LIMITS ON WEAK ANNIHILATION IN INCLUSIVE CHARMLESS SEMILEPTONIC B DECAYS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Thomas Oliver Meyer

January 2005

LIMITS ON WEAK ANNIHILATION IN INCLUSIVE CHARMLESS SEMILEPTONIC B DECAYS

Thomas Oliver Meyer, Ph.D.

Cornell University 2005

Theoretical predictions for the weak decay $b \to u\,\ell\,\nu$ are subject to the contributions of higher-order terms that must be neglected in making practical phenomenological calculations. One such term that arises at order $(\Lambda/M_B)^3$ in the non-perturbative expansion describes "weak annihilation" (WA). While the contribution of this term to the total rate is expected to be small, it is predicted to have a non-trivial distribution across phase space, concentrated near the maximal value of q^2. The significance of WA relative to the leading-order rate in this restricted region is thus greatly enhanced, and values for the CKM element $|V_{ub}|$ extracted from inclusive measurements of $b \to u\,\ell\,\nu$ made in the high q^2 regime are subject to a corresponding but unknown correction from WA. In this thesis, we analyze 9.7 fb^{-1} of e^+e^- collision data collected by the CLEO detector at CESR during a running period from 1990–1998. We identify $B \to X\,\ell\,\nu$ decays with high-momentum leptons and the proven technology of neutrino reconstruction, and fit the reconstructed q^2 spectrum for contributions from the dominant $b \to c\,\ell\,\nu$ background and the $b \to u\,\ell\,\nu$ signal. The signal shape includes, for the first time, a component representing the contribution of weak annihilation. We constrain the magnitude of WA relative to the total inclusive $b \to u\,\ell\,\nu$ rate and, consequently, the associated uncertainty on extracted values of $|V_{ub}|$. We find the magnitude of weak annihilation effects are not statistically significant and present limits for a series of different models of weak annihilation kinematics.

BIOGRAPHICAL SKETCH

Thomas Oliver Meyer was born the younger and heavier half of a set of twins outside the city of Chicago on May 3, 1975. He and his brother quickly developed a deep, long-lasting bond which led them, among other things, to conspire to "read" the family encyclopedia at the age of eight months, one torn page at a time. His reticence around people other than his brother eventually prompted his admission to Columbia Developmental Pre-School. Soon, although he still had difficulty counting by rote past ten, he was able to raise his hand when he needed to use the bathroom.

After two years of pre-school, Tom had learned to stop crying when separated from his brother and was ready for kindergarten on his own at nearby Willard School. Even at this tender age, he demonstrated a budding interest in marathoning by walking the mile home on one occasion, after having missed the bus. During these early years, his running included participation in the Special Olympics.

His elementary education continued at Clinton and Lords Park Schools before a family move across town brought him to Haines Junior High in St. Charles. At Haines, Tom struggled with indirect objects and algebraic proofs, but ultimately prevailed and was able to sidestep his freshman year of high school. In 1989, he joined other students from around the state at the Illinois Mathematics and Science Academy in Aurora, IL to participate in the early years of an innovative, publicly-funded residential high school. There he first encountered the fearsome "ϵ-δ" proof, managed to solve the notorious "stoplight dilemma" problem in his physics class, and ultimately learned how much could be accomplished in the last eighteen hours before an assignment was due. He graduated on schedule in 1992.

For his undergraduate education, Tom traveled westward to the California Institute of Technology in Pasadena, CA, where he enjoyed the aridity of the southern

California desert and the rigorous training of an intense physics program. There he met his future wife Janis Chang on move-in day, as he carried in his precious filing cabinet. He also developed a deep passion for trail-running, born on the the steep, rocky trails of the Arroyo Seco behind JPL, and furthered with three years of running college cross-country and four years of track.

During the summers of these years, he transitioned from the prep counters of Burger King to the air-conditioned trailers of Fermilab, where his sandwich-wrapping skills gave way to fascination with the PAW analysis package and other particle physics marvels. In 1996, he graduated from Caltech with a B.S. in physics, with honors.

Despite having previously visited Ithaca, NY in the depths of a Siberian winter, he departed the sunny shores of California less than a month after graduation, beginning graduate study in experimental particle physics at Cornell University in the fall of 1996. Almost immediately, he became deeply involved in the prototyping, commissioning, and testing of the new CLEO III detector, a project that absorbed the first four years of his graduate work almost exclusively. In October 1999, he and Janis were married on a perfect fall day in upstate New York, and a week later, he ran his first marathon in Chicago, finishing in under three hours. He and Janis have yet to take a proper honeymoon.

In the fall of 2000, he passed his candidacy ("A") exam and turned his attention to research on the elusive $b \to u\,\ell\,\nu$ decay under the expert guidance of Lawrence Gibbons. He gradually surrendered his many detector "care-and-feeding" responsibilities to younger, abler students, and worked to hone his ability to read others' C++ and to say the words "inclusive charmless semileptonic B decay" ten times fast.

Finally, on a cold day in December 2004, having spent more than eight years in graduate school, run a half-dozen marathons, and overseen the demolition and subsequent resurrection of a new house, he defended his dissertation on weak annihilation in inclusive $b \to u\,\ell\,\nu$ decay.

He plans next to head out into the world of quantitative finance and derivative securities, where his analytical, coding, and modeling skills will hopefully leverage him into a challenging and successful future.

Ozymandias

I met a traveller from an antique land
Who said: 'Two vast and trunkless legs of stone
Stand in the desert. Near them, on the sand,
Half sunk, a shattered visage lies, whose frown,
And wrinkled lip, and sneer of cold command,
Tell that its sculptor well those passions read
Which yet survive, stamped on these lifeless things,
The hand that mocked them and the heart that fed.
And on the pedestal these words appear –
"My name is Ozymandias, king of kings:
Look on my works, ye Mighty, and despair!"
Nothing beside remains. Round the decay
Of that colossal wreck, boundless and bare
The lone and level sands stretch far away.'

– *Percy Bysshe Shelley* (1792-1822)

ACKNOWLEDGEMENTS

Few people ever stay to read all of the credits that roll at the end of a movie, and I suspect far fewer will make it through the end of this section, even though it appears near the beginning of the work. However, despite the possibility of an empty theater at this point in the show, I cannot neglect the support, guidance, and occasional pleading of those that have helped make this entire production possible. Their efforts are immortalized in the mere existence of this work, even if their names are not specifically mentioned here.[1]

The cast and crew, in no particular order:

The rain-or-shine core of the High Noon group: Adam Engst, John Whitman, Rick Hoebeke, Derek Dean, Tim Ingall, Bob Weathers, Geoff Hutchison, Blake Jacquot, Sergey Kiselev, Casey Carlstrom, Boris Dzikovski, Herb Engman, Jay Hubisz, John Hylas, Bruce Roebal, Charlie Fay. You've kept me sane when sometimes it seemed that in spite all of my efforts, the Ph.D. at the end of the tunnel seemed to be receding. Running trails without Tim just ahead wouldn't have been half as fun; I hope you've enjoyed our rivalry at least as much as I did. Adam's optimism and consistent performances always keep me coming back for more, especially when that includes pancakes early on a Saturday morning. Geoff and Bob helped bring back the runner in me, as well as the energetic little kid at times. Jay, Derek, Blake, and Sergey, as companions on the graduate school adventure, have helped remind me it is possible to smile and do science at the same time. Rick, Charlie, JBW, Herb: you made it clear that sometimes, yes, it's just better to be older. Even if you look younger, Bruce! And JTH—I owe what little speed I have and years of injury-free training to you; for that matter, I wouldn't have any thesis data at all without you. Go High Noon!

[1]Even if only as a reader, you, too, deserve credit for accepting the rather ridiculous challenge of reading through this text.

The Finger Lakes Running Club, roadrunner logo and all. In particular: unstoppable president Lorrie Tily, who will surely brush aside any immovable wall she should find; trail master Joe Dabes, whose patience and dedication keep the trail series as hard-core as ever; and again Tim Ingall, whose unlimited energy and volunteerism keep the track and trails accessible and competitive, fun for all.

Scott Roberts, whose succinct yet detailed thesis formed the starting point for my own analysis. I only met Scott once, when visiting Ithaca as a prospective, and I asked him but one question: Where are the trails? His enthusiastic response makes me doubly indebted to him for both my research and my running. I only wish I could write as concisely as he did.

Lawrence, who, perhaps more so than I, was tested sorely in the completion of this project, yet never gave up hope.

Ed Thorndike, the advisor *superioraire* who kept common sense and physics intuition as ingredients in the project at every stage.

Jim Alexander, who fearlessly plies the interface between faculty and students, while also being one of the most down-to-earth particle physicists around.

Véronique Boisvert, my officemate and physics confidant, a fixed point as the analysis whirlpool developed into a real maelstrom. Without you, most of this—and a large part of the lab—simply wouldn't exist.

Gregg and Jana Thayer; together, they represent most of what kept me in the field. Whether it was commiserating over field theory with Gregg, camping in the Adirondacks with them before the onslaught of the TQT gain change, or obscure `tar` and `cvs` arcana, their continual support (peppermint-flavored or otherwise) throughout our many years in the same cage has kept me (largely) sane. Among the lessons we learned together is that you don't *really* understand a problem until you've been burned by it at least once. Literally. (ECL chips can run extremely hot!)

Matt and Katie Shepherd: Matt proved to be the Herculean savior ready to shoulder responsibility for the entire drift chamber when I was looking for a place to set it down; at the same time, he was ready and willing to collaborate on the Brave New World of suez access to old roar data. Outside those arenas, Matt and Katie's dinners and wine-sampling sessions provided the respite we needed before getting back up to slam ourselves into the wall again. And again.

