, is not bound to the quark but can be carried by gluons as well, in contrast to charge and photons in QED. The capacity to carry color charge infers gluons are capable of self-interaction. This is found in the expanded second term which contains a gluon field interaction term, resulting in this property entirely unique

to this force. The tensor index also allows gluons to have up to a four-field vertex versus the maximum three in QED.

Furthermore, the Lagrangian in Eq.1.2 has no flavor dependence and therefore all quarks also couple through only color, following Gell-Mann's prediction. Simultaneously, this first term also implies that colors are confined to the quarks, meaning that final states must be color neutral while individual quarks may carry any color. The confinement property carries through to the coupling constant calculation

$$\alpha_Q(Q^2) = \frac{12\pi}{(33 - 2n_f)Ln(Q^2/\Lambda^2)}, \tag{1.3}$$

where n_f is the degrees of freedom and Λ is the QCD scale. The coupling constant diverges as Q^2 or energy approaches zero and over large distances. The requisite energy to separate a quark would be greater than the energy to produce a quark-antiquark pair out of the vacuum. Instead, the energy used to separate quarks would generate a baryon-meson pair, thus maintaining the free-particle color neutrality, while making isolated quarks impossible.

The other implication of Eq. 1.3 is that the coupling becomes weaker at short distances or energies uniquely (thus "freeing" the particle)[7]. This coupling relationship produces a natural theoretical basis for asymptotic freedom, where constituent quarks can be treated like free particles at high Q^2. Asymptotic freedom lends itself to a method of calculating physical quantities directly through perturbation. In the "Bjorken" limit and small α_s, one can now use perturbation theory on quarks and gluons, adding Feynman diagrams as can be done in QED. This method is called perturbative Quantum Chromo Dynamics (pQCD).

1.1.3 Valence and SeaQuarks

Perturbative QCD had success in predicting the outcomes of proceeding experiments, especially when coupled with parton distribution functions at high x defined by measured constants from previous measurements. Comparing earlier findings and QCD, the simplis-

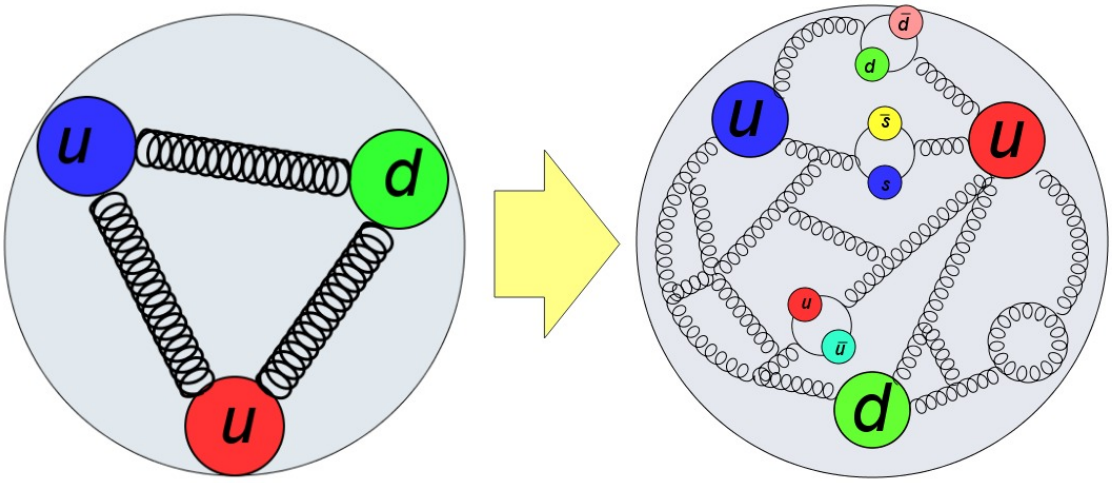


Figure 1.2: The left cartoon represents early proton structure, which is valid at high Q and high x. This model becomes the right cartoon with the addition of the quark-gluon sea. The springs represent gluons.

tic valence quark picture of nucleons illustrated on the left in figure 1.2, becomes a high Q approximation, appearing as just an electromagnetic distribution in the Bjorken energy limit. Now, the interior of the proton takes on a more fluid form at other energies, containing interactions between constituent quarks mediated by gluons carrying color charge. These gluons may interact and produce $q\bar{q}$ pairs capable of interacting with themselves and other quarks[5].

Given the low mass of the valence quarks compared to the proton rest mass, most of the energy in the system must be in the binding energy. The high binding energy allows for an extremely dynamic proton "sea" rather than three simply bound quarks. The quark "sea" consists of $u, d, \bar{u}, \bar{d}, c, s, \bar{c}, \bar{s}$ quarks and gluons constantly in flux, effervescing from the nucleon volume. When averaged over there still must be the original three valence quarks: up, up, down, but any measurement must also take into account these temporary seaquarks and gluons, all known as partons. As a result, physicists must treat the system using PDFs, or as a probability of interacting with a quark of a given momentum fraction x at energy scale Q. pQCD is not capable of resolving the low-x regime of PDFs on its own due to the coupling becoming significant. Instead, PDFs must rely on measurements to determine

their functional forms. However, pQCD has still been utilized to help determine the Q^2 evolution of PDFs and be applied in conjunction with the PDFs to scattering solutions.

First, the Factorization Theorem separates a cross-section or decay into the long distance and short distance interactions [8]. These functions are then composed in terms of the non-perturbative PDFs in scattering and perturbative diagrammatic components. Next, the DGLAP (Dokshitzer–Gribov–Lipatov–Altarelli–Parisi) evolution equation bridges the perturbative and non-perturbative energy ranges by describing how the PDFs themselves vary with energy [9].

1.2 Flavor Asymmetry

Implementing tools from pQCD and the PDFs, more information about nucleon structure can be extracted from experiments.

1.2.1 Structure Functions and Sum Rules

Yet again, DIS proved to be one of the most important probes into this complex underlying structure of the nucleon, with a scattering cross-section of

$$\frac{d^2\sigma}{d\Omega dE'} = \frac{\alpha^2}{4E'^2 sin^4(\theta/2)}\left[\frac{1}{(\nu)}F_2(x,Q^2)cos^2(\theta/2) + \frac{1}{M}\right], \tag{1.4}$$

using initial proton mass M, initial/final energy E/E' and $\nu = E - E'$. This relates to PDFs directly through the structure functions [10]

$$F_1 = \frac{1}{2}\sum_{i=u,d,s,c} e_i^2\big(q_i(x) + \bar{q}_i(x)\big), \tag{1.5}$$

$$F_2 = \sum_{i=u,d,s,c} x e_i^2\big(q_i(x) + \bar{q}_i(x)\big). \tag{1.6}$$

Eqs. 1.5 and 1.6 exhibit seaquarks and gluons contributing directly to the cross-section in

Eq. 1.4.

As the namesake suggests, the structure functions F_1 and F_2, relate scattering results directly to the PDFs and nucleon structure. Hadron structure properties can thus be expressed via physically verifiable sum rules contingent on these functions. In the case of the neutrino structure functions W, at constant ν/Q^2, the limit

$$\lim_{Q^2->\infty} MW_1(Q^2,\nu) = F_1(-x), \nu W_2(Q^2,\nu) = F_2(-x) \tag{1.7}$$

can be used to derive the Alder Sum Rule,

$$\int_0^\infty \left(W_2^{\bar{\nu}}(Q^2,\nu) - W_2^{\nu}(Q^2,\nu)\right) \simeq \int_0^1 \left(F_2(x)^{\bar{\nu}} - F_2(x)^{\nu}\right)\frac{dx}{x} \tag{1.8}$$

$$\simeq \int_0^1 \left(u(x) + \bar{d}(x) + \bar{s}(x) - d(x) - \bar{u}(x) - \bar{c}(x)\right)\frac{dx}{x} \tag{1.9}$$

$$\simeq (N_u - N_d) \simeq 2, \tag{1.10}$$

effectively stating that the number of up quarks exceeds the down quark number by two in the proton[11]. This result closely follows what pQCD would predict in this limit.

Sum Rules may also represent an integral deviation from a naive QCD prediction. The Gross-Llewellyn-Smith sum rule compares the quark and anti-quark PDFs through the structure functions

$$\frac{1}{2}\int_0^1 \left(F_3^{\bar{\nu}}(x) - F_3^{\nu}(x)\right)\frac{dx}{x} \simeq \int_0^1 \sum_{q=u,d,s,c} q(x) - \bar{q}(x)dx \simeq 3. \tag{1.11}$$

In this case there wasn't a large deviance from QCD. The heavy quarks can only be produced in tandem with an antiquark with momentum totalling that of the parent gluon, thus they do not contribute to the integral. The seaquark contributions were found to be much smaller than the valence quarks contributions. Therefore, $\int_0^1 (u - \bar{u})dx \simeq 2$ and $\int_0^1 (d - \bar{d})dx \simeq 1$ [12]. Measuring seaquarks thereby requires a high degree of precision;

although, seaquarks still impact hadronic products and lepton scattering.

1.2.2 Gottfried Sum Rule Violation

One of the sum rules apparently sensitive to the seaquarks was the Gottfried Sum Rule. The Gottfried Sum Rule (GSR) describes differences between the structure functions of the proton and neutron. Based on the success of pQCD so far, the GSR was expected to be

$$S_G = \int_0^1 \Big(F_2(x)^p - F_2(x)^n \Big) \frac{dx}{x}. \tag{1.12}$$

The first step to solve equation 1.12 theoretically is to expand the structure functions of the proton and neutron. Charge symmetry is well respected in strong interactions, giving the equalities $u_p(x) = d_n(x), d_p(x) = u_n(x), \bar{u_p}(x) = \bar{d}_p(x)$, and $\bar{d}_p(x) = \bar{u_p}(x)$. Taking the proton and neutron heavy quark PDFs to be equivalent as well leaves

$$S_G = \int_0^1 \frac{1}{3} \Big(u(x) - d(x) - \bar{d}(x) + \bar{u}(x) \Big) dx. \tag{1.13}$$

Rearranging equation 1.13 leads to

$$S_G = \frac{1}{3} \int_0^1 \Big(u(x) - \bar{u}(x) \Big) dx - \frac{1}{3} \int_0^1 \Big(d(x) - \bar{d}(x) \Big) - \frac{2}{3} \int_0^1 \Big(\bar{d}(x) - \bar{u}(x) \Big) dx. \tag{1.14}$$

Applying the Gross-Llewellyn-Smith sum rule in expression 1.11 to 1.14 yields

$$S_G = (2)\frac{1}{3} - \frac{1}{3} + \frac{2}{3} \int_0^1 \Big(\bar{d}(x) - \bar{u}(x) \Big) dx. \tag{1.15}$$

Finally, it was assumed that $\bar{d}(x) = \bar{u}(x)$. The primary generation mechanism of seaquarks was thought to be gluon splitting into a $q\bar{q}$ pair based on pQCD. Looking at the Lagrangian in Eq. 1.2, we observe that there is no flavor based interaction. This means that the gluon splitting probability is only dependent on the relative mass of the quarks. As the up and down quark masses are very similar, with a mass ratio between 0.38 and 0.58

[13], neither should be significantly favored in production by gluon splitting. Therefore, with no flavor asymmetry expected, the Gottfried Sum Rule is [14]

$$S_G = \int_0^1 \left(F_2(x)^p - F_2(x)^n \right) \frac{dx}{x} = \frac{1}{3}. \tag{1.16}$$

Early Flavor Asymmetry Measurements

The NMC experiment was one of the first to attempt to measure the GSR fairly directly. Due to their higher rest mass, muons replaced electrons in this DIS experiment. Muons enabled the NMC experiment to reach higher total energies and subsequently spacial resolutions. DIS was performed on hydrogen and deuterium nuclei serving as proton and neutron sources. NMC specifically measured the quantity $F_2(x)^p/F_2(x)^n = F_2(x)^d/F_2(x)^p - 1$ [13], but could resolve the GSR as

$$F_2(x)^p - F_2(x)^n = 2F_2(x)^d \frac{1 - F_2(x)^p/F_2(x)^n}{1 + F_2(x)^p/F_2(x)^n}. \tag{1.17}$$

A large deviation from the GSR was detected, with the GSR measured at 0.235 ± 0.026 as shown in the figure 1.3. This violation would consistently be encountered in latter measurements as well. If this deviation was due to the antiquark difference, then $\int_0^1 \left(\bar{d}(x) - \bar{u}(x) \right) dx = 0.148 \pm 0.039$!

As no potential mechanisms to explain this asymmetry existed, alternatives had to be ruled out. Potentially, contributions to the integral below the lowest measured bin at $x = 0.004$ could dominate, making the extrapolation down to zero invalid. Experiments E665 and NA51 attempted to probe the seaquarks in the low-x regime of $10^{-3} < x < 0.3$ to investigate this. Both collaborations concluded that the results were in agreement with NMC, as can be seen in the graph of E665 in figure 1.4. Any observed discrepancies were found to be consistent with nuclear shadowing in deuterium[16, 17]. The GSR was definitively violated, insinuating flavor asymmetry exists in the nucleon sea. This seaquark

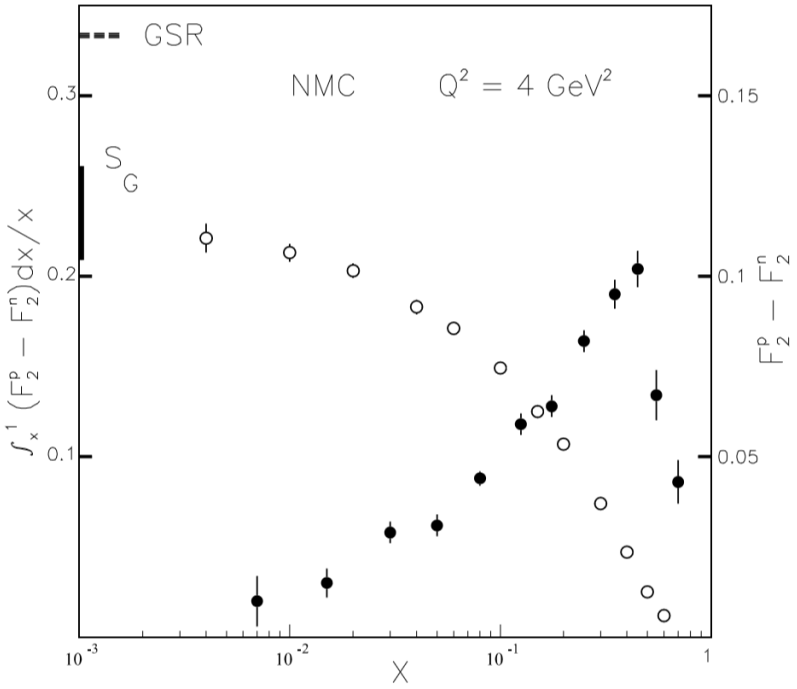


Figure 1.3: NMC measurement of the structure function difference is shown as a function of x and fixed Q^2 by the solid circles. Overlaid is the integral of the structure functions divided by x represented by the open circles. GSR extrapolations are shown by the bar on the left side. The figure was taken from Ref. [15].

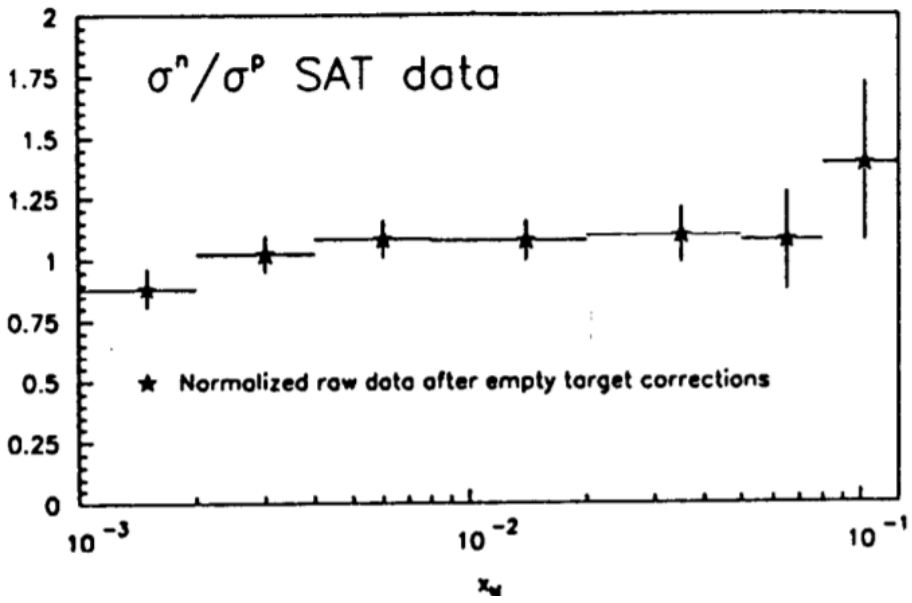


Figure 1.4: The neutron to proton cross-section ratio vs x as measured by E665 after target corrections. The figure was taken from Ref. [17].

flavor asymmetry signals potential strong force physics not yet predicted by pQCD.

1.2.3 Drell-Yan Process

One of the challenges DIS measurements face is a lack of flavor sensitivity. In DIS a virtual photon is exchanged with the nucleus shown in the Feynman diagram in figure 1.5. The photon exchange cannot decisively distinguish between a quark or an anti-quark. Therefore, DIS cannot provide a direct measurement of the seaquarks. The flavor insensitivity may be partly amended by measuring any hadronic products. These products are produced whenever the high energy probing lepton in DIS can transfer enough energy to a parton to cause hadronization. The measurement is then reclassified as Semi-Inclusive DIS (SIDIS). Unfortunately, SIDIS measurements still have limited flavor sensitivity and are far more complex than standard DIS.

However, by interchanging the time and space axis in the DIS Feynman diagram in figure 1.5, one exchanges a space-like virtual photon for a time-like one. The new interaction becomes directly sensitive to the anti-quarks, as displayed by the resulting diagram shown

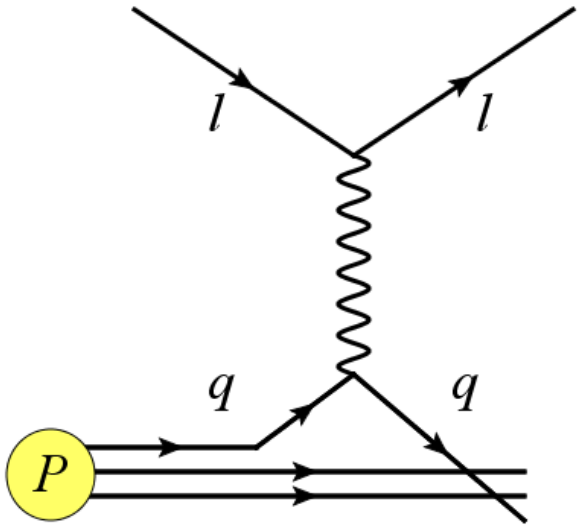


Figure 1.5: Feynman diagram of a DIS event is shown. Leptons are shown scattering off proton valence quarks.

in figure 1.6. Christened the Drell-Yan process, this interaction is defined by the annihilation of a quark anti-quark pair producing a virtual photon or Z boson, which decays into a lepton pair [18]. In a proton-proton collision, the anti-quarks used in the Drell-Yan process will be entirely seaquarks while the quarks would most likely come from the valence quarks of the other proton.

The leading order cross-section follows the same form as a standard lepton QED pair-annihilation/ production fermion diagram. Factors containing PDFs and the colors of incoming particles are added to account for the differences between using leptons and quarks. Therefore, the leading order (LO) cross-section is

$$\frac{d\sigma^2}{dx_1 dx_2} = \left(\frac{4\pi\alpha^2}{3M^2}\right)\left(\frac{1}{3}\right)\sum q_i(x_1)\bar{q}_i(x_2) + q_i(x_2)\bar{q}_i(x_1), \tag{1.18}$$

which was the cross-section that Drell and Yan originally proposed [18]. Direct measurements of sea anti-quark may hence only necessitate measurements of a lepton pair. Another advantage of this formulation is that the kinematic quantities are relatively simple to derive and observe experimentally. The differential angular distribution is expected to follow from the lepton QED $1/s^2$ annihilation/ pair-production form, thus resulting in a Drell-Yan

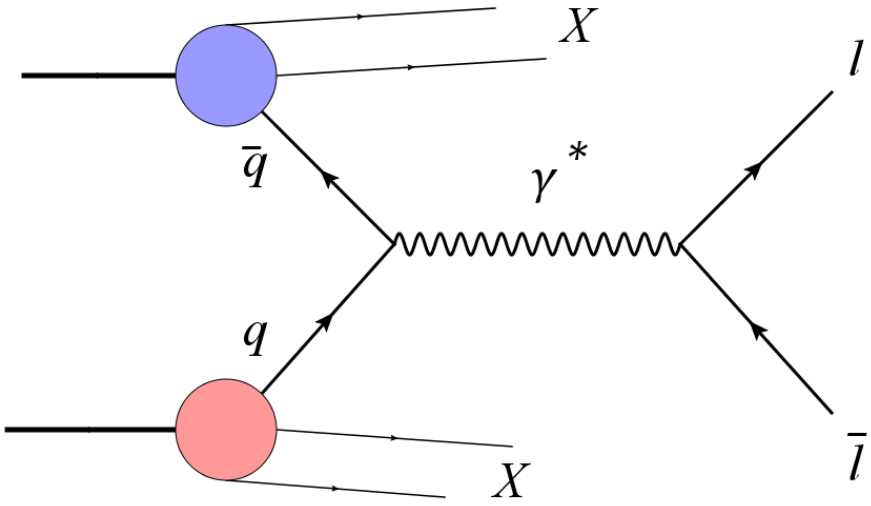


Figure 1.6: This Feynman Diagram shows a lepton-antilepton pair produced from quark-antiquark annihilation in a Drell-Yan event.

angular dependence of $1 + cos^2(\theta)$.

Experimentally, this angular form was roughly observed, but with some dileptons carrying more transverse momentum, p_T, than expected. Next to leading order terms (NLO) are responsible for this difference and also lead to almost doubling the total cross-section[19].

Multiplying Eq. 1.18 by a k factor and defining θ in the Collins-Soper frame account for the NLO corrections while still expressing the cross-section in the same straight-forward form. The cross-section in Eq. 1.18 doesn't explicitly contain any Q^2 dependence. This can be a remedied by solving for the mass in terms of s or Q via $M^2 = Q^2 \approx x_1 x_2 s$ and allowing the PDFs to evolve with Q^2[20] according to the DGLAP equations, making the cross-section

$$\frac{d\sigma^2}{dx_1 dx_2} \approx \left(\frac{4\pi\alpha^2}{9sx_1x_2}\right) \sum q_i(x_1, Q^2)\bar{q}_i(x_2, Q^2) + q_i(x_2, Q^2)\bar{q}_i(x_1, Q^2), \tag{1.19}$$

showing the aforementioned $1/s$ relation. The cross-section's direct sensitivity to the anti-quarks, makes the Drell-Yan process an ideal probe to measure seaquarks.

The Drell-Yan mechanism is not without its challenges, though. The QED cross-section

is suppressed by the hadronic interaction channels, requiring large statistics to perform any significant measurements. A proton beam in a fixed target experiment is capable of providing these measurements, as targets will have more interactions with the beam than another beam would in a collider experiment. In order to be sensitive to the flavor asymmetry, the neutron-proton Drell-Yan cross-sections must be compared as discussed in section 1.2.2 on the Gottfried Sum Rule. Utilizing a deuterium target as a stand-in for the neutron allows for a stable neutron target to be constructed with minimal nuclear effects.

1.2.4 Early Drell-Yan Experiments

Experiment Na51

CERN was the first laboratory to attempt to use a Drell-Yan experiment to observe flavor asymmetry in the NA51 experiment at $Q^2 = 25 - 30GeV^2$. This experiment created Drell-Yan events by colliding the $450GeV$ CERN-SPS primary proton beam into liquid hydrogen and deuterium targets. NA51's kinematics allowed for the approximation that $x_1 \approx x_2$ or beam protons' Bjorken $x \approx$ target nucleons' Bjorken x. This granted the simplified cross-sections

$$\sigma_{pp} \propto \frac{8}{9}u(x_1)\bar{u}(x_2) + \frac{2}{9}d(x_1)\bar{d}(x_2), \tag{1.20}$$

$$\sigma_{pn} \propto \frac{5}{9}u(x_1)\bar{d}(x_2) + \frac{5}{9}d(x_1)\bar{u}(x_2). \tag{1.21}$$

Ignoring less significant massive quark contributions and the small nuclear effects, the cross-section ratio becomes $\sigma_{pd} \approx \sigma_{pp} + \sigma_{pn}$.

Unfortunately, the kinematics involved when $x_F \approx x_1 - x_2 \approx x$ are not as sensitive to the sea quarks. Insufficient events were recorded to effectively span a large range in x, so the data was averaged into a single point at $x = 0.18$. What NA51 did find appeared to be in agreement with the DIS measurements on the violation of the GSR.

Experiment E866

Na51's limited results motivated another experiment with both higher statistics and a different range of kinematics. The experiment E866 utilized liquid deuterium and hydrogen targets again but appended nuclear targets onto the experiment in order to measure potential nuclear effects. E866 used the higher energy $800GeV$ Tevatron proton beam as opposed to NA51's $450GeV$ beam. Perhaps the most important difference was the use of dipole magnets to vastly expand the kinematic acceptance in the region $x_F > 0$ at $Q^2 = 54GeV$. In this phase space, anti-quarks are far more likely to originate in the targets than the beam[21].

E866's phase space also simplifies the cross-section ratio expression by taking $x_1 \gg x_2$. Essentially this limit restricts the cross-section ratio to measuring target seaquarks while ignoring the beam seaquark contributions. The resulting hydrogen and deuterium cross-sections are proportional to the PDFs as

$$\sigma_{pp} \propto \frac{4}{9}u(x_1)\bar{u_2}(x_2) + \frac{1}{9}d(x_1)\bar{d}(x_2), \tag{1.22}$$

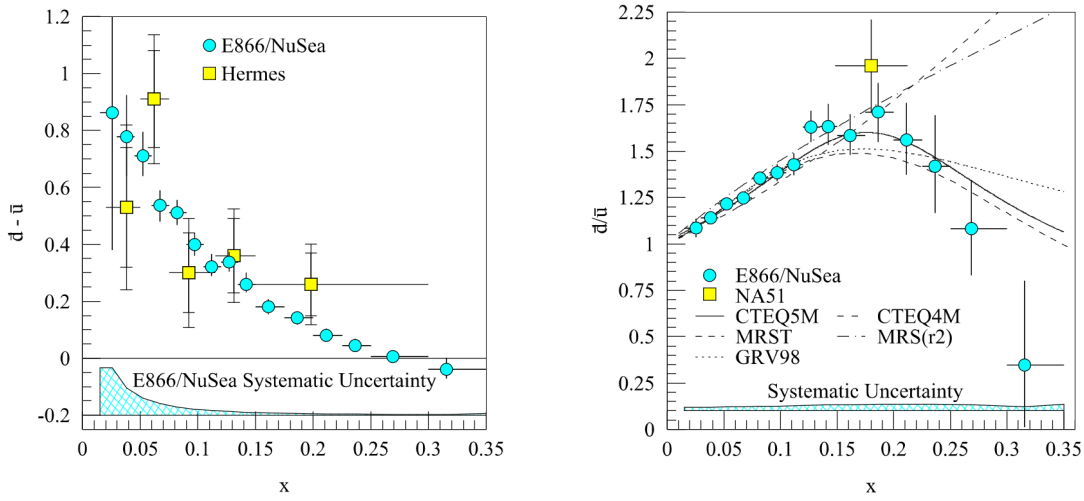
$$\sigma_{pn} \propto \frac{4}{9}u(x_1)\bar{d_2}(x_2) + \frac{1}{9}d(x_1)\bar{u}(x_2). \tag{1.23}$$

Assembling these equations and giving the deuterium cross-section a factor of two due to nucleon counts produces the approximation

$$\frac{\sigma_{pd}}{\sigma_{pp}} \simeq 1 + \frac{\bar{d}(x_2)}{\bar{u_2}(x_2)}. \tag{1.24}$$

Eq. 1.24 highlights the particular sensitivity to the sea quarks in this regime.

The E866 findings represented the first direct measurement of the seaquark ratio and confirmed the flavor anti-asymmetry, shown in the $\bar{d} - \bar{u}$ graph in figure 1.7 (a). Importantly, E866 achieved enough statistics to build a trend in x, while extending the asymmetry



(a) $\bar{d} - \bar{u}$ vs x as measured by E866 and Hermes.

(b) $\bar{d}/\bar{u}$ ratio vs x in NA51 and E866.

Figure 1.7: E866 results are shown as a function of x in terms of both the difference (left-panel) and ratio (right-panel) of $\bar{d}$ and $\bar{u}$. Some PDF predictions are also shown in the right-panel. The figures were taken from Ref. [9].

measurement the furthest in x_2 so far, with an upper bound of $x_2 = 0.35$ on the last point.

The points roughly agreed with the DIS experiments in the low x range, but also found an unexpected trend where the ratio dropped below unity shown in figure 1.7 (b). As the seaquark energy fraction must go to 0 as x approaches 1, the ratio must also later return to unity. The implication is multiple processes contribute to the ratio at different x values, each process dominating in different x regimes. According to E866, the max value of the ratio is 1.6 at about x 0.185. No model at the time was capable of explaining the distribution projected by E866. However, there are large error bars on the last couple of points in the ratio, prompting additional study to investigate this phenomenon. With additional points in this higher-x regime, asymmetry predictions of models can be directly tested against new data as well. Consequently, PDFs could be updated and further constrained.

1.3 Explanatory Models

Initially, the seaquark flavor asymmetry was attempted to be explained using Pauli blocking. Conceptually this was the extra up quark in the proton suppressing the $g \rightarrow u + \bar{u}$

channel. While there remains some debate on the topic, the calculations did not bear the experimental observations. Pauli blocking was found to have little contributions to the asymmetry, and only affects second-order pQCD calculations[9]. Lack of a satisfactory perturbative explanation motivated several candidate non-perturbative theories to explain the GSR discrepancy or $\bar{d} - \bar{u} > 0$.

The non-perturbative models still possess pQCD contributions. In all models, pQCD effects dominate in the low and high x regimes. The gluon PDF asymptotically increase as x approaches 0, and the gluons generate virtual quark pairs sending the ratio to one. At high-x, scattering becomes hard and thus nucleons act as their valence quarks. The anti-quark distributions must converge to 0 at a much greater rate than the valence quarks. Non-perturbative contributions plummet to zero as quarks radiate gluons and momentum is shared amongst produced quark pairs. The gluon PDF drops slower than the anti-quarks in this limit. As x goes to one, anti-quarks can be described by a scattering pole where the derivative ratio approaches unity. Therefore, all models follow the PDFs, which send the anti-quark ratio to one. Theoretical models models may dominate similar x regimes between $0 < x1$, but varying in max value and location. In total, four major types of theoretical models are reviewed followed by a computational model for predicting a GSR discrepancy with theoretical implications.

1.3.1 Theoretical Models

Instaton Model

Owing to the success of instanton like fields in other areas of physics, a QCD instanton field was proposed. QCD instantons would be associated with quantum tunneling in the vacuum and represent a potential source for chiral symmetry breaking of the strong force[22]. Instanton liquid model modifications of the QCD vacuum contribute to an effec-

tive QCD Lagrangian, consisting of the chiral quark terms

$$\mathcal{L}_{eff} \, \alpha \, \bar{u}_R u_L \bar{d}_R d_L + \bar{u}_L u_R \bar{d}_L d_R. \tag{1.25}$$

The effective Lagrangian component in Eq. 1.25 follows t'Hooft's SU(2) formulation, the L and R terms referring to the helicity, directly incorporating spin. The magnitude of the instaton term depends on the instanton density and average size. This instanton term carries into the instanton induced quark cross-sections and nucleon-quark cross-section. After accounting for these new interaction modes, a violation of the Gottfried sum rule (GSR) is predicted at $0.11 \pm .02$ in agreement with the NMC measurement[22].

Dorokhov and Kochelev also predicted the quark-quark cross-section anomalously increasing with energy or Q^2 and gaining a p_T dependence as well [22]. Simultaneously the spin and matrix elements are directly associated, as

$$\bar{d}_I(x) - \bar{u}_I = \frac{3}{5}\Big[\Delta u_I(x) - \Delta d_I(x)\Big]. \tag{1.26}$$

Equation 1.26 uniquely alludes to an increase in the valence quark flavor asymmetry with an increase in the seaquark flavor asymmetry, other models expecting the opposite [9]. The instanton approach reduces the total spin carried by quarks in the nucleon, potentially resolving the proton spin crisis. Additionally, this model is capable of explaining several different SIDIS and DIS experimental results, lending the theory empirical merit independently [22].

Chiral Quark Model

However, this was not the only model that attempted to simultaneously explain the spin crisis and sea flavor asymmetry by adding chiral interaction terms to the effective Lagrangian. Chiral Quark Models (CQM) and Chiral Quark Soliton Models (CQSM) are effective field theories. As their namesake suggests, CQMs employ the direct interactions

of chiral quarks to explain hadron structure and low energy QCD. CQMs parameterize QCD with degrees of freedom from the Goldstone Boson, gluons, and quarks. Chiral symmetry breaking in low-energy QCD has negligible contributions from gluons in CQM, and instead arises from valence quarks fluctuating into quarks plus Goldstone Bosons. Therefore, the QCD Lagrangian must be modified to an effective Lagrangian to accommodate these new interactions with

$$\mathcal{L}_{eff} = \bar{\psi}(x)\big(iD_\mu + V_\mu\big)\gamma^\mu\psi(x) + ig_A\bar{\psi}A_\mu\gamma^\mu\gamma_5\psi, \tag{1.27}$$

where V_μ and A_μ are the vector and axial currents respectively, reflecting the Goldstone Boson (Π) coupling $\mathcal{L}_{q\Pi} = -\frac{g_A}{f}\bar{\psi}\Pi\gamma^\mu\gamma_5\Psi, f \approx 93MeV$[23]. The Goldstone Boson coupling terms become a meson coupling matrix Π. The axial and vector currents A_μ and V_μ in Eq. 1.27 are thus expansions of the generator $O(\Pi/f)^2$, but A_μ possesses an extra $i\partial_\mu\Pi/f$ term. The quarks can be rewritten as a Fock decomposition of meson and quark couplings through the Goldstone Boson states, with

$$|U\rangle = \sqrt{N_u}|u_0\rangle + a_{\pi^+}|d\pi^+\rangle + \frac{a_{K^+}}{\sqrt{2}}|sK^+\rangle, \tag{1.28}$$

$$|D\rangle = \sqrt{N_d}|d_0\rangle + a_{\pi^-}|u\pi^-\rangle + \frac{a_{K^0}}{\sqrt{2}}|sK^0\rangle. \tag{1.29}$$

The a terms in Eqs. 1.28 and 1.29 originate from the corresponding probabilities of the associated state and N is the overall normalization constant. Both equations can accommodate a potential η meson term too. The Fock decompositions in Eqs. 1.28 and 1.29 imply contributions to the sea through processes such as $q_i \rightarrow \pi^{\epsilon_{i,j}} + q_j$, where i=1 corresponds to an up quark. Hence, a virtual meson cloud binds directly to the valence quarks [24]. The virtual meson cloud leads to a modification in the PDFs and a resulting GSR

violation

$$S_G = \frac{1}{3}\int_0^1 dx[u(x) + \bar{u}(x) - d(x) - \bar{d}(x)] \approx \frac{1}{3}(1 - 2a_{\pi_+}) = 0.278, \tag{1.30}$$

by taking the pion, up and down quark masses to be equal and ignoring the strange states [25]. The seaquark flavor asymmetry is therefore produced by these meson clouds bound to the chiral quarks. Consequently the quark states are redefined, which mandates a strange quark asymmetry in as a function of x.

The CQSM uses a similar Lagrangian to the CQM but parameterizes it in terms of the strange and non-strange quarks' mass difference. As in the CQM, the now SU(3) meson fields and quark fields are not independent, but instead, have a chiral coupling through a sort of rotating baryon "soliton". The effective Lagrangian becomes [25]

$$\mathcal{L}_{CQSM} = \bar{\psi}(x)\big(i\delta + MU^{\gamma_5}(x)\big)\psi(x) - \bar{\psi}(x)\delta_s P_s \psi(x) \tag{1.31}$$

. The first term in Eq. 1.31 is the symmetric SU(3) Lagrangian with the Chiral field $U^{\gamma_5}(x) = e^{i\gamma_5\pi_a\lambda_a/f_\pi}$ and is all that is required for an SU(2) model. Instead of a Fock state truncation, an infinite Dirac Sea of potential pion orbitals is allowed. Nucleon mass is then dependent on the soliton field energy. The pion field is uniquely obtained by minimizing the non-linear nucleon mass functional, thus giving the theory its moniker as the "soliton" model[26]. No free parameters are generated in this field, and extending it to SU(3) only requires one constrained, adjustable $\Delta s \approx m_s$ parameter. This results in the symmetry breaking in the second term and leaves the theory only coupling to the three pion fields instead of eight mesons as in CQM[26]. The theory then becomes particularly attractive due to its lack of parameters, infinite degrees of freedom and have solid phenomenology. CSQM predictions agree with PDF and TMD (Transverse Momentum Dependence) results so far within $(10 - 30)\%$[27]. The SU(3) extension even solves the NuTeV issue and

potential $s - \bar{s}$ asymmetry connecting it to the $\bar{d} - \bar{u}$, all while not being heavily tunable to the data[28].

Meson Cloud Model

However, instead of coupling a meson or pion cloud to the quark fields, the proton wavefunction can be composed of meson Fock states containing a base valence state in addition to baryon+meson states. Consequently, the entire nucleon couples to a virtual meson cloud illustrated in figure 1.8. The proton becomes

$$|p\rangle = \sqrt{N}|p_0\rangle + \sum_{mesons,baryons} c_{m,b}|mesons, baryons\rangle \tag{1.32}$$

integrated over the longitudinal momentum fraction and transverse momentum using a normalization factor N[29]. As a result, the proton would have a large contribution from the $n\pi^+$ state, generating an excess of $\bar{d}$ quarks since $\pi^+ = u\bar{d}$. Mass and energy coupling suggests that the $\Delta\pi$ state contributes next most to the sea. The πn and $\pi\Delta$ vertices represent a large source of variance between otherwise equivalent models, due to poor experimental constraints. However, the Goldberger Trieman relation implies $MG_A(t) = f_\pi g_{\pi N}(t)$ in Feynman t, and as $(\frac{g_{\pi N}}{g_{\pi\Delta}})^2 = \frac{72}{25}$ by the quark model, the model vertex coupling can still carry theoretical constraints, capable of potential future independent measurements from outside experiments[30].

Regardless, a simple approximation can be constructed from the first terms utilizing $c^2_{\delta^{++}\pi^-} = \frac{2}{3}c^2_{\delta^+\pi^0} = 3c^2_{\delta^0\pi^+} = \frac{\alpha}{2}$ and $2c^2_{p\pi^0} = c^2_{n\pi^+} = \frac{2\beta}{3}$. Dominant valence pion contributions to quark PDFs via

$$\frac{\bar{d}(x)}{\bar{u}(x)} \approx \frac{[\frac{5}{6}\alpha u^\pi(x) + \frac{1}{3}]\bar{d}^\pi(x)}{[\frac{1}{6}\alpha + \frac{2}{3}]\bar{u}^\pi(x)}, \tag{1.33}$$

lead to a sea flavor asymmetry in this simplified version of the model[31]. DIS scattering would now include contributions from the bare nucleon, intermediate pion and intermediate baryon.

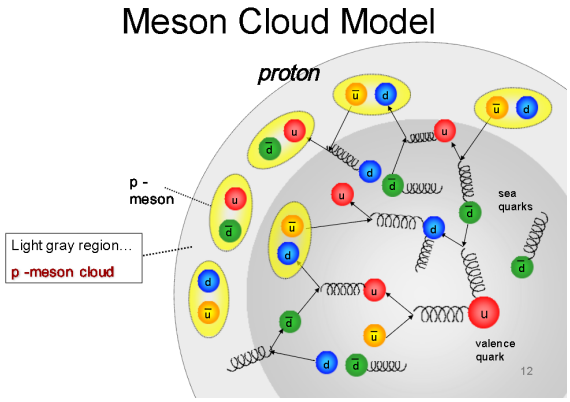


Figure 1.8: Nucleon meson cloud model illustration, showing an additional meson orbital added to the nucleon. The figure was taken from Ref. [32].

The meson cloud model has immediately accessible implications for other hadrons. The elegant formalism of this model can be most readily incorporated into Compton scattering and excitation predictions of higher mass hadrons. For example, a meson cloud model predicts the DIS electromagnetic transitions $\gamma^* + \Lambda \rightarrow \Sigma$ and $\gamma^* + N \rightarrow \Delta$ cross-sections using the Δ decay width. Those interactions and the associated decays are in agreement with experiment[33]. Also, the meson cloud model and C(S)QM have potentially measurable predictions for hadronic magnetic moments and charge radii[26, 34].

Statistical Model

The Statistical Model similarly incorporates a Fock decomposition to represent the proton as a probability of being in a given state. This method is perturbative in nature though, relying on standard QCD quark and gluon interactions. However, gluon production or splitting composes different states. The simplest model of the proton is $\mid p\rangle = \sum_{ijk}\{uud\} \mid \{ijk\}\rangle$ with probability being in a given state $\rho_{ijk} = \mid c_{ijk} \mid^2$ with i being the number of $u\bar{u}$, j is the number of $d\bar{d}$, and k is the number of gluons. Thus, accounting for standard quark-gluon interactions, the proton state becomes

$$\frac{\rho_{ijk}}{\rho_{000}} = \frac{2}{i!(i+2)!j!(j+1)!}\prod_{n=0}^{k-1}\frac{3+2i+2j+n}{(3+2i+2j)(n+1)+\frac{n(n+1)}{2}}. \tag{1.34}$$

The result of the relative term comparison predicts asymmetry at low energy[35]. However, Eq 1.34 lacks Q dependence. A natural extension to the classical statistical analogue involves treating the partons like a gas at equilibrium restricted to the nucleus volume. The partons then have a Fermi-Dirac/Bose-Einstein distribution $(e^{(x-k_p)/\bar{x}} \pm 1)^{-1}$, $+$ for quarks, $-$ for gluons[36]. Given that the statistical model becomes fundamentally perturbative when incorporating these energy states, the model permits transitions from low to high energy QCD without additional theoretical complexity. Consequently, the statistical model lends itself to augmenting different non-perturbative theorems, forming new hybrid models with the combinations. Compatible non-perturbative interactions, capable of being described by Fock state expansions, can potentially be combined with the model as well.

1.3.2 Model Comparison

One striking aspect is that all of the non-perturbative theories discussed inadvertently connect the polarized PDFs to the seaquarks, generally attempting to explain the proton spin puzzle simultaneously. All have predictions with the longitudinal and transverse momentum-fraction dependencies as well as relations to TMDs and polarized PDFs such as $\Delta\bar{u} - \Delta\bar{d}$.

Comparing the model sea flavor asymmetry predictions to one another and the E866 data as in figure 1.9, all models clearly over-predict the sea at $x > 0.3$ as measured by E866. The error bars are quite large in this region though. In fact, only Szczurech's Chiral Quark model doesn't over-predict at $x > 0.23$, but the model is too soft, diverging acutely at low x from not only the E866 results but also Na51's. More recent CQMs do show better agreement with the E866 data at $x < 0.3$, but under-predict near its max and disagrees with Na51 [37]. Regardless, the highly constrained CQSM of Pobylitsa reproduces the x fairly well except at the large error points and low x. It shouldn't be ruled out entirely. $\bar{d}/\bar{u}$ ratio measurements at high x are required to remove the large errors and rule out whether or not the sea ratio drops below unity. These measurements would increase the resolution

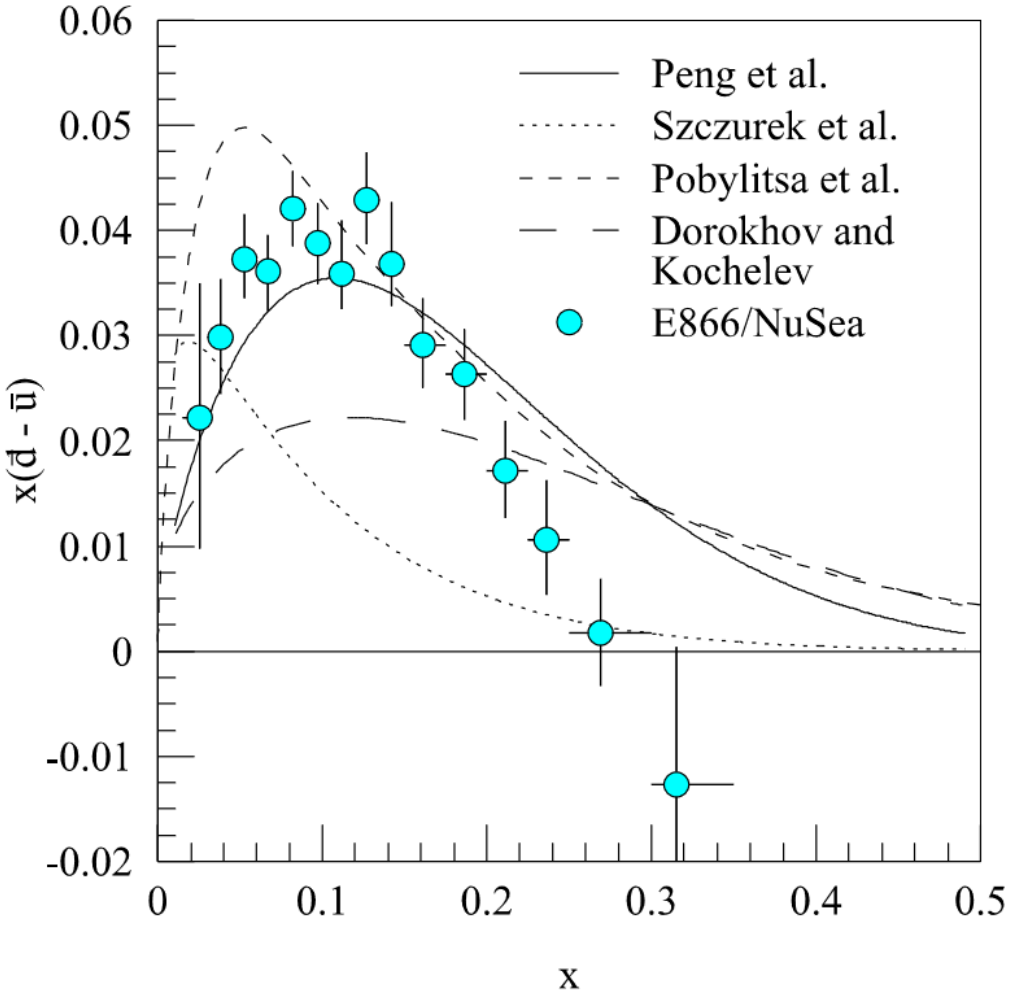


Figure 1.9: Various sea flavor-asymmetry predictions of models as a function of x are compared to E866 data depicted by the solid circles. Pobylitsa represents a CSQM model, Szczurech a CQM, Dorokhov the instanton model, and Peng a meson cloud model calculated with both delta and pion vertices. The figure was taken from Ref. [9].

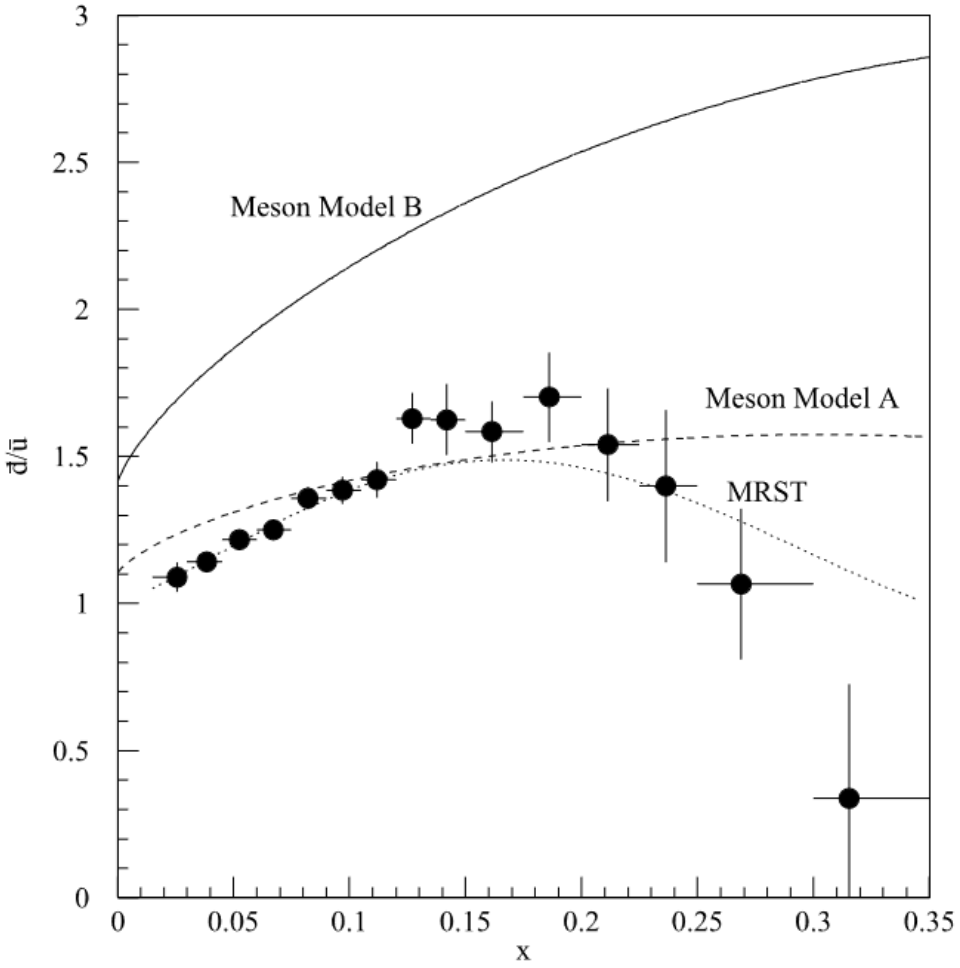


Figure 1.10: Two Meson Cloud Models and the MRST PDF are compared with E866 data, indicated by the solid circles, in terms of the $\bar{d}/\bar{u}$ ratio as a function of x. The meson models differ only in $\pi N\Delta$ form factors. Model A has $\Lambda = 1.0$ GeV and B has $\Lambda = 0.8$ GeV. Model A agrees better only when perturbative contributions are considered. The figure was taken from Ref. [21].

between models as most diverge during the x evolution. For instance, the instaton model ratio is expected to increase to 4 at high x, while meson models can take high x ratio values between 1.4 and 5, as seen in figure 1.10[9].

One caveat to the model comparison is that non-perturbative techniques do not describe the entire source of proton seaquarks. Gluon splitting contributions always compose a certain fraction of the seaquarks. Therefore, over-predictions in the ratio of sea flavor may be explained by the poorly constrained gluon component in the nucleon, bringing the ratio closer to unity. The $\bar{d} - \bar{u}$ measurement is not impacted by gluon splitting, but the anti-

quark $\bar{d}/\bar{u}$ ratio is more directly measured by Drell-Yan experiments. The ratio provides more control over the systematic error, and the impact of gluon splitting drops sharply in x. These effects make Drell-Yan experiments even more revelatory at high x, supporting the need for another experiment further. Therefore, it is concluded that another measurement is needed to make any conclusions about flavor asymmetry in the high-x region before ruling theoretical models out. These theoretical models will all be considered when comparing any new flavor asymmetry measurements.

Newer techniques, potentially capable of explaining this drop below unity have been developed. These employ computational methods, which do not necessarily require a new non-perturbative model altogether. However, these techniques aren't mutually exclusive with the aforementioned theoretical models and can be thought of as a parallel check rather than a complete alternative.

1.3.3 Lattice QCD

Recent computational methods are capable of modeling low energy QCD and hadron structure. These models were developed independently and not specifically designed to explain a violation in the GSR. This method, called Lattice QCD (LQCD), is derived directly from numerical calculations of first principles. Therefore, LQCD is one of the only fundamental descriptions of QCD physics. LQCD accomplishes this by placing particles on a lattice, forming a discretized space-time. A lattice of spacing a results in a natural ultraviolet cutoff of π/a and scale. While taking the limit as a approaches zero, the resulting infinities on the lattice may be renormalized. Therefore, only fundamental gluon and quark degrees of freedom remain, leaving the only tunable elements as the quark masses and the coupling constant. Hadronic states and matrix elements are then simulated using techniques similar to those employed by statistical mechanics. LQCD has so far found success in hadronic mass and cross-section estimations[38]. Like all models, the lattice construction inherently introduces estimable systematic and statistical errors, while the computation

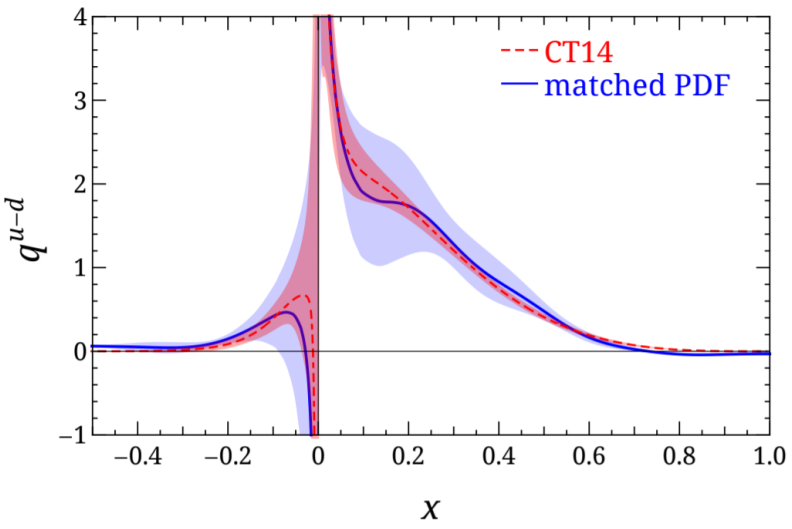
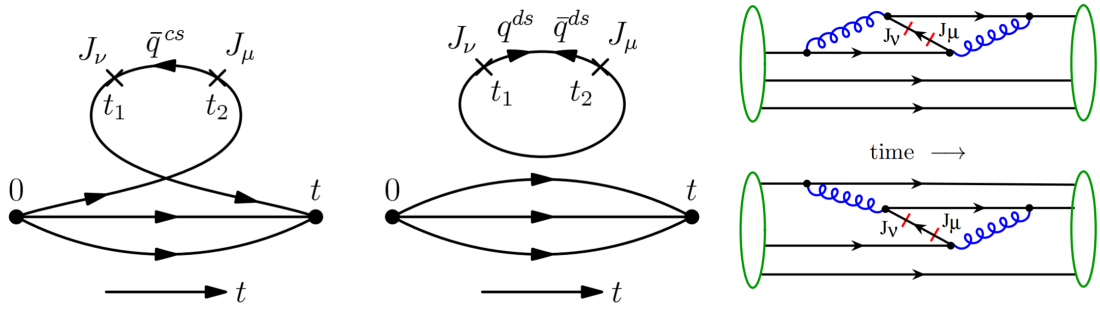


Figure 1.11: Shown is the Lattice QCD $u - d$ distribution as a function of x after re-normalization. The CT14 PDF is graphed as well for comparison purposes. The figure was taken from Ref. [40].

is difficult to complete.

Hadron structure is even more challenging though. Only "Quasi-PDFs" of quarks have been built through the large-momentum effective field theory (LaMET) technique. Starting with the rest frame of the nucleon, parton physics corresponds to light-cone correlations, equivalent to time-independent spatial correlations. From these correlations, one may construct Quasi-PDFs through nucleon matrix elements on the lattice[39]. By definition, these functions cannot be Lorentz invariant, but can be related to the true PDFs through a factorization formula, such as those GDPs (see section 1.4.1) use, at nucleon momentum $P \gg M_N, \Lambda_{QCD}$. Connecting the low x region to UV regime, polarized and unpolarized PDFs can be resolved. The $-x$ contributions come directly from $\bar{u} - \bar{d}$, due to the light cone identity $\bar{q}(x) = -q(-x)$ as seen in figure 1.11. The integral agrees with previous measurements of the Gottfried sum rule of 0.16(3) at $Q^2 = 2.3GeV^2$ and predicts a polarized sea asymmetry[39]. The current results are believed to show the correct qualitative form based on data, but require additional re-normalization and much better control on systematics before being able to employ true quark mass.

As LQCD is fundamental, it can place restrictions on other models, such as CQM or



(a) Connected and Disconnected Sea Diagrams (b) Connected Diagram Interactions

Figure 1.12: Connected and Disconnected Sea diagrams are compared in a) with the CS on the left and DS on the right. No interior quark lines connect to the DS as it is produced by the gluons. The two connected diagrams in b) involve one and two valence quark and can produce anti-quark asymmetry. The figures were taken from Ref. [41].

provide estimates on parameter values like $\delta - \pi$ and $n - \pi$ vertices in the meson cloud[30]. It may impact or change the interpretation of results over time, giving new context to current work.

Connected and Disconnected Sea

Returning to first principles, one may construct the origins of a nucleon sea in the hadronic tensor path integral formalism through three topologically distinct diagrams shown in figure 1.12. A connected sea diagram corresponds to a valence quark propagating either forward or backward in time, and a disconnected sea diagram corresponds to the vacuum polarization. LQCD and experimental data may separate the contributions and make predictions using this model, which would otherwise not converge using pQCD calculations. In fact, the defining three-point functions can be explicitly incorporated into LQCD[42]. The connected sea results in the asymmetry and has the important feature that it allows the ratio to drop below unity in agreement with the E866 results due to the quantum fluctuations of valence quarks. However, the short lifetime and thus higher energy implies that the connected sea would not overcome meson cloud or chiral contributions being compatible with both theories[41]. As a fully perturbative allowed diagram in QCD, connected and disconnected seas were suggested to be separately accommodated in an extended QCD

evolution to facilitate the global fitting[42].

1.4 Global Fits

There are more indirect impacts extending this measurement may have on our understanding of particle and nuclear physics. In modern particle physics, PDFs which physicists use are generated by what is known as global fits, which are built using a phenomenological model. An example of several globally fit PDFs by the MSTW collaboration are shown in figure 1.13 [43]. There is no analytical derivation of Global PDFs from first principles, so instead, they use a large, diverse set of data from thousands of data points to generate functions and their errors. The goal is to build continuous functions from discrete sets of points, split among LO terms and more convoluted NLO and NNLO terms. The latter depends on complete integrals of the previous order terms applied to a given PDF. Discrete empirical points may be connected on the grounds of the factorization functions and DGLAP evolution, filling gaps between hard scattering and DIS experiments[43]. A global Chi-square is built from all of the PDFs, while local ones are constrained to be within tolerance. The error used for these Chi-square parameters originate from measurements directly and error sets of these measurements. These are evolved using theoretical and computational models like Monte Carlo and Hessian methods with the DGLAP equations. The evolution predictions themselves can add their own systematic and theoretical errors on top of experimental constraints. The problem still seems fundamentally ill-posed. Fine-scale structure in x at a Q_0 tend to be smoothed, but these regions tend to not necessarily be areas of interest or measurement[44]. In fact, PDF changes from measurements can be predicted locally without a full global analysis utilizing the Hessian method[43].

Collider and other hadronic experiments then rely on these functions in extracting results and making predictions. One of the consequences is that error in measurement may then impact subsequent measurements by changing boundaries on these functions. In fact, PDF error contributions were dominant in the 2012 measurement of Higgs production[43].

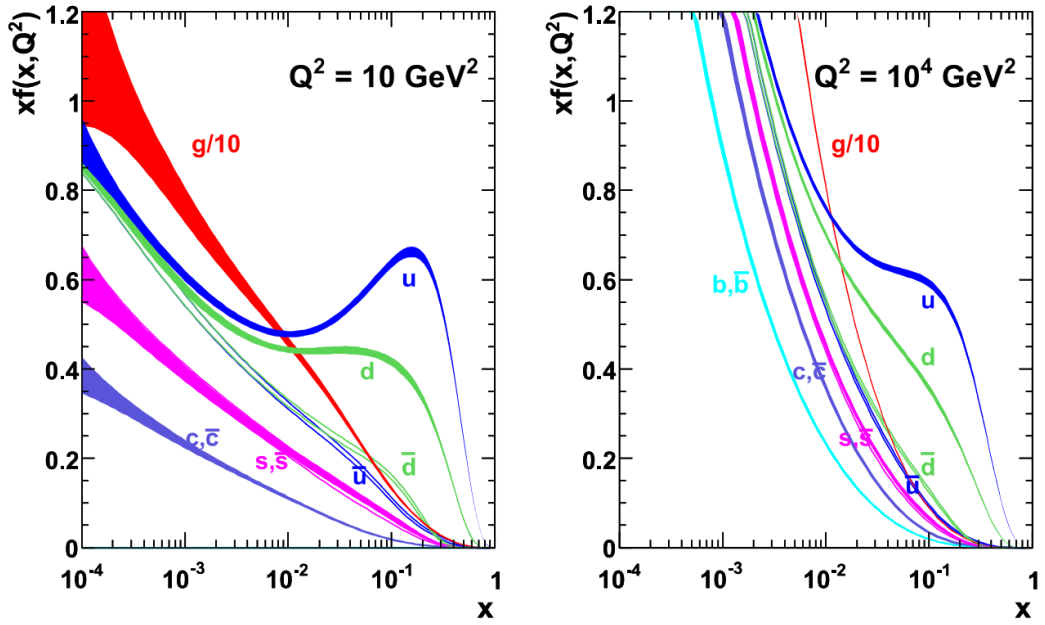


Figure 1.13: Parton Distribution functions built by the MSTW collaboration through global fitting are shown at two different energies, the left at $Q^2 = 10GeV^2$ and the right at $Q^2 = 10^4GeV^2$. The line width expresses the 68% confidence interval. The figures were taken from Ref. [43].

The $\frac{\bar{d}}{\bar{u}}$ ratio most clearly impacts $W^{\pm}$ bosons through $\frac{W^+ - W^-}{W^- + W+} \propto \frac{\bar{d}}{\bar{u}}$, but will also impact any high sensitivity proton-proton or proton-nucleon collision experiments. DIS and SIDIS measurements are all directly impacted by potential contributions and direct PDF calculations. E866 updated and provided important constraints to the seaquark distributions used in CTEQ and MRST. CTEQ5+ directly incorporated the trend including the last point falling below unity and extends this to the $x = 1$ limit.

1.4.1 Generalized Parton Distributions

The current studies may also have an impact on GPD's, as they are sensitive to DIS and Drell-Yan functions. GPDs combine transverse momentum (thus TMD's) and angular momentum functions with standard PDFs describing the nucleon in full 3-D space. This can be done by constructing form factors on the light cone of the nucleon, in effect adding an additional transverse ζ parameter to the standard PDF's[44]. SeaQuest is not as directly sensitive to TMDs such as the Sivers asymmetry, due to lack of polarization. While

SeaQuest does attempt to measure the Boer-Mulders, the primary link to the current results is that GPD's contain descriptions for $q - \bar{q}$ pairs and implications of meson formation in the nucleon. This makes them sensitive to seaquark meson degrees of freedom, reaffirming the aforementioned link between the polarized states and sea-asymmetry through non-perturbative theory links.

An improved $\frac{\bar{d}}{\bar{u}}$ can have an echoing impact on both theoretical and experimental physics through revealing the underlying nucleon structure. Consequently, SeaQuest is tasked with extending this measurement to high-x. The flagship question this new measurement and this thesis will attempt to answer is "does the seaquark flavor ratio really drop below unity?" If so, many of the aforementioned theoretical models would require heavy updates or become invalid as a description of the strong force, disfavored based on their predictions in high x. As there are multiple versions of models with similar underlying concepts, it becomes vital to restrict or eliminate these versions to compare future measurements.

There are practical implications to other experiments aside from the potential theoretical meaning. The E906 SeaQuest experiment is linked to the rest of modern collider physics through PDFs. The theoretical models separately link the seaquark flavor asymmetry to the polarized sea and the transverse structure of the proton. This Drell-Yan data may even be revisited by more modern formulations in the future, further justifying a new measurement in a higher-x region beyond the immediate model implications. Figure 1.14 includes the originally projected data given the promised beam and normal functioning. The data shown is assumed to follow the E866 data trends and helps show the projected value of the current measurement.

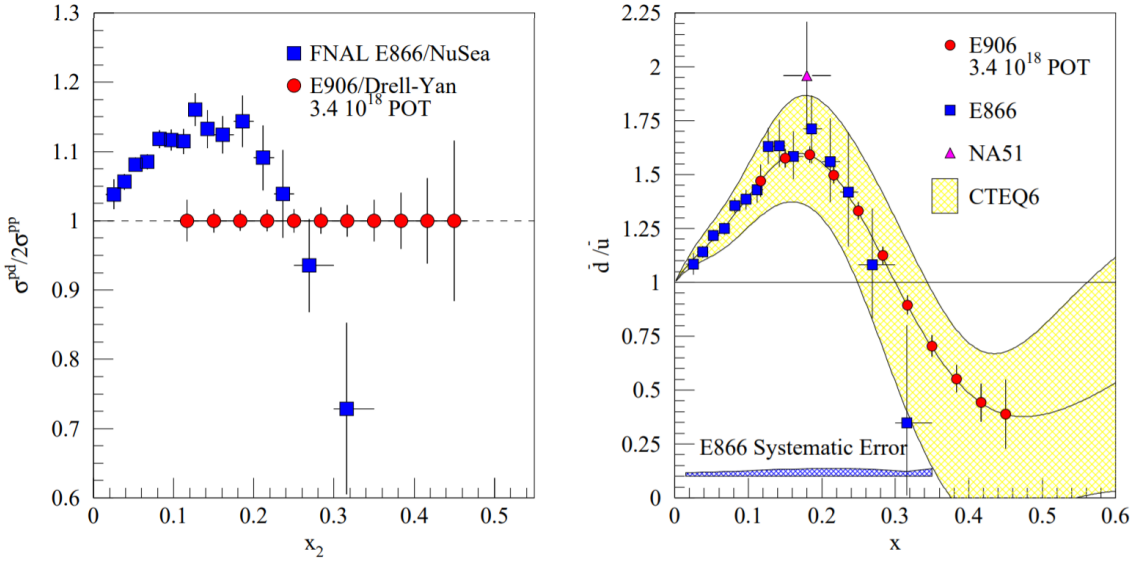


Figure 1.14: Projected E906 Data exhibits extension of the measurement in Bjorken-x. The left plot shows a projected cross-section statistics compared to E866, the data was set to unity for illustration purposes. The right plot displays a potential $\bar{d}/\bar{u}$ vs x, where the E906 data points are placed according to the CTEQ6 PDF, so the points can be compared to E866 and NA51 measurements. The figures were taken from Ref. [45].

Chapter 2

The Seaquest Spectrometer

The E906 apparatus, like E866, measures coincident dimuon pairs generated in proton beam collisions with stationary targets.[46] Specifically, the SeaQuest spectrometer targets consist of liquid hydrogen and deuterium and solid nuclear targets, to collect data on Drell-Yan events. An empty liquid flask and air only settings are also included to control for background. However, the beam colliding with the targets represents one of largest changes from E866. SeaQuest uses the Main Injector beamline in lieu of the now-closed Tevatron proton ring, allowing for much greater statistics.

The 120 GeV extracted beam's luminosity is monitored live by a Cherenkov counter upstream of the target. Downstream of the target, the remaining beam collides with the iron dump. The spectrometer downstream of the target including the dump is shown in figure 2.1. [46] The dump stops the beam while absorbing the hadronic products outside of neutrons produced in the last few cm of the dump. Only muons and neutrinos pass through the dump at any measurable intensity. The dump is surrounded by a dipole magnet, which divides particles by charge and momentum before they enter a series of tracking stations. A muon pair is shown traveling through the spectrometer in figure 2.2. The magnet and subsequent stations are tuned to increase high target Bjorken-x or x_2 acceptance. The magnet bends large opening angle events into the spectrometer where detectors are specifically arranged to capture the tracks. Before the last 2 stations, there is an additional air-core dipole magnet, which acts as a momentum analysis device.

This "analysis" magnet gives an additional kick along the same axis as the previous

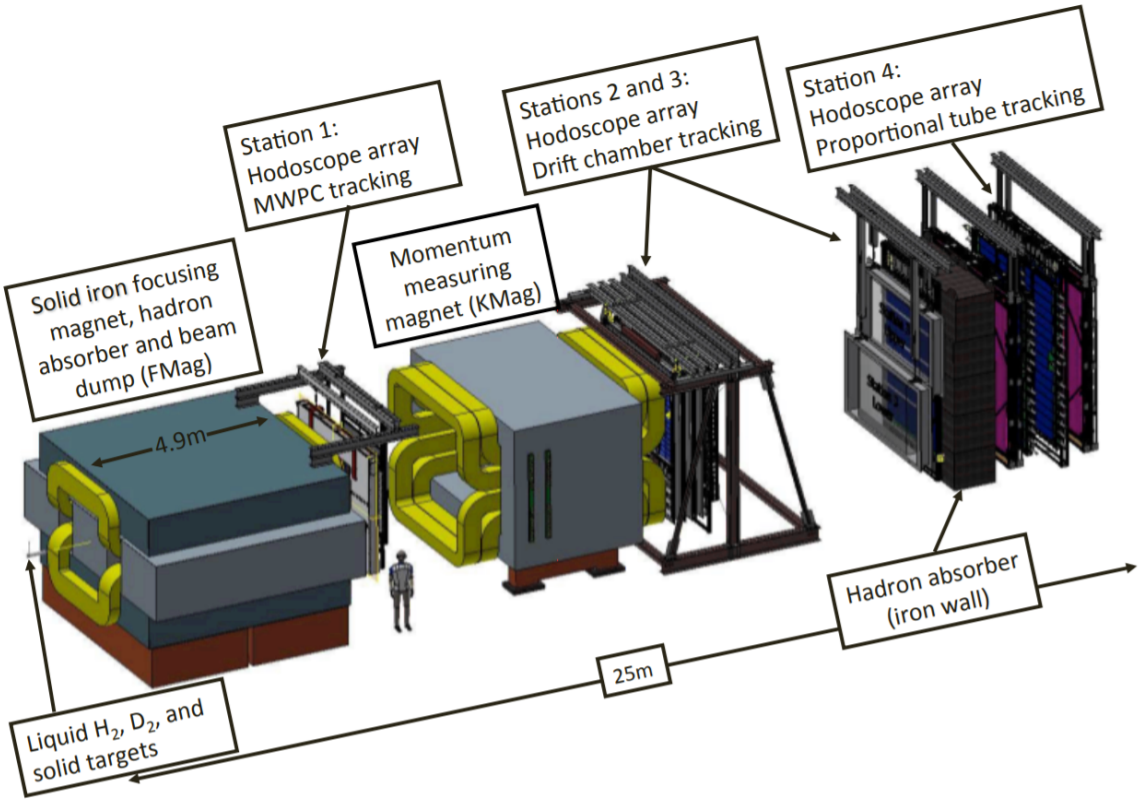


Figure 2.1: A SeaQuest spectrometer 3D model is shown to scale with a person for reference. Different components are labelled with the spectrometer oriented such that the beam moving enters from the left. The figure was taken from Ref. [46].

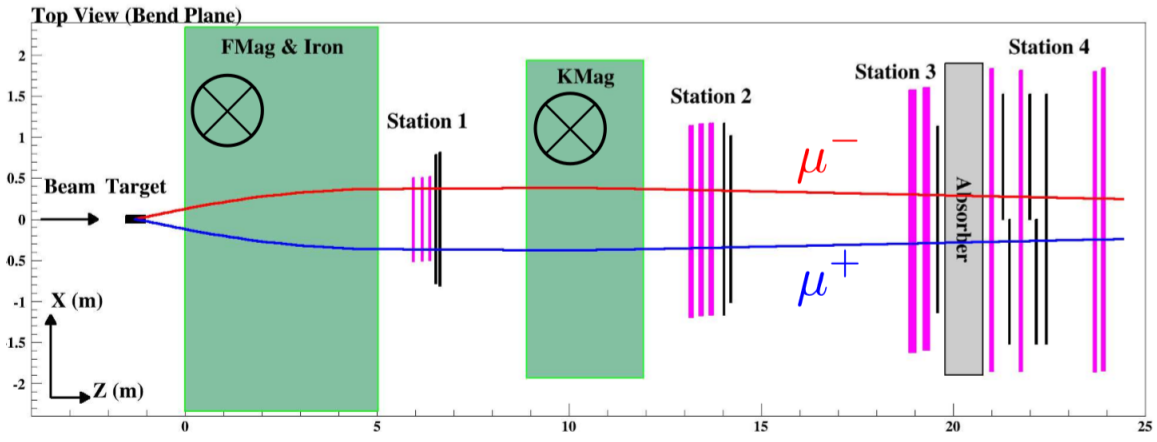


Figure 2.2: A diagrammatic representation of the spectrometer with muon track laid on the system where they would travel. The figure was taken from Ref. [47].

magnet, which enables the spectrometer to discriminate momentum and particle origin simultaneously. The tracking stations consist of a series of hodoscopes for triggering, drift chambers for tracking, and proportional tubes in the last station for charge. The detector TDC times are recorded by the Data Acquisition (DAQ) for tracking and eventually analysis.