Cornell: my office and lab mates, including the early-morning visits from lab peripatetic Charles Plager, the hours wincing over "other people's code" with Merlin Meyer-Mitchell, the stabilizing, hard-working presence of officemate Andreas Warburton, the quiet yet independent dedication displayed by Nadia Adam.

Longtime friends Dan and Cindy Sinars, who provided many a good meal followed by rousing games of euchre. I'm only sorry that Dan and I didn't quite manage to pull out of our five-year losing slump before you guys set yet another good example by graduating.

At Wilson, chief project architect and one-man army Chris Bebek who put me where I was needed, and where I needed to be. I can't think of the calorimeter without remembering who actually made it work. Similarly, John Dobbins as an indispensable advisor, resource, and colleague throughout the CLEO III upgrade. Chris Jones for maintaining the vision of suez out deep into the Fortran wilderness, and supporting those of us that were drawn, albeit slowly, into the light of the new world. And all the lab members and resources, who in one way or another, made this project possible, as hard as it was, as fun as it was, and as perfect as it was.

My mom, who in some ways was the single executive producer, always behind the scenes, making it all possible, time and time again.

My brother Tim and his wife Emily. Many a night I would console myself that graduate study in physics couldn't be all that bad if my brother had consented to it as well. Together, they provided the relief, support, and stability to make this long ride have a happy ending.

My wife Janis deserves more credit than I can possibly express, for love, encouragement, acceptance, and the occasional necessary rebuke when my priorities became tangled. For you, to you, and with you: this is all that I am.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1
Sketch of the Analysis

We try to orient the reader by providing a preview of the rest of this document, including a brief survey of the relevant physics topics and an overview of the analysis itself.

1.1 Particle Physics

The Standard Model provides a nearly complete description of physics at the scale of elementary particles, describing three of the four fundamental interactions in the framework of relativistic quantum field theory, the combination of special relativity and quantum mechanics appropriate for the regime of high energies and small distances.

In this framework, matter is made of particles known as fermions, which are divided into two categories: quarks (which do interact via the strong interaction) and leptons (which do not). Essentially all of the naturally-occurring matter on our planet (and in the universe) is composed from only two kinds of quarks (u and d) and one kind of lepton (e). But our understanding of the fundamental forces has greatly benefited from studying the entire set of fermions, which in turn tests our theories of ordinary matter.

The predominance of the lightest quarks and leptons that we observe in the universe today has not always been the case. In fact, most cosmological models predict the creation of equal amounts of matter and antimatter in the Big Bang, a proportion that astronomical observation has shown no longer holds today. One ingredient in the explanation for the disappearance of antimatter is a mechanism called CP violation, but in the Standard Model, the predicted magnitude of this effect is not sufficient to account for the observed deficit. By creating and studying the lesser-known quarks and leptons and their antiparticle partners in the laboratory, particle physicists hope to learn the answers to this mystery and other equally fundamental questions.

1.2 The $b \to u \ell \nu$ Decay

One feature of the Standard Model is a phenomenon called "flavor change," in which a quark of one particular type (or flavor) can change into another by emitting the (charged) force carrier of the weak interaction. Fig 1.1 illustrates an example of one such transition, where a b quark decays into a u quark and a W boson. The focus of this analysis are those decays where the W subsequently decays into an $(\ell\ \nu_\ell)$ pair, as shown.

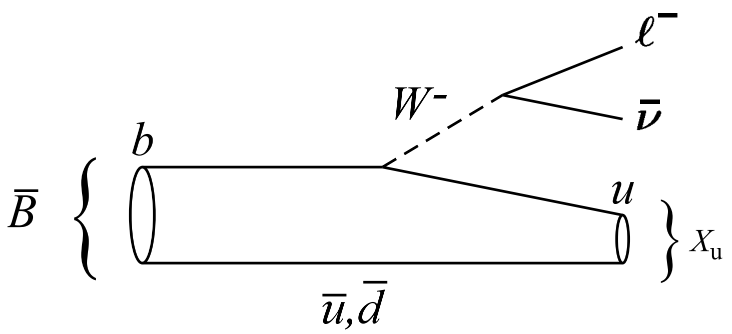


Figure 1.1: Feynman diagram for $b \to u\,\ell\,\nu$ decay. A b quark embedded in a $\overline{B}$ meson decays weakly into a charged W^- boson and a light u quark. Subsequently, the W may decay into a lepton and neutrino. In this Feynman diagram and others like it, time runs from left to right, and the orthogonal coordinate describes spatial separation.

The amplitude for $b \to u$ decay is proportional to the size of a parameter known as V_{ub}, one element of the Cabibbo-Kobayashi-Maskawa (CKM) matrix that governs the general changing or "mixing" of quark flavors. Many experimental and theoretical constraints allow for the implicit determination of the nine elements of this matrix with reasonable accuracy, but at present, V_{ub} remains one of the parameters that is least well-constrained by direct measurement. An excellent avenue for improving knowledge of its magnitude, $|V_{ub}|$, is through direct study of $b \to u$ decays as they occur in the the B meson system, particularly when the short-lived W decays leptonically as shown above. As a convenient short-hand, we label such decays as simply $b \to u\,\ell\,\nu$. An outstanding challenge in these kinds of experimental measurements, however, is the identification of the $b \to u\,\ell\,\nu$ decays amongst the backdrop of an enormous number of other channels for B decay. A final hurdle is that subsequent extraction of $|V_{ub}|$ itself is further plagued by large theoretical uncertainties, some of which are fundamentally unquantifiable.

The chief complication is that quarks are bound by the strong force into particles known as hadrons, and do not occur "bare" in nature. The binding involves the exchange of very "soft" or low-momentum gluons, the force carriers of the strong interaction. At CLEO, for instance, the b quark is always bound to another "spectator quark" ($\bar{u}$ or $\bar{d}$) to form a particle called the B meson. Similarly, this property of quark "confinement" forces the u quark born in the b decay to bind with other quarks to form a complex, generally multi-particle hadronic system, denoted generically as X_u in the figure. The $b \to u$ decay of interest is thus deeply entangled with—even obscured by—the strong physics of quark binding. These bound-state effects greatly complicate the theoretical analysis of $b \to u\,\ell\,\nu$ decays, and makes the problem of extracting $|V_{ub}|$ from experimental measurements nearly intractable.

Recent progress on the non-perturbative strong physics has been made with the development of a rigorous framework known as Heavy Quark Effective Theory (HQET). This analysis provides a systematic way of organizing calculations into an expansion in powers of $1/M_B$, with unknown parameters introduced at every order. It is generally believed that these parameters are universal across all B decays, and so can be determined with input from other experimental measurements. The formalism benefits from tying predictions to the underlying theory of QCD in a well-defined way, and provides for the first time a structure for evaluating many of the uncertainties typically overlooked or ignored when using phenomenological models.

1.3 Relevance of Weak Annihilation

Within the HQET framework, the $b \to u\,\ell\,\nu$ decay can be expanded as a series of operators that contain various combinations of the quark fields and parameters that encapsulate the non-perturbative physics. While the infinite expansion is formally sufficient for the calculation of all inclusive observables,[1] practical treatments are necessarily reduced to including only the first few terms. The neglected terms are then treated as corrections to the leading or next-to-leading predictions for decay rates and spectra.

One of the operators that arises at third order ($1/M_B^3$) in the expansion describes an effective interaction known as "weak annihilation," a contribution roughly analogous to the purely leptonic decay $B^\pm \to \ell^\pm\,\nu_\ell$. In that simpler decay, the two partner quarks inside the B meson annihilate into a W which again decays into a lepton and neutrino. In the more complicated hadronic environment of a $B \to X_u\,\ell\,\nu$ transition, however, the contributions from annihilation processes are more difficult to control theoretically, due to irreducible strong interactions between the "annihilating" pair of quarks and the remaining quarks and gluons, making evaluation of the weak annihilation term an outstanding problem.

Traditional calculations of $b \to u\,\ell\,\nu$ neglect the contribution of weak annihila-

[1] An inclusive measurement focuses only on the details of the lepton and neutrino, summing over all possible X_u final states. The theoretical motivation for this concept is the idea of quark-hadron duality, the notion that when a sufficient number of hadronic final states are included in the measurement, calculations at the quark level should accurately describe the real physics, albeit with non-perturbative corrections as specified by the HQET expansion. The irreducible bound-state effects of each particular resonance X_u become insignificant—in some sense, they "average out" in the consideration of a nearly exhaustive list of possible final states.

tion by truncating the operator expansion at lower order. Predictions for experimental spectra and even methodologies for the extraction of $|V_{ub}|$ are thus subject to an additional uncertainty due to the neglect of this term (and indeed, many others) in the heavy quark expansion. Since the uncertainty is itself difficult to quantify on either the phenomenological or experimental fronts, measurements of $|V_{ub}|$ in B decays are subject to an unknown uncertainty related to weak annihilation effects.

The goal of this analysis is to take one small step on the road toward a precision measurement of $|V_{ub}|$. Our contribution is an attempt to quantify or otherwise constrain the magnitude of weak annihilation effects, especially with regard to the inclusive measurements of $b \to u\,\ell\,\nu$ used to determine $|V_{ub}|$.

The general kinematic features of weak annihilation make it clear that its contribution will be concentrated near the kinematic endpoints of the lepton energy and q^2 spectra. We explore this region for evidence of weak annihilation effects by comparing the spectrum observed in data with a carefully crafted cocktail of simulated data samples that includes a standard leading-order description of the $b \to u\,\ell\,\nu$ rate but that also, for the first time, allows for a possible contribution from weak annihilation. Since there is no clear prediction for the size or extent of weak annihilation effects, we instead attempt to map out the *relationship* between the standard $b \to u\,\ell\,\nu$ rate as typically modeled and potential contributions from weak annihilation. Ultimately, the true $b \to u\,\ell\,\nu$ rate is given by the sum of both contributions, so we quantify our results as a bound on the error made by neglecting the weak annihilation term.