2.1 Timeline

The data was collected over 5 years segmented into 6 "runs". Throughout running, the detector evolved. Runs were generally intersected by Fermilab shutdown periods, during which major improvements and maintenance was performed on the spectrometer. Consequently, shutdown periods conveniently represented significant changes in geometry as well as chronological leaps.Therefore, all runs were chosen to be bounded by either a shutdown period or a decisive geometry change.

Run history and associated spectrometer changes are summarized in table 2.1, where the roadsets refer to trigger logic changes. Over the course of data taking, runs tended to improve based on knowledge from previous runs. The first run was a commissioning run, which ensured components were functional and met resolution requirements, while determining settings for running and improvements. This run is not in a comparable condition to other data runs, thus will not be included in analysis. Run 2 represents the first true data taking period while Run 3 represents the longest continuous set of data recorded in stable running conditions. However, Run 4 was discarded, due to several components becoming dysfunctional while attempting to use the new Drift Chamber (DC) 1. Despite there not being a shutdown, Run 5 was separated from Run 4, as the spectrometer was returned to fully operational condition.

The original DC1 was returned to the spectrometer in tandem with the new DC1 added in Run 4. The original DC1 was renamed DC0 for runs 5 and 6. Thus, Run 5 is the first usable run implementing the new DC1, which should massively improve the acceptance in

Run	Configuration Change	Date
1	Comissioning Run	3/12-4/12
2	New lower DC3 and H1,2 PM base swap	
	Roadset 57	06/25/2014 to 08/20/2014
	Roadset 59	08/20/2014 to 09/03/2014
3	Standard Maintenance Only	11/13 to 9/14
	Roadset 62	11/08/2014 to 01/14/2015
	Magnet Polarity flipped	01/14/2015
	Roadset 67	01/25/2015-06/19/2015
	D1 and H1 moved	05/13/2015
	Roadset 70	06/19/2015 to 07/03/2015
4	DC 1 added, old one removed	11/2015 to 03/2016
	Various Geometries and work	11/2015 to 03/2016
5	DC1 repaired, old DC1 returned as DC 0	03/2016 to 07/2016
	Roadset 77	
	Roadset 78	
6	DAQ Upgrade	11/2016 to 07/2017
	Roadset 78	11/2016 to 07/2017

Table 2.1: A table of important dates and changes over the course of an experiment.

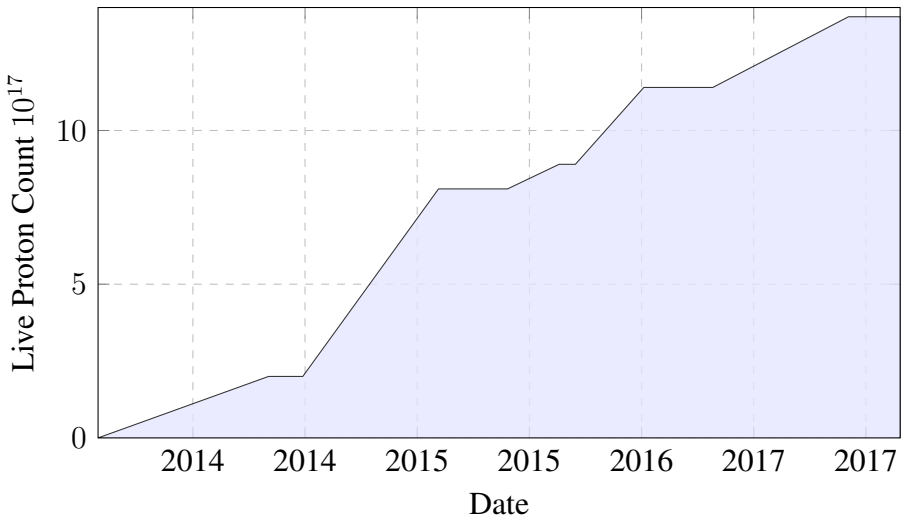


Figure 2.3: Graph depicting the approximate live or usable protons delivered to SeaQuest over the course of the experiment as determined by the QIE sum. The first run displayed is Run 2 and the last Run 6. The slopes are determined by the numbers of good spills and beam intensity.

high x_2 and mass due to the chamber's larger size and greater precision. Runs 5 and 6 also employed high purity P, $P > 99.9\%$, D_2 with improved purity confidence. Specifically, the ultra-pure $P > 99.999\%$ D_2 used in Run 2 provides an important check on the rest of the runs with less pure D_2 ensuring correct contamination treatment. The last run, Run 6, utilized an upgraded DAQ system, which drastically reduced deadtime, but has not been fully decoded or tracked as of this thesis. Therefore, the current analysis will be based on data from Runs 2, 3, and 5.

Most of the live protons were delivered during run 2 and 3 as shown in figure 2.4. However, the data in the latter runs tends to have a higher fraction of useful events. The intensity in run 5 and 6 was much lower with a slightly smaller duty factor, or beam quality, shown in figure 2.4. The trigger upgrades improved the high mass dimuon purity while the chambers increased the acceptance. Paired with the continued optimizations, data taking was cleaner and more Drell-Yan dense in time.

As the fundamental spectrometer was based on the design and remaining components from the E866 spectrometer, improvements over time favors E906's data taking ability in extending the measurement.

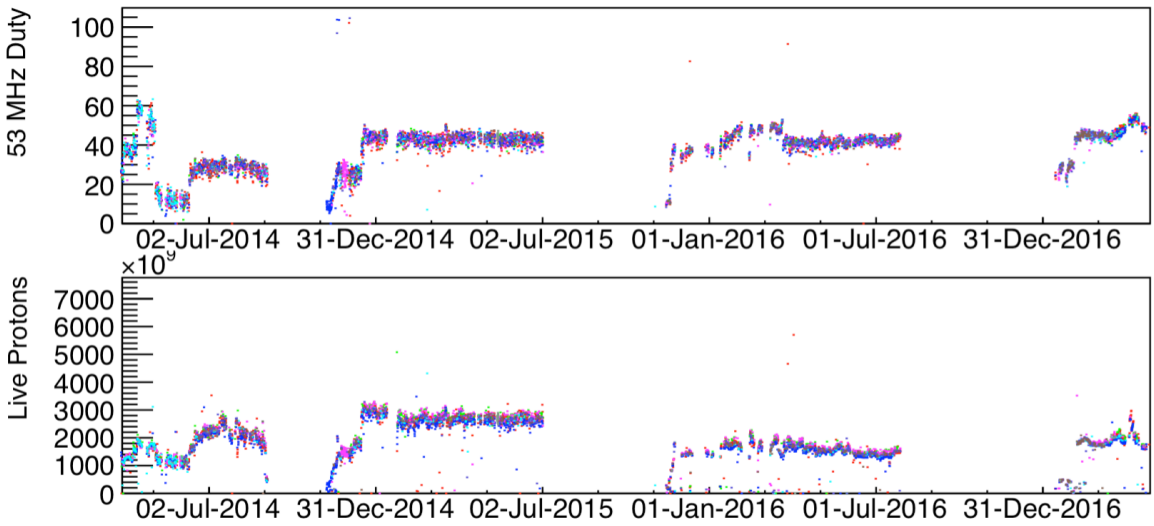


Figure 2.4: The duty factor and live protons are shown on a spill basis over the duration of experiment except the last month of running and starting with run 2. The colors represent various targets, the liquid targets being red and blue. The lower live proton count in latter runs is due to lower intensity, while the duty factor improves from run 2. The start of every run has a lower duty factor and live proton count as the beam is tuned. The figure was taken from Ref. [48].

2.2 Proton Beam

Throughout the entire experiment, SeaQuest held an advantage over the NuSea experiment in the form of its beam. E866 used the 800 GeV Tevatron beam, while E906 used the 120 GeV beam from the Main Injector. As the Drell-Yan cross section is inversely proportional to s and as $1/s \approx 1/E_{beam}$ in fixed Bjorken-x, lower beam can improve the statistical power of the high-x region. In fact, the cross-section is approximately 7 times higher. Our principle charmonium backgrounds, $J/\psi, \psi'$ scale as s as well, thereby reducing this background by the same factor for this beam energy. Similarly, the radiation hazard decreases with energy by about the same factor. These two factors permit a potentially greater instantaneous intensity. The luminosity delivered was approximately 6×10^{12} protons over 5 seconds each minute versus E866's 2×10^{12} over 20 seconds per minute. The intensity was 12 fold suggesting up to a 80 fold increase in statistics for a given x per second than in E866, rather than the initially estimated 50 fold increase. However, a lot of the high intensity data was inhibited or contained bad events making a 50 fold increase

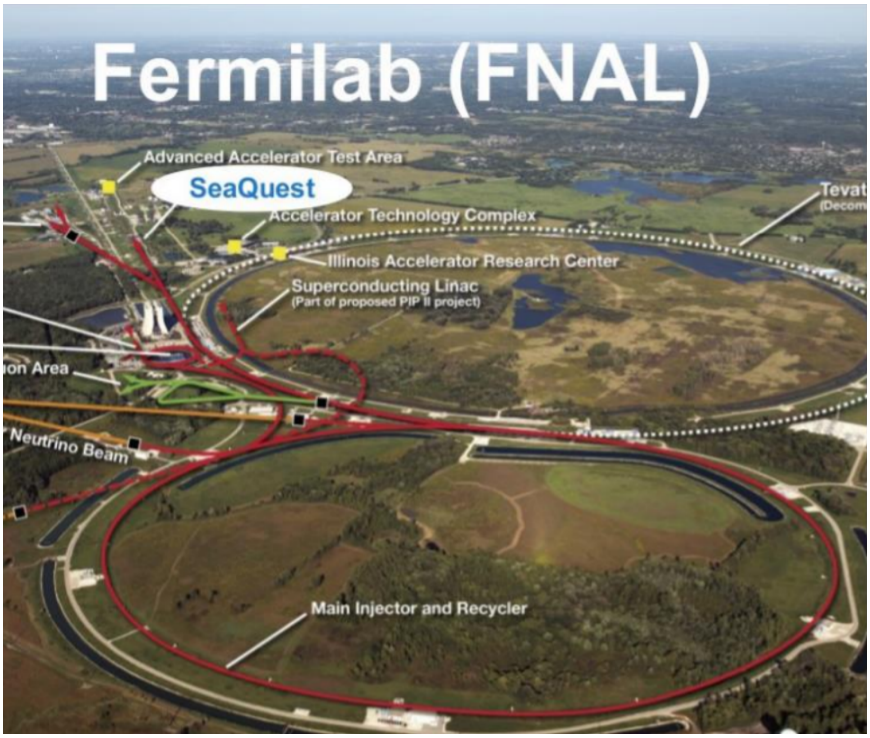


Figure 2.5: An overhead view of the Fermilab accelerator complex shows the two different accelerator rings and the beam path travelled. The figure was taken from Ref. [49].

unrealistic. The data per spill would then be just less than 15 fold from the beam alone, but still could have an order of magnitude increase in statistics for the same run period.

The extracted beam is distributed by the septa to either SeaQuest at NM3 or the muon test beam line from the Main Injector as seen in figure 2.5. This beam was delivered by a 53.1 MHz RF resonant extraction, in the form of 1 cm diameter x 20 cm long proton bunches called "RF buckets" or just buckets with approximately 10^4 protons each. These RF pulses also acted as the clock for the trigger and data acquisition (DAQ) systems in the spectrometer.

The beam contains 588 RF buckets per "turn", with 96 left empty to allow for 8 GeV proton injections from the kickers. There are 369,000 turns in a spill, but this proton number greatly varies due to the poor overall extracted beam quality. This was measured in terms of a "duty factor" $\frac{<I>^2}{<I^2>}$, meaning that the instantaneous intensity is not stable. As mentioned earlier the duty factor generally improved over the course of the experiment starting as low

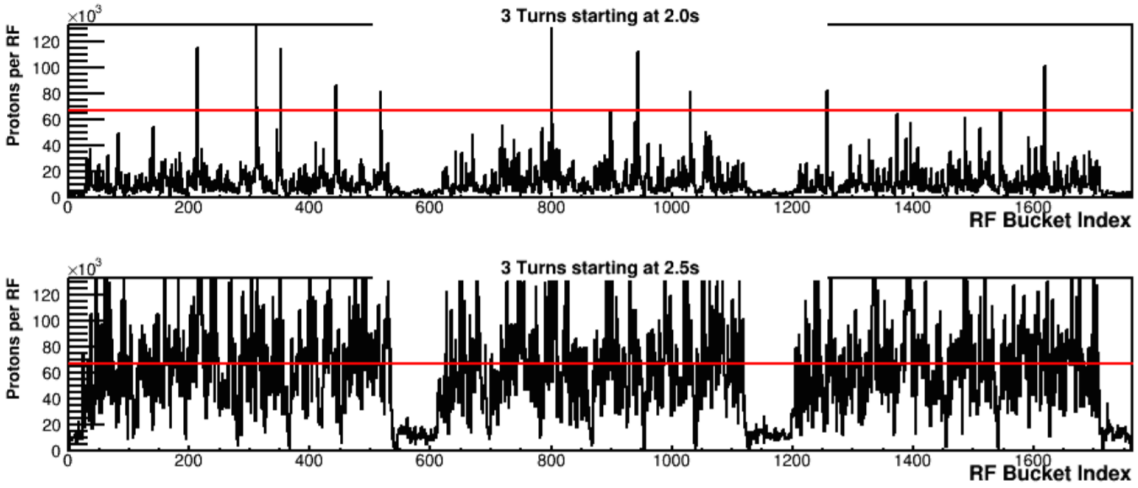


Figure 2.6: Charts taken directly from the Beam DAQ measure the protons per bucket from the BIM readings during the same spill. The top shows a low duty region and the bottom a high duty region with almost entire turn inhibited, the threshold represented by the red line. The figures were taken from Ref. [46].

as about 28%, averaging 43% over the end of Run 2 and during Run 3, exceeding 55% in 2016.

Some of these high-intensity buckets filled nearly the entire detector with hits, and were dubbed "splat events" according to their appearance. While splat events have a high probability of triggering the system, muon tracks cannot be reconstructed from the detectors. Discovery of these events necessitated the addition of a Beam Intensity Monitor (BMI) upstream of the spectrometer.

2.3 Beam Intensity Monitor

The beam intensity monitor's purpose was to give a live picture of the beam intensity in each bucket. Live measurements of intensity not only impact the analysis of the events but also inhibit the data acquisition for a ± 9 bucket window around high-intensity RF buckets. Data acquisition is halted whenever the intensity goes above a threshold, set between 65-95 thousand protons. The inhibit prevents the highly variable beam from filling the data with unusable events while blocking good events with dead time.

The BIM measures this intensity using a Cherenkov counter filled with 80% Argon and

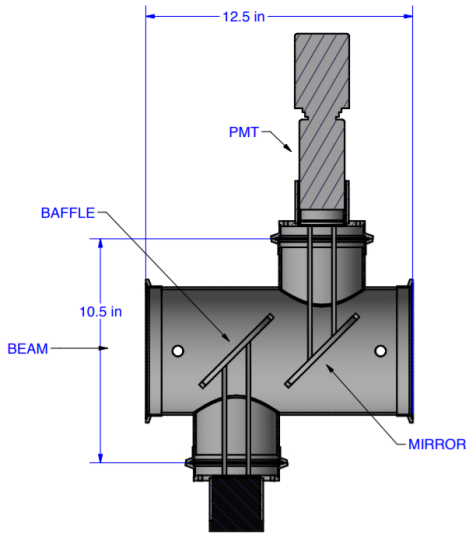


Figure 2.7: A 3D CAD model of the Beam intensity monitor is depicted. Note that the beam passes directly through the mirror used to reflect Cherenkov Radiation. The figure was taken from Ref. [46].

20% Carbon Dioxide gas mixture at atmospheric pressure. A schematic of the BIM apparatus is shown in figure 2.7. A Kapton mirror reflects the light into a Photomultiplier Tube (PMT) connected to counting and digitization electronics. A custom Charge Integrator and Encoder (QIE) circuit, originally built for the LHC, form a highly linear system with a time resolution of 1% QIE bin size uncertainty over a dynamic range of order 10^5. These properties enable the system to provide a snapshot of the beam intensity close in time to the trigger for 16 buckets before and after the triggering bucket despite extremely large variations or spikes in intensity.

The actual reading has to be normalized using the secondary emission monitor upstream in the G2 beamline building, dubbed G2SEM. G2SEM returns the integrated number of protons for each spill and was calibrated by Fermilab applying the activation of a thin foil. The normalization factor drifts over time as the mirror degrades. The mirror reflectively deteriorated so much that it had to be replaced during Run 5 and 6.

The integrated QIE sum was not just employed to normalize the response, but also to determine the inhibited and dead proton count. The inhibited count is defined as the measured number of protons impacting the target when the DAQ was inhibited by the BIM.

The dead proton count is the number of protons delivered while the DAQ was recording data and unable to measure new muons. These counts were needed to record the number of protons on each target which the spectrometer was able to record. This count became the primary means of measuring the protons colliding on each target known as the live proton count. It was required to normalize the cross-section extractions.

2.4 Target

After the instrumentation package, the beam impinges upon the target system, where the sought dimuon pairs are created. The targets are just upstream of the beam dump, set-up in the beamline, and surrounded by shielding. As 90% of beam interactions occur in the dump, the dump itself has a thin 5x25 cm cylindrical hole excavated along the beam axis. This hole improves the vertex separation of events generated in the dump from those originating in the target.

The targets consist of a series of different targets which may be divided by physics goal and type. As shown in Figure 2.8, these are:

1. Solid targets made of iron, carbon, and tungsten to measure nuclear effects
2. Cryogenic liquid targets of hydrogen and deuterium to measure the flavor asymmetry
3. An empty flask and a no target setting in order to estimate non-target background contributions to the target dimuon measurements

All of these targets sit on a large movable table. The table determines the current target in the beam, regularly moving the targets to partition beam time according to the targets' relative density as the experiment is measuring cross-section ratios between them. The liquid target spill counts are kept in parity in terms of expected number of interactions to minimize statistical error. Consequently, LH_2 receives twice the number of spills per cycle as deuterium has twice the nucleon number and about twice the number of interaction

lengths. The nuclear targets have spills based on their physics goals as well. The Carbon target has combarable interaction length numbers to Deuterium as shown in table 2.2. Meanwhile, the Iron and Tungsten targets have similar number of interactions lengths and thus each get one spill. The regular rotation also prevents changes or drifts in the spectrometer instrumentation or beam changes from disproportionately impacting one target. Potential corrections for such changes will then generally cancel in the final result when taking the ratio, even if those changes are unknown.

Target Material	Density (g/cm^3)	Length (cm)	Number of Interaction Lengths	Spills/Cycle
Liquid H_2	0.071	50.8	0.069	10
Empty Flask	–	–	0.0016	2
Liquid D_2	0.163	50.8	0.120	5
No Target	–	–	0	2
Iron	7.87	1.905	0.114	1
Carbon	1.80	3.322	0.209	2
Tungsten	19.30	0.953	0.096	1

Table 2.2: Target materials are shown in the table along with their associated density, interaction lengths and the number of spills the target receives per table motion cycle. The targets are ordered corresponding to their relative position on the target table.

The table position is varied by a stepper motor driving a lead screw along a set of rails. The stepper motor is capable of performing a single $2.54\mu m$ step, to ensure that the system precisely and reliably positions the targets. Telemetry of the table positioning employs a large number of analogue and digital input/output channels monitored by a Siemens APACS+ programmable logic controller (PLC). The target positions were monitored by a magnetic proximity switch, to prevent net drift of the table positions. The table did require some occasional maintenance over the course of the experiment. This required the removal of the entire target system and repositioning the sensors. A precision camera was implemented to guarantee precise functionality after resetting the table.

Autoradiography of the titanium windows and solid target disks has shown that the beam was positioned within $5mm$ of the target center. The beam fell well within the target

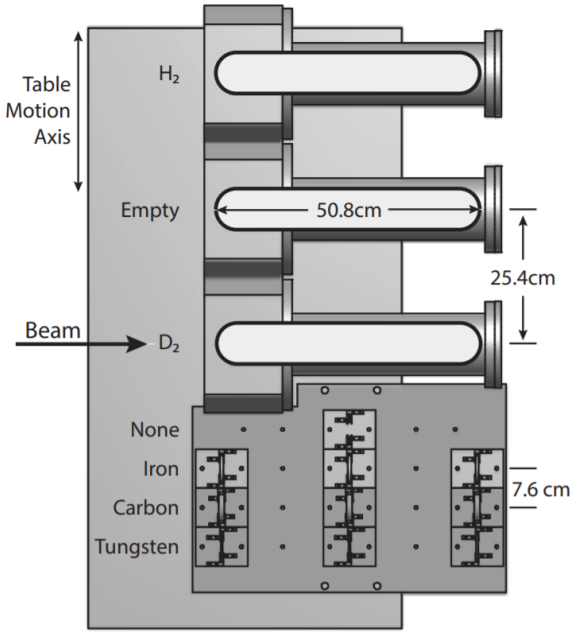


Figure 2.8: A 3D scale model of the target table shows the outer flask dimensions and the arrangement of the targets. The solid targets each use 3 disks. The figures were taken from Ref. [46].

area for all targets, as the solid disks and liquid volume exceeded the most extreme beam drifts possible.

The targets were designed to be comparable in terms of interaction length and average dimuon vertex. The nuclear targets were split into 3 disks, each 1/3 of the target thickness listed in table 2.2. The disks were set $25.4cm$ apart to approximate the spatial distribution in the liquid targets, except for the iron disks during the second run, which were $17.1cm$ apart instead. The rest of the target properties are shown in table 2.2, including the cryogenic targets.

The cryogenic targets necessitated specific manpower and monitoring to be allowed to run. These targets represent one of the primary but unavoidable safety concerns in the experiment. Not only were they filled with explosive hydrogen, but a rapid expansion of the 2.2 liters of LH_2 contained in the 50.8x7.62 cm flasks would be the equivalent of approximately 8 kg of TNT.[50]

Therefore, Fermilab mandates people to monitor the cryogenic targets during running, in addition to several back-up systems and regular maintenance. The back-up systems in-

volve passive redundancies of the main control system, which vent the target under certain conditions. These systems are checked regularly when not running and at least once after filling the targets to ensure proper functionality. The last line of defense is akin to a giant pop-off valve, which opens when the pressure exceeds $10\ psi$. However, the valve brings the volume in thermal contact with the air, subsequently causing the gas to continue to warm and expand safely until the entire target flask is empty. This relief valve was exchanged between run 5 and 6 to increase the capacity which was an involved process due to space restrictions and radiation exposure limitations.

The other safety measures involve the use of automated interlocks to stabilize the target before it has to be completely evacuated. The automated system applied readings from a series of sensors and was installed on the same PLC as the table. The sensors included a series of hydrogen certified Serta pressure gauges, several thermometers on and around the cryogenic flask, and resistance based level sensors inside the target cells. All of these sensors were associated with at least one interlock. The prioritized purpose was emergency warming procedure, and warning, but the system also had some programmed responses should certain conditions be met. For example, if the flask pressure exceeded a set level, the system would release small amounts of gas and turn down the heater until the pressure returned to safer levels. Because the target is a vital component with an electronically controlled safety measure, the PLC and other target electronics were powered on an independent line supplemented by large capacity UPC's. The system is capable of warming safely under electronic observation even when power is lost. Small lapses of power won't even force the system to warm or leave running condition.

The interlocks were intended to be contingency measures. The system was primarily maintained through a feedback loop on the PLC, adjusting itself based on the current measurements. The temperature and pressure readings from the flask determine the average current supplied to three 500 Ohm resistors. These resistors heat the target flask, and constantly have a small current applied as the cooling exceeds the environmental warming.

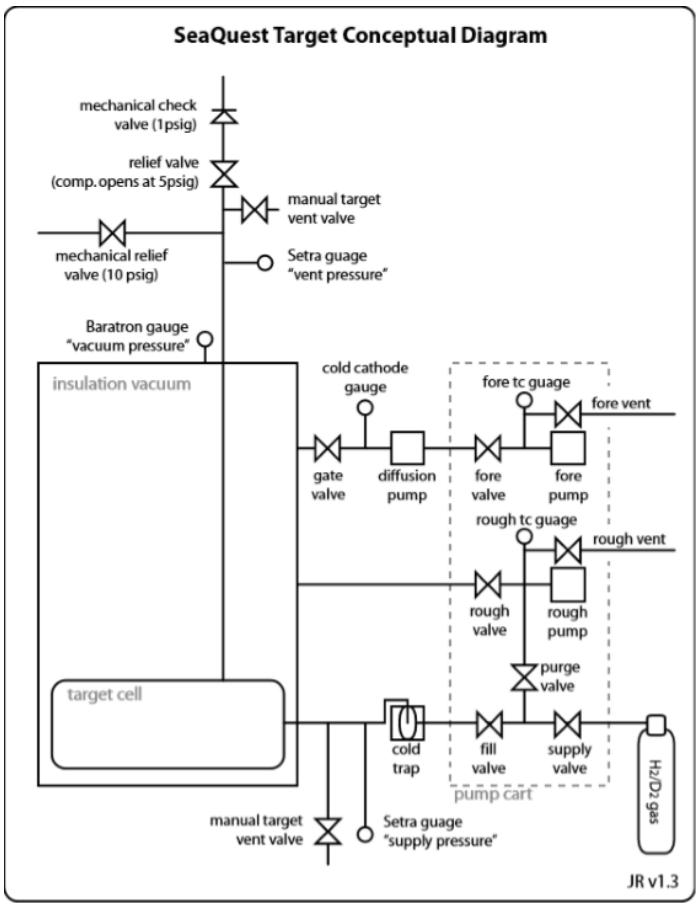


Figure 2.9: A simplified conceptual flow diagram of a SeaQuest cryogenic target shows the target cell and insulation lines connected to the pumps and gas lines. Notice the separation between the two different parts of the system. During running, the roughing lines are isolated from the insulation vacuum and flask volume. The figure was taken from Ref. [51].

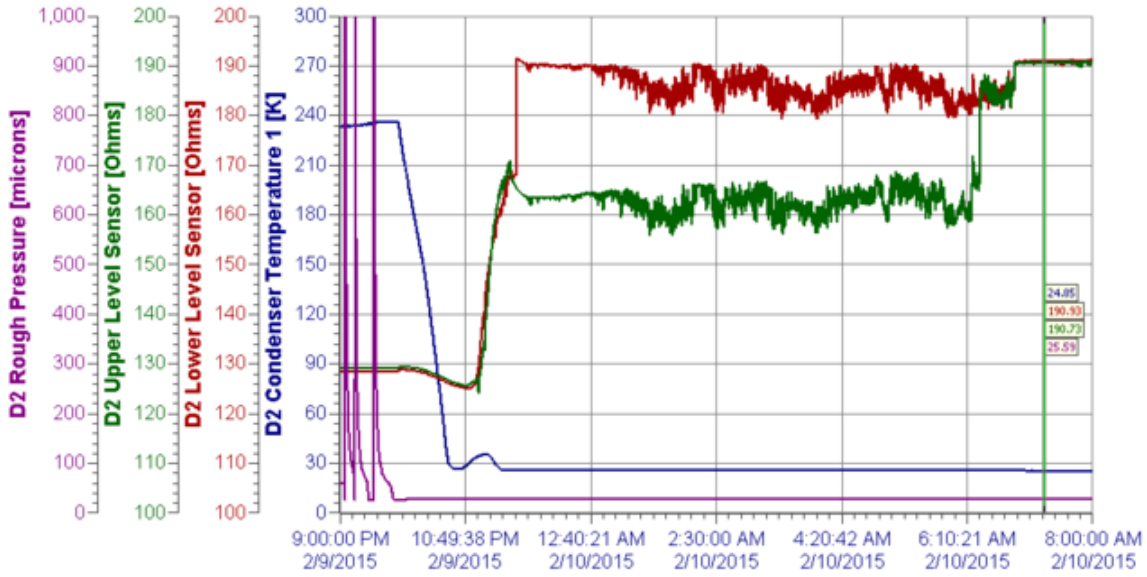
The feedback system maintains a stable temperature of about 24 and 22 Kelvin for D_2 and H_2, respectively, and a pressure of 700-900 Torr. The pressure is kept slightly above one atmosphere to preserve the target purity during venting, though contaminants should freeze before reaching the actual liquid. The pressure and temperature readings are also sent to a control computer with an iFix workspace GUI. The workspace enables shift staff to control valves or heating to respond to issues or even build historical plots. A sample of the live plots is shown in figure 2.10. The control computer has an alarm system which signals for human intervention before interlocks are activated.

If the liquid targets do evaporate, the entire system has to be flushed and re-cooled. Cooling down is an extensive process, taking between 12 and 18 hours after the system

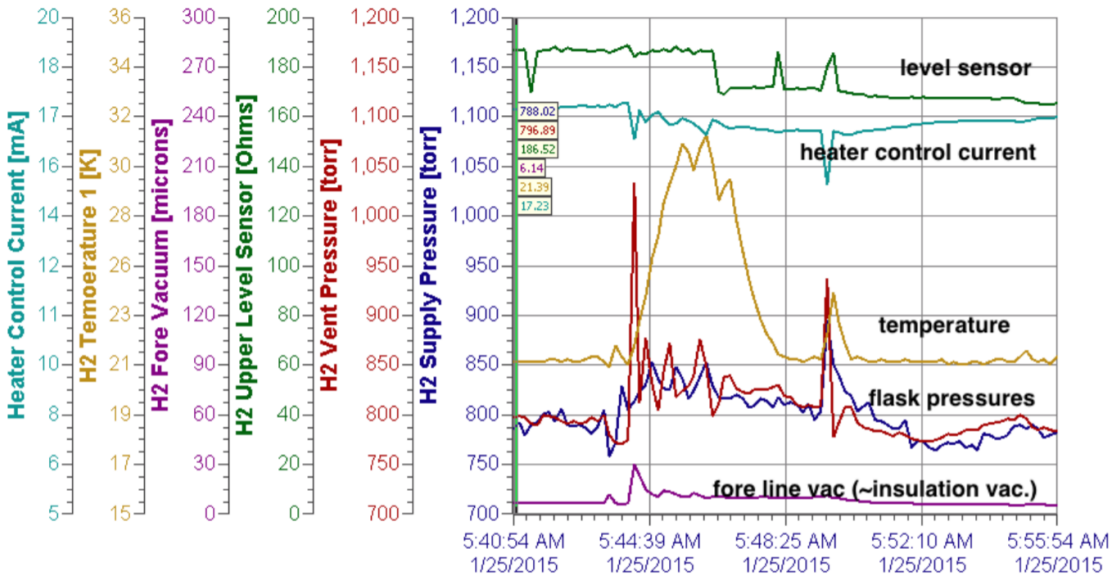
is completely warmed. The process begins with pumping out the air and other gasses in the system using roughing pumps shown in figure 2.9. Next, a series of H_2/D_2 gas flushing and purging cycles clean the system. After the last flush is sufficiently vacated, the diffusion pumps are activated. These pumps combust at atmospheric pressure, but are capable of pumping the insulation vacuum encompassing the target flasks to below 10^{-5} Torr if engaged at sufficiently low pressure. The target required the insulation vacuum to thermally decouple it sufficiently from the environment, in order to reduce heat load and maintain liquid temperatures. The H_2 and D_2 gasses pass through a LN_2 cold trap to freeze out impurities, before they flow into the actual target flask. After that, the flasks are cooled with a helium refrigerator, capable of 25 Watts of cooling power at 20 K. The refrigerator employs a water-cooled compressor pumping on a cold head outside of the target area. The cryogenic cooling system needs to be monitored with installed instruments to ensure normal functionality. Now, the slow process of cooling followed by liquefaction begins, requiring target specialists to monitor the system continuously. Resistance level gauges assist in determining when the flask is fully filled, and ready for running. The sensor responses of a full filling procedure are shown in figure 2.10(a).

The target is operated along the vapor-liquid saturation curve, from which the average liquid density over time can be calculated. Variations in the temperature and pressure estimate the associated uncertainty. Despite active efforts to sustain cooled targets, the targets had to be regularly shut down and warmed up during official beam shut down periods for maintenance.

Maintenance was made more complicated due to radiation safety hazards. This work included relatively mundane tasks such as constant leak and electric fault checking to rebuilding the gauge electronics and pump systems outside of the cave. In fact, most of the system had to be exchanged or repaired at some point during the course of the experiment. The pump systems were constantly running during data collection. Therefore, they had to be craned out for work on a couple of occasions. All of the gauges and pop-off valves were



(a) Target During Cooldown



(b) Target During Running

Figure 2.10: Plots created by the control computer during the experiment display temperature, pressure and level measurements of the cryogenic targets. Plot a) shows an example of the system variables during a flask filling procedure as reported by the Target Control GUI. Plot b) shows graphs of the parameters during normal running reported by the same GUI. A small pressure spike occurs which was quickly resolved.

(a) Target system still in the cave set for running. (b) Targets on table were taken out for decommissioning.

Figure 2.11: Two photographs of the cryogenic target system. The left panel has the target in the cave during maintenance as it appears while in use. The right panel shows the target craned out and prepared for decommissioning, which makes the system more visible.

regularly calibrated and replaced as well. The workload was exacerbated due to out of production components and minimizing certain replacement parts to decrease radioactive waste.

However, the most challenging work was performed in the "target cave", shown in the left panel of figure 2.11, where the beam strikes the target. Work here was highly time restricted, due to direct beam exposure making this the site of greatest radioactivity and radioactive contamination in the experiment. Tasks had to be planned far in advance and warranted additional Fermilab personnel's assistance. Before a job was started, a full write-up of the job had to be completed, all materials gathered, and personal protective gear planned. Practice runs were performed to ensure quick and efficient deployment in the very confined target cave, which had poor accessibility.

Maintenance inside the cave included many of the same relatively simple tasks as outside the cave. However, the target system also required more advanced work, including cold head maintenance and replacement of heating resistors. These projects necessitated hauling the entire target system down with only a hand crane and working in the enclosed space while the target was in place. When opening target components to air, projects must be completed before water absorption impedes or prevents reassembly. The last undertaking in the target cave was decommissioning. The components were carefully disassembled before the shielding was removed for potential re-use. Afterwords, special craning and tear down of the target to separate components by activity level and purpose while preserving the system as much as possible. The target after removal from the cave grants a rare unobstructed view of the target components, and is shown in the right panel of figure 2.11. Future Drell-Yan experiments could apply similar targets potentially with parts of this experiment.

2.5 Data Acquisition System

The actual measurements of the target dimuons and subsequent recording of data implements a multifaceted detector and data relay system. The integral components can be considered by their chronological application in data acquisition. When a dimuon passes through the detector, it generates a pattern of hits in the detectors spread among four stations. The hodoscope hit information reaches the trigger system prompting a decision. If the event passes, the trigger signal induces the TDCs and Read Out Controllers (ROCs) to record timing information from the drift chambers. The drift chamber information is assembled into tracks in the detector by a tracker, which uses the tracks to derive the mass and momentum of the dimuon pair.

2.5.1 Hodoscopes and Trigger

However, the primary background is almost always present in the spectrometer in some capacity. This background comes in the form of single muons originating from particle decays, primarily charged π and K decay in particular. The event recording process generates between 150-200 μs of dead time. This means that if every track or hit was recorded the data would be flooded by almost exclusively useless background events. Therefore, the trigger must select target dimuon pairs accurately, while deciding in a fraction of a single RF bucket period. Such a trigger must arise out of a position detector capable of reading out extremely quickly.

Hodoscopes

The hodoscopes, which are built from fast response plastic scintillators, fill this role. These detectors are organized into two planes. The X plane hodoscopes are arranged vertically to measure bend plane or X position at the station, and the Y plane hodoscopes are arranged horizontally to measure non-bend plane or Y position. The Y plane is divided in the center along the X-axis. Stations 1 and 2 each have single $X-Y$ planes each with its own PMT, station 3 has a single X plane, and station 4 has two sets of X and Y hodoscopes with shared 2-inch PMTs on both ends. The panel bars were situated with a slight overlap in each plane. The first two stations experience large hit rates and possess a special voltage divider to keep the system response linear with signal while handling all of the incoming data. The modifications of station 1 achieved the most efficient hodoscopes. The plane average efficiencies of hodoscopes vary from 0.959 to 0.978. TDCs digitize input from each hodoscope in four phases to form special fast clocks from the accelerator 53.1 MHz RF frequency. The fast clock enables a 1.117 ns hodoscope time resolution, the TDC only recording the latest hit.

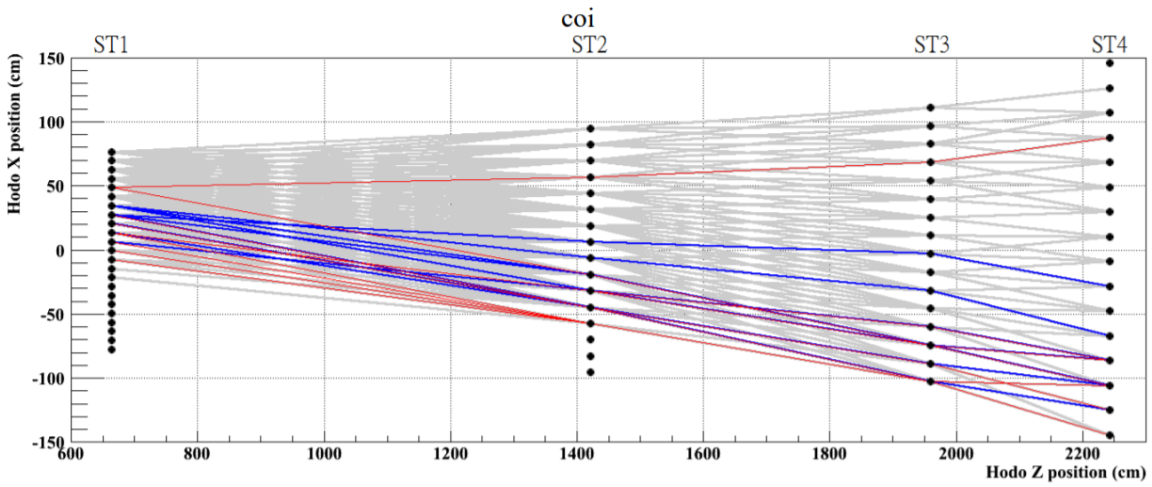


Figure 2.12: An illustration of the trigger roads across several stations is shown. A hit in each station forms a group of potential roads muons would trigger, and the coloring or intensity of the lines represent the expected frequency. The figure was taken from Ref. [52].

Trigger

The hodoscope inputs are also sent to the trigger system separately from the TDCs. In particular, the X hodoscopes form the basis of the trigger. Hits in each plane composed a road. The road represents the rough track a single muon may have travelled. A conceptual example of roads can be seen in figure 2.12. Each plane forwards the binary hit information into a FPGA, or Field Programmable Gate Array, corresponding to its orientation and detector-half. These fast-acting processors determine whether to fire a trigger by applying a logic matrix to a look-up table of these roads. The FPGA signal processing and arrangement can be seen in figure 2.13. Roads are determined on the lower level, while the firing decision is made on the second level utilizing a logic matrix operating on the roads.

The main Drell-Yan trigger relies on a coincidence between at least two potential paths that a positive and negative muon could travel from the target. The trigger dictates one track in the top and another in the bottom half of the detector before firing, without any veto. However, hits from a target Drell-Yan dimuon pair can appear identical to the hodoscopes for some background pairs. Rather, the roads are determined by the analysis of Monte Carlo and background hits. The look-up tables were generally meant to select for high

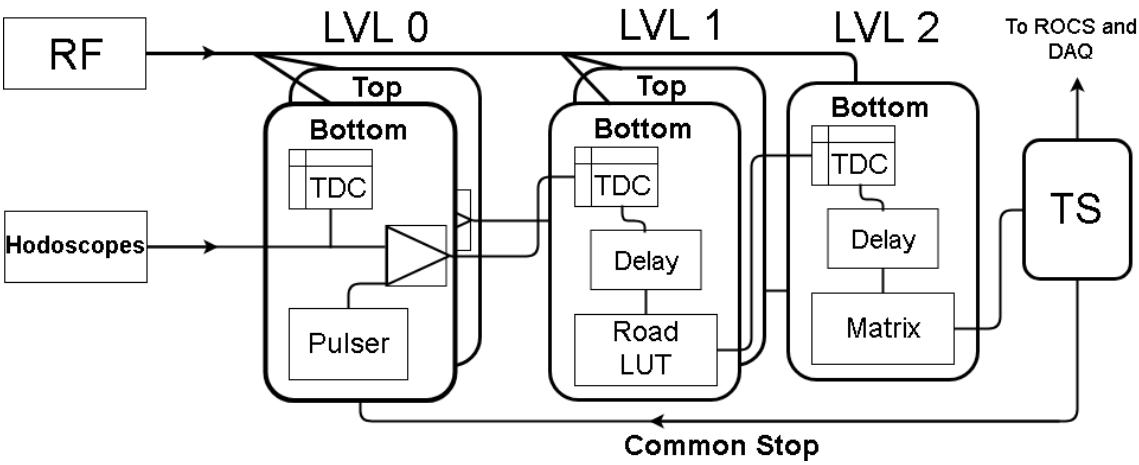


Figure 2.13: A logical flow diagram that attempts to show the FPGA signals passing information between "levels". The level 1 shown only includes the x-axis the system triggered on for the data being analyzed.

mass Drell-Yan events and against "hot" roads with high single track activity. The resulting look-up tables are the aforementioned roadsets from the detector timeline.

There were a few other additional matrices including one for events which occur in just one half of the detector, one with momentum cuts, and one with just single roads for combinatoric background studies. There are also a series of NIM triggers, for cosmic rays tests, and one to take random data to understand the beam and events in the spectrometer at any given time.

Later, special triggers, specifically an experimental Boosted Rule Fit trigger, were developed. This trigger and the same plane road trigger could only be tested and implemented in the last run. These were the first attempts at the implementation of the Level 1 Y-hodoscopes and had a potential estimated 43% increase in acceptance and 75% reduction in the background. Developments such as that show that the trigger was not only a crucial spectrometer component, but a tunable one. The trigger was designed to be flexible enough for reasonable adjustments to the geometry of the spectrometer over the course of the experiment.

Regardless of how it is generated, the trigger signal is sent to the trigger supervisor along with an identifier for the associated logic matrix. The trigger supervisor contains

pre-scale factors, and these factors determine how often a trigger type should be accepted. Pre-scale factors allow the DAQ system to record background triggers without the system being swamped by dead time. During the last run, pre-scaled rates were implemented to record different signal matrices at different rates simultaneously. If the trigger supervisor accepts the trigger signal, the trigger supervisor produces a common stop. The common stop causes all the TDCs and aforementioned QIE boards to hold their current values, while the common stop signals the VMEs to start the readout of the current data. The timing data is passed to the readout controllers which transfer the information to the decoding server and DAQ computer for further processing. In parallel, there is a scalar and beam DAQ system recording the current spill's beam information. The beam DAQ records the beam intensity information from the BIM as well as the information provided by Fermilab's Accelerator division including the G2SEM. The scalar DAQ records the conditions under which the data were taken including the spill, trigger counts and trigger settings. These systems portray vital information about an event, but have large dead-times caused by the readout and data transfers. Therefore, the recording process only initiated with a trigger, which in effect decreases the dead time.

Improving the live time and data quality is crucial to effective data taking, but the hodoscopes are not precise enough for meaningful position measurements. Consequently, the hodoscope and trigger systems cannot derive Bjorken-x or confidently calculate the momentum. SeaQuest requires complementary high precision detectors in order to take these measurements.

2.5.2 Proportional Tubes, Drift Chambers and Tracking

Drift Chambers

Unlike the hodoscopes, the drift chambers measure position with high precision, which enables the determination of the muon's path through the spectrometer as well as the origin of a dimuon pair. Drift chambers rely on the ionization of $Ar : CH_4 : CF_4$ or $Ar : CF_4 :$

$C_4H_{10} : C_3H_8O_2$ gas in a strong electric field, causing an electron cascade in the gas. By measuring drift time, the chambers can resolve position resolution to below 250 μm. The corresponding max drift time is 250 ns, with a single plane efficiency above 90%. In fact, the dimuon mass resolution is constrained not by the detector resolution, but by multiple scattering and energy loss in FMag.

SeaQuest measures the invariant mass in terms of the momentum of the muons. The momentum corresponds to a position in part due to magnetic deflection between the target and the first chamber. Chamber 1, later renamed as 0, and the new chamber 1 are positioned just downstream of the dump. Approximately 90% of the beam will interact with the nearby dump, with some events appearing as if they emanated from the target. In order to resolve the spawning vertex of the dimuons, station 2 was placed right after KMag, which is located between stations 1 and 2. The magnetic field in KMag lies in parallel with that in FMag. The second kick allows simultaneous discrimination of the vertex and momentum information using the tracking information provided by the chambers. Further downstream, chamber 3 measures the path the muon is traveling after the magnet. Additional information from the chamber 3 further contributes to resolving the muon track. The muons diverge further after KMag, spreading over a larger area. Therefore, two chambers are used at chamber 3 to improve the acceptance with one positioned above the other, at the same position along the z-axis.

A single plane of wires can only measure the distance of an event to the closest wire. Each of the chambers instead relies on six wire planes perpendicular to the beam. All chambers had an X plane with wires aligned on the y-axis to measure the x-position, followed by U and V planes offset by ± 14 degrees, respectively. There are three additional "prime" planes needed to resolve "left-right ambiguity" with each one possessing wires parallel to the first three U,V, and X planes. The left-right ambiguity refers to the fact that measuring the drift distance cannot discern if the muon traveled to the left or right of the activated wire. The wires in these additional prime planes are offset by 1/2 of the wire

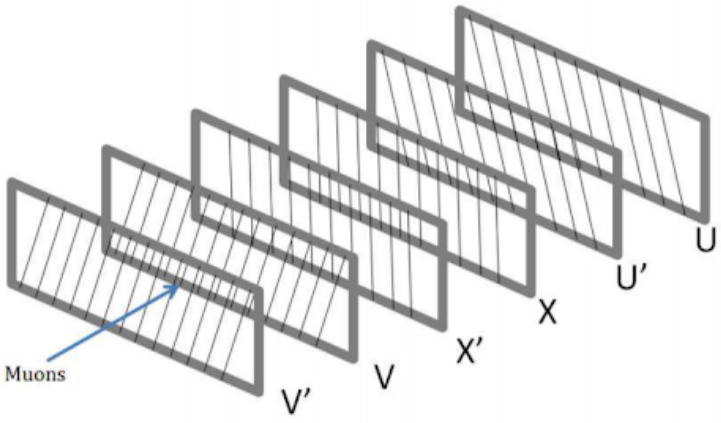


Figure 2.14: Cartoon of the different wire plane orientations used in the drift chambers. The figure was taken from Ref. [47].

spacing on the axis perpendicular to the wires but within the plane. Therefore, a wire in the prime plane is to the left or right of every non-edge wire in the non-primed planes. If at least one wire-plane and its corresponding prime plane fire, the left-right ambiguity may be resolved. Offset prime planes also act as redundancies making the chamber more efficient- if at least 4 hits are made in the planes with at least one in each wire orientation, the hit can be resolved. Redundancies are important as the efficiency decreases with occupancy or activity in the spectrometer. The six wire planes in the drift chambers are thus highly effective at resolving individual muon hits.

Tracking

Connecting resolved hits can estimate the path or track that the muon followed. This is done through specific tracking software. The software has to consider that the detector frequently contains backgrounds. In order to combine tracks, different combinations of points are attempted. The first step is selecting only "good" hits and removing artificial clusters, or additional hits found in a small region with perhaps one corresponding to a muon. Some hits originated outside of a TDC time window and were formed either from a different event, or from after pulses from echoes or electronic ringing. Artificial clusters are primarily generated by one of the following effects:

1. Edge hits cause an additional wire to fire, so the higher drift distances are removed, while the others are kept.

2. Electronic noise over multiple wires produce hits which are removed.

3. Muon induced delta rays are created when a muon transfers a large amount of momentum to an electron. The electron itself causes cascades which act like other charged particles in the same vicinity. In this case, hits on the edges can be kept, the others are removed.

The tracker builds all potential "tracklets" between chamber planes in chamber 2 and 3. Combinations of tracklets roughly facing one another are combined to form tracks. Tracks are required to agree with hodoscope hits on a plausible target trajectory. Then using the magnetic field saggita ratio of hits, the program searches for hits in the first drift chamber in a small radius around the plausible track paths. An example of this is shown in Figure 2.15. Lastly, a Kalman fitter floats the track origin along z trying to find where it best fits with the beamline. This location is used as its origin. Next, it looks for 2 tracks that come from similar origins, performing goodness-of-fit cuts and quality cuts on the tracks. If two compatible tracks are present, the tracker uses the Kalman fitter again by fixing the origin to calculate kinematic properties like momentum and mass. The particles producing the tracks are also identified to ensure they come from muons.

The proportional tubes performed charge and final position measurements, replacing the drift chambers in the last station. They were placed behind a 1m thick iron wall with the station 4 hodoscopes. The small momentum-dependent deflection before and after the iron absorber was the cardinal signature of a muon. All charged particles passing through the detector were confirmed to be muons, and thus the tracker didn't perform particle identification.

The system was calibrated with cosmic rays and J/Ψ events, finding that the proportional tubes had more than sufficient resolution to identify muons. All of these detectors

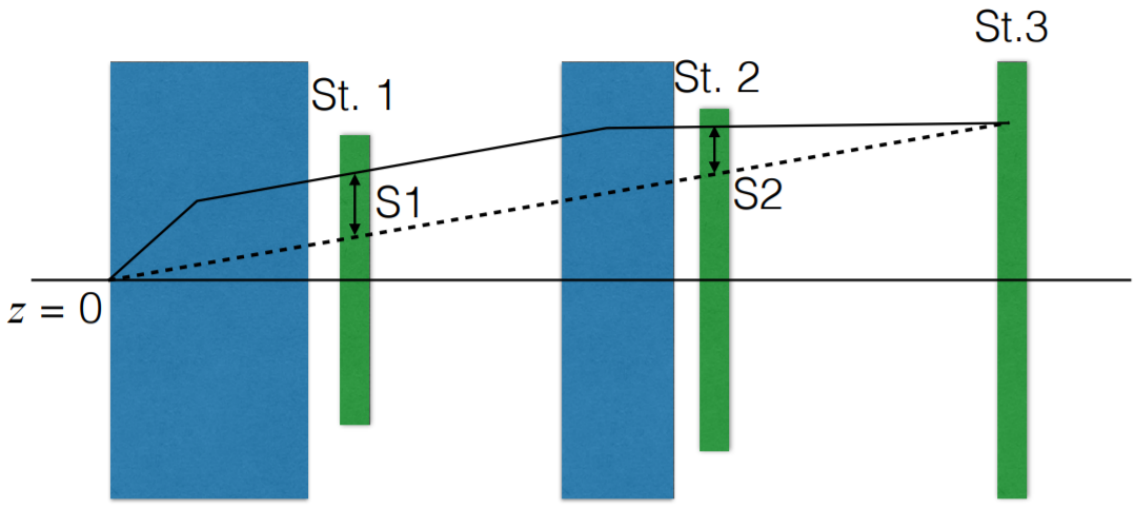


Figure 2.15: Schematic diagram representing the tracker that attempts to connect the drift chamber fits to form full tracks in the detector. The figure was taken from Ref. [47].

enable the spectrometer to select candidate dimuon pairs produced by the targets and measure them.

Chapter 3

Analysis and Preparation

The dimuons measured in the SeaQuest Spectrometer can now be implemented in the calculation of the cross-section as a function of target quark Bjorken-x, or x_2. The cross-section ratios of the targets compose the primary experimental physics results from which the asymmetry can be derived. Specifically, the experiment measures the ratio of the cross-sections. By keeping most spectrometer parameters the same, many factors contributing to the target cross-sections and associated errors cancel in the ratio. SeaQuest accomplishes the cross-section measurements by counting the number of dimuons with a particular energy produced by a given target. The ratio is determined primarily by the different target counts, when the counts become the true dependent variable. In essence, SeaQuest becomes a counting experiment.

Anything impacting the counts directly, outside of the target dependent difference in Drell-Yan production, is a source of error. While the primary single muon backgrounds are ignored during acquisition, the vast majority of recorded events are still not Drell-Yan dimuons produced in the targets. Therefore, the data requires cuts to remove background and "bad" events. The term bad event is used to describe any event that becomes unusable for any non-physical reason. An example is loss of beam information or poor data integrity. There will always be a certain fraction of inextricable events left in the data, that can be accounted for by subtracting off the empty target data and extrapolating on the beam intensity dependence of the data. The cross-section ratio can be corrected for any target differences by considering the measured total amount of beam on each target, the purity of the targets,

interaction lengths and any other potential differences which may contribute to the error budget.

Once the cross-section ratio is derived as a function of x_2, the $\frac{\bar{d}}{\bar{u}}(x_2)$ can be extracted.

3.1 Cross-section Ratio Derivations

The etymology of particle cross-sections originates from an early conception of particles as hard spheres instead of a quantum treatment. In a gas of hard spheres, the probability of collisions is proportional to the cross-sectional area of the constituent spheres. The gaseous particle cross-section is

$$\sigma = 1/\lambda n \tag{3.1}$$

with mean free path, λ or the average length the gas travels before collisions and a sphere density of n. Unlike directly measuring total probability, cross-section does not scale with the absolute number of particles due to the density term. As a result, the cross-section corresponds to the probability of individual particles interacting.

The terminology survived the gaseous treatment of particle physics and has no direct correspondence to a particle's physical area in quantum mechanics. The dependency on the probability of interaction does carry over from the classical analogue into the particle physics cross-section though. The cross-section does not depend on the target particle density or beam luminosity and is applicable to different accelerator experiments regardless of target or beam. Instead, interaction cross-sections are defined as the scattering amplitude of the interaction integrated over the solid angle or

$$\frac{d\sigma}{d\Omega} = |f_{i\to f}(\theta, \phi)|^2. \tag{3.2}$$

The scattering amplitude of an interaction refers to the probability of finding the defined final product states f, at a given θ, ϕ, when beginning from the initial state i. The cross-section is the probability of finding the final products at any angle. Dependencies of the

scattering matrix that do not explicitly relate to the outgoing solid angle, such as Bjorken x, may carry through to the total cross-section function. Therefore, $\sigma(x_2)$ is a valid expression of the cross-section as a function of x_2.

The SeaQuest spectrometer is only capable of counting the number of dimuons within its acceptance region and not the cross-section directly. The detector acceptance region defines the solid angle the probability is integrated over. The probability of an interaction can be estimated from the total interaction divided by the number of trials or $P = N_{success}/N_{total}$. The number of trials is the number of potentially accepted interaction opportunities measured by the number of beam protons multiplied by the average number of target nucleons with which they are capable of interacting. Taking the beam area to be within the target area, the interaction volume is the beam area multiplied by the target thickness. The averaged active target nucleon count is the number of atoms in the active target volume and can be parameterized by the target density and length of the target interacting.

Therefore, the cross-section in a thin volume of target as measured by the detector would take the form

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}(\theta,\phi) = \frac{N_{out}(\theta,\phi)}{nz\Delta\Omega N_{trial}}, \tag{3.3}$$

where z is the thickness, $\Delta\Omega$ is the accepted solid angle portion, and n is the number density with $n = \rho N_A$, N_A being Avogadro's number and ρ is density.

As the beam particles interact with the targets, the beam attenuates in a dz portion of the target according to

$$\frac{d\Phi}{dz} = -n\sigma\Phi, \tag{3.4}$$

with beam flux Φ, target density n, and total cross-section of all interactions σ. The beam flux as a function of z is then

$$\Phi(z) = \Phi_0 e^{-n\sigma z}. \tag{3.5}$$

The mean interaction vertex is solved from the first moment of Eq. 3.5 and represents the effective target length or nuclear interaction length. Dimuons can be treated as originating

from the mean interaction vertex. However, the mean interaction vertex is different in the hydrogen and deuterium targets. Therefore, these terms do not cancel in the ratio and must be considered.

Compiling these components yields the cross-section ratio

$$\frac{\sigma_{LD_2}}{2\sigma_{LH_2}} = \frac{\epsilon_H N_D \rho_H \lambda_H P_H}{\epsilon_D N_H \rho_D \lambda_D P_D}, \tag{3.6}$$

with efficiency ϵ, nuclear interaction length λ, number of protons P, dimuon count N, and the indices are H for liquid hydrogen and D for liquid deuterium. $\Delta\Omega$ is almost equal in the numerator and denominator. The small vertex difference creates a small aberration in the solid angle of the detector, but this difference is far smaller than the detector resolution. Therefore, any difference will be absorbed into the relative efficiencies, as other effects will tend to dominate it. The ρ and λ variables are determined by the known target properties. This leaves the efficiency and counts per proton as the actively measured variables in the experiment. The counts are separated by their x_2, measured during data acquisition.

3.1.1 Dimuon Yield

The Drell-Yan dimuon yield is the principle quantity being measured in the experiment. Ensuring this is accurate entails the use of cuts and fits to remove the large quantity of dimuons outside of the target Drell-Yan category. These backgrounds track as dimuons, vastly outnumbering the signal events. These events will not have the same seaquark sensitivity, and instead will be dominated by the relative intensity of the two cross-sections.

Quality Cuts and Run Removal

The "raw" tracked dimuon yields are mostly background events and have no guarantee that the event is usable. Some of the tracked dimuons may be afflicted by data acquisition faults or possess nonsensical values making the corresponding measurements unreliable or

unusable. The tracker does not access statistics of the beam, targets or any error or issue which may have occurred. Thus a series of quality cuts on data taking "blocks" and spills have to be performed. This particular work will adopt the "block" nomenclature, defined by a block of data which is split by CODA reset, a DAQ program which records the raw data to prepare it for decoding. These must be constrained in size due to fundamental system limitations. Other SeaQuest publications may re-use the term run for blocks, but this was selected against for clarity.

Changes to the spectrometer state, firmware, or trigger require a DAQ reboot. Each block has spectrometer conditions recorded for that block. The DAQ components are re-booted in between blocks and DAQ faults. Error procedures require the reboot of the system as the primary troubleshooting step, building natural splits between spectrometer good and error states. The first cut requirements are there to ensure events exist in good blocks. Within the cuts, the first check is that all the data was successfully recorded and/or decoded without error flags (miscellaneous error warnings) thrown by the program or applied manually by a collaborator.

Quality cuts are carried out on both the spill and block level, where anomalous or missing values for magnetic current, run settings or other values are removed for integrity. Duplication of those values entails a decoding or recording error for example. A list of these cuts is shown in table 3.1.

Some of these errors were recovered by simple re-tracking or decoding issues, while others had individual investigations to resolve bespoke issues for individual blocks. An example was timing errors where simply adding an offset to the recorded tdc times allowed for tracks to be recovered. Recovered data sets still had to pass these cuts to be included.

Furthermore, some runs were considered "bad" based on run conditions recorded during data taking or anomalies in the data. With controls on multiple levels bad or simply suspicious events could be effectively removed ensuring data integrity before analysis.

Table	Value	Good range
Spill	targetPos	[1, 7]
	spillID	$\neq 0$
	spillID	no duplicates
Target	TARGPOS_CONTROL	Spill targetPos
	spillID, name	no duplicates
Scaler	TSgo, EOS	[1,000, 8,000] [100, 6,000]
	acceptedMatrix1, EOS	[1e3, 8e3] [100, 6e3]
	acceptedMatrix1, BOS	Entry exists
	afterInhMatrix1, EOS	[1e3, 3e4] [100, 1e4]
	acceptedMatrix1/afterInhMatrix1, EOS	[0.2, 0.9] [0.2, 1.05]
	spillID, scalerName, spillType	no duplicates
Beam	S:G2SEM	[2e12, 1e13]
	F:NM4AN	> 1000
	spillID, name	no duplicates
BeamDAQ	QIESum	[4e10, 1e12]
	trigger_sum_no_inhibit	[4e9, 1e11]
	inhibit_block_sum	[4e9, 1e11] [4e9, 2e11]
	dutyfactor53MHz	[15, 60] [10, 60]
	spillID	no duplicates
kTrack	# of rows	> 0

Table 3.1: Spill quality cuts applied to the data are listed. These cuts remove the entire associated spill from the data.

3.2 Data Anatomy

The quality cuts in section 3.2 do not address the background issues though. Constraints on background were developed through multivariate analysis before the derivation of the cross-section. The analysis is impossible to perform without such work and has been the subject of much of the total man-hours spent on derivation. Upgrades to this component is what largely differentiates this work from previous SeaQuest results.

3.2.1 Modelling

Signal and background samples are prerequisites to several facets of multivariate analysis. In particular, machine learning requires representative samples for training which becomes the basis for subsequent classification. There are two principal sources for samples: real data recorded during the run and gun Monte Carlo simulations.

GMC -basics

The gun Monte Carlo (GMC) is integral to the data analysis stream. It is supposed to represent a clean or true physics event. The GMC answers the key question "If an event with given properties was generated in a collision, what would it look like in the spectrometer?"

The operating principle of SeaQuest's GMC is randomly to throw events uniformly in the Collin-Soper Frame [19, 53]. The precise frame definition is shown in Appendix B. Each event is then given a weight corresponding to theoretical and experimental cross-section results. This framework holds advantages in that it provides additional counts for rare low rate events, and computationally downplays the more common ones by giving smaller counts with higher weights. As a result, the low x_T regime events are not given as many counts as it's cross-section suggests, while the high x_T regime events of interest have larger counts than expected. The Gun Monte Carlo throws either a beam proton or dimuon

pair. The latter is used for both background and signal samples generated by physics events. The former case represents a background model where individual beam proton interactions can be studied as it passes through different materials. The GMC employs geometry files containing the dimensions of all materials within the spectrometer. The geometry files are fed into the Geant 4 Monte Carlo simulation. These geometry files are constructed from regular surveys of the spectrometer to ensure the spectrometer material volumes are accurate. The muons are tracked as they pass through detectors and the magnetic fields with both true values and tracking measurements taken.

A "messy" Monte Carlo is fabricated utilizing the same set as before, but with supplemental random effects representing detector inefficiencies. Additionally, prescaled "NIM3" random events are inserted into the GMC hit tables. A NIM3 event is taken during a spill determined by randomized time values rather than detector activity to represent real events. These events may even be empty, but do have a corresponding intensity. The "real" effects are derived from efficiency measurements and special data taking runs for this purpose. Detectors were read as if real, allowing the simulated muon tracks to act as real data with the same variables available. The final messy Monte Carlo well represents muon readings as they would appear in real data.

The Gun Monte Carlo simulations were initially directly implemented to determine the trigger roads and to cross-check the tracker's ability to accurately deduce the muons kinematic properties and efficiency. GMC's are separated by parent particle species and generation location which determines the initial position assigned to the dimuon. Properties of thrown or generated muons are determined for a given physics event according to either an assigned distribution or the relevant cross-section. Assigned distributions diverging from real cross-section kinematics are assigned corresponding weight factor distributions. When events are multiplied by their corresponding weights, the distributions then map to the true cross-section or real distributions. The weighting scheme allows for a simple correction technique, where data taken from the current experiment can be directly applied to the

Monte Carlo.

3.2.2 Drell-Yan GMC

The signal sample used in the analysis derives from a target Drell-Yan GMC. Drell-Yan events are thrown with a uniform mass distribution between 2 and 10 GeV and a uniform x-Feynman distribution

$$x_F = \frac{2p_z\sqrt{s}}{1-\frac{M^2}{s}} \approx x_1 - x_2. \tag{3.7}$$

Dimuons are assigned an outgoing probability implementing the $1+cos(\theta)^2$ form present in the cross-section, thus throwing uniformly random in $cos(\theta)^2$. The events are then assigned a weight based on relative cross-section in the selected kinematic range with the k factor discussed in section 1.2.3.

The p_T distribution generates a weight factor according to

$$w(p_T) = \frac{p_T}{(1+(p_T/(2.8^2 GeV/c))^6 GeV/c} \tag{3.8}$$

based on current knowledge of the cross-section. However, there is known to be an experimental dependence of p_T on x_F [54]. A simple correction can be employed based on the measured data in the experimentally measured functional form. The cross-section dependence on p_T now follows

$$\frac{d\sigma}{dp_T} = \frac{p_T}{(1+(p_T/f(x_T))^2)^6} \approx \frac{p_T}{\left(1+(p_T/\frac{\delta p_T}{\delta x_T})^2\right)^6 GeV/c}. \tag{3.9}$$

The $\frac{\delta p_T}{\delta x_T}$ parameter is derived from the slope parameter m in the fit $p_T = mx_T + b$ with $m = -0.9 \pm 0.1\ GeV/c$ and $b = 2.49 \pm 0.06$. Finally the weights considered the beam attenuation and interaction with the targets when comparing vertices.

3.2.3 Background Separation

As discussed, SeaQuest suffers from a high background rate as a consequence of the relative rarity of the sought events. The Drell-Yan cross-section depends on a QED process involving seaquarks, while the primary scattering products found in the detector were the result of QCD valence interactions. The relative production ratios are proportional to $\alpha^2_{QCD}(s)/\alpha$ and square PDFs. The result is Drell-Yan dimuons occurring at approximately $1 : 10^7 - 10^8$ the rate of other muon sources. Furthermore, 90% of the beam interacts in the dump instead of the targets themselves.

The trigger attempts to filter as much background as possible and the tracker tries to restrict the contributing dimuons to within the target region and single muon contributions. Most of the tracked data is still either bad or background events. Background events are primarily separated from good target dimuons through physics and machine learning cuts. There are multiple ways to categorize populations useful to classification. Three main background classes chosen for convenience and naturalness. These are

1. Dimuon events from processes other than Drell-Yan events,

2. Dimuon events originating outside of the target, including true Drell-Yan events,

3. Combinatoric events made of two single muons passing through the detector simultaneously.

The primary dimuon sources in the spectrometer's acceptance phase space are the J/ψ and ψ' charmonium ($c\bar{c}$) mesons. This is due to the J/ψ muonic decay channel, which has an unusually high branching fraction of 5%[55]. J/ψ hadronic decay channels are heavily suppressed by the OZI rule, resulting in this relatively high lepton branching fraction [1]. Consequently, this charmonium state possesses an extremely tight decay width of 92.9 ± 2.8 keV around its approximate mass of $3.096900(6)$ GeV [55] as well as being a large dimuon source.

The related ψ' particle is an excited 2s charmonium state with a greater mass of 3686.097 $\pm 0.025\ MeV$ and a decay width of $294 \pm 8\ keV$ [55]. This width is several times greater than the J/ψ state, but still relatively narrow. The ψ' particle almost always decays into the non-excited J/ψ state plus a photon, yet can still decay into $e^- + e^+$ or $\mu^- + \mu^+$. Muons and electron each have a branching ratio just under 1%. As the ψ' particle's mass is closer to the kinematic range of interest, it has a higher acceptance rate than J/ψ. The J/Ψ peak is still larger, but the relative proportion of ψ' accepted is greater making the peak appear abnormally large. Consequently, ψ' is more difficult to differentiate from signal events, making it more troublesome charmonium backgrounds.

The impact of these backgrounds may be estimated employing the GMC again. However, the J/Ψ and Ψ' dimuons are generated differently than Drell-Yan dimuons. The simulation instead employs the particle invariant mass values and the dimuons are thrown uniformly in θ. The physical properties such as the branching ratio and mass are sourced from the particle data book [55], while the applied p_T weight is taken from an external empirical distribution [56].

Relative peak positions of J/Ψ and Ψ' were used to derive the beam position and ensure the magnetic fields are correct. The relative peak positions determine any required corrections to the magnetic field. When the field is correct, the beam position on the y-axis is derived from the mean measured y-position of the muon vertex. The background/signal representations the GMC provides is a vital component to making cuts and differentiating dimuon pairs.

The simplest and most decisive way of separating these background sources is a mass cut at 4.2 GeV. Only dimuons with mass above this boundary are considered in the final count. Cuts like this mass cut, motivated by physical reasoning are labeled physics cuts. Based on the detector resolution, this cut removes all but an estimated $3\% \pm 1\%$ of J/Ψ and $9.5\% \pm 0.5\%$ ψ', respectively, which can be seen in figure 3.1. The ψ' particle has a dramatically greater bleed into this higher mass region than J/Ψ due to its greater width

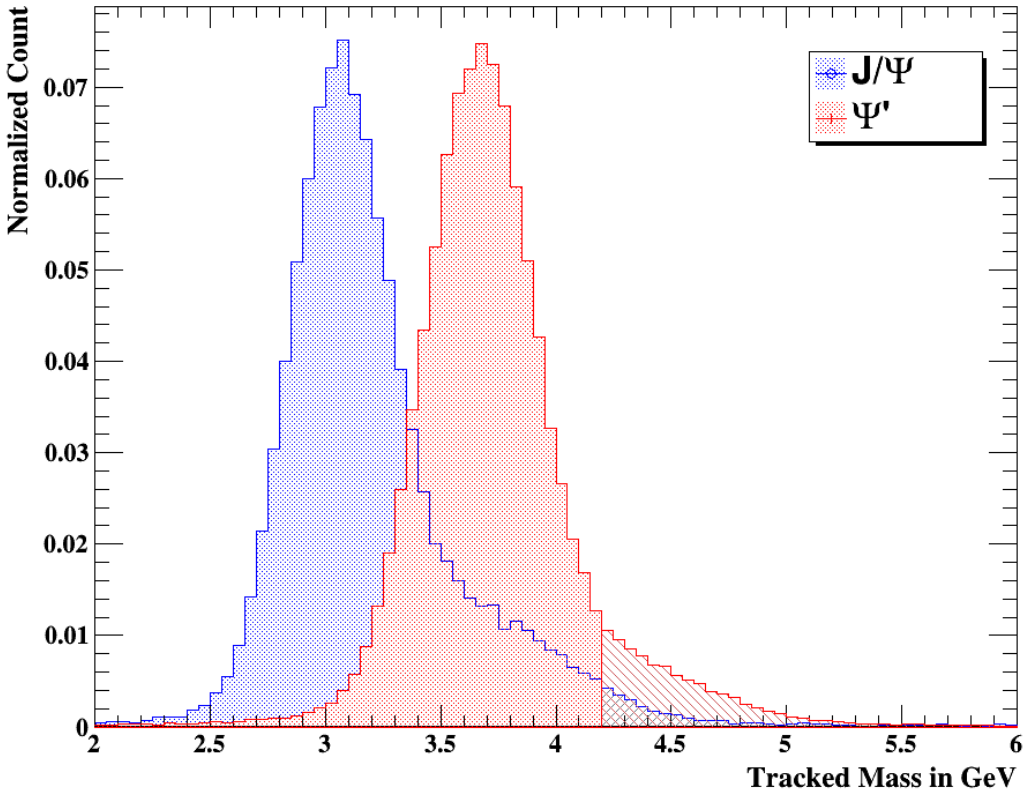


Figure 3.1: Simulated J/Ψ and ψ' mass distributions are estimated by the tracker. No cuts have been applied to these distributions with the mass cut at 4.2 GeV represented by the greyed region. Each distribution was normalized by the its integral to form the normalized count, thus the relative peak sizes and counts are unphysical.

compounded with the detector resolution.

There still exist Drell-Yan events in the mass region below the 4.2 GeV cut. However, as mass is proportional to x_2, few carry sufficient x_2. The potentially useful events that are present have high mutual information or overlap with background sources. There were attempts to recover events, but the resulting misclassification error out-weighed the gain in dimuon yield.

3.2.4 Non-target dimuons

The same sorts of physics events occurring in the target occur in the nucleons of materials in the aforementioned volumes as well. Dimuons forming outside of the target contribute to the count and vertex to the target erroneously. True dimuons originating in the BIM are meters away from the target region, thus they are not discussed due to the spectrometer's resolution in z. However, the iron beam dump is located near the target and the cryogenic target flask completely surrounds the target. Therefore, these dimuon sources merit evaluation and control.

These events are target independent impacting the ratio as $\frac{D+X}{H+X}$, taking H and D to be hydrogen and deuterium counts respectively with X as these additional non-target dimuons.

This potential error must either be compensated for or estimated. One important check is the dump-target separation along the z-axis. The tracker explicitly considers the possibility of the vertex originating from the dump and performs z-calculations with that possibility left open. The GMC results are shown in figure 3.2

Here the dimuons are resolved into two groups centered around different vertices with a small overlap. The group overlap is further reduced by taking cuts. Specific Monte Carlo events are generated for the purpose of testing this separation. These Monte Carlo sets include all the particles as the target. The results are similar to data in terms of the capacity to resolve the different particles. Cutting on vertex tracking location is the simplest way to remove the majority as evidenced by figure 3.3, but the simulated productions are also considered when making machine learning cuts.