1.4 The Analysis

The analysis uses a sample of B decay events culled from an enormous sample of B decays collected from electron-positron (e^+e^-) collisions at the Cornell Electron Storage Ring, located at the edge of the Cornell University campus in Ithaca, NY. Special criteria devised for this analysis are used to select the events that are likely to have proceeded specifically along the $B \to X_u\,\ell\,\nu$ channel. The selection is designed to be as inclusive as possible, including all possible hadronic systems X_u, and focuses chiefly on the properties of the characteristic lepton and neutrino.

Although we observe the lepton from the B decay directly, we are limited to deducing information about the neutrino based on other observables in the event, since it has vanishingly small probability for interacting with any of the active elements of our detector. Although the neutrino escapes entirely undetected, it carries away some portion of the initial momentum and energy and is thus made visible by its absence. By carefully examining everything else in the event that was detected, we compute the "missing" energy and momentum, and thus infer the neutrino direction and energy.

In order to study and understand the physics and backgrounds of this analysis, we need an accurate and complete simulation of $b \to u\,\ell\,\nu$ decays. Inclusive calculations are correct only in an average sense, however: their faithfulness to the real world only holds when looking on sufficiently coarse scales, above the $\mathcal{O}(100)$ MeV level. In answer to this need, we've developed a "hybrid" Monte Carlo generation package that combines aspects of known exclusive states with the more general requirements imposed by the inclusive theory. This software simulation tool gives us a realistic picture of how the signal decays might appear in the real world.

Although there are no detailed phenomenological predictions for weak annihilation, we employ a simple model that evinces the broad features expected of the process. By widely varying the parameters of this model, we generate a series of possible manifestations of weak annihilation in the $b \to u\,\ell\,\nu$ data sample. Armed with knowledge of the different sources expected to contribute to an observed spectrum, we apply a fitting procedure to determine the relative contributions from the dominant $b \to c\,\ell\,\nu$ background, the $b \to u\,\ell\,\nu$ signal as ordinarily modeled, and our model of weak annihilation. In this fashion, we develop a data-driven estimate of the size of weak annihilation effects, over a broad range of possible weak annihilation scenarios.

Ultimately, we measure a contribution from weak annihilation that is statistically consistent with zero, and set an upper limit on the relative importance of these effects in modeling (and measuring) inclusive $b \to u\,\ell\,\nu$.

1.5 Organization of the Document

The chapters that follow break the analysis into smaller pieces, and are organized as follows. Chapter 2 tries to provide a context for the analysis program as a whole, and then briefly reviews the Standard Model, the physics of the weak sector, and the decay of the B meson. This chapter, rather conversational in style, marks the transition from armchair generalities about particle physics to the more detailed discussion in subsequent chapters of the particulars of this analysis.[2] Chapter 3 provides a more in-depth discussion of the phenomenology of $b \to u\,\ell\,\nu$ decays, with an emphasis on what the current theoretical challenges are in this area. Chapter 4 describes in broad strokes the experimental facilities with which the data were collected.[3] The next chapter (Ch 5) outlines the analysis technique in detail, focusing on the preparation of a data sample expected to be rich in $b \to u\,\ell\,\nu$ with well-understood efficiencies and errors. Chapter 6 begins with a description of a model for weak annihilation, and then describes how the data

[2] As such, Ch 2 may exhaust the casual reader, to whom I address the following: it won't get any better if you keep reading, and no, we didn't see anything interesting.

[3] The CLEO-aware reader is encouraged to skip these three chapters (Ch 2–4) and head directly to Ch 5; do not pass Go, do not collect $200.

were analyzed for the extraction of a limit on the contribution of this process to other measurements of $b \to u\,\ell\,\nu$. Chapter 7 covers sources of systematic error and how the sensitivity to each was translated into a quantitative uncertainty estimate. The last chapter wraps things up with a terse summary of the analysis and a look at things to come. Finally, a few appendices in the back explain more fully some of the tools and algorithms developed for the analysis.

Chapter 2
Background and Context

In this chapter, we provide a cursory overview of the particle physics enterprise as a whole and touch on the relevance of the particular work described in this thesis. Setting the stage in this fashion requires the introduction of new pieces and characters, and a context in which they seem real. Rather than developing the background in excruciating detail, however, we will concentrate on the objects in the foreground that are most essential to the story line. If the subsequent narrative is less than compelling, blame the author for not seeding the tale with the right elements in the beginning, and try to find those passages that still ring authentic and true.

2.1 Introduction

Experimental work often takes place in the context of some model, picture, or other kind of understanding about the actual physics that drives a particular process. The experiment then typically extends the validity of the theory by confirming some previously made prediction. Or perhaps the experiment reveals some unforeseen behavior that requires re-interpretation of the theory or even the construction of a new one. In either case, the theory provides a prediction which is then confirmed or refuted by experiment.

In this particular project, however, the experiment–theory connection is used in a slightly different way. Although particle physicists have an incredibly successful and well-tested theory of matter at the smallest scales (called the Standard Model), there are still certain calculations—and so predictions—that lie beyond the reach of our current computational tools. Such is the case with the charmless, semileptonic decay of the b quark, written schematically as $b \to u\,\ell\,\nu$. A difficult, tedious, and ultimately only approximate calculation can be made to predict the frequency with which this decay occurs, but even more work is required if one desires quantitative details about the various particles in the aftermath of the quark-level decay. The kinematic properties of those daughters, such as the spectrum of energies of the charged lepton, or the range of masses of the hadronic system that forms around the u quark, are, quite bluntly, beyond reach of calculation. Both of these examples are instances where a single, indisputable theoretical prediction simply cannot be made because the phenomenological computation cannot be broken into pieces that are calculable. Instead, modern phenomenologists are reduced to looking for connections between this particular decay process and other decays of the b quark. By applying powerful mathematical analogies, they can transform the experimental measurement of a different b decay process into a prediction about the $b \to u\,\ell\,\nu$ decay. The work described in this thesis is a small example of such a step taken in this "backwards" direction: data is used to essentially "create" a prediction in an area where traditional theory is silent.

The relevance of this particular decay (and so this work) lies in its great utility in measuring a fundamental parameter of the Standard Model called $|V_{ub}|$. This number controls how often a b quark will decay weakly into a u quark and the W particle (described by the reaction equation $b \to uW$), but it plays a more important role in our understanding of the balance between matter and antimatter in the universe today. The fact that this particular decay happens at all (*i.e.* that $|V_{ub}| \neq 0$) makes the Standard Model more interesting because it allows for a mechanism generically called "CP violation," one of the necessary ingredients in any story explaining why the universe seems to have so much matter—and so little antimatter—today. By measuring this key input to the Standard Model, we can better understand the amount of CP violation accommodated by the model, and so help answer the question of whether the Standard Model mechanisms can possibly account for the large imbalance—or asymmetry—that we observe today.

The weight of evidence is already against that possibility: it seems unlikely that the Standard Model can explain the enormous excess of matter over antimatter. In fact, it is clear that the Standard Model doesn't even apply to the extreme energies, densities, and temperatures of the early, evolving universe.[1] What theory, then, does apply? By more deeply exploring the data to which we have access in present-day particle accelerators, we hope to find an answer. There are clues all around that point toward an extension or replacement of the Standard Model that can explain—even predict—the appearance of the universe today, but the challenge is to find enough of these clues to see, even if dimly, what this new picture of the physical world will be. One exciting place to begin this search is in the weak decays of quarks, and in the decay of the b quark in particular, since we have good reason to believe that the new physics will show up quite clearly in this arena. Naturally, we have to be clear about what it is we already do understand so that we can identify the data that don't fit—the clues that can point the way to the future theory. In another context, then, the work described in this thesis is one small step on that road to pinning down what we do know, pressing against the boundaries of what we don't.

The crux of the computational difficulties surrounding $b \to u\,\ell\,\nu$ is the fact that the b quark is never found in isolation in nature. In our context, it is always embedded in a larger particle called the B meson. There the b quark is paired with another (anti-)quark in a bound-state that greatly complicates the simple picture of the $b \to uW$ transition. The quark and its decay cannot be studied on its own;

[1] The Standard Model is often abused as well for its lack of explanatory power, a matter discussed below. Among other things, it doesn't explain the patterns (and patterns within patterns) of the masses, charges, and couplings of the elementary particles.

it must always be disentangled from the arms of both its partner quark in the meson and an ever-present but hazy cloud of other "virtual" particles that blurs the definition of exactly what and where the b quark itself really is.

So, in the laboratory, we are limited to study of the broader category of B meson decays, and must cope with all of the complications this implies. Some decays do occur via the $b \to u\,\ell\,\nu$ channel at the quark level, and identifying these decays is itself a challenge. But there are other difficulties as well: namely, other B decay processes can mimic the $b \to u\,\ell\,\nu$ decay and frustrate efforts to make a clean measurement of the decay rate and so eventually extract an estimate of $|V_{ub}|$. Another difficulty is our limited ability to predict the detailed dynamics of $b \to u\,\ell\,\nu$ decay. Any technique more sophisticated than a naïve free-quark calculation must somehow account for the implicit strong dynamics that envelops the quarks, even if the calculation is built from terms combining the quark fields. The current framework for analyzing $b \to u\,\ell\,\nu$ provides a formal structure for understanding the decay in such a fashion, but practical and theoretical limitations require simplifying the general analysis to just a few terms. One of the neglected terms corresponds to a theoretical mechanism known as "weak annihilation," a process in some ways analogous to the direct annihilation of the two quarks in a meson into the carrier of the weak force, the W boson.

Compared to the total $b \to u\,\ell\,\nu$ rate, the contribution of weak annihilation is expected to be small, but complete description of its kinematics—*where* it contributes—is notoriously difficult to evaluate. The significance of weak annihilation lies in what little is known about its relative importance across the full phase space of possible $B \to X_u\,\ell\,\nu$ decays. In analogy to the monochromatic purely leptonic decay of a meson, the contribution is expected to be effectively limited to a small region of phase space, and for experimental reasons we will explore shortly, it is in this same region that most observations of $b \to u\,\ell\,\nu$ are made. Thus experimental measurements that employ a truncated rate calculation will inevitably neglect the presence of a feature of the decay to which they may be acutely sensitive.

A complete understanding of the $b \to u\,\ell\,\nu$ rate requires knowing the relative size of the weak annihilation term(s) in comparison to the few leading terms retained in the phenomenological calculation. A correct decomposition is clearly essential to interpreting experimental measurements of $b \to u\,\ell\,\nu$, since some portion of the observed, true rate arises from the weak annihilation graph but is not accounted for in current calculations. Our ignorance of weak annihilation thus directly limits our understanding of $b \to u\,\ell\,\nu$, and so the precision with which we can measure $|V_{ub}|$. This can ultimately derail our hopes to test the Standard Model as a viable explanation for why the universe appears as it does today.

The goal of this analysis is to use the data we do have to "inform the theory," to constrain our ignorance of weak annihilation with an experimental measurement. We try to answer the questions: *Just how big can this weak annihilation process* *be? Given that we know (approximately) where it should occur, can we see it?* Do we see it? Can we use data to help determine or at least constrain the details of this process? Can we learn something from the data even in the absence of a robust prediction for weak annihilation?

The key to this puzzle is that even though we don't have a concrete description of the detailed physics of weak annihilation,[2] we have a good handle on where it is important in the the space of B decay configurations, and where it is not. By carefully comparing the data in different regions against itself, we can actually derive a data-driven limit on just how large this unknown process can be, and thus a limit on how much it can impact standard $b \to u\,\ell\,\nu$ measurements. More broadly, our knowledge of $b \to u\,\ell\,\nu$ and the related $b \to c\,\ell\,\nu$ decays has seen significant progress in just the last few years. Such excellent understanding of the semileptonic B decay process as a whole, combined with a careful scrutiny of the data, allows us to tease out some valuable information to feed back into the broader theoretical picture of the general $b \to u$ transition.

With these preliminaries out of the way, let us now turn to a more careful survey of the backdrop for the analysis.

2.2 The Standard Model

The name "Standard Model" is the label applied to the wildly successful picture that particle physicists have developed and refined over the past fifty years to explain the content and interactions of matter at ridiculously tiny distance scales and enormously high energies. What concerns us here is only the empirical content of the theory[3] and a survey of the features of the Standard Model's design that lie behind its success. (For a more in-depth review of the Standard Model, including its development, history, and limitations, consult any of Refs [1, 7–9, 16] or the references listed therein.)

2.2.1 Quarks and Leptons

The Standard Model posits that all matter is made of particles known as fermions, which are divided into the two categories of quarks and leptons. The essential distinction between the two types will become clear below. Fig 2.1 provides some basic information about these elementary particles, listing their names, electric charge, and masses, as well as indicating a grouping into "families" that

---

[2]The ignorance here is best described as two-fold: theorists can't say much about the process, and I can't understand even what little they do say.

[3]Let's not digress with a discussion of the distinctions between a theory and a model. Whoops, I guess we almost did, anyway. . . .

reflects a pattern in nature that the theory itself cannot explain in any deep way. The canonical picture is that these particles are truly fundamental, indivisible, and point-like.[4] The particle physics enterprise for the last several decades has been directed toward first identifying these particles, and then understanding the relationships between them. That assembled knowledge is fully captured in this simple list of what it is that constitutes the observable "stuff" in the universe.[5]

FERMIONS matter constituents spin = 1/2, 3/2, 5/2, ...

Leptons spin = 1/2			**Quarks** spin = 1/2		
Flavor	Mass GeV/c^2	Electric charge	Flavor	Approx. Mass GeV/c^2	Electric charge
ν_e electron neutrino	$<1\times10^{-8}$	0	u up	0.003	2/3
e electron	0.000511	–1	d down	0.006	–1/3
ν_μ muon neutrino	<0.0002	0	c charm	1.3	2/3
μ muon	0.106	–1	s strange	0.1	–1/3
ν_τ tau neutrino	<0.02	0	t top	175	2/3
τ tau	1.7771	–1	b bottom	4.3	–1/3

Figure 2.1: The fermions of the Standard Model. Quarks (which can interact via the strong interaction) appear on the right, leptons (which do not) on the left. Essentially all of the naturally-occurring matter on our planet (and in the universe) is composed from only two kinds of quarks (u and d) and one kind of lepton (e). These three particles combine to form atoms, the chemical building blocks of all matter. Figure reproduced by permission [10].

These universal ingredients[6] do not exist idly on their own, unaware of the existence of other particles. They can influence each other through a set of equally fundamental interactions that make up the rest of the Standard Model.

---

[4] As we shall see later, however, this description can be viewed as nothing more than a convenient fiction that applies under certain circumstances, with no necessary ontological commitment.

[5] The word "observable" is used to qualify this claim since it is now clear that the universe consists largely of so-called dark energy and dark matter, new forms of matter-energy not accommodated by the Standard Model. These phenomena are described as "dark" precisely because they are not directly observable, yet actually seem to dominate the composition of the universe today.

[6] Pun intended. Stay tuned. Later, there's a joke about a traveling salesman and the farmer's daughter.

2.2.2 Interactions

The interactions between the particles in the Standard Model can all be described as the result of only four basic forces, described briefly in Fig 2.2. Gravity is the weakest of the four, and is ignored in the particle physics setting because the masses of the elementary particles are so small. Its effects at these scales are always dwarfed (by many orders of magnitude) by those of the other forces.

PROPERTIES OF THE INTERACTIONS

Property \ Interaction	Gravitational	Weak (Electroweak)	Electromagnetic (Electroweak)	Strong: Fundamental	Strong: Residual
Acts on:	Mass – Energy	Flavor	Electric Charge	Color Charge	See Residual Strong Interaction Note
Particles experiencing:	All	Quarks, Leptons	Electrically charged	Quarks, Gluons	Hadrons
Particles mediating:	Graviton (not yet observed)	W^+ W^- Z^0	γ	Gluons	Mesons
Strength relative to electromag for two u quarks at: 10^{-18} m	10^{-41}	0.8	1	25	Not applicable to quarks
Strength relative to electromag for two u quarks at: 3×10^{-17} m	10^{-41}	10^{-4}	1	60	
for two protons in nucleus	10^{-36}	10^{-7}	1	Not applicable to hadrons	20

Figure 2.2: The four fundamental interactions in the Standard Model. Their relative importance at the subatomic scale is shown, making it clear why gravity is never considered when analyzing the behavior of elementary particles. Figure reproduced by permission [10].

Due to the short distances across which Standard Model particles typically interact (10^{-10}–10^{-15} m), they must be treated with the tools of quantum field theory, the combination of special relativity and quantum mechanics appropriate for this regime of high energies and small distances. In field theories, interactions are described not with mysterious action-at-a-distance forces, but as the direct result of some force-carrying particle impacting the particle being acted upon. Thus each of the forces in Fig 2.2 is listed as being "mediated" by another new particle; more details about these particles (known as bosons) are listed in Fig 2.3. Associated with each particle and interaction is a "coupling strength" that characterizes the strength with which a particle "feels" the interaction.

Note, for instance, that leptons do not interact via the strong force, and only electrically charged particles interact via the electromagnetic force. Neutrinos, in particular, being uncharged massless[7] leptons, are subject only to the weak force.

The second half of the Standard Model thus appears to be simply a second list of particles, this time of ones that carry forces, along with a set of rules about

[7]Recent observations of the phenomenon of neutrino mixing (see references cited in Ref [4]) make it clear that these particles must necessarily have non-zero mass, but other direct experimental limits suggest they must still be extremely light. Given that gravity will not enter into our discussions again, the claim still stands that neutrinos are subject only to the weak force.

BOSONS force carriers spin = 0, 1, 2, ...

Unified Electroweak spin = 1		
Name	Mass GeV/c^2	Electric charge
γ photon	0	0
W$^-$	80.4	–1
W$^+$	80.4	+1
Z^0	91.187	0

Strong (color) spin = 1		
Name	Mass GeV/c^2	Electric charge
g gluon	0	0

Figure 2.3: The bosons of the Standard Model. Rather than being constituents of matter, these particles are really the agents that relay the fundamental forces between all the various particles in the theory. Figure reproduced by permission [10].

the effects they have on the fermions (and each other). To appreciate the deeper content of the theory, however, we need to pause and review how interactions are introduced in the development of the Standard Model. This will, in turn, give us new insight into the rather stunning success of the theory.

As is clear from the fermion table in Fig 2.1, not all elementary particles are alike. They are distinguished by various characteristics such as mass and electric charge, and the curious label “spin” referred to at the top of the chart. In particle physics, mass is really a derived quantity, something that is determined by all the other characteristics of the particle;[8] these other properties are generally called the particle’s “quantum numbers.” These salient traits summarize (for the physicist) exactly how the particle behaves under any given circumstance. For instance, the spin quantum numbers describes how the particle reacts to a magnetic field, and the charge conjugation quantum number C describes the relationship between the particle and its antiparticle. The technical description of the Standard Model thus amounts to an enhanced table of the fermions with many more columns, specifying all of the quantum numbers in explicit detail. With this in mind, let us turn to how the notion of a “gauge charge” for each particle (and with respect to each interaction) arises within the Standard Model.