Regardless, there is a fundamental cross-entropy error caused by an inextricable fraction of the non-target events. Drell-Yan events originating in the flask walls, in particular, are fundamentally indistinguishable from Drell-Yan events produced in the target at the spectrometer resolution. Therefore, the remaining events are subtracted from each target via the normalized empty target dimuon count as these events are by definition target independent. Empty flask measurements thus allow for the removal of non-target dimuon

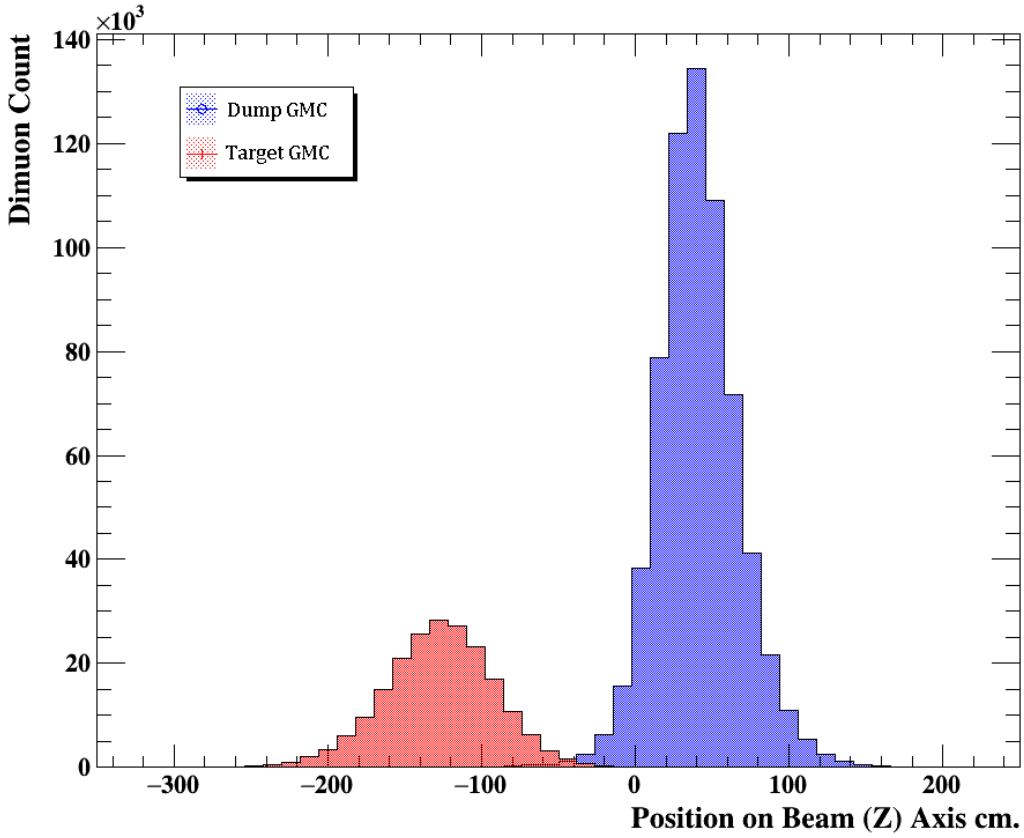


Figure 3.2: Target and Dump Drell-Yan events with mass above 4.2 GeV events are simulated as a function of the tracked dimuon origin along the z-axis. The tracker introduces powerful dump-target separation on its own.

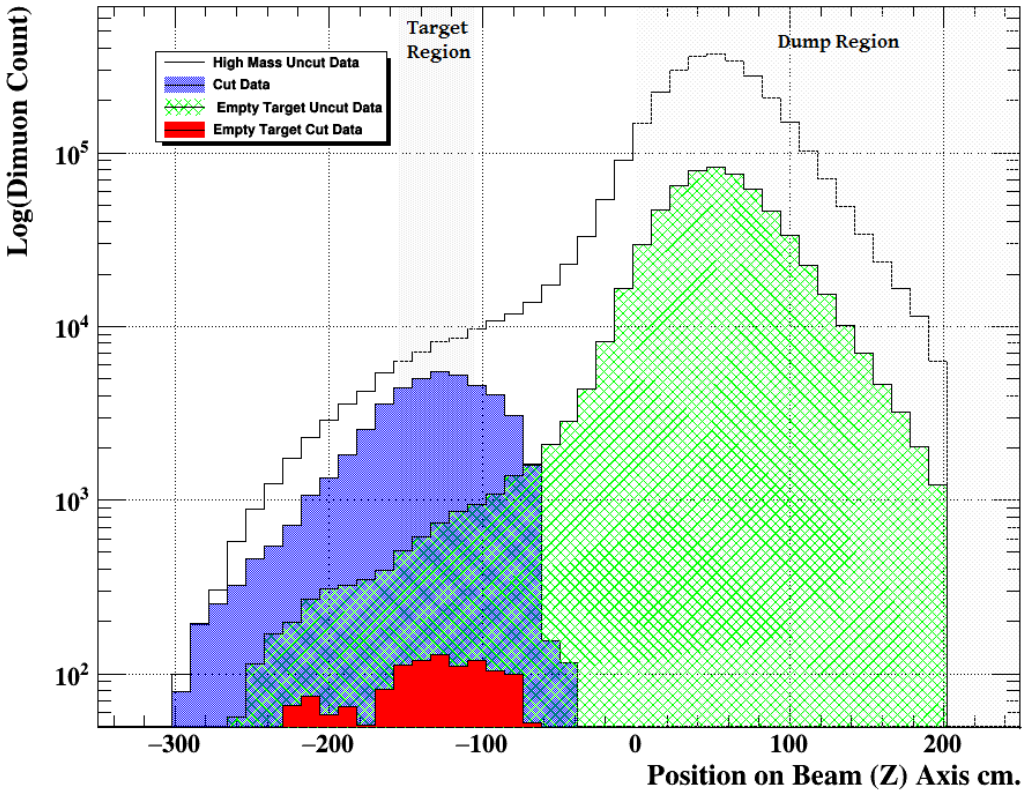


Figure 3.3: Real LH_2 data tracked z-origin is compared with samples from the dump dominated Empty Target and cut versions of the Empty Target and LH_2 data. The effective control in origin shows the effective removal of dump events with the inseparable combinatoric left in the Empty Target.

events statistically as a function of x_2, and provide a measurement of the resulting error. The empty target position also supplements evidence of the successful target-dump separation, as the empty target counts are $64.4\% \pm 0.4\%$ of the target counts.

The dump and empty flask materials also produce single muons that may contribute to the background count by forming an apparent dimuon pair with another single muon. Combinations of flask and dump single muons would be removed with the empty target subtraction. These backgrounds are treated as combinatoric background, rather than non-target dimuon sources as the behaviour and separation techniques starkly differ.

3.2.5 Combinatoric Background

Combinatoric background refers to the situation when an apparent dimuon consists of tracks arising out of two or more simultaneous muon production events. The events must be capable of triggering the spectrometer and be tracked back to the target. Virtually all of the untracked events and the majority of tracked events are combinatoric background of two or more individual physics events. Due to sheer magnitude of the difference in cross-section magnitude, the event count is far greater than the Drell-Yan cross-section, with many events outside of the mass region of J/Ψ and Ψ'.

The probability of a combinatoric event is the combined probabilities of the constituent events as they are in effect completely independent. The largest source of combinatoric background is two simultaneous single muon tracks. Returning to discussions which lead to Eqs. 3.2 and 3.3, measured event probability is $P = N_{success}/N_{total}$ and is proportional to the cross-section. The cross-section acts as a probability per incoming proton. Hence, inverting the relationships to derive the probability without an outgoing event count requires the use of $P = \sigma N_{total}$. N_{total} is simply the number of proton interaction opportunities which may be derived from the integrated beam luminosity L or

$$\frac{dP}{dt} = \sigma L. \tag{3.10}$$

The cross-section here is written to include the beam attenuation in Eqs. 3.4 and 3.5. The individual physics events generating a track are essentially independent, thus single track events "A" and "B" have combined probability $P(A|B) = P(B|A) = P(A)P(B)$. If events A and B have probabilities $P(A) = \sigma_A I$ and $P(B) = \sigma_B I$ with intensity I, the combined probability follows $P(A+B) = \sigma_A \sigma_B I^2$. Therefore, the dominant combinatoric rate dependence is proportional to I^2, while dimuons are proportional to I. High-intensity dimuon events inevitably possess a greater cross-entropy with combinatoric events and are less efficiently tracked. An intensity cut stops the combinatoric background error, which inevitably outstrips the usefulness of the potential statistics provided.

Higher order effects make the probability relation less accurate at greater intensities. When event interference and saturation becomes significant, the events are no longer independent. Instead, there exists a stochastic relationship between events. Three or more track events begin to comprise a larger fraction of events as well. Before cuts, $95.8\% \pm 0.05\%$ of events contained up to two tracks. The fraction increases further when considering events that actually track to a single dimuon coming from the target, which passes quality cuts. Of those, none were found with 3 individual tracks that are equally capable of forming a dimuon. This reduces the effective situation to a two-track plus interference model. The interference is treated similarly to hits causing inefficiency as a function of intensity. True events contaminated by additional hits in the detector are treated as an intensity effect, not as a background. These additional hits derive from leptons which only pass through part of the spectrometer, in particular, drift chamber 1. The inefficiency takes the form of incorrect kinematic information or the loss of an event altogether. True dimuons that track with single muon tracks are combinatoric background if the alternative track does not directly interfere with the true track directly. In this case, it is no longer a deviation from the true track, but truly a different dimuon.

Intensity-dependent inefficiencies impact combinatoric events through interference the same way as true dimuons. Therefore, the significant combinatoric background can be

treated as dimuons as a function of I^2. Additional intensity effects are treated separately due to following a unique functional form.

Starting with single muons, estimates for the combinatoric background may be modelled. A unique GMC was the first model for the combinatoric background by simulating the single muon background. This Gun Monte Carlo was fabricated from simulated beam proton collisions in the detector. Each was fired one at a time and subsequent proton interaction products were tracked. By tracking all particles with adequate momentum that stay within the spectrometer, all possible sources of muons are revealed with subsequent detector hits. There are two dominant sources of single muons: $\pi^{\pm}$ decays at $40.7\% \pm 0.2\%$, and $\kappa^{\pm}$ decay at $7.54\% \pm 0.05\%$. The kaons are responsible for at least some of the measured charge asymmetry, due to κ^{-} being 2.088 ± 0.002 times more likely than κ^{+}. Using the GMC to track these single muons reveals potential single muon source vertices and how they might travel through the detector.

The three major locations where the beam interacts are the target, dump, and BIM. The tracked originating vertex is not necessarily the location of the original interaction. The pions and kaons travel some distance before decaying, thus the muon may not even originate from the beam axis to which the tracker attempts to fit the vertex. True origins of single muons are estimated by the GMC shown in figure 3.4, but without forming a dimuon vertex. Consequently, dump events are less likely to pass as they decay later, but upstream muons are more likely to pass. Single target muons appear to still be the most likely to track back to the target. Target events may combine with other tracks and be more likely to track as a dimuon. Relative target rates approximate combinatoric background vertex origins experimentally. Events with combinations originating in the dump and BIM will also occur in the no target position. Despite only 10% of muons originating in the target, the LD_2 event rate is almost twice as large as the empty target. The vast majority of these events are combinatoric background. The tracker thus has some success at separating even single non-target muons, but the GMC wasn't capable of predicting the tracked dimuon

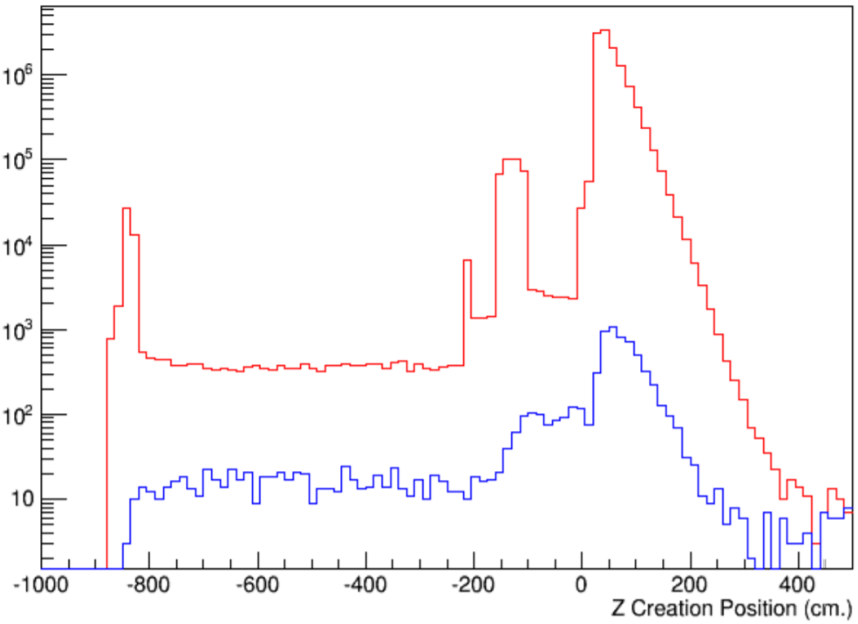


Figure 3.4: Single muon origins are simulated by the GMC, which tracks child particles starting from the beam proton collision. The origins of these muons do not represent the origin of muons forming combinatoric events.

sources.

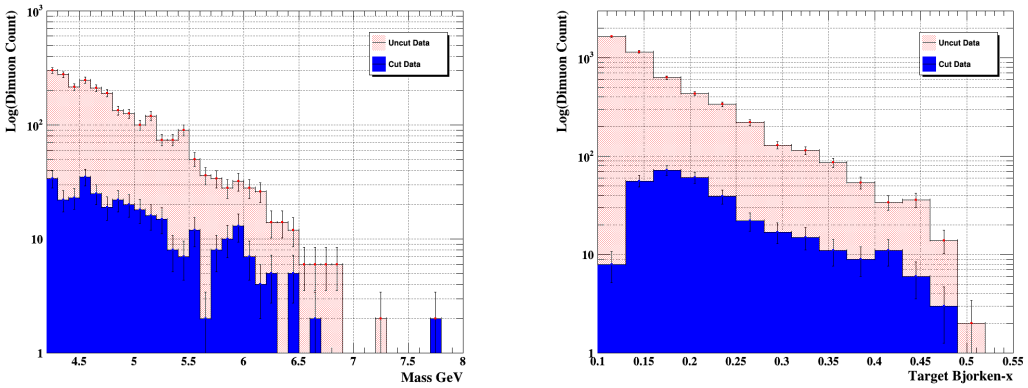
If only 1 in every 10^3 single muon events needs to be able to track as good events to outnumber the Drell-Yan events by over an order of magnitude due to the low QED cross-section. Therefore, normally small branching ratios or insignificant deviations from true distributions may contribute a large combinatoric background. As a result, extremely small deviations of the Monte Carlo model from real events can compound when forming a combinatoric background representation, and thus become significant. As rare events which track as good dimuons are challenging, the computational efficiency is extremely small and prone to errors. Moreover, as the events are thrown one at a time, no intensity effects are present. A messy set is not sensible to build, as the GMC is simulating those very events used to build the messy sets.

However, another estimation method was employed to model the background. Real data provides a better picture of what occurs in the detector at any given time, and has millions of potential samples to draw from. In order to form a representation of the combi-

natoric background, individual tracks recorded with special triggers were mixed with one another. Tracks for mixing are roughly grouped by intensity by way of occupancy (discussed more in Section 4.1). The resulting mass and x distributions are shown in figure 3.5. As these events are derived from real data, the tracks include only muons which pass through the entire detector. The special trigger requires the track to form a road capable of triggering with another muon. Essentially, track generating hit patterns from one half of the detector are mixed with track producing hit patterns on the other half. These constructed combinatoric dimuons include real intensity effects and any relevant spectrometer data acquisition impact. The resulting distributions should follow the actual muon combinations in the detector. Mixing any independently formed simulations means variance in the sample compounds as bias in mixing, which is further exacerbated by binomial variance. The intensity mixing groups ensure more realistic combinations but are still imperfect due to the descretizing that aspect of intensity. This background still represents the best background model and should be representative of combinatoric backgrounds tracked by the detector. Physically, these combinations should represent the behaviour of real data. Therefore, this sample is used as a background during training a machine learning classifier.

By employing an understanding of the background, additional controls may be added to ensure the accuracy of the final cross-section results. The two basic controls for the tracked combinatoric backgrounds are the empty target subtraction and machine learning cuts.

The empty flask subtraction is one of the key controls of the combinatoric background. Single muon events composed only of tracks from the dump, upstream of the target, and/or the flask walls would contribute to the empty target events about the same as when the target is in place. All events are impacted by rate dependence and spectrometer inefficiencies equally, so subtracting the empty target normalized by proton count is valid. The empty flask does not account for the total potential impact of the dump and flask. However, due to sheer quantity, there is a large chance of these single muons combining with single muons originating from the target volume and successfully tracking as a target dimuon pair. These



(a) Mixed Data Mass

(b) Mixed Data x_2

Figure 3.5: Mixed background estimates simulate key potential kinematic distributions with mass on the left and x_2 on the right. The data has two cut sets applied, the Uncut set has some basic cuts applied showing that this group is non-trivial to seperate. The results of Machine Learning cuts is shown in the cut data set. The left and right graphs also prove that the combinatoric background is capable of generating high mass and x_2 events. The choppy mass distribution relates to both how BDT's make cuts over all variables and the low counts in the cut curves.

events along with target single muon pairs form a large portion of the combinatoric background. Thus, as all pure dump originating combinatoric background is contained in the empty flask measurements. Combinatoric background subdivided by muon origin combinations may thus be studied through the background targets. The dump-dump combinatoric background is less than the target-dump and target-target combinatoric background as the average Empty target's trigger count per spill is 2550 ± 8 versus the LD_2 target's 3950 ± 10.

Events with at least one track originating in the target cannot be subtracted and must be treated as their own subspecies during cuts. The primary differentiation in terms of origin becomes the presence of at least one target muon composing a dimuon. These events may also have unexpectedly high mass due to the parent particle traveling and emitting the muon at an angle causing further displacement on the axis. Using the mixed background samples and real data sets, machine learning cuts were built capable of removing some of these events. Events inextricable by cuts efficiently can then be handled by taking advantage of the intensity dependence. The preference is to remove any background that is removable over using the intensity, as this minimizes the error. The machine learning does apparently

have sufficient classification to allow for extrapolation.

3.3 Machine Learning and Multivariate Analysis

Once the signal and background populations are at least partially defined, events are classified via employing recorded variables to devise a function separating the populations. In many experiments, clearly identifiable peaks simplify this process, but E906 does not possess a discernible peak or resonance apart from J/Ψ. Outside of their mass regime, other contributions from different species are present in the spectra. In fact, across all remaining major kinematic variables, no readily identifiable splits between signal and background remain. Therefore, maximizing usable information to optimize classification should be prioritized.

One approach to classification is to build rectangular physics cuts by assembling potentially useful variables using known properties and choosing the best apparent separations. This carries the advantage of using external physics knowledge or information to find helpful variables or cuts. Humans may also possess exceptional feature identification. However, humans are biased, potentially imposing patterns and approaching problems with personal expectations. Notwithstanding, building these sorts of cuts requires a great deal of manpower as variable numbers increase, and cannot necessarily be simply adapted to other data sets.

Data-driven techniques do not contain the same bias but may be liable to mistakes. The classification choices are entirely reproducible from the procedure and can have a consistent methodology. There is no ready interpretation of the results and the systems are more vulnerable to errors in the production of Monte Carlo or background. Rectangular cuts are themselves believed to incapable of classifying as efficiently as other learners or classifiers [57]. Thus, machine learning is selected for the scientific preference for reproducible work and potential unseen gains in the classification power.

3.3.1 Variables

One of the first steps in this procedure is to create a variable list. The number of independent variables is not completely determinate but derives from the measured detector points, charge, and intensity. This information is translated into the derived values during tracking using detector parameters consistent between events. This restricts the order of magnitude of potential independent variables to be less than one hundred. In total a dimuon may have 32 drift chamber measurements, a couple of useful proportional tube measurements, hodoscope roads, and some beam information from which all information or variables must derive. External information employed through the Monte Carlos allows for the actual discrimination. The good variables for discrimination are not necessarily the kinematic or detector variables either. In fact, variables like mass more closely relate to the square of variables such as x-position which is roughly proportional to p_z.

Initially, the data was downloaded as dimuons containing the full track information from both tracks. Then, once downloaded, the dimuon was converted into the two kTracks that shared the same dimuon information. From here a pairwise product and division of all of the kinematic variables were done. This large set of possible variables served as the starting position for the analysis.

Usable variables are combined in terms of multiplication and division to form a massive master list from which the employed list takes its variables. Addition and subtraction are ignored as such linear combinations are already within the parameter space. Two main procedures are utilized to compose a variable list.

The first is principal component analysis (PCA), a procedure reminiscent of an empirically determined Gram-Schmidt Orthonormalization applied to a set of data. Using an overly determined set, Gram-Schmidt Orthonormalization will generate a spanning orthonormal set with a unique transformation matrix. The first step in building the transformation is zeroing the means of the set. Starting with the variable responsible for the greatest variation in the set, the vectors are normalized by their variance. The covariance

replaces the inner-products. As variables are added to the set their contribution to the total variance decreases. The correlations are subtracted out, and the final set contains only decorrelated, normalized variables. The first variables are the principal components of the set. The last variables in the list may be removed by either restricting the count or "score" represented in variable length and reducing the set to significant information contributing variables only. Even if a variable is a linear combination of two others, due to random variation it would appear to contribute some uncorrelated variation. This procedure can be performed in conjunction with other methods discussed later. PCA thus simplifies the list but doesn't necessarily find the best classification variables or rank them properly.

The classification power of variables can be directly estimated. The most straight forward technique to accomplish this, is based on Fisher's Linear Discriminant Analysis (LDA). LDA attempts to derive a quantity maximizing variation between sets divided by the total variance within the set or $F = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}}$. This is the eigenvalue problem maximizing

$$F(\vec{x}) = \frac{\vec{x}^T \Sigma \vec{x}}{\vec{x}^T (C_S + C_B)\vec{x}}, \tag{3.11}$$

where x_i are values of variable i, Σ being the between variation matrix, and C_S/C_B are the signal/background covariant matrices, with their sum being the "within" variation matrix. The maximizing eigenvalue solution yield Fischer coefficients which can classify data or determine classification. The Fisher coefficients are simply

$$F(\vec{x}) = \frac{\sqrt{N_S N_B}}{N_S + N_B} \sum_{i=1}^{n_{var}} (C_{S;j,i} + C_{B;j,i})^{-1} (\bar{x}_{S,j} - \bar{x}_{B,j}), \tag{3.12}$$

where N is the number of events in the signal/background samples. Adding over the coefficients multiplied by the value gives the discriminant values. Fisher discriminant provides another way to transform data, while applying the same calculations for the aforementioned decorrelation techniques.

Another method to estimate classification power was developed later to more directly

align with analysis. Boosted Decision Trees were employed to separate the data using the similar architectures as with classification. The Boosted Decision Trees determine the ranking by the appearance frequency and squared gain of each variable splitting node weighted by the node's event count.

Random sets of 50 variables were ranked based on Linear Discriminant Analysis or BDT's. The top 20 variables in each group were then passed onto the next phase of selection. The "linear" original kinematic tracker variables acted as a basis of comparison between sets.

Cross-comparisons were implemented with normalized ranking scores and required the new variables to outperform standard variables before assembling them into a singled ranked 150 variable list. The ranking scores were further divided by an averaged communality or averaged covariance score. Communality is derived from the covariance matrix of the set [56]. The communality of a variable is a measure of how much of the variance of a variable is due to selected principal components; e.g if there were 10 components used and if the mass had a communality of 70 then 70% of the variance of the mass is accounted for by those 10 components [56]. Conceptually, these factors attempt to maximize new information added, by efficiently compensating for how much of the classification power is contained by other listed variables.

Several lists were composed with which the classifiers were trained, and the final performances compared between sets. Three main ranking lists with a control list were developed: independent, BDT dependent, and Fischer dependent. The control list was simply the linear variables. The training was tested with and without PCA transformation deployed on the variable sets.

3.3.2 Machine Learning

The main selection is driven by machine learning (ML). As the sets are being actively discriminated based on test data fed into the system, only supervised techniques will be

considered. The classifiers will only rely on tracked and detector variables, but not the detectors themselves. Multiple ML techniques were attempted with various architectures including linear discriminators, automated rectangular cuts, neural networks, support vector machines, Probability Density Estimators Foam (PDEFs) and Boosted Decision Trees. Deep Neural Nets were tested but perform poorly due to the relative independent input number, training sample quality, and training sample size. The strongest classifiers were the BDTs and the PDEFs. BDTs are also uniquely suited due to their training and testing speed, large variable number, and robustness to weak variables. Additionally, BDTs are conceptual "white-box" classifiers, meaning the actual logical mechanics behind decisions is available to the experimenter. In contrast, Neural Nets are black boxes, where neural nets connections have no clear or immediate relationship. Therefore, boosted decision trees will be explicitly explained, the white-box nature making the underlying logic applicable to an individual classifier.

3.3.3 Decision Trees

Data is divided into a training sample and a testing sample. A sample of events in the training set is selected to become the root node of one of the trees. That sample is then divided based upon some splitting criteria into two daughter nodes. The criteria take the form of a cut on a particular variable's value. The strongest discriminator is found for this purpose by scanning through variables and solving for a cut z-value which optimizes the separation of the signal and background training set. The splitting is accomplished with a signal/background symmetric function such as

1. Gini Index: $p \cdot q$
2. Cross entropy: $-p \cdot lnp - q \cdot lnq$
3. Misclassification of error : $1 - max(p, q)$
4. Statistical significance : $\frac{S}{\sqrt{S+B}}$,

where p and q are the fraction of the events that are signal(S) and background(B), respectively. The functions estimate the information gain as well.

The two sub-populations are fed into daughter nodes. Then those daughter nodes are subsequently split; this process continues until the requisite depth or an over-training constraint, such as a purity limit, is reached. In a single tree, all bottom nodes are defined as signal or background.

3.3.4 Random Forest and Ensemble Learners

Ensemble learners are assembled from a series of weak learners into a single strong learner. A series of rules can form the basis of decisions through direct logical assembly or boosting. Formations of ensembles are made throughout the analysis phase with cuts and other means. Generally, there are three standard ways of building these ensembles: bagging, boosting, and stacking. Stacking is performed by tacking on subsequent classifiers trained on the results of the first. Forests, or ensembles of decisions trees, are assembled using the first two methods, though the third was attempted.

The response of the forest is a weighted voting or boosting of the constituent trees.

Bagging

Bagging is short for bootstrap aggregating and splits the data into multiple sets. This algorithm resamples the data, by either explicitly picking different random samples from the total set with a replacement or reweighing events randomly. The derived subsets are treated as statistically independent and the decisions are aggregated by adding them in the case of BDTs. The result is more robust and has less variance. Bootstrapping is also important for generating individual cuts when scanning through variables, cross-checking, and performing variance measurements.

Adaptive Boosting

Boosting fits the models iteratively, re-weighting the data based on previous classifiers. Misclassified data is prioritized in preceding learners, while well-understood pieces are diminished. Trees are generated to work on increasingly difficult events, making this process particularly sensitive to smaller sub-populations. Consequently, groups of trees tend to be sensitive to particular sub-populations within the data. The statistical power and re-weighting calculations provide an initial weight for each tree. These weights are then used as initial values in a final fit on all the constituent tree weights optimizing the classification power.

Gradient Boosting

After a tree is constructed, a loss function between the predictions and input data are generated and a gradient of this function over this space is performed. A regression tree is then created whose output nodes are matched to the average value of the gradient in each region taking the structure of decision trees. In effect, the system is performing an adaptive boost, but considers all the trees before it and where the boosting was trending. In effect gradient boosting is similar to "velocity" in other machine learning algorithms. The trees build hypercubes of cuts over successive iterations. The loss function is then updated with a weight multiplied by the new regression tree and this process iterates. The weights are constantly being adjusted and calculated during the process, making the algorithm more computationally expensive but should converge faster while being potentially more accurate. Some of the sensitivity to sub-populations is lost, but the algorithm is more resistant to overtraining on deviations between sets based on random variance.

The sum of weighted trees over a data set creates a total response function represented in figure 3.6.

In all methods, there is less bias in the total ensemble and the resulting system is robust to missing or incorrect data. The system is still regressed or classified taking account

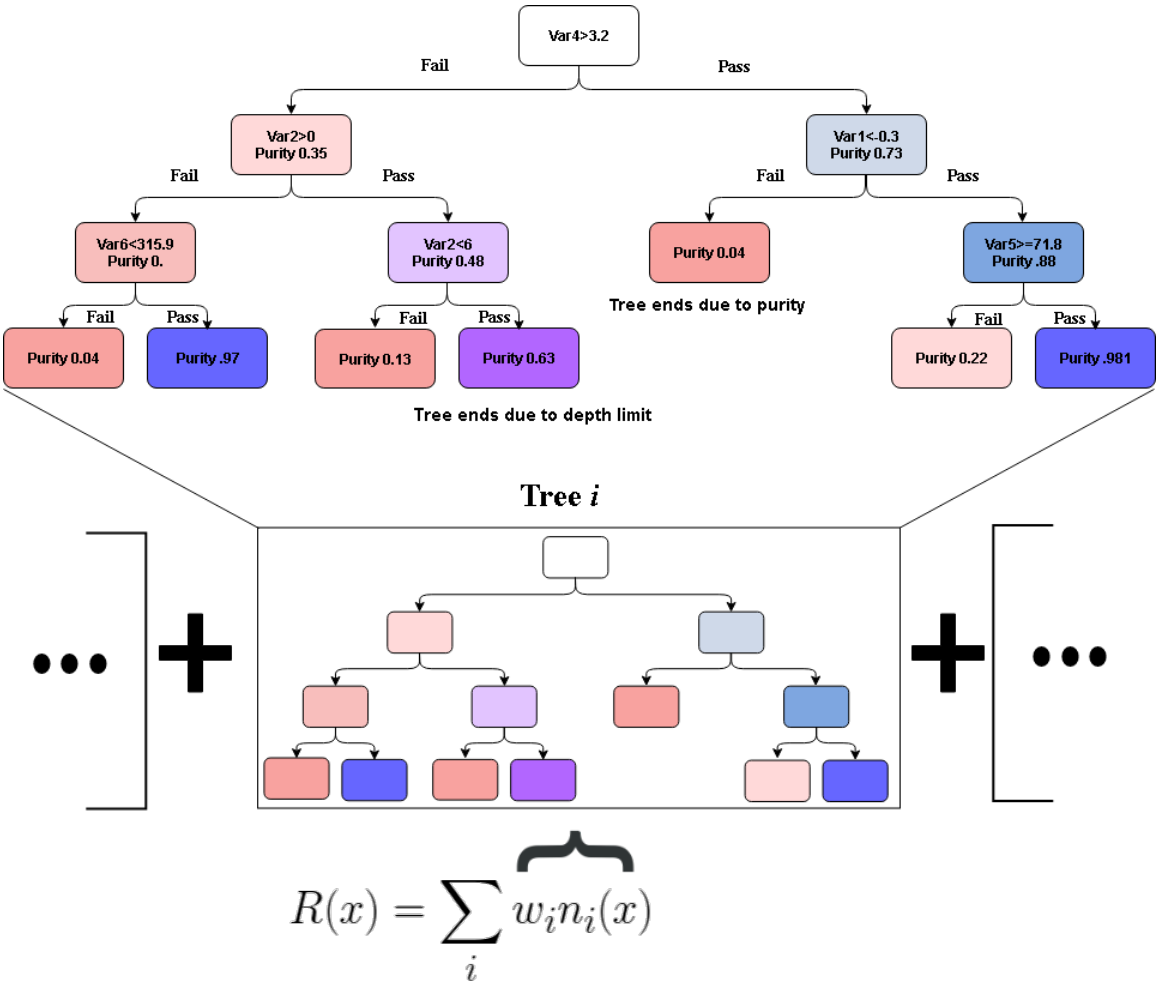


Figure 3.6: The diagram attempts to show how Boosted Decision Trees function. Each tree is given a weighted and a response function which tends to simply be -1 or 1 based on a signal or background classification. The tree is magnified above, displaying the individual cuts on the representative background. The purity ($\frac{N_{signal}}{N_{signal}+N_{back}}$) increases with depth on the right, with greater signal purity shaded blue and greater background purity shaded red.

only the currently present features. Less confident cuts have decreased impact, later trees receiving less weight. The BDT response function is also a tunable cut, with response curves tending to center around particular data features based on method. In contrast, neural nets tend to have peaks at the signal and background response extremes. This tuning variable allows for error optimization and keeping both target's symmetric in acceptance.

The BDT architectures vary in several listed ways, but perhaps depth and number of trees most clearly form an optimization problem. Depth is a bias-variance trade with the number of trees. Overtraining becomes a serious risk if the depth or tree number is overly large. Other architectures tested include the boosting method, the splitting optimization, and variable bagging. The number of trees in the forest and the maximum depth are initialized before classification and the resulting ML results are stored in a weight file. This weight file contains the blueprint of the information learned by the ML algorithm and is employed when evaluating new events.

3.3.5 Cut Results

As the variables and classifiers have been established, how the signal and background samples are actually implemented must be demonstrated. The backgrounds are considered separately and combined into a single "ensemble learner". All BDTs are trained with the messy Drell-Yan Monte Carlo as the signal with different background samples. The background samples include the aforementioned messy J/Ψ Monte Carlo, messy Ψ' Monte Carlo, dump Monte Carlo, mixed background, and real data.

The last background seems paradoxical but relies on the relative rarity of target Drell-Yan events next to background events. A small, estimable amount of cross-entropy error exists in the final samples due to signal contamination in the background sample. BDTs, in particular, mitigate the potential issue as the cross-entropy shifts the response towards the signal side of the response curve. The cut location is simply adjusted with the distribution shifted closer to the background side, being treated as cuts particularly vulnerable to over-

training in the high purity regime. However, greater signal cross-entropy error decreases the effective response range and increases the local error derivative, thus accounting for any error contribution directly. The error contribution is still heavily constrained by the fractional percentage of the background and the use of other classifiers classifier help compensate for this weakness.

Lastly, while the initial results included a stacked BDT model, this was simplified to a cut followed by a BDT model. Initially, the stacked BDT was trained to perform the sweeping high purity cuts using all backgrounds simultaneously. Then several BDTs were individually trained on backgrounds surviving the first BDT, and applied together to the subset of data passing the first BDT. Conceptually, the initial easy to separate background species dominate the boosting process. Consequently, while the easy to perform tree cuts are successful, harder to remove species closer in number to Drell-Yan events appear trivial. Response level cuts become almost impossible to estimate, and boosting fits are completely dominated by the first few trees. Separating these trees is the simplest solution. However, tuning the response of the first BDT had little impact on the final results.

Therefore, standard rectangular cuts were attempted. The final signal passing rate was held at 98% of the Monte Carlo, the same as the BDT. The rectangular cuts were less effective, thus outside of the BDT cut region. Hence, they couldn't contribute to additional errors in terms of signal cuts. Using these rough cuts had no impact on the final passing rates of signal and background, and were thus kept for simplicity. The cuts were built using a simple purity gradient steps, in variables ordered by the BDT derived estimated variable strength and making cuts subsequently. The strength is estimated by eliminating any cuts on the variable and comparing the results. When the desired signal efficiency is reached the variables are held and the last couple of numeric steps are removed. The exact value or purity is not important for these cuts as long as they perform some rough reduction. The rectangular cuts utilized can be found in the glossary.

The reduced data and signal/background samples were then passed to the BDT builds.

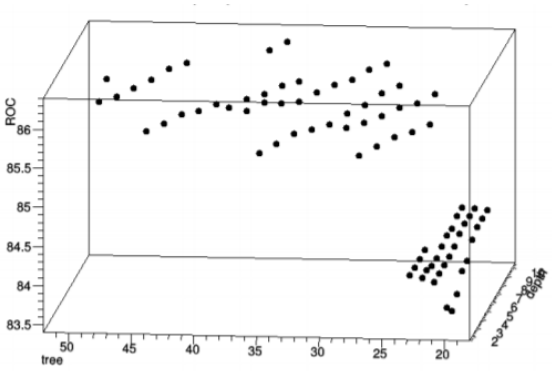


Figure 3.7: Various measured response operator characteristic (ROC) integral curves are displayed as a function of depth and tree count. The ROC integral is a measure of the efficiency of the a classifier, the function itself measuring background removal efficiency against signal efficiency. The closer to 100 the function is the better.

Generally, the optimal BDT architecture for the chosen variable set was the same. The optimal architecture was found to be 35 trees with a depth of 4, splitting on Cross-Entropy and utilizing adaptive boosting. An example of relative tree architecture performance in a background is shown in the figure 3.7.

It should be noted that BDT performance with just the dimuons yielded worse results than using just the track variables, suggesting that the ML is more sensitive to bad tracks than bad dimuons. Early tests showed the J/Ψ and dump backgrounds were trivial when the other background BDTs are applied. The Ψ' data also had low numbers and was combined with J/Ψ into a single charmonium background set before retraining. Alternatively, the dump background sets are not directly trained on but tested against. Dump events are included in the general data background, and mixed background making it less surprising that they don't require explicitly require training.

Each listed background sample is testing against independent species, which can be considered mutually exclusive save the full real data set which should contain all of the others. Therefore, if one independent BDT rejects an event as background, it should be rejected. The real data background has a relatively conservative cut. However, when only a pair of classifiers are tested, the result reduces to only one of the trees except at the intercept where it becomes a "union" statement. The response cut location of each BDT determines

the weight ratio. Therefore, it too is equivalent to a simple logical union between classifiers when tuned. The optimization problem becomes determining cut points.

The cut points are roughly optimized according to the error contribution function and retaining target symmetry. Acceptance between sets must stay close to unity or it contributes to error through a correction factor. Response tuning opens more options than rectangular cuts to retain symmetry. Note there is no known physics reason for deviation in kinematic variables between cuts aside from backgrounds.