The incredible theoretical beauty of the Standard Model is that the force-carrying bosons, and the interactions they represent, need not be externally invoked. Instead, they are an almost inevitable result of applying a general[9] principle called *gauge invariance* to the system constructed so far. The forces arise rather

[8] In the full Standard Model, the masses of the fundamental particles is itself generated dynamically, meaning that interaction with yet another force-carrying particle (called the Higgs boson) is what provides a particle with the property called mass.

[9] And considerably sacred

naturally by simply requiring that the fermions "behave reasonably" under certain abstract symmetry operations. This means that aside from the fermion content postulated at the beginning, the entire Standard Model is really a deductive consequence of just a few very natural assumptions. For example, the requirement of electromagnetic gauge invariance amounts to requiring that electrons be insensitive to the local phase of their quantum field; following this statement to its logical conclusion, one finds that an electromagnetic force must exist, and will interact with electrons in a very particular way.[10] This interaction or "coupling term" between the two fields appears with a scaling factor that sets the strength of the interaction. In the case just described, the new term is said to represent a particle of the electron field interacting with a particle of the newly-introduced electromagnetic field, the photon, at some particular point in space and time. The coupling constant turns out to be proportional to the electron charge e. To complete the introduction of this interaction, every fundamental particle in the theory is then assigned a "charge" that describes its susceptibility to the electromagnetic interaction.

In a similar way, the application of other gauge symmetries leads to the identification of the other fundamental interactions and the assignment of "weak" and "strong" charges to each particle. Thus the final, working version of the Standard Model is a list of the elementary particles, their quantum numbers, and their gauge charges with respect to each interaction.[11] The interactions themselves arise during the construction of the theory itself.

Thus we have a simple recipe for deriving the Standard Model, which schematically can be depicted as follows:

$$(\,(\ell,\nu_\ell)_i, q_j) \;\oplus\; (\textsf{gauge symmetries}) \Longrightarrow (\gamma, Z, W^\pm, g) \;\otimes\; (\textsf{interaction rules}).$$

The complete Standard Model requires the specification of only 18 parameters on the left-hand-side to fully determine the right-hand-side, from which predictions can be readily made. These parameters amount to basically an enumeration of the coupling strengths of each of the three relevant interactions, the masses of the fermions,[12] and a few additional parameters that turn up in the development of the full theory.[13]

---

[10]This description, in short form, is how one arrives at the theory of quantum electrodynamics (QED) from a simple armchair exercise.

[11]Whether one counts gauge charges as quantum numbers is perhaps an issue of semantics rather than physics, for the concepts are quite entangled.

[12]In the dynamic mass generation scheme referred to previously (known as the Higgs mechanism), these masses are replaced by couplings to a new scalar field called the Higgs boson.

[13]One enumeration of the parameters of the Standard Model is as follows: six quark masses; three lepton masses; three couplings for the electromagnetic, weak, and strong forces; four quark-mixing parameters; and the Higgs mass and vacuum expectation value.

This discussion perhaps begins to make clear the widespread appeal of the Standard Model and why the theory as a whole is such an efficient and effective tool: if the inputs are valid, the resulting theory will be *deductively* correct. It *has* to apply to the real, physical world, if we believe in the symmetry properties we've imposed. But this same "equation" makes the limitations of the Standard Model framework equally clear. The fermion content of the theory, and the symmetries that are to be respected, are all external inputs. The Standard Model cannot explain *why* the fermions appear with the properties they do, nor *why* the coupling strengths for the different forces fall into the unusual hierarchy that they do. It only offers explanations in terms of its own inputs; it can clearly do no more. Those 18 parametric inputs have to be determined experimentally before any quantitative prediction can result, and those predictions will never explain the seemingly contingent values of those parameters.[14]

The power and limits of the Standard Model as a predictive and explanatory framework should now seem clear. By next understanding the Standard Model as nothing more—and nothing less—than an *effective* field theory, we will see that while it may well be just an approximate description of entirely new physics, it is appropriate in an extremely powerful way for describing *any* phenomena at the distance and energy scales of quarks and leptons. We've argued so far that Standard Model has to be correct (given certain assumptions, of course). We now explore the claim that the world couldn't possibly be any other way—up to a point, quite literally.

2.2.3 Effective Field Theories[15]

To the physicist, size is just about everything. The word *scale* is usually preferred, since it captures the broader notion of a typical, natural, or otherwise characteristic dimension associated with some physical process or object. And in physics, scale is intimately connected with the idea of relevance: It is an unspoken physical law that, in most circumstances, the only physical effects that matter in the analysis of some problem are those that are at the same scale as the problem itself.[16] Most physical intuition is really based just on a grasp of what the

[14]One of the goals of theoretical particle physics is to develop a larger, more inclusive "grand unified theory" that not only unites the three fundamental forces into a single one at sufficiently high energies, but that also predicts or otherwise justifies the values of the parameters of the Standard Model. The task is a challenging one.

[15]While this may appear to be the name of a management seminar for young field theory interns, the word *effective* is used here in the sense of being adequate or sufficient for accomplishing some purpose.

[16]The cases where this isn't so, such as the study of critical phenomena where short-range interactions can suddenly lead to long-range order, are ever the more interesting because of their exceptional character.

relevant scale for a particular problem or process is. Thus, for instance, though quantum mechanics is required to describe the motions of atoms and molecules, ordinary Newtonian mechanics can be used for automobiles, airplanes, and planets, because the scale of those problems (10^{-5}–10^{10} m) is so much larger than that of the quantum world (10^{-8} m and smaller). This notion of "separation of scales" is the same concept at work when we say that Einstein's special theory of relativity is really only useful at "high speeds." Here, the scale is the speed of light, c. When an object's velocity is within, say, a factor of ten of the speed of light, a proper treatment requires the use of special relativity. At slower speeds, neglecting these corrections will have negligible impact on the result.[17]

In particle physics, these notions take on new importance. Historically, until experiments were developed that showed otherwise, it was completely adequate to speak of atoms as the basic constituents of all matter—until one wanted to explore the interactions between individual atoms, *i.e.* until one started looking at the size scale of the atom itself (10^{-10} m). Then quantum mechanics is required, and (famously so), the separate notions of "wave" and "particle" become rather contrived and artificial. Similarly, quarks form bound states with sizes of order 10^{-15} m, and masses in the 0.1–1 GeV range; on these scales, the "quark" notion of the Standard Model is exactly the right physical concept for understanding the physics of mesons and baryons.

Conceptually, the idea is that although a different theory may apply at very small distance scales (complete with new particles, forces, or even concepts of space and time), it is possible to describe the long-range effects of that theory with an "effective theory" that captures the dominant low-energy physics to (almost) arbitrary accuracy. The low-energy theory replaces the details of what's happening on smaller scales with new "effective" interactions that very practically describe the effects that we observe. The new particles or interactions may not correspond to something objectively "real",[18] but for computational purposes, these low-energy "relics" fully describe the observable physics. At the same time that the effective interactions substitute for the "real" high-energy content in a way that preserves the low-energy physics, they also obscures the details of what's going on at higher scales. The high-energy degrees of freedom have been basically "integrated out"[19] and so lost, but what remains is a low-energy theory completely and rigorously adequate for making predictions at the low-energy scales. The theory is also left with an implicit cutoff, beyond which the high-energy physics can no longer be

[17] Some would say that the extent to which this statement that "small effects lead to small corrections" holds is a measure of just how much of the world is linear to first approximation.

[18] We leave it as an exercise for the diligent reader to consider what it means for something to be "real" in this context. To quote the B side of a famous mix tape, "Reality is the only word in the language that should always be used in quotes."

[19] Because we're not sensitive to the high-energy degrees of freedom, our measurements essentially average over them.

ignored.

Thus, the constructs we call quarks and leptons and "fundamental" interactions can be considered as low-energy or large-scale approximations to some higher-energy theory. And until we can actually make observations at those scales, we are left with a guessing game, trying to speculate what ultimate (or at least higher-energy) theory might reduce to the Standard Model in the appropriate low-energy limit.

Viewing the Standard Model as an effective theory is appealing in several ways. First, it neatly transfers the burden of explanation to another, more general theory expected to hold at higher energies. As a mere descendant from a more ultimate theory, the Standard Model is not required to "answer" all of the fundamental questions. Secondly, some of the structure of the Standard Model does become more natural against the backdrop of a larger theory [17].[20]

In conclusion, then, we've seen that the notion of the Standard Model as a renormalizable, effective field theory really means that it consists of interaction terms that arise naturally from the integration over the additional degrees of freedom of a high-energy theory. From this point of view, the fundamental interactions are simply an insightful identification of terms that would appear in the low-energy approximation to an arbitrary high-energy theory. *Rather than being a description of the contingent way the world happens to be, the Standard Model is a complete description of the way the world must look at relatively small scales, no matter what physics is really present at much smaller ones.*[21]

[20]For instance, the low dimension terms that appear in the Standard Model Lagrangian are simply those that dominate at scales much less than the scale M of the high-energy theory. As such, the theory is automatically renormalizable. The appearance of light fermions, and the vector-like and parity-violating couplings of the gauge bosons become natural consequences of chiral and local gauge symmetries of the more primitive theory.

The scheme eventually founders, however, on what is called the "hierarchy problem," the fact that if the Standard Model is an effective field theory, the Higgs and other scalars naturally should have masses at the scale of the new physics, instead of being light ($M \sim 1$ TeV) as is currently expected. The best candidates for resolving this and related problems propose new properties such as supersymmetry, compositeness, or large extra dimensions.

[21]A shorter summary might be this: The physics most relevant for describing the phenomena observed at a particular scale is determined by that scale; that is, the constructs we imagine to exist at one scale may just be the low-energy artifacts of a much higher-energy theory.

2.3 The $b \to u\,\ell\,\nu$ Decay

We now turn to the $b \to u\,\ell\,\nu$ decay itself, with an aim to touch lightly on some of the more interesting bits and pieces, rather than provide an authoritative and exhausting review of the literature on this topic.[22] The tour begins with a look at the phenomenon of "quark mixing," the mechanism behind the $b \to u$ transition; this diversion clarifies the role and relevance of the parameter $|V_{ub}|$. The discussion then turns to semileptonic B decay, inclusive versus exclusive measurements, and concludes with a recognition of the experimental "facts of life" that make the study of $b \to u\,\ell\,\nu$ decays such a challenging and enduring enterprise.

2.3.1 Quark Mingling

At the heart of the $b \to u\,\ell\,\nu$ decay is a simple $b \to u$ transition. This is an instance of quark "flavor-changing" or "mixing", where a quark of one flavor[23] transforms into a quark of a different flavor. In the Standard Model, this change of quark flavor is limited to charged weak interactions, *i.e.* those mediated by a charged $W^{\pm}$ boson. Fig 2.4 shows the lowest-order charged weak current diagram involving the b quark.

The mechanism behind this mixing process is the quantum mechanical principle of superposition. The "real" quarks are eigenstates of the Hamiltonian, and are diagonal in the so-called mass basis, but as it turns out, the weak interaction is not diagonal in this representation. The terms in the Standard Model Lagrangian that describe the weak interaction thus introduce a coupling between the different mass eigenstates. As expected in a (complete) linear vector space, however, there exists a unitary transformation that relates the weak and mass eigenbases. By convention, this transformation is expressed as a 3×3 unitary matrix called the Cabibbo-Kobayashi-Maskawa (CKM) matrix [18] that operates on the charge $-e/3$

[22] For excellent and thorough reviews, see *e.g.* Refs [11, 20, 21, 76].

[23] Recall the discussion of quantum numbers in Sec 2.2.2. For historical reasons, one can also associate with each quark a quantum number measuring the presence of that particle in any other, elementary or not. Thus, for instance, associated with the b quark is a quantum number that measures the b-ness of a B meson or a Λ_b baryon, or even the s quark (where the assignment would be zero). For even more arcane reasons, these quark-content quantum numbers are said to describe the "flavor" of a particle. Thus a b quark changing to a u quark is a change in flavor. (This nomenclature could leave a bad taste in your mouth.)

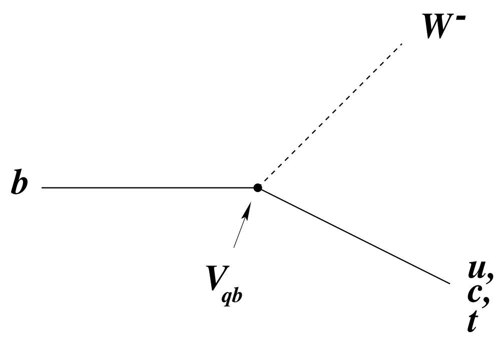


Figure 2.4: The tree-level diagram describing the weak decay of the b quark. At the weak decay vertex, the b quark can decay to a u, c, or (kinematically forbidden but possibly virtual) t quark. In the Standard Model description of the weak interaction, the coupling constant at the vertex is proportional to the the CKM matrix elements $|V_{qb}|$ $(q = u, c, t)$. This matrix is a direct result of the fact that the weak interaction is not diagonal in the quark mass eigenbasis.

quark mass eigenstates d_i and rotates them to the weak eigenbasis d_i':[24]

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{2.1}$$

2.3.2 $|V_{ub}|$

The Standard Model parameter $|V_{ub}|$ essentially measures the probability amplitude for a lone b quark to decay into a u quark mass-eigenstate via the weak interaction. Hence it governs the magnitude of all $b \to u$ transitions, and its precise value is important for making predictions within the context of the Standard Model. This property is of course shared generally by all of the CKM matrix elements, but the element V_{ub} has particular significance. It has been billed as a key ingredient in the incorporation of CP violation into the Standard Model [34] and elsewhere as a "gaping hole" in our understanding [48].

To understand the role of V_{ub} in the Standard Model, it is useful to explore the structure of the CKM matrix in more detail. There are several possible parameterizations, but here we quote the approximation of Wolfenstein [19], which

[24]The weak interaction terms in the SM Lagrangian have the general form $\bar{\psi}_i\gamma^\mu(1-\gamma^5)\psi_i W_\mu$. For the quark fields, $\psi_i \equiv (u_i, d_i')$ for each of the three generations. When the quark mass eigenstates q_i are introduced, the interaction term breaks into pieces of the form $\bar{u}_i\gamma^\mu(\sum_j V_{ij} d_j W_\mu)$, displaying the coupling between each "up-type" quark and all three "down-type" quarks quite clearly.

is sufficient to emphasize the natural (but unexplained) hierarchy of the various elements:

$$V \simeq \begin{pmatrix} 1-\frac{1}{2}\lambda^2 & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\frac{1}{2}\lambda^2 & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix}. \tag{2.2}$$

Here the parameters A, ρ, and η are real numbers intended to be of order unity, and $\lambda \approx 0.22$ is the "small" expansion parameter. (Historically, λ was originally introduced to explain mixing among only two quark generations, and is given by the sine of the Cabibbo angle, $\sin\theta_c$). V_{ub} lies in the upper right-hand corner of the matrix, one of only two terms with a complex phase in this representation. This irreducible[25] phase is the sole mechanism by which the Standard Model explains CP violation; the appearance of this complex number in the Lagrangian essentially requires CP violation to occur in order to preserve overall CPT symmetry. In addition, the imaginary part proportional to η directly sets the scale for such violation according to the usual Jarlskög determinant

$$J_{CP} \simeq A^2\eta\lambda^6. \tag{2.3}$$

Hence both the magnitude and phase of V_{ub} are important in constraining the amount of CP violation predicted by the Standard Model.

The CKM matrix must satisfy unitarity by construction. Because of its ubiquitous role in the description of weak decays, there are numerous experimental avenues for measuring most of its elements, either individually or in various combinations. By checking that independent measurements of the various elements are consistent with each other and the overall requirements of unitarity, we probe for inconsistencies in our current understanding of the Standard Model, ultimately hoping to expose new physics processes beyond it.

The familiar "unitarity triangle" is a graphical way of depicting such consistency checks. The unitarity constraints on the CKM matrix are effectively orthogonality conditions on the rows and columns of the matrix, and each constraint can rather naturally be represented as a triangle in the complex plane. Of the six triangles possible, only the one constructed from the d and b columns has all three sides of comparable size. If the triangle is appropriately normalized, the CKM element V_{cb} specifies the length of the base and V_{ub} determines the location of the apex; see Fig 2.5. Since B decays are readily accessible to experiment, this triangle is a useful tool for representing theoretical and experimental progress on constraining the elements of the CKM matrix. With the wealth of experimental data now available, the triangle is in fact significantly over-constrained, and provides a powerful method for testing the internal consistency of the Standard Model.

[25]The fact that there is a complex phase in the matrix is unavoidable if there are at least three quark generations coupled to the charged weak current in the manner previously described (left-handed doublets). Where the phase occurs in the matrix is

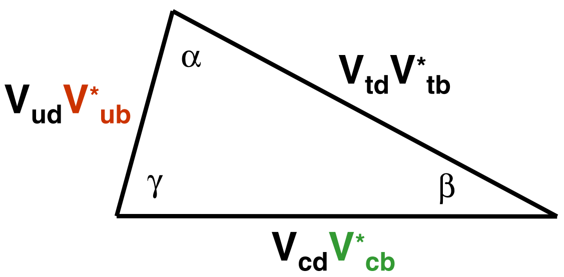


Figure 2.5: The standard unitarity triangle, a geometric construction representing the requirements of unitarity as applied to the CKM matrix. This is the only one of six possible triangles that has a non-trivial appearance, that is, with all sides of the same order (λ^3). Note that $|V_{ub}|$ figures into the height of the triangle.

The measurement of $|V_{ub}|$ plays an important role in this picture, not because of its direct sensitivity to "new" physics,[26] but because the precision on its measured value affects the utility of independent constraints on the triangle's other sides and angles. (It is important to note the distinction between an inconsistency in the experimental measurements of the CKM parameters and a true breakdown of unitarity. The latter is unlikely in sensible physical scenarios, but it is possible (even hoped) that conflicting experimental constraints will expose a new sector of physics beyond that described by the Standard Model. Once the new features are incorporated into the theory, and the experimental data is re-interpreted in light of the new physics, however, the consistency of the triangle will be restored. At no point will the true unitarity of the CKM basis transformation be in doubt.)

It is also evident from Eqn. 2.2 that $|V_{ub}|$ is down by a factor of $\lambda\sqrt{\rho^2+\eta^2}$ relative to $|V_{cb}|$; numerically this factor is approximately 0.08, making it an experimental challenge to even measure the decays $b \to u$ in the face of much larger $b \to c$ backgrounds. This issue of *Cabibbo-suppression* is one of the chief complications to the experimental determination of $|V_{ub}|$; we will return to it again (and again).

---

dependent both on the representation (basis) chosen for the matrix and on the order of the approximation, if any is made.

[26]Measurements of $|V_{ub}|$ can potentially be affected by some types of "new physics," but the deviations are likely to be small. Physics processes with "loops" in their Feynman diagrams, such as $b \to s\gamma$ and neutral B mixing, are much better candidates for hints of new physics. In a loop, a (virtual) new particle can enter directly, conferring enhanced sensitivity to non-Standard Model particles.