Potential deviations in the weighting factor from the Monte Carlo are expressed by deviations in the relative tree weights. The error of this system thus derives from the local derivative of the response curve. The most efficient cut (MEC) on the response is located there. Starting from those cuts, the cuts are deviated searching for symmetrical acceptance or minimizing deviance in acceptance particularly in the x_2 bin. A mass fit employing the Monte Carlo mass distributions can estimate the relative surviving populations of species. Prioritization of backgrounds roughly follows

$$\frac{\sigma}{N_S} = \sqrt{(\frac{1}{N_S} + \frac{N_B}{N_S^2})}. \tag{3.13}$$

The only change the error calculation suggested was loosening the PSIBACK BDT response cut to allow more signal. It was loosened until no significant additional signal passed and then made symmetric. The cut error was still insignificant compared to the cut error of the other two. Remaining Ψ' is an insignificant source of error in this cut region, which is estimated to be below 0.1% of the total data.

The mass fits were formed by passing the major Monte Carlo samples, mixed background data and real data through the BDT weight files with the mass cut removed. These fits, shown in figure 3.8 ensure the sensibility of results. Species included were target charmonium, mixed background, target Drell-Yan, and a fixed Empty target. Low mass charmonium peaks stabilize the fits as a decisive feature, but their associated MC does

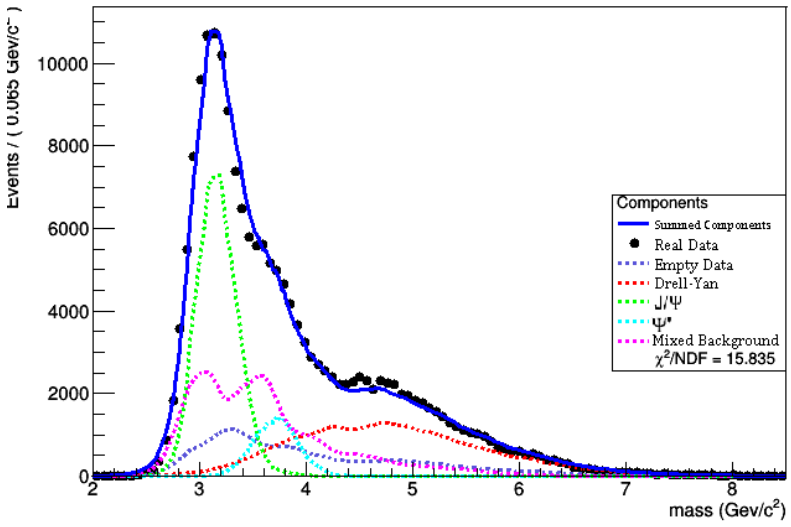


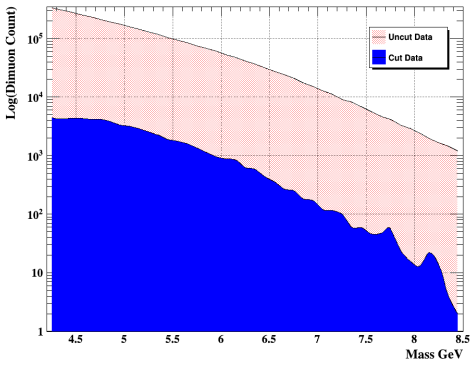
Figure 3.8: GMC components fit together to approximate the total cut LH_2 data set. The Drell-Yan and mixed sets are estimated to be of similar size.

contribute error to the fit due to being heavily cut. The problem is a multivariate linear regression fit. Passing signal and background sample mass distributions are combined according to a weighted summed. The weights are fit according to $W\vec{m} = Y$ where W is the diagonal weight matrix, $\vec{m}$ is the model mass distributions, and Y is the mass distribution of the data.

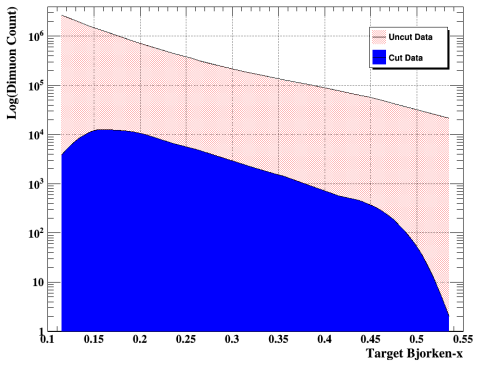
The error cited only includes the statistical error contributed by the number of events and some of the mass error. There is not a contribution from outside species or compensation for the rate dependence which impacts the mass distribution in an important, but unclear manner. The total graph is simply a sort of average intensity plot. Mistakes like that can have effects, which can be up to a factor of two in Drell-Yan or mixed background.

Important to the flavor asymmetry and cross-section measurements are the x distributions. The final real data's x_1 and x_2 distributions before and after cuts are shown in figures 3.9 and 3.10. Additional distributions showing the full impact may be viewed in the appendix.

The remaining distribution does not deviate far from the GMC but does vary as expected. The resulting mass and x distributions seem reasonable and the relative target

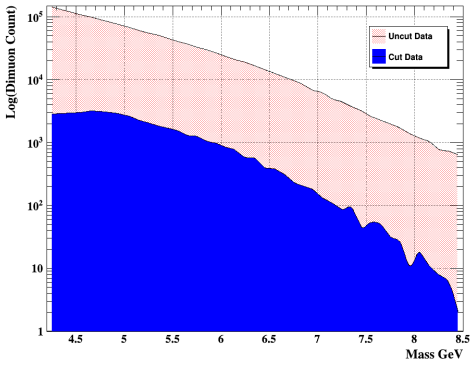


(a) LH_2 Cut Mass

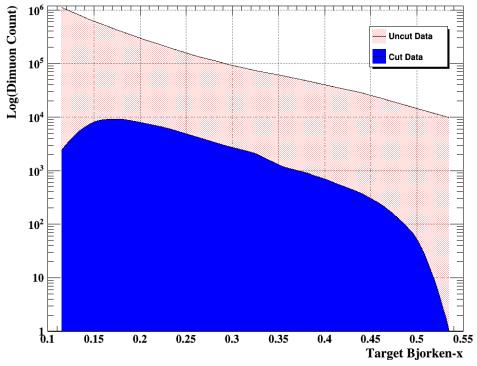


(b) LH_2 Cut x_2

Figure 3.9: Important kinematics of the real LH_2 data sample, with the uncut sample representing the real data after the first simple cuts are made, and the cut sample representing the set after machine learning. Mass is on the left and x_2 is on the right.



(a) LD_2 Cut Mass



(b) LD_2 Cut x_2

Figure 3.10: Important kinematics of the real LD_2 data sample, with the uncut sample representing the real data after the first simple cuts are made, and the cut sample representing the set after machine learning. Mass is on the left and x_2 is on the right.

production is being sought. Therefore, the distributions may be accepted. A summary of the cuts from raw data down to analysis selection is shown in figure 3.11

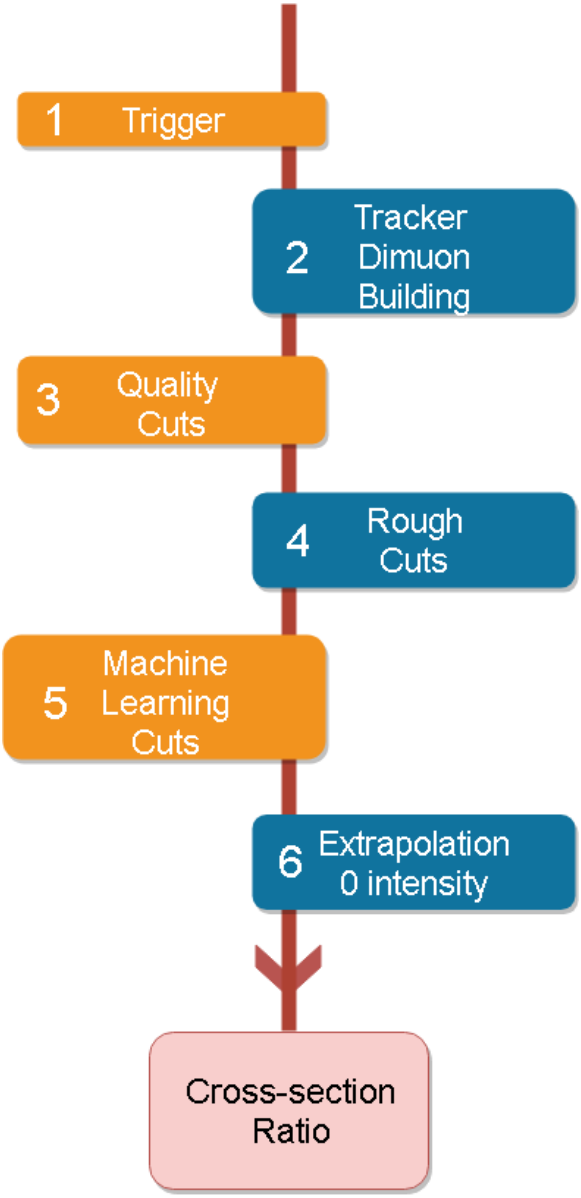


Figure 3.11: A flow diagram listing all the data reduction phases the data goes through until finally deriving a physics result.

Chapter 4

Results and Extraction

4.1 Rate Dependence

While true cross-sections are not dependent on luminosity, the apparent event count and associated kinematics were observed to depend on beam intensity. The impact of the instantaneous beam intensity or number of concurrent physics events is dubbed the rate dependence. Tracking upgrades and data cuts remove the highest rate events to alleviate the issue. However, the remaining rate dependence must still be addressed, as it potentially impacts the final results. Despite beam intensity being independent of target position, the targets vary in rate dependence. This is due to differing nuclear interaction lengths, as deuterium contains twice the number of nucleons per volume as hydrogen at equal molecular density. In effect, the deuterium experiences almost twice the detector activity as hydrogen for the same luminosity.

The "rate" is measured in two major ways: intensity and occupancy. Intensity is a measurement of the beam luminosity, consistent with past usage in section 2.2. The integral of beam intensity over a period of time is the proton count. On the other hand, the occupancy refers to the number of hits detectors experience over a period of time.

In section 2.3, the BIM was explained to measure the instantaneous intensity in this way with a Cherenkov Counter, but required outside normalization. The BIM RF00 data is transformed into an averaged intensity measurement over a 18.8 ns bucket. The transfor-

mation follows the form

$$I_{trig} = (RF00 - Pedestal)QIEsum/G2SEM, \tag{4.1}$$

where the QIEsum/G2SEM is the normalization factor and the Pedestal is the average amount of charge read when there is no beam. This measurement was dubbed the trigger intensity.

Another related variable is chamber intensity. However, chamber intensity attempts to compensate for noise in drift chambers caused by nearby RF buckets. Instead of directly using RF00, it forms a linear combination of earlier RF buckets, weighted empirically by drift chamber hit measurements. These RF bucket measurements are transformed into proton counts according to Eq. 4.1. The final expression appears as a correction term on RF00, with well-over 50% of the total intensity originating from the RF00 bucket.

How many particles actually entered the spectrometer is not explicitly measured. Occupancy does just that though. It records what the spectrometer sees by measuring the number of hits in detectors. While higher luminosity produces more events on average, the actual number of events passing through the detector is more directly responsible for effects like saturation. As a result, occupancy is strongly correlated to trigger intensity as shown in figure 4.1, and is even better predicted by chamber intensity.

Occupancy may be further broken down into the number of hits in the hodoscopes and the number of hits in each of the chambers. These variables were employed in quality and tracker cuts to prevent saturation effects. Average occupancy is target dependent unlike trigger intensity. However, rate dependent efficiency as a function of occupancy should be target independent, unlike it would be for other intensity functions.

Each rate variable is associated with a form of rate dependence for which it is the best variable. The rate dependence manifests in three major forms, which are

1. Detector Efficiency,

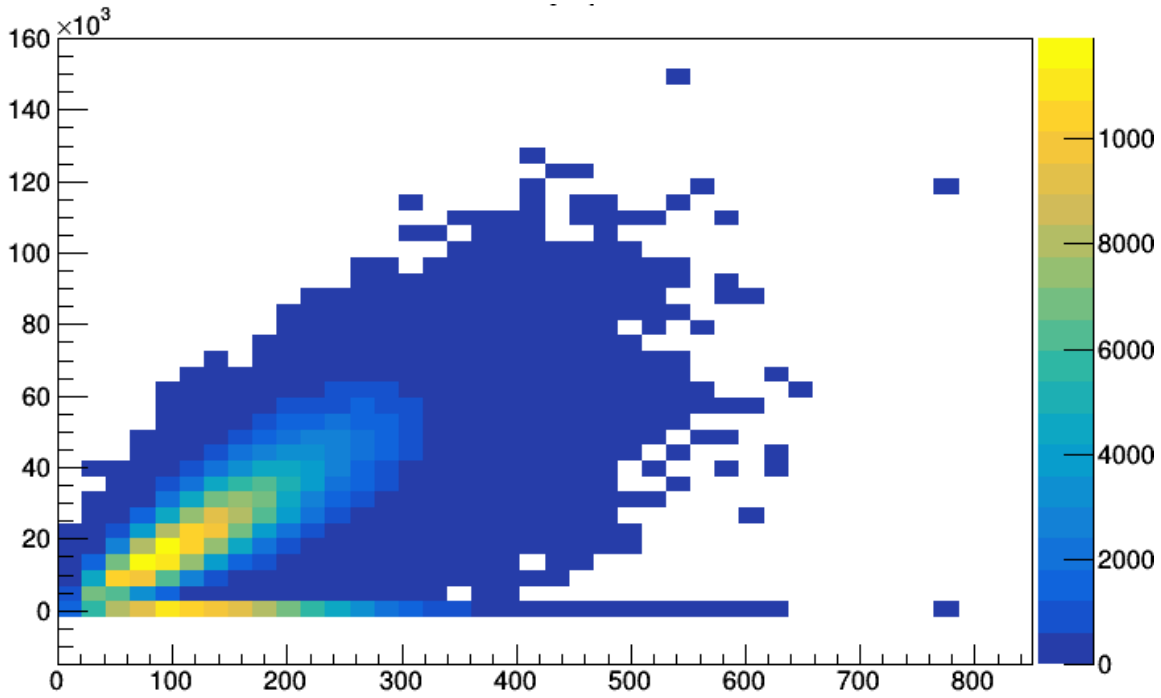


Figure 4.1: Occupancy versus intensity shows two highly correlated intensity variables with Occupancy on the x-axis measured in DC1 hits, and trigger intensity on the y-axis. The line on the bottom is activity during empty events.

2. Tracking Efficiency,

3. Combinatoric Background.

The first and second are similar in functional form, while the last behaves uniquely. The detector efficiency derives from local dead time variation, ringing, saturation, and overall probability of successful electron cascade. Essentially, the chambers are less likely to respond to an event as more events occur. The saturation effects are cut out by requiring trigger intensity to be less than 64,000. Redundancies in the chamber allow the system to function even when hits are lost making the spectrometer more efficient. Generally these effects are weaker when considering relative ratio contributions in the considered data fraction.

Tracking efficiency refers to the capacity to track dimuons accurately when noise is present. This can cause hits to disappear, be moved, or add hits deviating from the true track. It decreases the chance of a true hit being chosen, but due to how tracks are made

in parts, the effect is limited. The result is a small potential deviation in terms of measured momentum, and loss of events or even rare pick-up events. The latter is the largest contributor to the ratio error, as it directly impacts acceptance. The tracking efficiency is measured by comparing the messy Monte Carlo sample against the clean Monte Carlo. The extra hits and randomized hit dropping function simulate effects of occupancy on dimuon events.

The chamber intensity was designed to consider these effects. However, it wasn't sufficient to determine the efficiency on an event by event basis. Instead, the drift chamber occupancy determines in part both the tracker and detector efficiencies. More hits in chambers increase the probability of an incorrect choice in tracks and decrease the precision in tracklet choice. Hit interference on the electronics is more likely to occur as occupancy increases. Therefore, occupancy is the best parameter to model efficiency values.

Yet still, the combinatoric background is perhaps the dominant form of rate dependence. Combinatoric background probability increases proportional to I^2 and is thus capable of quickly overtaking the effect of any inefficiency. This makes the rate dependence highly cut sensitive. The kinematic distributions of the combinatoric backgrounds were also found to be highly rate dependent. As the background is better controlled, the relevant kinematics are reduced closer to those of the Drell-Yan, and the high intensity event counts decrease.

The rate dependence reduces the accuracy of the mass spectrum fits, thus making them unreliable for extraction. Recently improved cuts in the current thesis allow for the current extraction to rely on an extrapolation technique. The cross-section ratio is measured as a function of intensity as well as x_2. The rate dependence is then controlled by extrapolating that function to 0. The cuts sufficiently control the rate dependence impact such that the fit parameters become useful and more precise than accounting for each rate dependent effect separately.

4.2 Cross-section Extrapolation

In the limit of zero intensity, it follows that no intensity impacts remain. In order to illustrate the intensity impact in this limit, a hypothetical beam of individual protons can be imagined. The rate can be defined to be such that every proton arrives well outside the detector relaxation time. In this thought experiment, dead time and local detector saturation become negligible compared to real physics events. Similarly, the tracker will not see any additional hits from other protons, and should thus have no intensity driven effects. Potential detector hit loss and other failures may still come into play, but would be entirely symmetric between the denominator and numerator. Combinatoric background events become exceedingly unlikely due to coming from the same event. This follows from the I^2 dependence as well. Yet, the probability of a Drell-Yan event would remain the same. In this low intensity case the rate dependence is essentially eliminated. The counts become the true dimuon counts, and can be transformed directly into cross-sections. Therefore, this no-intensity count would be ideal for making measurements. It follows that the limit as intensity approaches zero becomes an ideal measurement.

At zero intensity Drell-Yan events are technically impossible, but the limit of the count function ratios still exists. While $\lim_{I\to 0} N(I) = 0$ makes the ratio an indeterminate form. By employing l'Hopital's rule, the cross-section can be extracted. Based on the cross-section, the yield is simply $N = I\sigma\lambda/P$. Therefore, by l'Hopital's rule, the count ratio at zero is proportional to the cross-section ratio. Employing the previous argument, the rate dependent effects are eliminated in this limit as the true cross-section is independent of luminosity.

The rate dependent cross-section ratio derived from the counts can thus estimate the true cross-section ratio by extrapolating the intensity to zero. The zero intensity values can then be extracted from fitting the measured rate dependent cross-section ratio as a function of intensity. The intercept becomes the zero intensity limit of the function, and the rest of

the function represents the deviation from the true cross-section as a function of intensity.

The primary rate dependent effect is the combinatoric background. Considering the combinatoric background, the measured cross-section can be expanded in a Taylor series in terms of intensity as

$$\frac{\sigma_{dp}}{2\sigma_{pp}}(I) = \frac{DI - C_D I^2}{2(HI - C_H I^2)} \approx \frac{D}{2H} + \frac{HC_D - DC_H}{2H^2} I - \frac{C_H(HC_D - DC_H)}{2H^3} I^2, \quad (4.2)$$

setting $\int D, HdI = (N_{D,H} - N_E)$ and C representing the combinatoric background with a count $N_C(I) = CI^2$. Conceptually, as the intensity decreases to 0, all of the intensity effects drop out. The combinatoric background production goes as I^2, dropping faster than dimuon production while approaching I. The limit at zero intensity is the constant term containing no combinatoric background.

Efficiencies can be considered functions of the form $1 - \epsilon I$, and thus do not change the intercept value. Ostensibly, the efficiencies due to occupancy are target independent, and thus the actual cross-section ratio shouldn't be sensitive to this factor. The occupancy is strongly correlated with the intensity, and therefore occupancy based intensity effects should correlate strongly with intensity as well. Therefore, the efficiency ratio is expected to robustly extrapolate to unity at zero intensity. The tracking efficiency intercept is shown to go to unity within $\pm\sigma$ in figure 4.2. This extends to the individual bins.

Building the function requires the counts to be divided into separate x_2 sets. The exact bin boundaries used are (0.1,0.13 ,0.16 ,0.195 ,0.24 ,0.29 ,0.35, 0.45,0.58).

Next, each x_2 bin is transformed into a histogram with respect to the intensity. Multiple bins and variables are tested, but regardless of organization, the cross-section ratio is developed with the independent target counts. The points are constructed from the individual intensity dependent cross-sections, including the empty target subtraction with related

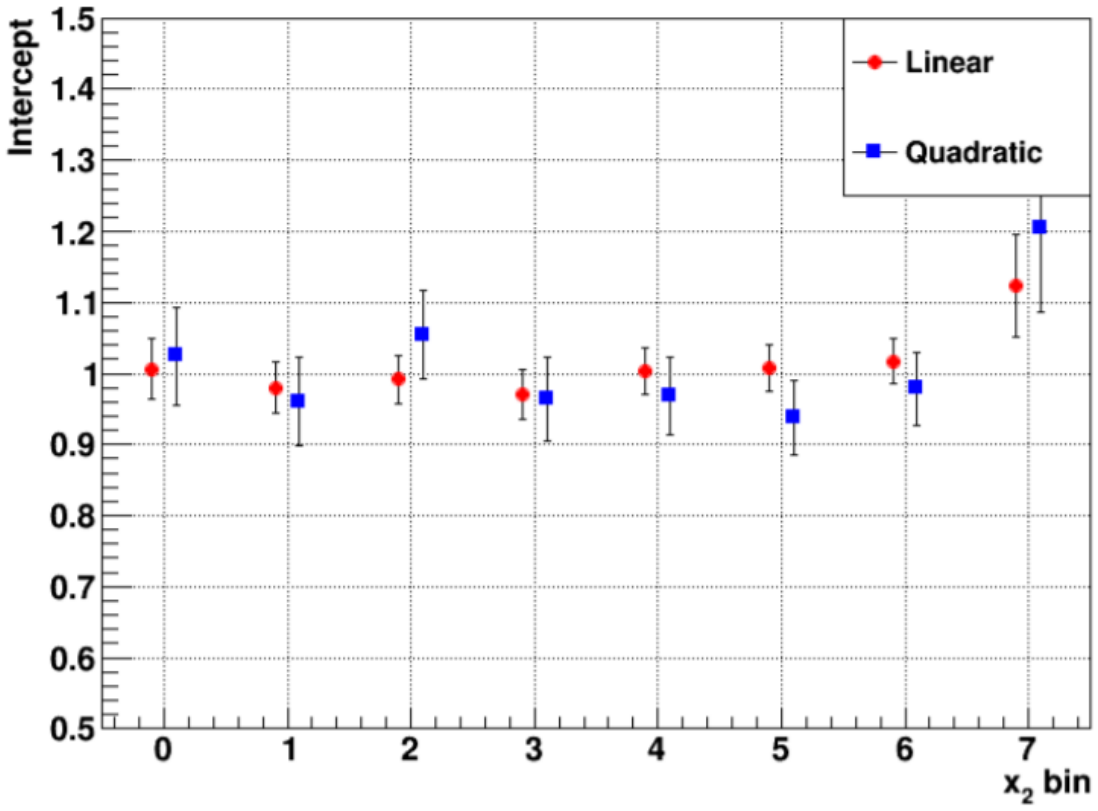


Figure 4.2: Tracking efficiency ratio of hydrogen and deuterium as a function extrapolated to zero shows strong agreement with one except in the unreliable last bin. This has been improved in the simultaneous fits. The figure was taken from Ref. [58].

variables denoted by the E subscripts according to

$$\sigma_{H,D} = \frac{\epsilon_{H,D}(I)\Delta_{H,D}(N_{H,D}(I) - N_E(I)\frac{P_D}{P_E})}{\rho_{H,D}\lambda_{H,D}P_{H,D}}, \tag{4.3}$$

where N is the counted values, P is a proton count, λ effective target length, Δ acceptance, and ϵ is the efficiency. The empty target must be subtracted from the results. The efficiency effects depending on intensity as described by luminosity are target dependent. Combinatoric background is not shown here, as the function needs to extrapolate to this equation.

The proton measurement used for the normalization was the live proton count. A large fraction of the protons are still delivered during dead time. Dead time itself depends on intensity. As dead time of an event spans more than a single event, the raw proton count depends on intensity of previous events. The live count simply ignores additional protons delivered to the spectrometer while the inhibit signal is active. Both the live and raw counts are taken from the BIM as discussed in section 2.3.

With the normalization finalized, both targets are combined to form the point-wise cross-section ratio following Eqs. 3.6 and 4.2 such that

$$\frac{\sigma_{LD_2}}{\sigma_{LH_2}} = \frac{\epsilon_H N_D \rho_H \lambda_H P_H - N_E(I)\frac{P_D}{P_E}}{\epsilon_D N_H \rho_D \lambda_D P_D - N_E(I)\frac{P_D}{P_E}}. \tag{4.4}$$

However, data in runs 2 and 3 are only $95.8\% \pm 0.2\%$ deuterium with the rest being hydrogen. The hydrogen takes the form of an $8.2\% \pm 0.2\%$ HD molecular contamination in the target. First order corrections are made by attributing a fraction of the deuterium yield to both atoms equally. This fraction is determined by the effective average target lengths λ_{DH} ratio with λ_{D_2}. However, interaction length and total attenuation become different as the attenuation Φ depends on both species. The new total attenuation may be defined by first setting $\lambda_{eff} = \rho_H\sigma_H + \rho_D\sigma_D$. The average attenuation over the target length L according

to Eq. 3.5 is thus

$$\Phi_{eff} = \int_0^L e^{-\lambda_{eff} z} dz = \frac{L-1}{\lambda_{eff}} e^{-\lambda_{eff} z}. \tag{4.5}$$

The effective thickness of each species must be derived by defining $\tau = L\rho/V$ with volume V. The molecular size of the DH molecule is smaller than D_2 by a factor of 1.094. The total volume of Deuterium is a combination of the effective volumes of each Deuterium species. If the proportion of Hydrogen in Deuterium is F, then the τ of each species may be written

$$\tau_D = (1 - F/2)\frac{\rho_D L}{V} = \frac{\rho_D L(1 - F/2)}{1 - F + 1.094F}, \tag{4.6}$$

and

$$\tau_{DH} = \frac{\rho_D F L}{2V} = \frac{\rho_D L F}{2(1 - F + 1.094F)}. \tag{4.7}$$

The cross-section ratio compensating for the contamination Deuterium may be rewritten as

$$\frac{\sigma_{LD_2}}{2\sigma_{LH_2}} = \frac{\epsilon_H \tau_H m_D}{2\tau_D m_H \epsilon_D} \left(\frac{\frac{N_D}{P_D \Phi_D} - \frac{N_E}{P_E}}{\frac{N_H}{P_H \Phi_H} - \frac{N_E}{P_E}} - \frac{\tau_{DH}}{\tau_H} \right), \tag{4.8}$$

where m is the atomic mass replacing the density $\rho = \frac{2m_A}{V}$. A table of the related values originating from the particle data group is shown in the glossary. Run 5 data simplifies with the DH fraction set to zero.

4.2.1 Model Selection

The best fit functional form over the fit region is not known from the onset. When fitting, unnecessary higher order terms can cause poor extrapolations or have strange behavior. Low order fits of low complexity are preferred. The higher order terms are smaller by factors of C/H, which could not be explicitly well constrained. However, based on Eq. 4.2, the linear and quadratic parameters should have opposite signs.

Regardless of overall functional form, each x_2 bin was independently extrapolated to zero in trigger intensity. Various models to parameterize the problem were attempted. Valid

models were considered on the basis of physics justification, goodness of fit testing, and information criterion or parameter significance. They are expounded on in the following list.

1. Many potential functional forms may parameterize the problem, but without a physical interpretation the results don't carry merit. The physical extrapolation is based on a Taylor expansion and thus the space of fitting functions should contain subsets consisting of nested polynomial sets, each allowing the intercept to vary with x_2.

2. Goodness of fit is the requirement that a statistical test must not reject the data as originating from the underlying fit function. χ^2 and Kolmogrov-Smirnov were the two primary tests considered. χ^2 was preferred when making final decisions.

3. A solitary χ^2 parameter fails to represent a figure of merit for a functional form. Overfitting becomes a real risk as the distribution doesn't penalize or check the significance of additional parameters.

The first two points are fairly standard, though the third has multiple valid approaches. Extraneous parameters, determined to be insignificant while the fit still possessed other significant parameters, are removed. Significance here is defined by a t-test of the parameter against zero. This approach lends itself to a hierarchical fitting technique but the forms tested against are more varied here as multiple variables must be considered. Hence, this work will focus on the Akaike information criterion (AIC) and likelihood parameters to determine the best fit amongst various sets of valid nested fits.

The final determining factor is the aforementioned AIC. The advantages of AIC include large statistical precedent, objective or repeatable methodology of selection, and ease of computation or development of error as compared to bootstrapping entire fits[59]. The model was based on using Kullback–Leibler divergence to minimize information loss[60]. Therefore, this method of model selection is not a heuristic, but instead constructed on

information theory. AIC is calculated according to

$$2k - ln(\mathcal{L}), \tag{4.9}$$

where k is the number of parameters and $\mathcal{L}$ is the likelihood. Therefore, in the ideal Gaussian case the criterion reduces to $2k - \chi$, which also provides a useful approximation. The best model has the minimum value in this framework. The criterion thus penalizes models for extra parameters while considering the likelihood or probability associated with different models. However, unlike χ^2, the absolute values lack meaning. Only the relative values amongst models are relevant.

In the case of low event samples a correction term can be added that considers the total number of points n following

$$AIC_c = 2k - ln(\mathcal{L})\frac{2k^2 + 2k}{n - k - 1}. \tag{4.10}$$

The model error is constructed through Burnham's method. This technique employs the test parameter from the associated likelihood-ratio test, which is the relative likelihood $e^{AIC_{min} - AIC_i}$ for nested or related models. The final parameters are constructed from a weighted average of top models or taken from the minimum model. The model error is thus derived from the weighted variance [59]. This may only be applied to the same data sets, so can only be employed after selecting a variable and binning.

The fits were performed with the χ^2 and ROOT's binned log-likelihood method. They were found to converge to the same values. Unfortunately, the unbinned maximum likelihood was unavailable for fitting due to the necessity of constructing bins to form a ratio. Fits were tested with all intensity variables listed in section 4.1, including simultaneously as a multivariate fit. Uniform binning in intensity was selected at 8 bins for all x_2 bins except the last x_2 bin due to its poor statistics. While variables bin size may minimize error, it may introduce bias [61]. The model was independent of total bin number for bin number

between 6 and 15. Some of the comparisons are shown in tables 4.1 and 4.2.

Variable	χ^2	AIC
Chamber Intensity	9.33	4.59
Occupancy	12.1	4.96
Trigger Intensity	7.87	4.01

Table 4.1: This table summarizes some of the variable fitting results. A linear fit was done over every x_2 set and the average of the values are listed for simplicity.

Bin Number	5	6	8	10	12	14
χ^2	2.00	3.12	6.15	5.48	8.80	8.08
AIC	4.03	4.10	4.19	4.21	4.49	4.50

Table 4.2: This table summarizes some of the different bin fitting results. A linear fit was done over every x_2 set and the average of the values are listed for simplicity. Note that the "best" set was selected against for higher order fits which had better χ^2 values. The bin range listed is shown to be quite insensitive in the AIC.

In the end, the trigger intensity variable was chosen to be fit over 8 bins, except for the last x_2 bin. Due to low event counts, it was only divided into 5 bins and required a full Poisson treatment of the range.

A simultaneous fit was also tested and contrasted with the individual fits. The simultaneous fit function combined the x_2, I bins to perform the intensity fit, but allowed the intercept parameter to vary. Computationally, this can be performed through a regression on the polynomial parameters followed by a weighted average for each x_2 bin. The cumulative χ^2 parameter can then be minimized through iteration. This computation was cross-checked to direct Minuit χ^2 minimization and is one to one in results. The last bin contained minuscule counts and was not used in the calculation of the common parameters. Thus, only the intercept was fit with the data.

The degrees of freedom of the simultaneous fit isn't directly definable, as it is not a simple linear regression. The constant terms are not linear in parameters as they are not defined over the entire range of intensity and x_2 values. Instead, they are piece-wise continuous step functions in x_2, $f(x_2) = \delta_{x_2,i} p_0$ with $\delta_{i,j}$ being the Kronecker delta function.

The union of these functions x_2, $g(x_T) = \bigcup_{x_2} f(x_2) = \sum_{x_2} f(x_2)$, successfully describes the function at all intensities and x_2. This function does not parse linearly in x_2 and intensity, and thus does not have a strictly defined degree of freedom. In terms of the intensity function it appears as an intercept shifting in each x_2 bin.

A more complex spline fit could have solved this, but requires resolution in x_2 and more information. However, a hypothetical spline can be built for comparison. A splined sigmoid function near the bin boundaries could be interpolated between the values. The interpolated spline starts at some $x - \epsilon$ and ends at $x + \epsilon$, with ϵ chosen such that any resulting deviation in the likelihood and parameter value is a fraction of the measured error. There are $2N$ knots in this system, with an additional equality constraint on each knot. The knots on the boundaries make the hypothetical spline act as a natural spline. Therefore the total number of degrees of freedom in this effective picture becomes N or the number of points. Compared to the individual fits, the intercepts are analogous, but higher order terms are different. The individual fits possess an additional $N - 1$ degrees of freedom than the individual fits due to variation on the points.

The results of all fits are shown in table 4.3.

Variable	χ^2	NDF	Parameter	AIC	Δ AIC
Linear	91.3	103	24	139.3	22.7
Quadratic	74.6	95	32	138.6	22
Sim. Linear	102.3	110	10	122.3	5.7
Sim. Linear w/ x_2	102	109	11	124	7.4
Sim. Quad	94.6	109	11	116.6	0

Table 4.3: Results are compared between the simultaneous fits and independent fits. The last column is defined as the AIC difference of the current fit against the fit with the minimum AIC.

4.2.2 Results

The final cross-section fit with the best AIC is shown in figure 4.3. The run 5 data is fit separately from the rest employing the same top fits. The two sets are combined

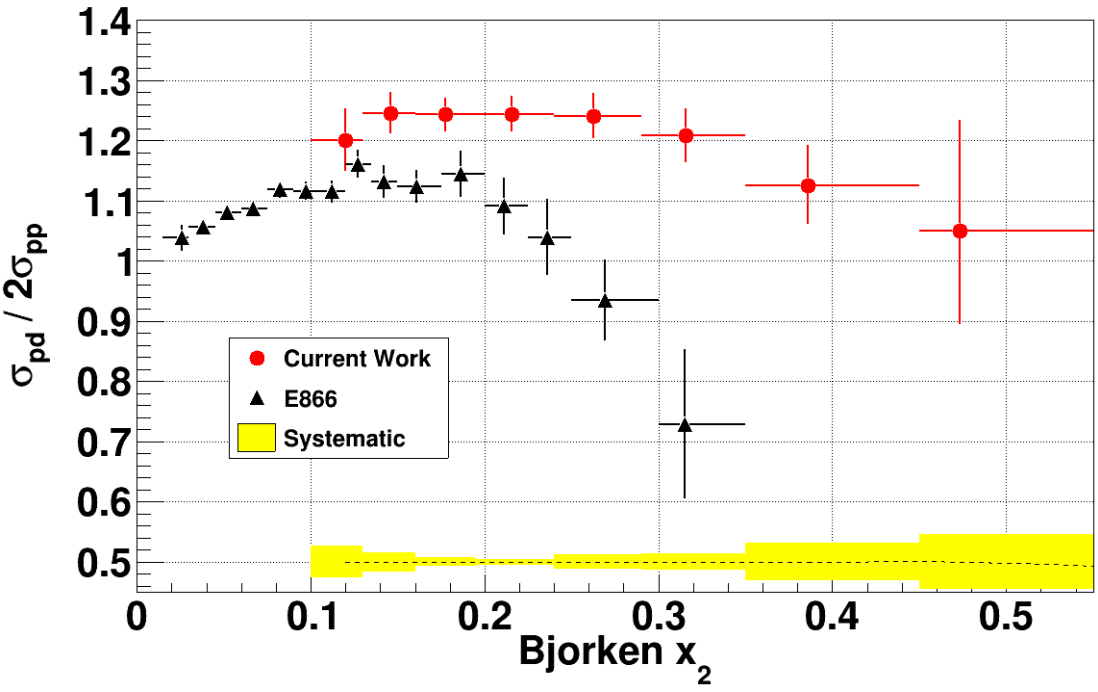


Figure 4.3: Final Cross-section Ratio of σ_{pp}/σ_{pd} plotted as a function of x_2. The E866 results are shown on the same plot for the purpose of comparison. The asymmetric systematic error is shown on the bottom with a dotted line separating the upper error from the lower error. Note the factor of two division on the y-axis.

statistically, after checking to ensure there isn't a statistically significant difference of at least the 95% level. These were combined to form the final cross-section. The specific points may be viewed in table 4.4.

The current data differs from E866's data slightly due in part to the difference in energies. Further analysis of this result is to be compared to past experiments in conjunction with the asymmetry extraction. However, a large divergence with E866 is apparent at high x_2.