2.3.3 Semileptonic B Decay

In a semileptonic B decay, the b quark embedded within the B meson decays weakly to a c or u quark, and the virtual W from that vertex produces a charged-lepton and neutrino pair. The amplitude for the flavor-change $b \to q$ is governed by the CKM element V_{qb}, making the decay rate $B \to X_q \ell \nu$ proportional to (among other things) $|V_{qb}|^2$. Thus these decays offer direct experimental access to V_{cb} and V_{ub}. Furthermore, the relatively large semileptonic branching fraction ($\Gamma_{\rm SL} \gtrsim 10\%$) guarantees that these decays will be numerous in any sample of B decays.

Fig 2.6 shows the standard Feynman diagram for semileptonic B decay. (The presence of both quarks and leptons in the final-state gives this *semileptonic* decay its name.) Note that the partner or "spectator" quark q ($= \bar{u}, \bar{d}$ in Fig 2.6) does not participate directly in the weak decay, but must still adjust in the aftermath to the new circumstances of a suddenly different partner quark (c or u). Often, the spectator quark will combine with additional quarks that emerge from the vacuum to make a more complex hadronic system of several mesons.

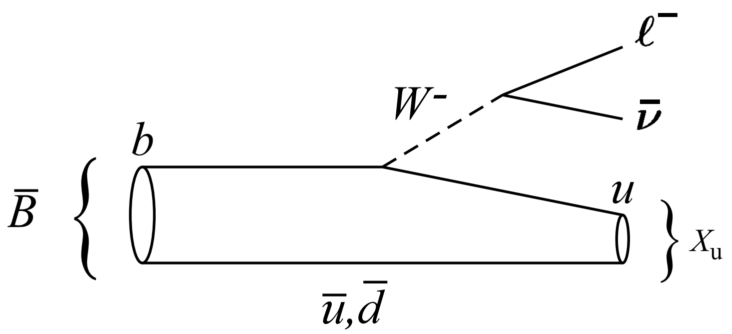


Figure 2.6: Feynman diagram for $B \to X_u \, \ell \, \nu$ decay. In the more general $B \to X \, \ell \, \nu$ decay, the final-state u quark resulting from the weak decay of the b can be replaced by the (heavier) c quark.

By collecting all of the final state hadrons into a single hadronic system X, the decay can be treated as a straightforward 3-body process $B \to X \, \ell \, \nu$. Additionally, it is useful to combine the two leptons (charged and neutrino) into the single virtual W that connects the $b \to u$ decay to the production point of the lepton pair. The variable of choice in the decay is then q^2, the magnitude of the four-momentum transfer to the lepton-neutrino system, also the invariant mass of the virtual W. Denoting the subsequent hadron as X with mass M_X, the following relations follow immediately from considerations of four-momentum conservation:

$$q^2 = M_W^2 = (p_\ell + p_\nu)^2 = M_\ell^2 + 2E_\ell E_\nu - 2|\vec{p}_\nu||\vec{p}_\ell| \cos\theta \qquad (2.4)$$

$$= (P - p_X)^2 = M^2 + M_X^2 - 2ME_X. \qquad (2.5)$$

Here, M represents generically the mass of the initial meson, a B in our case, and P is its momentum four-vector. Note that the second equation has been explicitly evaluated in the rest frame of the B, so there is no contribution from its own initial motion. Taking the neutrino mass as vanishing, the expression in terms of the leptonic variables becomes

$$q^2 = M_\ell^2 + 2E_\nu(E_\ell - |\vec{p_\ell}|\cos\theta), \tag{2.6}$$

which is useful when considering the allowed ranges of the various kinematic variables. For instance, one can easily derive the following bounds for q^2 as a function of the observed charged lepton energy E_ℓ

$$M_\ell^2 \ \leq\ q^2 \leq 2ME_\ell + \frac{2M_x^2 E_\ell}{2E_\ell - M}. \tag{2.7}$$

(Here, we've also taken the lepton ℓ as massless in deriving the upper bound.) It is useful to note here that q^2 attains its global maximum of $(M - M_X)^2$ when the hadronic decay products are at rest in the B rest frame and the lepton and neutrino recoil back-to-back with all of the remaining kinetic energy. In this configuration, the probability for the final-state quarks to form a hadron is quite high, and the form factors are at their maximum. In fact, theoretical predictions are often normalized to this point of *zero recoil*, where, to leading order, the form factors are simply unity at q^2_{max}. Similarly, the form factors reach their minimum when q^2 reaches its global minimum of M_ℓ^2, where the neutrino and lepton are produced collinear with each other and recoil against the final-state quark. Fig 2.7 illustrates these two extreme configurations.

2.3.4 A Hard Problem of Soft Gluons

One might expect that the element V_{ub} could be easily determined from the direct measurement of $b \to u\,\ell\,\nu$ decays.[27] The nomenclature $b \to u$, however, is slightly misleading, since it really refers only to the quark-level decay that occurs within the strongly-interacting environment of the initial- and final-state mesons. Accounting for the "gluon cloud" and light partner quark that surround the b quark within the meson is essentially an intractable calculation in QCD. This ever-present collection of low-momentum or "soft" virtual gluons and quarks is often termed the "brown muck" because of the obscuring effect it has on the underlying physics. Peering through this cloud to grasp the weak physics governing the decay is extremely difficult, because the strong physics cannot be simply disentangled from the weak.

Theorists, however, have tackled this dilemma with the introduction of various *models* and the discovery of several *symmetries* that appear in certain limits. The

[27] This is in contrast to other CKM elements whose values have to be inferred in a less direct manner, from interference or mixing processes, for example.

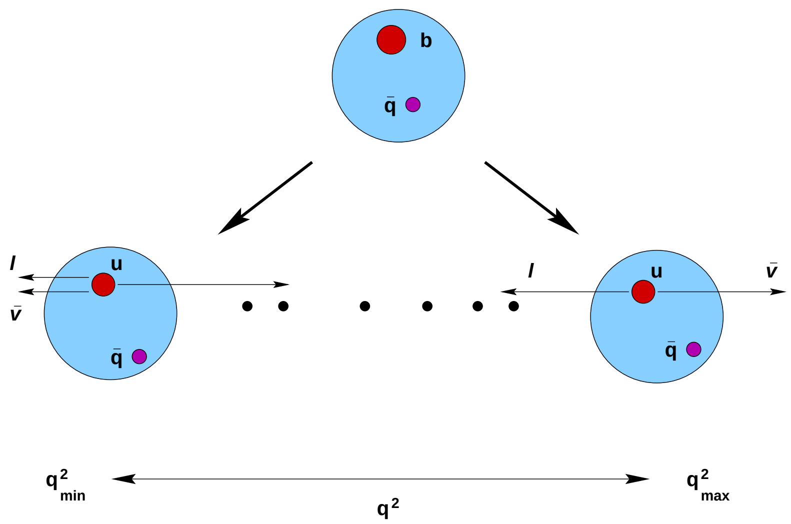


Figure 2.7: An experimentalist's rendition of the kinematic configurations for the semileptonic decay of a heavy meson such as the B. The blob at the top represents the meson before decay. The lower figures illustrate the range of final state configurations after the decay $b \to uW$, with $W \to \ell\overline{\nu}$. The kinematic variable q^2 measures the momentum transfer to the lepton-neutrino system, and spans a range from its minimum value when the lepton and neutrino are collinear (depicted on the right side) to its maximum when the u quark is produced at rest and the lepton and neutrino escape back-to-back (shown on the left). Figure modeled on one found in Ref [21].

semileptonic decay of the B meson is one area rich with the influx of both types of ideas. The fact that the weak decay of the b leads to the production of an $\ell\nu$ pair in the final state (instead of even more strongly-interacting quarks) also helps simplify the theoretical and experimental analysis.

The effects of the strong interaction can usually be disentangled from the well-understood weak physics because the hard parts are limited to the hadronic current, from which the leptonic current completely factorizes. As a result, the effects of the strong interaction can be rigorously parameterized in terms of a small number of *form factors*, which are (unknown) functions of the Lorentz-invariant quantity q^2. In this sense, the weak and strong physics can be separated, and there is some chance that the dynamics of particular decays can be understood.

For example, the partial width for the exclusive decay $B \to \pi\ell\nu$ is easily calculated to leading order as

$$\frac{d\Gamma}{dq^2}(B \to \pi\ell\nu) = \frac{G_F^2|V_{ub}|^2}{24\pi^3} p_\pi^3 |f_+(q^2)|^2 \tag{2.8}$$

where $f_+(q^2)$ is the single form factor that hides the unknown strong physics governing the re-assembly of the b decay products into a final-state pion. This function, however, cannot be predicted directly from the theory of QCD; rather, it is either calculated in the context of certain models of quark interactions, or is determined experimentally. Certain theoretical symmetries then relate the form factor for this particular decay to the form factors in other decays, returning some measure of predictive power. Difficulties remain, however, in assessing the "corrections" to the symmetry considerations, and in evaluating the accuracy of a particular model. Nevertheless, with recent progress on model-independent predictions from the lattice, interest in this avenue is surging, as it promises increased precision on both $|V_{ub}|$ and $|V_{cb}|$. For more discussion of these issues, particularly in the context of exclusive decays, consult Ref [76] and the references therein.

2.3.5 Inclusive Measurements

Put briefly, an inclusive measurement sums over several decay channels that share some common feature (*e.g.* a global quantum number such as lepton or baryon number). However, both experimentally and theoretically, this formal definition doesn't reflect how an inclusive analysis is actually performed, nor what the chief advantages of such an approach are.