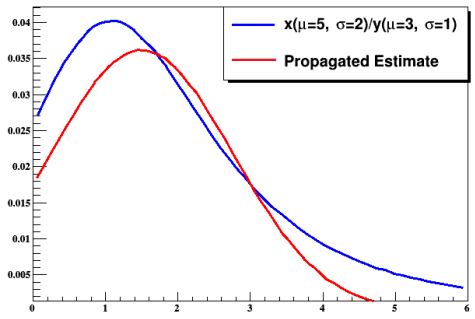
4.3 Cross-Section Errors

4.3.1 Error Propagation

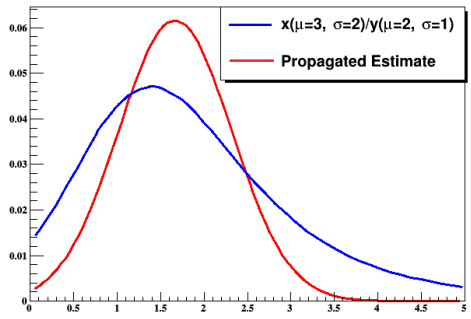
One of the fundamental issues in error propagation is that the formula is an approximation of the probability distribution's second moment. Normal Ratio distributions do

x_{min}	x_{min}	x_2	$\frac{\sigma_D}{2\sigma_H}$	Stat. (lower)	Stat. (higher)	Total Systematic
0.10	0.13	0.12	1.20	0.051	0.054	0.013
0.13	0.16	0.15	1.25	0.034	0.035	0.008
0.16	0.195	0.18	1.24	0.028	0.028	0.003
0.195	0.24	0.22	1.24	0.029	0.030	0.002
0.24	0.29	0.26	1.24	0.037	0.038	0.006
0.29	0.35	0.32	1.21	0.044	0.046	0.007
0.35	0.45	0.39	1.13	0.064	0.0672	0.016
0.45	0.58	0.47	1.05	0.16	0.18	0.023

Table 4.4: Cross-section ratio values are shown with symmetric systematic errors.



(a) Normal Ratio with large negative portion.



(b) Normal Ratio shifted by a standard deviation in both the number and denominator.

Figure 4.4: Two different example simulations of a positive definite normal ratio distribution are compared against the propagated normal distribution. The errors are held to be equivalent. The left panel has a worse estimate as it is closer to zero then the right. The distribution is skewed right and the mean estimate in the left panel is off.

not possess a defined second moment. Under specific circumstances, the distribution can be approximated by a Normal distribution. The variance in those cases may be estimated by error propagation. A larger range of values maps to a normal distribution when both numerator and denominator are 3 σ from zero [62]. However, the underlying probability distributions tend to be skew. An example of this is shown in figure 4.4.

Immediately three problems present themselves with the normal approximation. These are

1. Inaccurate propagation estimates,

2. Underlying skew distribution,

3. Strange behaviour near zero.

The strange behavior near zero exacerbates the skewness and propagation problems, and can even make the underlying probability distribution non-central. Counts should be positive definite though. There is a subtraction operation in the current extraction via the Empty Target, which has a chance of making the count estimate appear negative. This measurement is bounded at zero as the normalized subtraction relies on the empty target being a subset of the data. Under appropriate conditions, the probability of negative values in the ratio distribution becomes negligible. Within this regime, the first two points still present a potential issue when attempting first order error propagation.

One simple solution is to transform the variables to a function that is not as skew. $Log(1 + ratio)$ seems natural as it transforms the cross-section ratio into an addition problem as $Log(ab) = Log(a) + Log(b)$. The divergence only depends on the slow divergence of the log function, which tends to only be problematic close to zero. Adding a constant solves this. The disadvantage in this case is that the quantity of interest is no longer the quantity being measured. This in turn makes the measurement less useful.

A higher moment approximation could also make the error propagation formula far more accurate. The disadvantage is that even if transformed properly, the moment does not lend itself to showing the skew nature of the distribution and may be misinterpreted as normal error bars. The confidence interval no longer represents the Normal confidence interval, thus potentially leading to misinterpretation.

Therefore, another simple solution is attempted. The underlying normal ratio probability distribution function is simulated using random number generators for every point in the extrapolation fit. All relevant parameters are now available. In the parameter range the asymmetry corrections are not negligible, but also not large or dominant. Therefore, the study tests the data with asymmetric error bars calculated at the standard probability range. The bars thus represent the equivalent confidence intervals as normal distribution

error bars. One can then intuit the spread of the underlying distributions directly from the graph. The second σ value does not diverge far from the normal approximation. More importantly, the χ^2 estimator remains accurate, converging to the same minimum and has similar properties up through the selected $\pm\sigma$ range. The χ^2 parameter estimator now corresponds to the same probability within the $\chi^2_{min} \pm 1$ region. In total, this correction remains a relatively small term in the calculations, improving the results over the standard normal approximation.

Instead of error propagation, Monte Carlo errors will be preferred when possible. This just models the error using the underlying probability distributions of variables contributing to error. Normally distributed errors multiplied or added together propagate exactly. Therefore, the errors are propagated until the point by point ratios are constructed. Afterwards, the combined error probability distributions are employed in creating a Monte Carlo distribution for each point. These point by point errors are then utilized by the fit. The final cross-section errors are derived like fit parameter errors, using the $\chi^2 \pm 1$ definition. When fitting, the total impact of asymmetric error bars is mitigated by the Central Limit Theorem. Most points are barely influenced and have less total effect when extrapolating together.

4.4 Errors

The errors must now be defined and extracted. The systematic and statistical uncertainty definitions will be chosen to be consistent with probability distribution function groups. Random errors decrease with quantity of measurements number while systematic errors do not. Therefore, systematic errors become the limit of improvement and data taking. Consequently, systematic errors tend to be the largest contributor to PDFs as experiments are generally run to the limit of their precision. SeaQuest was designed to not differ too much in that respect. While the systematic errors do not dominate the random errors, the experiment is approaching the limit of improvement. With all the data included and properly tracked, SeaQuest may be able to push past that boundary.

4.4.1 Statistical Uncertainty

The statistical uncertainty is restricted to the statistical features of the extraction and relate to the dependent variables. The dependent variables in the experiment are the target dimuons and beam protons. These errors are fed through the fitting program in the bin by bin cross-section calculations and are restricted to the subsequent parameter error. The extraction also has some statistical error due to relying on real dimuons.

The most clear source of statistical uncertainty comes from the dimuon of the form $\sqrt{n}$, where n is the bin count. When n is greater than 20, Poisson errors are taken to be approximately normal. The empty target count contribution tends to be larger. Empty target counts have more proportional error than the liquid target counts, and are compounded with similar errors.

Associated with the dimuon counts are the cuts. Final dimuon counts directly depend on the cut applied. Therefore, counting error is highly correlated with cut error. Most impacts of the cut are associated with the relative errors in the relative efficiencies. As the scores aren't implemented in any calculations, the remaining impacts are the distribution and cut choice errors. While cut locations are definitive, an error is associated with the underlying distribution derived from data. The sensitivity is associated with $\frac{dN}{dr}$ of the response curve and ends up possessing a max error of 0.002%. The target counts are propagated together with the rest of the related variables predicated by Eq 4.8. The propagation equation is thus

$$\delta R = \delta \frac{\sigma_{LD_2}}{2\sigma_{LH_2}} = \sqrt{(\frac{\partial R}{\partial N_D})\delta N_D^2 + (\frac{\partial R}{\partial N_H})\delta N_H^2 + (\frac{\partial R}{\partial N_{DH}})\delta N_{DH}^2}. \tag{4.11}$$

As the error of each count follows the Poisson estimate of $\sigma = \sqrt{n}$, the precipitating

total statistical errors fractions are

$$\frac{\delta R^2}{R^2} = \left(\frac{N_H/P_H\Phi_H - N_E/P_E}{N_D/P_D\Phi_D - N_E/P_E}\right)^{-2} \left(\frac{N_D(P_H\Phi_H)^2}{(P_D\Phi_D)^2} + \frac{(N_D P_H \Phi_H)^2}{(P_D\Phi_D P_H)^2 P_H}\right)^2 + \left(\frac{(N_D/P_D\Phi_D - N_H/P_H\Phi_H)(N_D/P_D\Phi_D - N_E/P_E)}{P_E(P_D\Phi_D)^2(N_H/P_H\Phi_H - N_E/P_E)}\right)^2. \quad (4.12)$$

for Runs 2 and 3. By dividing the error by the cross-section, the error formula does not explicitly depend on T_{DH}, and thus Eq. 4.12 applies to Run 5 as well.

4.4.2 Systematic Errors

There are numerous systematic errors associated with different parts of the experiment. Before the extrapolation, the errors can be broken down with into groups. This paper will divide systematic errors come into two main varieties: normal form, and bounded form. The normal systematic errors are defined as errors which have an underlying normal distribution that aren't impacted by the number of measurements. They can be treated similar to statistical errors, and propagated in the same fashion as statistical errors. The bounded errors are propagated using the delta method, which is related to a standard propagation Taylor expansion. The change in the resulting parameter values form the approximate derivative. The method tends to overestimate the impact of independent errors, sometimes to paradoxical levels [63]. This is due to the variable contributions sometimes becoming correlated by the fitting process. However, the bounded errors don't contribute a large enough portion of the error budget for this to be a concern. Thus, while an overestimation, the delta method is chosen to be conservative.

Normal systematic errors are preferably added to the individual points and propagated via Monte Carlo when possible. Two techniques may be applied. When the impact and potential correlations are clear, simple propagation with the statistical uncertainty in computation may be used. A lot of normal systematic errors are associated with the beam properties. Broadly, the beam contributes to the error through intensity and proton count

measurements. The live proton counts normalizing the liquid target counts and empty targets are passed into the point by point error margins. Error is determined through the normalization through the G2SEM rather than the BIM's counting error. The error is derived from the spill-to-spill variation in the normalization. Proton counts are always implemented in a ratio of two proton counts, and therefore the normalization error is diminished. The error doesn't cancel entirely as the normalization variation is a feature of the BIM varying in time, not the actual factor. This error was estimated to be 0.16%. Due to the size of the error MC and delta methods were equivalent.

The pedestal tends to be stable for long periods of time, but does shift over the course of the experiment at least once. It can only be measured during empty buckets, and thus is not known for the bucket during which data is taken. Instead, averaged nearby empty bucket measurements are utilized. Variation in the values determine the pedestal error. The pedestal acts as a translation along the intensity axis. Consequently, the propagated pedestal error relates directly to the higher order fit parameter errors. This dependence results in the error behaving similar to a statistical error trend shown in Table 4.5.

The remaining systematic errors are associated with target properties. Properties of materials derived from external sources have much greater precision than local properties. The contamination error ($\pm 0.2\%$) of the percent hydrogen in the deuterium target uses the delta method of propagation. There is also an error in the target flask length inherited from E866 where the precise target length difference is known to $\pm 0.2\%$. The target volume depends directly on the length, thus the final percent error propagates almost exactly. The Drell-Yan events may also be contaminated by J/Ψ and Ψ'. However, the remaining contamination was estimated to be negligible within the scope of the error estimation. Run 5 does not possess any contamination error. Most of the systematic errors are not large compared to the statistical uncertainty of the fit. The final list of errors is compiled in Table 4.5.

x_2	$\frac{\sigma_D}{2\sigma_H}$	Statistical Uncertainty	Systematic Uncertainties					
			Length	Pedestal	Contam	G2SEM	Fit	Total
0.12	1.200	0.052	0.002	0.0122	0.0002	0.0012	0.0042	0.013
0.15	1.246	0.034	0.002	0.0070	0.0003	0.0020	0.0029	0.008
0.18	1.243	0.283	0.002	0.0025	0.0002	0.0001	0.0012	0.003
0.22	1.244	0.029	0.002	0.0012	0.0002	0.0000	0.0003	0.002
0.26	1.241	0.038	0.002	0.0047	0.0002	0.0001	0.0026	0.006
0.32	1.208	0.045	0.002	0.0047	0.0003	0.0001	0.0043	0.007
0.39	1.125	0.065	0.002	0.0128	0.0002	0.0001	0.0086	0.016
0.47	1.052	0.148	0.002	0.0205	0.0002	0.0002	0.0100	0.023

Table 4.5: Table of the cross-section values with all of the errors. The second column are the statistical, the rest pertain to systematic errors. Note, this includes runs 2,3, and 5.

4.5 Asymmetry Extraction

The asymmetry extraction is performed using the measured dimuon values and the MMHT14 NLO and CTEQ14 LO PDFs. Essentially, the program calculates what the cross-section ratio should be based on the kinematics of the found dimuons and of a given asymmetry. The derivation does not depend on the counts, but on x_1, x_2, and mass. Those values are fed into the PDF for $u(x), d(x), c(x), s(x), \bar{u}(x), \bar{d}(x), \bar{c}(x), \bar{s}(x)$ according to the leading order (LO) equations

$$\sigma_{H_2} = 4u(x_1)\bar{u}(x_2) + d(x_1)\bar{d}(x_2) + 8c(x_1)\bar{c}(x_2) + 2s(x_1)\bar{s}(x_2), \tag{4.13}$$

$$\sigma_{D_2} = \sigma_{H_2} + 4u(x_1)\bar{d}(x_2) + d(x_1)\bar{u}(x_2) + u(x_2)\bar{d}(x_1) + 4d(x_2)\bar{u}(x_1) + 8c(x_1)\bar{c}(x_2) + 2s(x_1)\bar{s}(x_2). \tag{4.14}$$

with additional interaction factors ignored as they cancel in the ratio[8].

The resulting cross-section is then compared to extrapolated cross-section ratios as a function of x_2. The asymmetry value is adjusted according to the difference and this process is repeated until the final cross-section is within a range of 10^{-4} of the true ratio. For

each iteration, the asymmetry is changed according to the derivative of the high x_F approximation Eq. 1.24. This approximation is used here because the precise difference is not as important as the time to converge. The approximate form from Eq. 1.24 can be rewritten as

$$\frac{\bar{d}(x_2)}{\bar{u}(x_2)} = -\frac{1 - \frac{\sigma_{D_2}}{\sigma_{H_2}} + \frac{d(x_1)}{4u(x_1)}}{1 + \frac{d(x_1)}{4u(x_1)} - \frac{d(x_1)\sigma_{D_2}}{4u(x_1)\sigma_{H_2}}}. \tag{4.15}$$

Taking the derivative with respect to the cross-section ratio yields

$$\frac{(1 + \frac{d(x_1)}{4u(x_1)})^2}{(1 + \frac{d(x_1)}{4u(x_1)} - \frac{d(x_1)}{4u(x_1)}R)^2}, \tag{4.16}$$

where R is the cross-section ratio. The asymmetry is thus incremented according to

$$\Delta R(\frac{(1 + \frac{d(x_1)}{4u(x_1)})^2}{(1 + \frac{d(x_1)}{4u(x_1)} - \frac{d(x_1)}{4u(x_1)}R)^2}), \tag{4.17}$$

with an optional velocity term for faster convergence. Convergence was found to be independent of starting position, as tested in leading order at a value of both 0.5 and 2. The dimuons fed into the calculations have a couple constraints. The tracked x_1, x_2, and mass calculations are used in the derivation of the cross-section. Primarily, these dimuons are from the same set as those used in the cross-section derivation, and thus they have the same cuts applied. However, to improve purity of the set an additional occupancy cut requires the trigger intensity to be below 20,000. In the lowest third of the data, small amounts of combinatoric background should remain. The dimuons are subdivided again by the x_2 value into the same bins as the cross-section value. The calculations are performed for each of these individual bins and are given their own error. The asymmetry results are shown in Table 4.6 and figure 4.5.

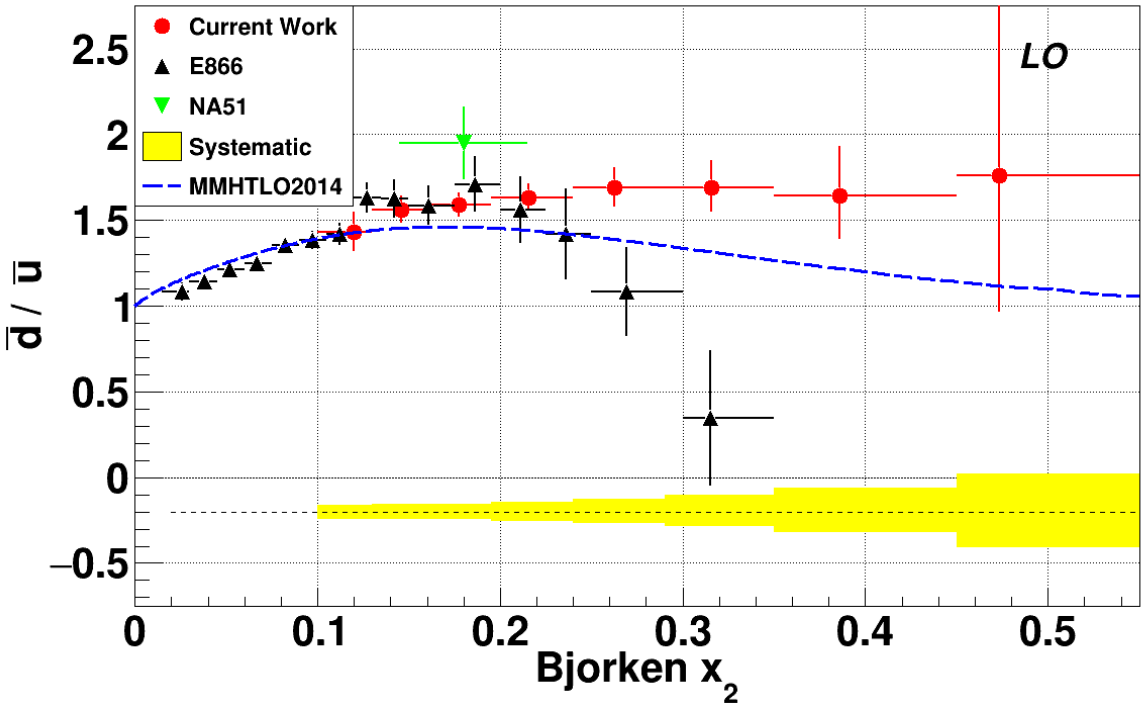


Figure 4.5: Leading order estimation of the proton sea asymmetry as a function of x_2. The asymmetric systematic error is shown on the bottom with a dotted line separating the upper error from the lower error. Experimental results from E866 and NA51 are displayed for comparison as well as the PDF MMHT2014lo68cl.

x_2	$\frac{\bar{d}}{\bar{u}}$	Statistical (lower)	Statistical (upper)	Systematic (lower)	Systematic (upper)
0.12	1.43	0.111	0.118	0.042	0.041
0.15	1.56	0.079	0.082	0.042	0.042
0.18	1.59	0.070	0.072	0.044	0.045
0.22	1.63	0.078	0.083	0.053	0.054
0.26	1.69	0.110	0.118	0.068	0.072
0.32	1.69	0.144	0.156	0.085	0.096
0.39	1.64	0.254	0.293	0.123	0.13
0.47	1.76	0.793	1.161	0.209	0.22

Table 4.6: LO $\bar{d}/\bar{u}$ points are listed with associated errors.

4.5.1 Results

Here, an additional choice was made to increase the rate of convergence of the set. The data was divided into mass and x_1 bins. The LO calculations tested the effects on a random sub-sample of 10% data and found the difference to be less than 1% of the statistical error in each x_2 bin. The version of the extraction program used in this thesis, decreased the run time from a day to less than a minute, while increasing the number of data points processed by a factor of 10 [19]. It is based on work contributed by Bryan Kerns.

There was also a NLO estimation. This estimation uses the MMHT14 NLO and CTEQ14 NLO PDFs to do some of the quark calculations as well as old Fortran CTEQ4 PDFs which control the NLO estimation direction. CTEQ4 did not have any impact from the E866 dataset. The k factor requires an integral over the entire range. The data is extrapolated to $x_2 = 1$ using a quadratic formula. The result possesses more cross-dependence between x_2 bins and large error on the subsequent corrections. The NLO $\bar{d}/\bar{u}$ results are shown in figure 4.6.

4.5.2 Extraction Error

The cross-section error is passed on into the final error calculations. The systematic error uses the delta method, with a couple of points added to more accurately estimate the

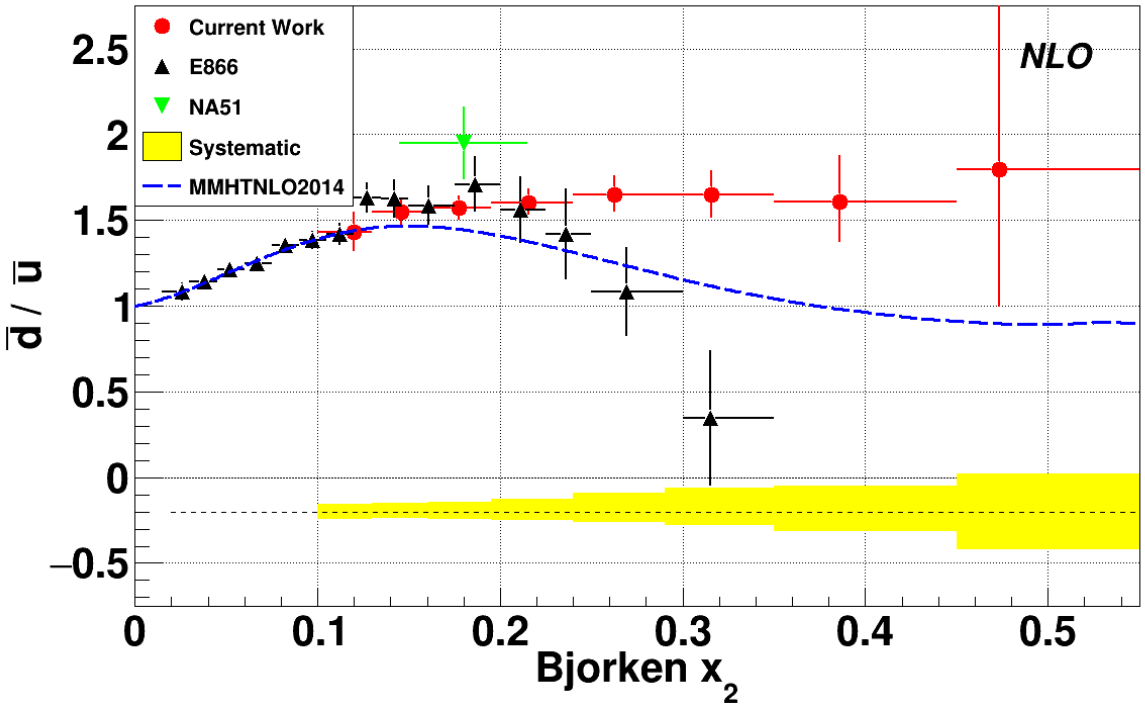


Figure 4.6: Next to leading order estimation of the proton sea asymmetry as a function of x_2. The asymmetric systematic error is shown on the bottom with a dotted line separating the upper error from the lower error. Large differences between NLO and LO are not apparent. The next to leading order PDF MMHT2014nlo68cl is shown for comparison as well as E866 and NA51 data.

derivative for propagation purposes. Errors are expensive in this way, as they multiply the computations required several fold. The statistical errors are propagated using a 10 point model with each passed onto the next level to estimate the asymmetry. No advantage was observed over the delta method, so NLO only used that method.

The extraction process introduces further systematic errors outside of the cross-section. Perhaps the most unambiguous of these is ascribed to the PDFs themselves. These values are derived directly from the PDFs as reported by the LHAPDF library. Another PDF dependent error is the "PDF Group" error, which is derived from the difference in results as derived from different PDF sets. The alternative PDF sets used to derive these values are CTEQ14 LO and abm11 LO. An important factor to be cognizant of is that these latter PDF sets were developed using the original E866 data, which causes them to drop below zero as seen in figure 1.13 in the projection of the E906 data. The central values depend on MMHT14LO as the PDF doesn't drop below unity. The group and PDF errors should be partly correlated as they are developed using similar data sets. The potential correlation is ignored, electing for more conservative errors for the extrapolation of results.

The calculation assumed that $\sigma_{pd} = \sigma_{pp} + \sigma_{pn}$ meaning that no nuclear effects were present. The full impact of nuclear effects are not known, but two different model simulations are explored which consider different potential contributions. The error due to nuclear effects are caused by pion contributions, Fermi motion of constituent particles, binding effects, nuclear motion and off-shell contributions [64, 65]. Kamano and Lee's model includes calculations for Fermi motion and pion exchange contributions [64]. Meanwhile, Ehlers, Accardi, Brady and Melnitchouk's model incorporates Fermi motion and off-shell changes to parton distributions in their computation, but not pion exchange. A physically motivated reason to select one model over the other was not found, so no correction was added to the asymmetry. Instead the difference was incorporated into the systematic error. The two papers predict opposite signed corrections, thus one is taken as the lower error bound, while the other is the upper error bound. The errors increase drastically as a

x_2	Lower Systematic Uncertainties				
	Ratio Sys	PDF Error	PDF Set Var.	Nuclear	Total
0.12	0.029	0.0034	0.0081	0.029	0.041
0.15	0.019	0.0041	0.011	0.036	0.042
0.18	0.009	0.0041	0.0067	0.043	0.044
0.22	0.007	0.0041	0	0.052	0.053
0.26	0.017	0.0041	0	0.066	0.068
0.32	0.022	0.0036	0	0.082	0.085
0.39	0.062	0.0023	0	0.11	0.123
0.47	0.13	0.0010	0	0.17	0.209

x_2	Upper Systematic Uncertainties				
	Ratio Sys	PDF Error	PDF Set Var.	Nuclear	Total
0.12	0.029	0.0081	0	0.029	0.041
0.15	0.019	0.0098	0	0.036	0.042
0.18	0.009	0.0097	0	0.043	0.045
0.22	0.007	0.0097	0.0031	0.053	0.054
0.26	0.017	0.0096	0.020	0.067	0.072
0.32	0.022	0.0084	0.039	0.084	0.096
0.39	0.063	0.0053	0.049	0.11	0.13
0.47	0.13	0.0022	0.0076	0.17	0.22

Table 4.7: List of systematic errors split into the upper and lower contributions. The PDF Error refers to the error as reported by the PDF itself. The PDF Set Var. corresponds to differences between the sets built by different PDF collaborations.

function of x_2. However, when propagated, these errors tend to diverge from the qualitative properties, so the conservative proximal percent error function $\sigma = (0.5 + 5x_2)\%$ is employed [66]. The final set of errors are shown in 4.7. Of the systematic uncertainties, nuclear errors tends to be the largest followed closely by the ratio systematics.

4.6 E866 comparison and Analysis

Reviewing the two experimental results, differences materialize. The cross-sections are higher in the current results, but that can be attributed to the difference in Q^2 between the two experiments. Furthermore, the two sets of $\bar{d}/\bar{u}$ points agree to within 1 σ until the last two E866 high-x_2 points. The two-sample z score of the last E866 point and the x_2 bin

between 0.29 and 0.35 is 2.66. Therefore, the null hypothesis, that these samples have the same mean, is rejected at $\alpha = 0.005$.

Next, the high x_2 regime between 0.24 and 0.45 may be tested against unity. The three ratio bins in this range are all at least 2 σ away from unity, and two of the three are over 3 σ from unity. The data trend shows no indication of decreasing to below unity. The current results contest the trend to below unity in the tested x_2 range and supports a stable high x_2 ratio.

The consistent $\bar{d}/\bar{u}$ at low x_2 appears to suggest only a small Q^2 evolution impact. One possible explanation for the difference in results is that different physical mechanisms come into play. However, SeaQuest is more sensitive to the contested x_2 regime, and E866 had poorly constrained final data points. Thus, a well-motivated statistical explanation is preferred.

The error propagation problem discussed in section 4.1 was not considered in the E866 result. The symmetric error bars suggest a significant portion of values lay below zero, despite the measurement being theoretically positive definite. An approximate range of corrections may be estimated for that point. Assuming the error contributions are approximately the same, the upper error bar may underestimate the spread by a factor of three.

Furthermore, as a first order expansion of error propagation is insufficient to model the variance, the next order terms should be included with the expectation value. Thus if $Ef(x) = f(\mu)$ with $f(x) = \frac{\bar{d}}{\bar{u}}(x)$ to first order, the new expectation value becomes $Ef(x) = f(\mu) + \frac{1}{2}\sigma_f f''(\mu)$. While it is impossible to accurately estimate this second derivative without a handle on the data, it may be concluded that the last points in E906 and E866 can disagree without additional physical justification.

Returning to the current results, the data appears to flatten at high x_2 at a ratio of about 1.6 to 1.8. As the first few points agree with E866 data, many of the comparisons discussed in section 1.3.2 apply to the current work as well. Unfortunately, the last points are not precise enough to estimate whether or not the distribution has peaked. Therefore, not many

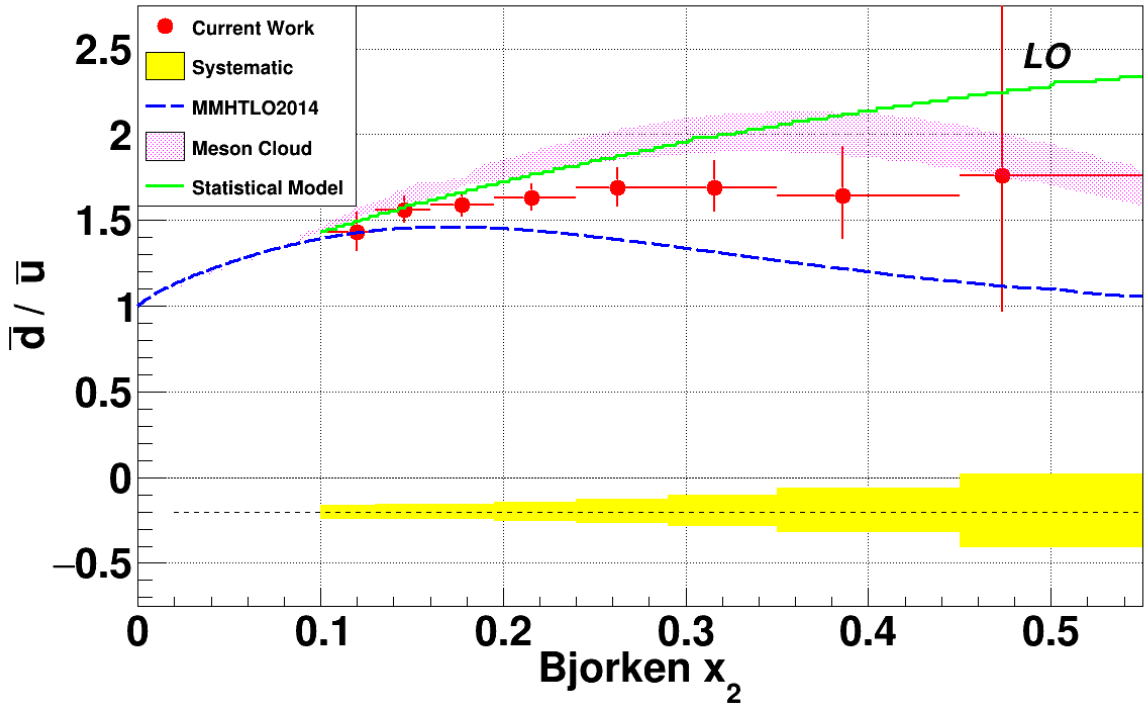


Figure 4.7: Leading order estimation of the proton sea asymmetry as a function of x_2 compared directly to two models at SeaQuest kinematics. One is the meson model[67], whose band size is determined by the vertex uncertainty, the other is the statistical model [68]. The MMHT2014lo68cl PDF prediction is displayed to assist cross-comparisons.

of the potential physics mechanisms can be completely discounted beyond those already in disagreement with the lower x_2 E866 data. However, the meson model A seen in figure 1.10 appears to have qualitative features that are close to the current data set. The Omega Meson model constructed by Alberg et. al. predicts a slowly decreasing $\bar{d}/\bar{u}$ between 1.4 and 1.5 in the x_2 range of 0.23 and 0.38 [67]. As can be seen in figure 4.7, this is within a 1 σ difference with the current result and thus is at least partly supported by the current data. The results don't disagree with limits of the statistical model either, the data . The data does seem to not drop off as steeply as the CQ(S)M models discussed in section 1.3.1 though. Regardless, the data does not imply that the ratio drops below unity in the current x_2 range, and instead supports a stable ratio.

Chapter 5

Conclusions

The SeaQuest experiment looked to extend the flavor asymmetry measurements of the proton sea as a function of Bjorken-x, in the range of 0.1-0.58. Previous studies discovered an anomalously high $\bar{d}/\bar{u}$ suggesting new physics through the violation of the Gottfried Sum Rule. This lead to the development of multiple non-perturbative models. However, E866 later found an unprecedented drop in the ratio to below unity at high x_2. Investigating the emergent behavior was the primary thrust of the current work. In this regard the objective was met. The results agreed with the E866 findings except at the high x_2 points. Furthermore, the $\bar{d}/\bar{u}$ showed no sign of trending towards unity at larger x_2 values. The ratio stabilized around 1.6 instead. The Q^2 values of the two experiments differ by about a factor of two, but the true explanation of the difference may be due to the handling of the computation of the large x data points. To this end, there may be merit in disregarding those E866 points. Most models attempting to explain the asymmetry cannot explain the ratio dropping below unity. These results thus support the development of non-perturbative QCD models. The high-x behavior is part of what discriminates models. Moreover, E866 data has been employed to derive PDFs, which impacts other experiments. Changing the accepted cross-section ratios may thus impact future particle physics studies and even decrease their error by constraining the anti-quark distributions. Unfortunately, the error in the last couple of $\bar{d}/\bar{u}$ points is quite large, making further elimination of non-perturbative models challenging.

SeaQuest still has data left to analyze and has yet to reach the limit set by its systematic

error. Beyond adding more data, future studies could improve upon the current work with the application of nuclear results to place limits on the nuclear effects. These grow rapidly in x and can quickly consume the error budget. As data is added, there may be enough information to resolve $\bar{d} + \bar{u}$ from the data. The results would then depend less on the PDFs, and the $\bar{d} + \bar{u}$ value is allowed to vary while determining $\bar{d}/\bar{u}$.

The machine learning cuts could also be improved by re-weighting the Monte Carlo based on the current results. If the mixed background sets were improved to have more reliable intensity information, the cuts and extraction processes could both be improved. This could greatly reduce the random error in the fits.

The extrapolation procedure this study relies upon doesn't consider the normalization of the empty target as a function of intensity. While underlying intensity distributions should be the same, there is no inherit control ensuring that. This effect would cause additional error in the fit.

Statistical methods and procedures employed by the current project provide an objective, repeatable basis designed for expansion. Current developments are extendable to other data sets. The techniques and related software will hopefully be implemented in future Drell-Yan or medium energy studies. In particular, a new experiment will upgrade the SeaQuest spectrometer over the next year and is the natural heir to this study. Called SpinQuest, the collaboration hopes to measure proton TMDs in the polarized sea. This can build further upon the current work using not just hardware and software experience, but also help place constraints on non-perturbative QCD models of the proton. Many of the models have predictions associated with angular momentum and spin, thus have consequences for transverse momentum distributions. SpinQuest is not alone in measuring TMDs and building global fits either. There is a large space for experiments and studies to constrain the space of possible models.

There are a few proposals for new experiments to study the proton sea on the horizon. In the future, the new J-PARC facility should be capable of extending the measurement,

probing the x regime between 0.3 and 0.7. J/Ψ production at the J-Parc beam energy of 30 GeV is dominated by quark/anti-quark annihilation [69]. This provides a unique way to probe the sea at high x. Another approach to probe the sea is through W^+ and W^- production, which are directly sensitive to the light quark sea. W/Z production can thus be measured at various collider experiments, and thus enable the measurements of the light sea. Proton structure remains a topic of strong interest, with potential discoveries on the horizon. The flavor of the proton sea may still provide insights on the underlying function of non-perturbative QCD.

Appendix A

Accepted Data Additional Distributions

The full kinematic distributions of the data after the cuts are displayed in figures A.1-A.10 for both the hydrogen and deuterium targets. The x_1 and trigger intensity distributions, shown in A.1, A.6 and A.5, A.10 respectively, may be of particular interest as these values were implemented directly in the cross-section extrapolation and asymmetry extraction. The other variables provide the angular information, showing spacial distributions of the dimuons. This information may be used to compare the current study against studies performed by other collaborations. Blue is used for the H_2 target data and red is used for the D_2 target data, both colored for immediate visual clarity. The sets are normalized, but the error bars displayed are derived from the numbers in each bin.

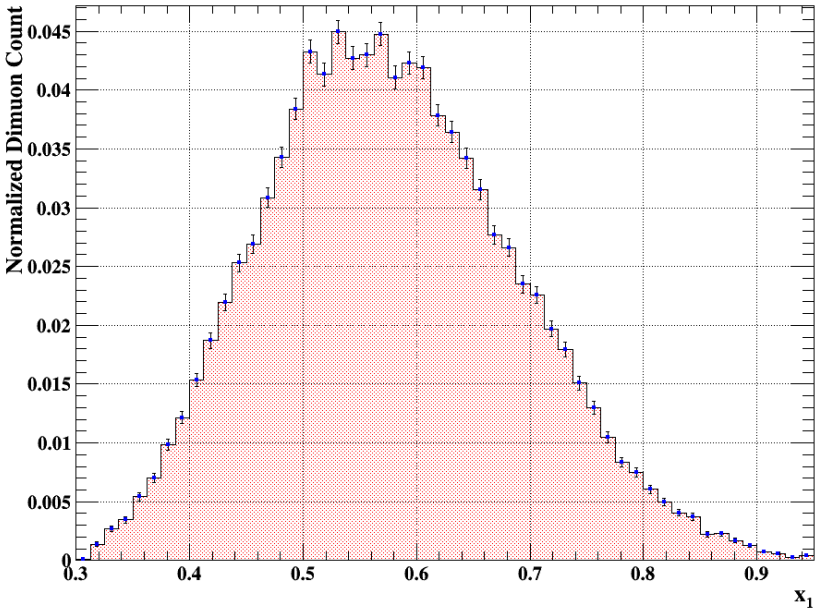


Figure A.1: The normalized deuterium x_1 distribution after cuts is used for calculating the flavor asymmetry.

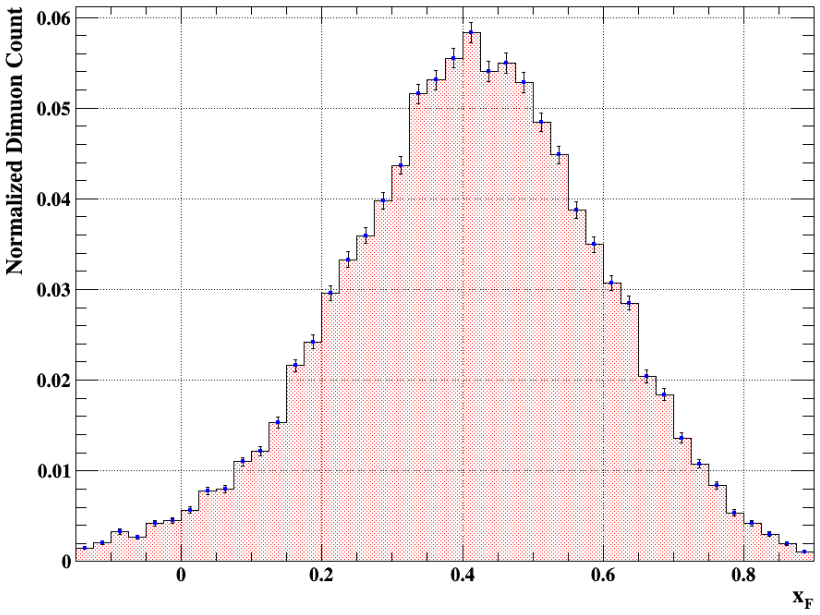


Figure A.2: The normalized deuterium x_F distribution after cuts gives insights on the relative contributions of target and beam anti-quarks.

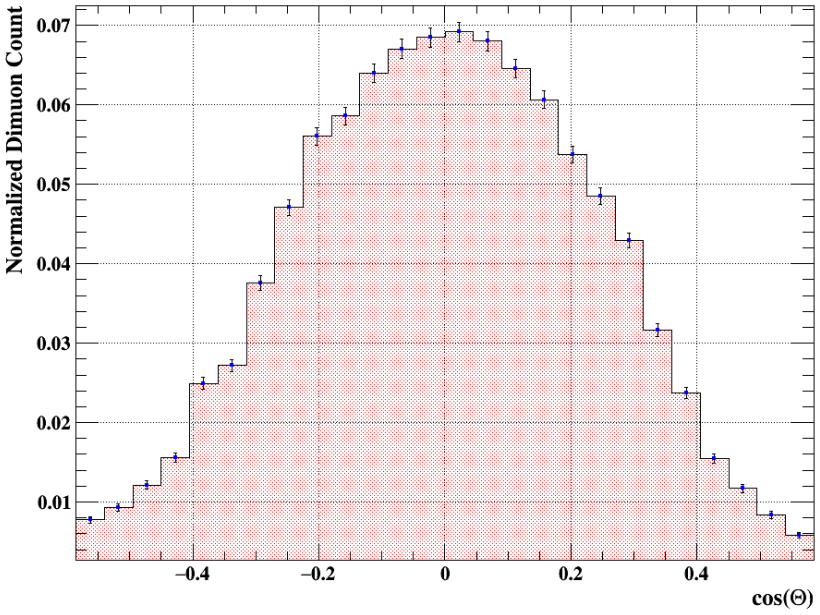


Figure A.3: The normalized deuterium $cos(\theta)$ distribution in the Collins-Soper Frame after cuts which gives information on the angular distribution components of the Drell-Yan cross-section.

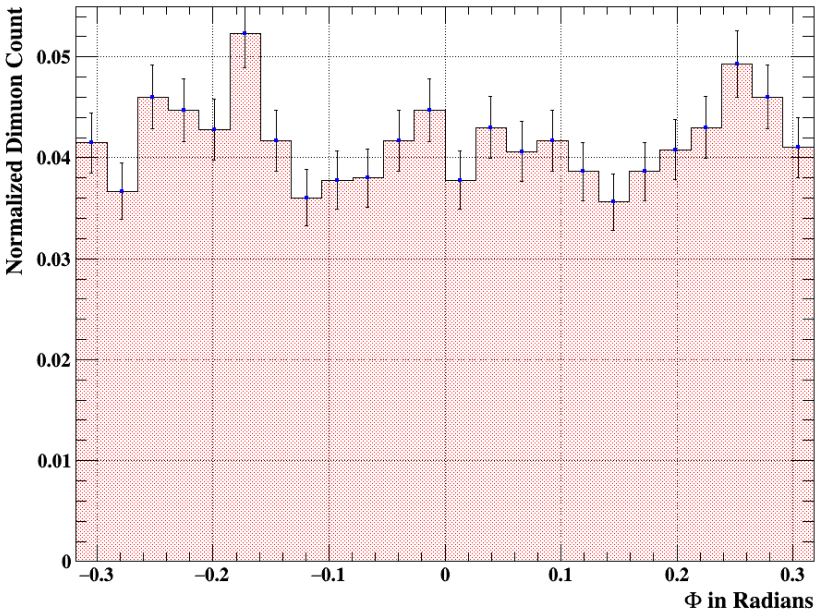


Figure A.4: The normalized deuterium ϕ distribution in the Collins-Soper Frame after cuts which gives information on the angular distribution components of the Drell-Yan cross-section.

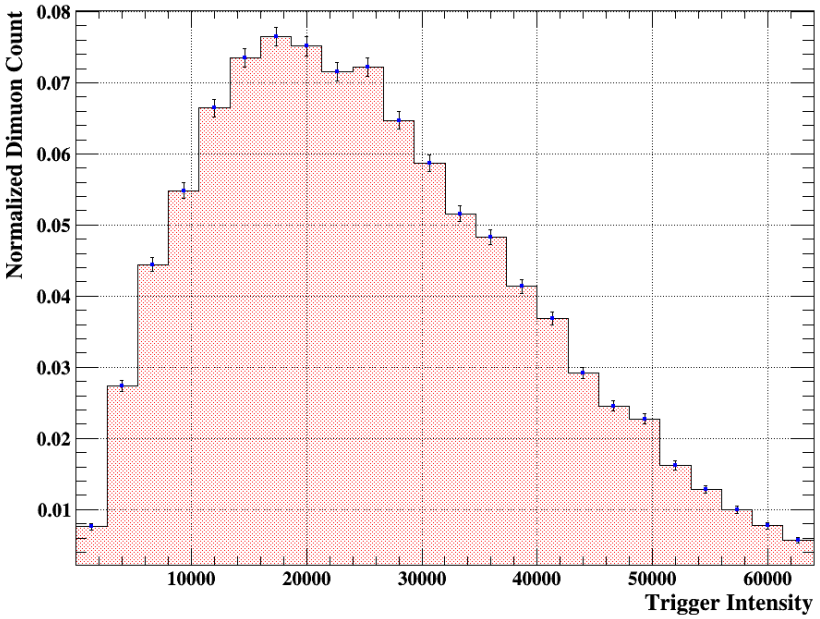


Figure A.5: The normalized deuterium trigger intensity distribution, which is used in extrapolation.

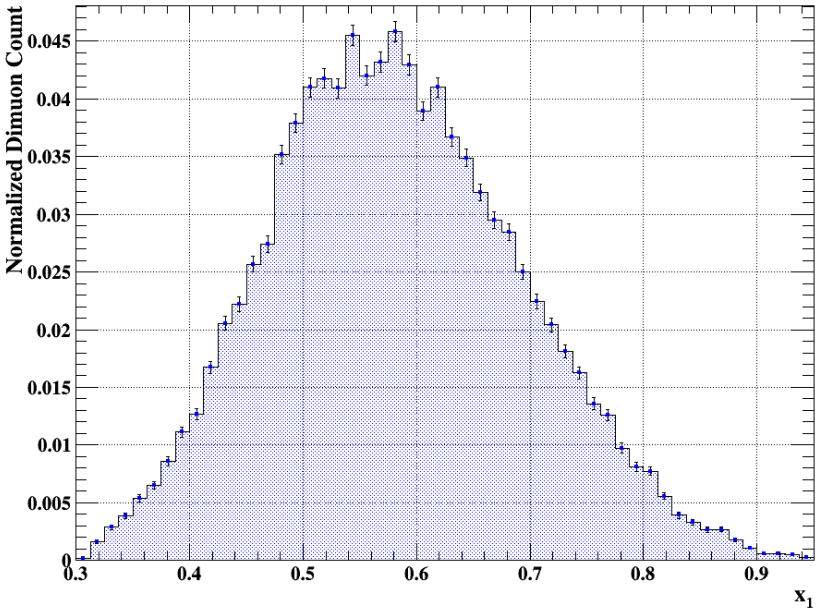


Figure A.6: The normalized hydrogen x_1 distribution after cuts is used for calculating the flavor asymmetry.

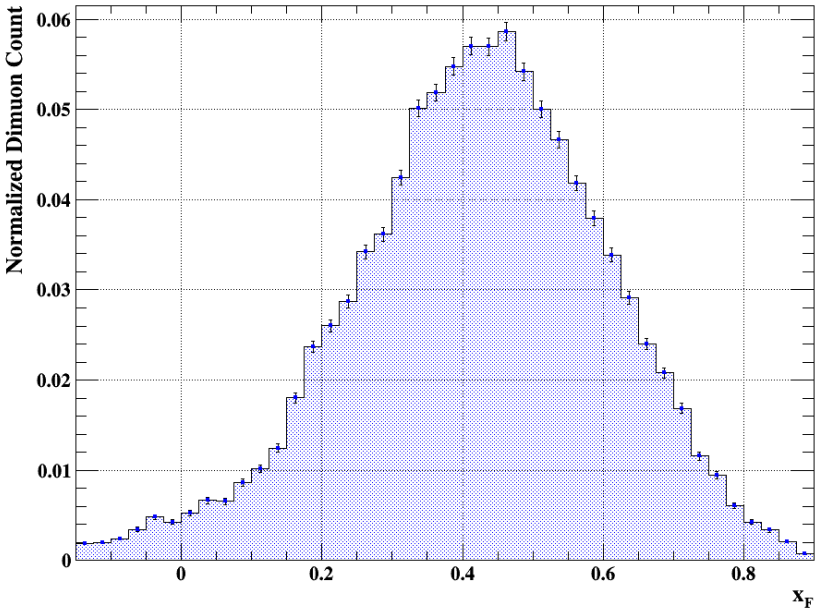


Figure A.7: The normalized hydrogen x_F distribution after cuts gives insights on the relative contributions of target and beam anti-quarks.

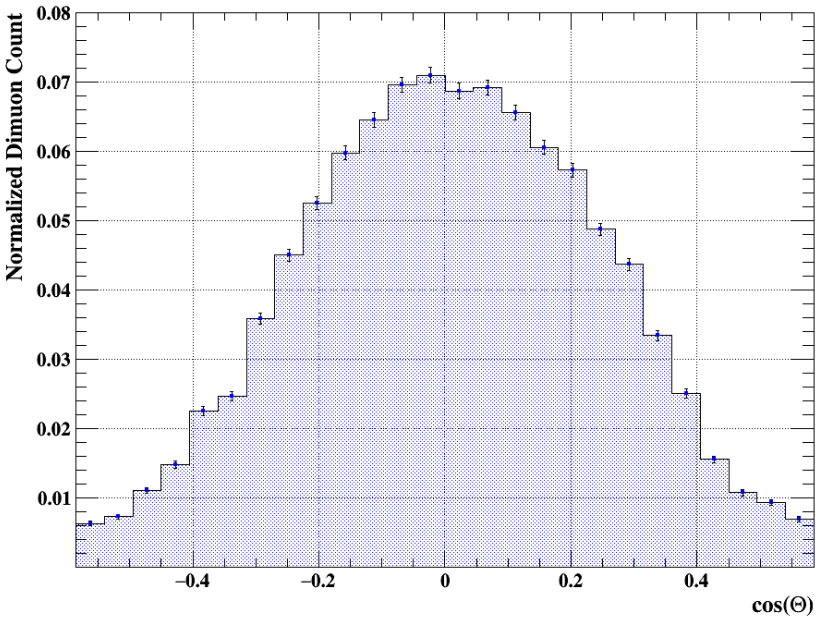


Figure A.8: The normalized hydrogen $cos(\theta)$ distribution in the Collins-Soper Frame after cuts which gives information on the angular distribution components of the Drell-Yan cross-section.

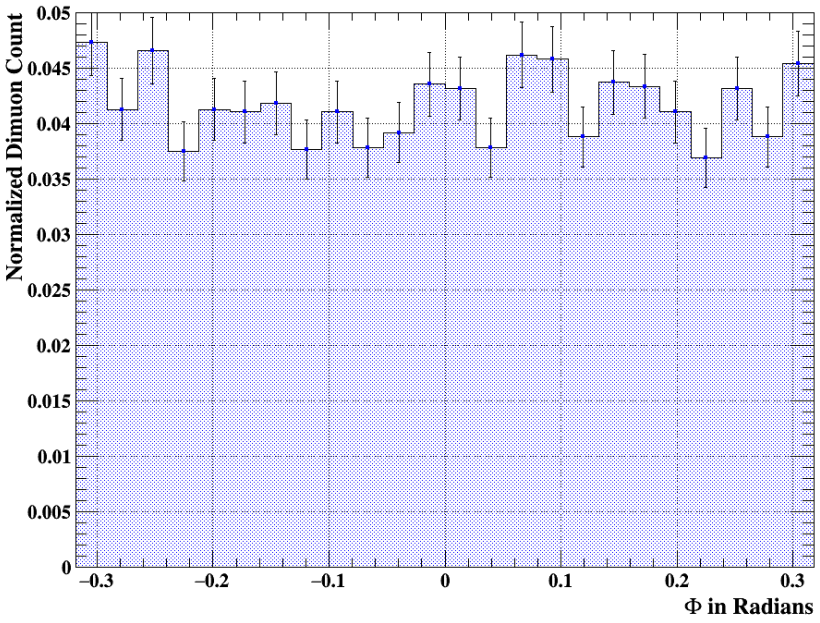


Figure A.9: The normalized hydrogen ϕ distribution in the Collins-Soper Frame after cuts which gives information on the angular distribution components of the Drell-Yan cross-section.

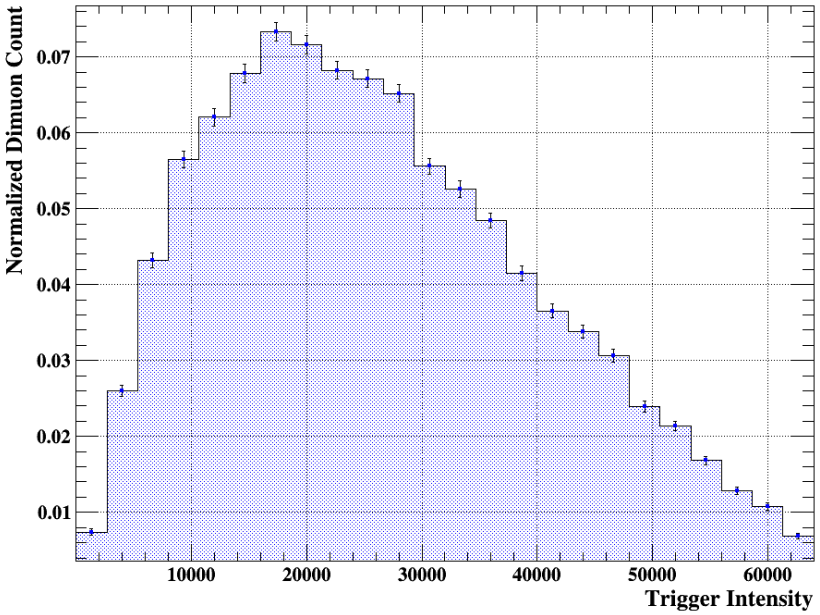


Figure A.10: The normalized hydrogen trigger intensity distribution, which is used in extrapolation.

Appendix B

Kinematic Variable Definitions

Dimuon momentum, mass, θ, and ϕ are all defined in the Collin-Soper frame. The Collin-Soper frame is a center of mass frame with z defined along the line of proton collision. The azimuthal angle is defined as θ and is chosen to be in the lepton plane. The remaining polar angle is ϕ and defines the hadronic plane.

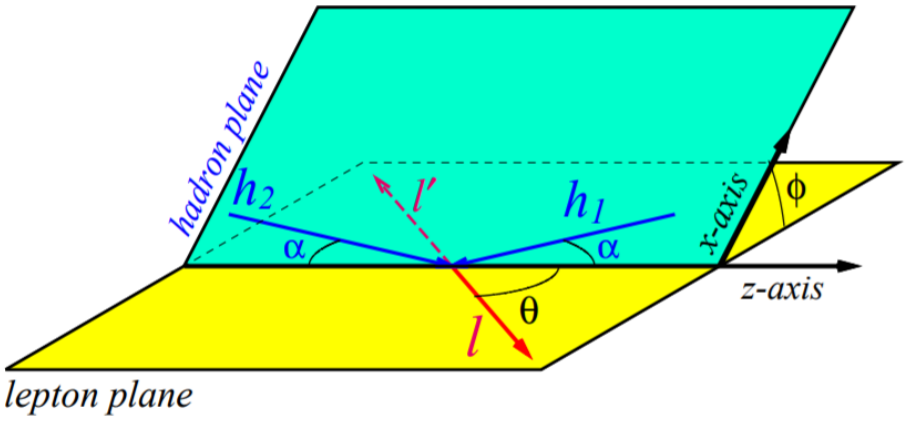


Figure B.1: Diagram defining the Collin-Soper frame variables during a Drell-Yan event. The figure was taken from Ref. [53].

Now Bjorken-x must also be defined in this experiment. Feynmann x, or x_F is defined as $q_{\parallel}/q_{\parallel max}$, where $q_{\parallel}$ is the longitudinal momentum of the virtual photon in this case. It can be written as

$$x_F = \frac{2q_{\parallel}}{\sqrt{s} - M^2/\sqrt{s}}, \tag{1}$$

where M^2 is the mass of the virtual photon and x_1, x_2 is defined by the ratio between the

parton's momentum and its proton's momentum when boosted longitudinally to an infinite momentum frame. Before boosting x may be defined as

$$x = \frac{\gamma(k_\parallel + \beta k_0)}{\gamma(P_\parallel + \beta P_0)}, \tag{2}$$

where $\vec{k}$ is the momentum 4-vector of the parton and $\vec{P}$ is the momentum 4-vector of the proton. Boosting to the infinite momentum frame, $\beta = 1$, yields

$$x = \frac{(k_\parallel + k_0)}{(P_\parallel + P_0)}. \tag{3}$$

However, $k_\parallel$ of a single parton is immeasurable when studying Drell-Yan. Therefore, an approximation had to be used. The common DIS approximation showed earlier is

$$x_i = \frac{Q^2}{2\vec{P}_i \cdot \vec{q}}, \tag{4}$$

which assumes negligible transverse momentum $p_\perp$. However, when there is a small $p_\perp$ the following form is more accurate

$$x_i = \frac{\vec{P}_j \cdot \vec{q}}{\vec{P}_j \cdot \vec{P}_i}. \tag{5}$$

These forms are equivalent and exact when $k_\perp = 0$. This is shown by expanding the x formulas in the small $k_\perp$ approximation. The formula now contains a correction term as a function of $k_\perp$. Using the variable p for the second parton's momentum, yields the formulas

$$\frac{Q^2}{\vec{P}_j \cdot q} \approx x_i + \frac{1}{P_j}\left(\frac{k_\perp^2}{4p_\parallel} + \frac{k_\perp \cdot p_\perp}{2p_\parallel}\right) \tag{6}$$

and

$$\frac{\vec{P}_j \cdot \vec{q}}{\vec{P}_j \cdot \vec{P}_i} \approx x_i - \frac{p_\perp^2}{4P_i p_\parallel}. \tag{7}$$

Equation B.7 has a smaller correction term in terms of $k_\perp$ and $p_\perp$. Therefore, $\frac{\vec{P'}_j \cdot q}{\vec{P'}_j \cdot \vec{P'}_k}$

was selected to calculate x_1 and x_2 in this work.

Appendix C

Cut Variable definitions

The variable definition or expansion for Table 3.1 is included here.

1. **acceptedMatrix** is the trigger matrix count which the trigger supervisor passes to the DAQ for recording for a particular matrix number.

2. **afterIhn(X)** is a prefix for variables in the database designating a count as only including uninhibited events.

3. **BOS** is a short hand name meaning Beginning of Spill used by Fermilab and subsequently all hardware systems. Added to another variable it means the value at Beginning of Spill.

4. **EOS** is a short hand name meaning End of Spill used by Fermilab and subsequently all hardware systems. Added to another variable it means the value at End of Spill.

5. **kTrack** refers to a MySQL data table that contains individual track data returned by the tracker dubbed "kTracker".

6. **Matrix1** is the primary dimuon trigger matrix.

7. **NM4AN** is a beam measure of the magnet's current.

8. **spillID** is the database ID corresponding to a particular spill. These are equivalent to the DAQ spill count.

9. **targetPos** is the database name for target position, each value corresponding to a specific target.

10. **trigger_sum_no_inhibit** is a proton count taken by the BIM, ignoring the inhibit signal.

11. **TSgo** is the raw live trigger count number as reported by the Scalar DAQ computer.

Appendix D

Filter Code

The mySQL query of the cuts used to gather data was

FROM kDimuon a, kTrack b, kTrack c,

WHERE

and a.targetPos>0 and abs(a.dx)<.36 and abs(a.dy-1.6)<.22

and (a.dz between -300 and 2) and abs(a.dpx)<2 and abs(a.dpy)<2

and (a.dpx*a.dpx+a.dpy*a.dpy<6.) and (a.dpz between 37 and 116)

and a.mass between 2.0 and 8.7

and (a.dx*a.dx+(a.dy-1.6)*(a.dy-1.6)<.17)

and (a.xF between -.15 and .95) and (a.xT between 0.05 and 0.5)

and abs(a.costh)<.58

and abs(a.trackSeparation)<345 and a.chisq_dimuon <60

and b.chisq_target<42 and (b.pz1 between 9 and 75) and b.numHits>13

and b.xT*b.xT+(b.yT-1.6)*(b.yT-1.6)<1350

and b.xD*b.xD+(b.yD-1.6)*(b.yD-1.6)<1200

and b.xD*b.xD+(b.yD-1.6)*(b.yD-1.6)>10 and abs(b.roadID)<55000

and b.chisq_target<2.7*b.chisq_upstream and b.chisq_target<11*b.chisq_dump

and b.charge=+1 and (b.z0 between -350 and 85)

and b.chisq/(b.numHits-5)<13

and c.chisq_target<42 and (c.pz1 between 9 and 75) and c.numHits>13

and c.xT*c.xT+(c.yT-1.6)*(c.yT-1.6)<1350

and c.xD*c.xD+(c.yD-1.6)*(c.yD-1.6)<1200

and c.xD*c.xD+(c.yD-1.6)*(c.yD-1.6)>10

and c.chisq_target<2.7*c.chisq_upstream and c.chisq_target<11*c.chisq_dump

and c.charge=-1 and (c.z0 between -350 and 85)

and c.chisq/(c.numHits-5)<13

and abs(b.px1-b.px3+.416)<.008 and abs(c.px1-c.px3-.416)<.008

and abs(b.py1-b.py3)<.008 and abs(c.py1-c.py3)<.008

and abs(b.pz1-b.pz3)<.08 and abs(c.pz1-c.pz3)<.08

and abs(b.chisq_target+c.chisq_target-chisq_dimuon)<17.5

and b.y1*b.y3>0 and c.y1*c.y3>0 and b.y3*c.y3<0 and b.numHits+c.numHits>28

and b.numHitsSt1+c.numHitsSt1>7

and least(b.pz1,c.pz1)-greatest(b.x1,-c.x1)/4.5>11 .

REFERENCES

[1] D. Griffiths, *Introduction to Elementary Particles*. Wiley-VCH, 2004.

[2] M. Gell-Mann, "The EightFold Way: The Theory of Strong Interaction Symmetry", 1961.

[3] R. A. Serway, C. J. Moses, and C. A. Moyer, *Modern Physics*. Thomson Learning Inc, 2005, ch. Elementary Particles.

[4] M. Gell-Mann, "A schematic model of baryons and mesons", *Phys. Lett. 8*, no. 3, pp. 214–215, 1964.

[5] B. Povh, K. Rith, C. Scholz, and F. Zetsche, *Particles and Nuclei: An Introduction to the Physical Concepts*. Springer, 2008.

[6] O. Greenberg, "Spin and unitary-spin independence in a paraquark model of baryons and mesons", *Phys. Rev. Lett. 13*, no. 3, pp. 598–602, 10 1964.

[7] H. D. Politzer, "Reliable perturbative results for strong interactions?", *Phys. Rev. Lett. 30*, no. 26, 1973.

[8] J. C. Collins, D. E. Soper, and G. F. Sterman, "Factorization of hard processes in qcd", *Adv. Ser. Direct. High Energy Phys. 5*, pp. 1–91, 1989.

[9] E906, "Drell-Yan Measurements of Nucleon and Nuclear Structure with the FNAL Main Injector", Tech. Rep., 2005, Funding Proposal Submitted to the DOE, pp. 1346–1349.

[10] K. Nakamura and others (Particle Data Group), "Structure functions", *J. Phys. G 37*, (2010) and 2011 partial update for the 2012 edition.

[11] S. L. Adler, "Adler sum rule", 2009. arXiv: `0905.2923 [hep-ph]`.

[12] J. T. Londergan and A. W. Thomas, "Additional corrections to the gross–llewellyn smith sum rule", *Phys. Rev. D 82*, p. 113 001, 11 2010.

[13] M. Tanabashi, K. Hagiwara, *et al.*, "Review of particle physics", *Phys. Rev. D 98*, p. 030 001, 3 2018.

[14] A. L. Kataev, "The Gottfried sum rule: Theory versus experiment", in *Particle physics in laboratory, space and universe. Intelligentsia and education: Proceedings, 11th Lomonosov Conference on elementary particle physics, and 5th International Meeting on problems of intelligentsia, Moscow, Russia*, pp. 194-200, Moscow Russia, 2003.

[15] M. Arneodo, A. Arvidson, *et al.*, "Re-evaluation of the gottfried sum", *Phys. Rev. D 50*, R1–R3, 1 1994.

[16] F. Steffens and A. Thomas, "On the interpretation of the na51 experiment", *Phys. Lett. B 389*, no. 2, pp. 217 –220, 1996.

[17] A. V. Kotwal, "Recent results from E665", in *The heart of the matter: From nuclear interactions to quark gluon dynamics. Proceedings, 6th Rencontres de Blois, Blois, France, June 20-25, 1994*, pp. 201-210, 1994.

[18] T. M. Yan and S. D. Drell, "The Parton Model and its Applications", *Int. J. Mod. Phys. A29*, p. 0071, 2014.

[19] B. Kerns, "A measurement of the x dependent sea flavor asymmetry at seaquest", PhD thesis, University of Illinois at Urbana-Champaign, Urbana Illinois, 2017.

[20] R. Vogt, "What is the real k factor?", *Acta Phys. Hung. New Ser. Heavy Ion Phys. 17*, p. 75, 2002.

[21] R. Towell, "Measurement of the antiquark flavor asymmetry in the nucleon sea", PhD thesis, University of Texas at Austin, Austin Texas, 2009.

[22] N. I. Kochelev, "Instanton contribution to the quark distribution functions", in *Possible measurements of singly polarized p p (polarized) and p n (polarized) collisions at HERA. Proceedings, 2nd Meeting, Zeuthen, Germany*, 1995.

[23] Z. X. Song H. and B. Q. Ma, *The European Physical Journal C 71*, p. 1542, 2011.

[24] H. Song, X. Zhang, and B. Q. Ma, "Isospin symmetry breaking in the chiral quark model", *Phys. Rev. D 82*, p. 113 011, 11 2010.

[25] R. Vogt, "Physics of the nucleon sea quark distributions", *Progress in Particle and Nuclear Physics 45*, no. Suppl. 1, Mar. 2000.

[26] M. Wakamatsu, "Flavor structure of the unpolarized and longitudinally polarized sea-quark distributions in the nucleon", *Phys. Rev. D 90*, p. 034 005, 3 2014.

[27] P. Schweitzer, "Sea Quark Flavor Asymmetries from Chiral Symmetry Breaking", in *The Flavor Structure of Nucleon Sea INT Workshop INT-17-68W*, University of Washington Institute for Nuclear Theory, 2017.

[28] D. I. Diakonov, "Chiral Quark-Soliton Model", NORDITA, Copenhagen, Tech. Rep. hep-ph/9802298. NORDITA-98-14, 1998.

[29] M. Alberg and E. M. Henley, "Light quark distributions in the proton sea", *Nuclear Phys. A 663-664*, pp. 301c –304c, 2000.

[30] M. Alberg and G. Miller, "Taming the pion cloud of the nucleon", *Phys. Rev. Lett. 108*, p. 172 001, Apr. 2012.

[31] M. Alberg and G. A. Miller, "Chiral Light Front Perturbation Theory and the Flavor Dependence of the Light-Quark Nucleon Sea", 2017.

[32] *Improved Measurement of $\bar{d}/\bar{u}$ Asymmetry in the Nucleon Sea*, Jinnouchi / Shibata Lab Colloquium, Tokyo Institute of Technology, 2009.

[33] G. Ramalho and K. Tsushima, "What is the role of the meson cloud in the $\Sigma*0\to\gamma\Lambda$ and $\Sigma*\to\gamma\Sigma$ decays?", *Phys. Rev. D 88*, p. 053 002, 5 2013.

[34] R. Bijker and E. Santopinto, "Spin and flavor content of octet baryons", *J. Phys. Conf. Ser. 322*, p. 012 014, 2011.

[35] M. Alberg and T. Matossian, "Perturbative and Non-Perturbative Origins of the Proton Sea, 1441", *AIP Conference Proceedings*, Sep. 2011.

[36] C. Bourrely, J Soffer, and F. Buccella, "A statistical approach for polarized parton distributions, 23", *European Physical Journal C*, Sep. 2001.

[37] W. C. Chang and J. C. Peng, "Flavor structure of the nucleon sea", *Progress in Particle and Nuclear Physics 79*, pp. 95 –135, 2014.

[38] R. Gupta, "Introduction to lattice QCD: Course", in *Probing the standard model of particle interactions. Proceedings, Summer School in Theoretical Physics, NATO Advanced Study Institute, 68th session, Les Houches, France, July 28-September 5, 1997. Pt. 1, 2*, pp. 83-219, Los Alamos National Laboratory, 1997.

[39] H. W. Lin, J. W. Chen, S. D. Cohen, and X. Ji, "Flavor structure of the nucleon sea from lattice qcd", *Phys. Rev. D 91*, p. 054 510, 2015.

[40] J. W. Chen, L. Jin, H. W. Lin, Y. S. Liu, Y. B. Yang, J. H. Zhang, and Y. Zhao, "Lattice Calculation of Parton Distribution Function from LaMET at Physical Pion Mass with Large Nucleon Momentum", 2018.

[41] J. C. Peng, W. C. Chang, H. Y. Cheng, T. J. Hou, K. F. Liu, and J. W. Qiu, "On the momentum dependence of the flavor structure of the nucleon sea", *Phys. Lett. B. 736*, pp. 411 –414, 2014.

[42] K. F. Liu, W.-C. Chang, H. Y. Cheng, and J. C. Peng, "Connected-sea partons", *Phys. Rev. Lett. 109*, p. 252 002, 25 2012.

[43] C. Schmidt, J. Pumplin, and C. Yuan, "Updating and optimizing error parton distribution function sets in the hessian approach", *Phys. Rev. D 98*, p. 094 005, 9 2018.

[44] *Parton Distribution Functions and QCD Global Fitting*, eRiken/BNL Workshop:High pT physics at RHIC, Brookhaven: MSU, 2003.

[45] J. Arrington *et al.*, "Drell-Yan Measurements of Nucleon and Nuclear Structure with the Fermilab Main Injector: E906", Fermilab, Tech. Rep., 2006.

[46] C. A. Aidala *et al.*, "The seaquest spectrometer at fermilab", *Nucl. Instrum. Meth. A 930*, pp. 49–63, 2019.

[47] K. Nagai, "Recent measurement of flavor asymmetry of antiquarks in the proton by drell-yan experiment seaquest at fermilab", PhD thesis, Tokyo Institute of Technology, 2017.

[48] P. Reimer, *Seaquest protons on target, so far*, SeaQuest Internal Memo, SeaQuest, 2017.

[49] D. Morton, "$\bar{d}/\bar{u}$ Ratio at SeaQuest", in *APS April Meeting 2017: Quarks to Cosmos*, SeaQuest Collaboration, 2017.

[50] R. Raymond, *Hydrogen/deuterium targets for e906*, SeaQuest Internal Memo, SeaQuest, 2009.

[51] D. Morton and P. Reimer, "Target starting procedure", Fermilab, Tech. Rep., 2017, Safety Documentation.

[52] S. H. Shiu, J. Wu, *et al.*, "Fpga-based trigger system for the fermilab seaquest experiment", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 802*, pp. 82 –88, 2015.

[53] B. Pasquini and P. Schweitzer, "Pion transverse momentum dependent parton distributions in a light-front constituent approach, and the boer-mulders effect in the pion-induced drell-yan process", *Phys. Rev. D 90*, 1 2014.

[54] J. S. Conway, C. E. Adolphsen, *et al.*, "Experimental study of muon pairs produced by 252-gev pions on tungsten", *Phys. Rev. D 39*, pp. 92–122, 1 1989.

[55] M. Tanabashi *et al.*, *Phys. Rev. D 98*, no. 3, p. 030 001, 2018.

[56] M. H. Schub *et al.*, "Measurement of j/ψ and $\psi\prime$ production in 800-gev/c proton - gold collisions", *Phys. Rev. D 52*, p. 1307, 1995.

[57] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, and A. Christov, "Tmva - toolkit for multivariate data analysis", *CERN-OPEN*, 2007.

[58] K. Nagai, *Ktracker-efficiency on extrapolation method*, SeaQuest Internal Memo, SeaQuest, 2018.

[59] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd. Springer-Verlag, 2002.

[60] H. Akaike, "A new look at the statistical model identification", *IEEE Transactions on Automatic Control 19*, pp. 716 –723, 6 1974.

[61] "Potential fitting biases resulting from grouping data into variable width bins", *Phys. Lett. B 735*, pp. 146 –148, 2014.

[62] G. Marsaglia, "Ratios of normal variables", *Journal of Statistical Software 16*, 4 2006.

[63] P. Sinervo, "Definition and Treatment of Systematic Uncertainties in High Energy Physics and Astrophysics", in *PHYSTAT2003, Stanford Linear Accelerator Center*, pp. 122-129, Stanford California, 2003.

[64] H. Kamano and T. H. Lee, "Pion-exchange and fermi-motion effects on the proton-deuteron drell-yan process", *Phys. Rev. D 86*, p. 094 037, 9 2012.

[65] P. J. Ehlers, A. Accardi, L. T. Brady, and W. Melnitchouk, "Nuclear effects in the proton-deuteron drell-yan process", *Phys. Rev. D 90*, p. 014 010, 1 2014.

[66] D. Geesaman, *Nuclear errors*, Private Email Correspondence, 2019.

[67] M. Alberg, E. M. Henley, and G. A. Miller, "Omega meson cloud and the proton's light anti-quark distributions", *Phys. Lett. B. 471*, no. 4, pp. 396 –399, 2000.

[68] E. Basso, C. Bourrely, R. Pasechnik, and J. Soffer, "The drell–yan process as a testing ground for parton distributions up to lhc", *Nucl. Phys. A948*, pp. 63–77, 2016.

[69] W. C. Chang and J. C. Peng, "Flavor structure of the nucleon sea", *Progress in Particle and Nuclear Physics 79*, pp. 95 –135, 2014.