Experimentally, one abstracts the salient features of the inclusive decay and ignores the particulars of each channel. An inclusive measurement of semileptonic decay typically makes a statement about all B decays that include leptons in the final state, and all specifics about the hadronic content are ignored. Adopting this technique has consequences on both statistical and systematic uncertainties. Summing over multiple decay channels means a larger data sample, but the lack of a reconstructed decay chain makes it harder to separate signal from background. Writing faithful inclusive Monte Carlo is non-trivial when the "signal" mode is so loosely defined, or equivalently, when many exclusive channels are to be included. Backgrounds have to be well-managed, or at least well-understood, so one is sure of properly selecting the signal modes.

On the theoretical side, an inclusive calculation within the framework of heavy quark effective theory is similarly performed without explicit reference to each possible hadronic resonance. The true benefit of the approach is that that the calculation can actually simplify into one solely at the quark level. The informal equation below captures the spirit of this connection between the physical resonances and the idealized quark-level decay.

$$\sum_i B \to X_u^i \, \ell \, \nu \approx b \to u \, \ell \, \nu \tag{2.9}$$

In essence, the left-hand side is what's actually observed experimentally, and the right-hand side is the underlying weak decay we're after. In the limit of quark-hadron duality, *i.e.* when enough exclusive channels are properly included, the irreducible hadronic physics idiosyncratic to each channel becomes irrelevant and the calculation can be handled at the quark level, albeit with the introduction of a series of effective local interactions and a collection of non-perturbative parameters. These parameters, which arise even at leading order (namely, the quark mass m_b), are to some extent universal across B decays, however, so a predictive framework is still possible. Essentially, the final-state hadronic physics is averaged out, leaving behind the bound-state effects in the B meson, but these can be abstracted and determined from other decays.[28]

Generally, inclusive measurements of decay rates help confirm our understanding of total widths and their decomposition into exclusive modes, and so can also be sensitive to new physics [28]. But the tradeoff between an inclusive and exclusive measurement is by no means clear-cut. The experimental measurement may be easier to implement mechanically and may be subject to different (sometimes independent) sources of error, but it can still require detailed understanding of backgrounds and detector effects. The theoretical prediction benefits from a firm connection to the underlying theory, and can, to a large extent, sidestep the problem of bound-state effects, without reference to a particular model. But the inclusive theory can only make statements about an idealized decay in the limit of quark-hadron duality. And, as we shall see in the sequel to this section, experimental realities can restrict the inclusive calculation in ways that can lead the theory to actually break down, leaving theorists to resort to other, less rigorous tools. Ultimately, progress on $|V_{ub}|$ requires active pursuit along all possible avenues.

2.3.6 In a Nutshell: The Challenge of $b \to u\,\ell\,\nu$

Inclusive study of the semileptonic decay $b \to u\,\ell\,\nu$ requires forging a careful compromise between the twin standards of particle physics analysis: high efficiency and low fake rates. For statistical reasons, it is vital to collect as many events as possible, but to ensure a high-quality result, it is equally essential to make sure that the event sample includes *only* what's intended. Balancing these two ideals is the general experimental challenge.

The snag in the experimental pursuit of $b \to u\,\ell\,\nu$ is singling out the semileptonic $b \to u$ decays from the numerous other B decay modes. Although the semileptonic decay route is one of the more common ones (the branching fraction is about 10.2% [5]), restricting interest to the *charmless* decays $B \to X_u\,\ell\,\nu$ is what makes the analysis difficult. The problem is that $b \to c$ transitions overwhelm $b \to u$ ones by an order of magnitude,[29] so tight experimental cuts are

[28] The non-perturbative parameters must be determined in a sensible and consistent renormalization scheme in order to be of application to another decay. Measurements made in one process cannot be applied to another blindly.

[29] The extent to which the charm contribution dominates depends on the range where one looks. As a whole, the charmless branching fraction is less than 5% of the total

necessary to eliminate the former decays from the event sample. The kinematics of the two semileptonic channels are similar, so most experimentally-accessible variables offer little discrimination—except near the ends of their allowed ranges, where the mass difference between the c and u quarks can become important. The experimental challenge is thus finding—and applying—a discriminating variable (or set of variables) that reliably distinguishes the $b \to u$ decays from $b \to c$ and that has a ready theoretical interpretation.

One historically popular technique for distinguishing $b \to u\,\ell\,\nu$ employs a cut on the energy of the charged lepton ℓ. While this is no longer the only avenue used to isolate $b \to u\,\ell\,\nu$ decays, it captures the essential difficulties that plague most approaches. In the B rest frame, the kinematic endpoint of the lepton energy spectrum can be calculated by considering the case where the (massless) neutrino is produced at rest and the lepton and hadronic system recoil against each other with maximum energy. In this limit, the decay can be treated essentially as a 2-body process, kinematically-speaking. The lepton energy in this circumstance is given by:

$$E_\ell = (M_B^2 + M_\ell^2 - M_X^2)/(2M_B). \tag{2.10}$$

Since the lightest X_c meson (the D at 1.87 GeV) is heavier than the lightest X_u meson (the π at 0.140 GeV), it is clear that the lepton energy spectrum for charmless decay extends beyond that for charm. Numerically, the endpoint for $B \to D\,\ell\,\nu$ is 2.31 GeV, while for $B \to \pi\,\ell\,\nu$, it is higher at 2.69 GeV. The idea behind the experimental cut is that by looking for leptons with energies above the kinematic limit for $b \to c$, one can be confident that the decay is genuine $b \to u$. As Fig 2.8 shows, however, finding these high-momentum leptons is hard because they're rare, and measuring their momentum precisely is critical to ensuring they're true signal. With an overwhelming $b \to c\,\ell\,\nu$ rate sitting just below the charm endpoint, even relatively small or infrequent measurement errors can cause the $b \to u\,\ell\,\nu$ signal region to be flooded by mismeasured $b \to c\,\ell\,\nu$. This situation, where small measurement errors can cause the ever-present $b \to c\,\ell\,\nu$ background to smear into an ideally background-free region and swamp the $b \to u\,\ell\,\nu$ signal, is a classic example of the difficulties on the experimental front. But there is a far more serious obstacle on the theoretical side.

As just noted, the kinematic window for cleanly observing $b \to u\,\ell\,\nu$ is only a few hundred MeV wide—a tiny slice of a spectrum that covers a range an order of magnitude larger—so only a small fraction of the total $b \to u\,\ell\,\nu$ rate is sampled above a cut at the endpoint for decays to charm. In order to extrapolate measurements from this region to the full charmless semileptonic rate, or to interpret even the partial rate in the context of some theoretical expectation, a detailed understanding of the spectrum in the region of the kinematic endpoint is necessary. But the physics in this small window above the charm endpoint is controlled by

semileptonic fraction, but in the region above the kinematic endpoint for charm, $b \to u$ is the only contributor, up to leakage from $B \to X_c\,\ell\,\nu$ due to detector smearing or bias.

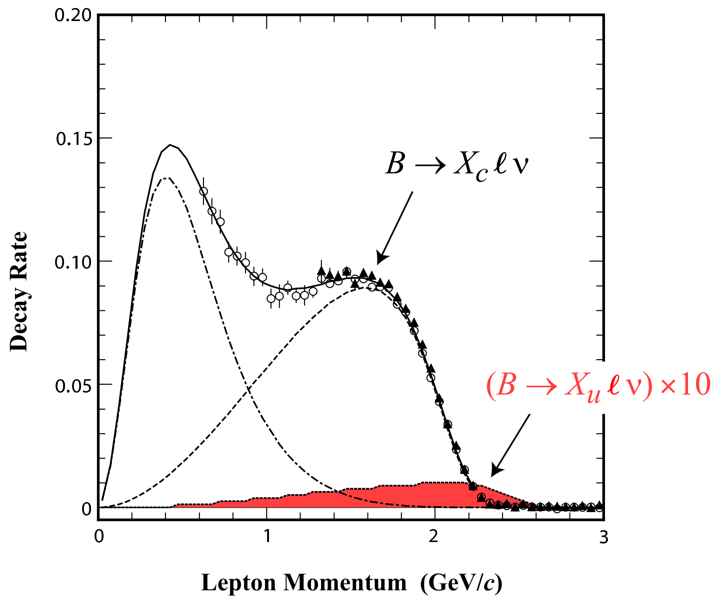


Figure 2.8: The lepton momentum spectrum from B decays as observed at CLEO. The points indicate the measured decay rate in each momentum bin, open circles for electrons, and solid ones for muons. The dashed curve shows the contribution expected from the dominant $b \to c\,\ell\,\nu$ process, and the shaded area at the bottom represents the contribution from the much rarer $b \to u\,\ell\,\nu$ process, which has been increased by a factor of ten for display purposes. The dash-dotted curve shows the contribution from so-called "secondary" leptons produced in the sequential "cascade" chain $b \to c \to s\ell\nu$. The solid curve represents the mathematical sum of all three sources of leptons, and agrees well with the measured spectrum. Figure modified by the author from one that originally appeared in Ref [11].

the same incalculable strong dynamics that binds the b quark into the B meson. To see this clearly, note that Eqn 2.10 reveals that the endpoint for the free quark decay $b \to u\,\ell\,\nu$ is about $m_b/2 \approx 2.3$ GeV (using $m_b \sim 4.65$ GeV), coinciding with the charm endpoint, while the physical endpoint is at $M_B/2 \approx 2.6$ GeV (using $M_B = 5.3$ GeV). It is *entirely* due to the fact that the b quark is embedded in the B meson that the charmless endpoint is pushed up to 2.6 GeV.[30] But the details of how the b quark behaves inside the B meson are fundamentally non-perturbative, and cannot be calculated with the traditional tools of perturbative field theory. So

[30]Another way of seeing this same problem is to note that Eqn 2.10 implies that the change in the lepton endpoint varies with parent mass M according to $\Delta E_\ell \approx M/2$. By changing from an idealized quark-level description with $M \to m_b = 4.65$ GeV to the actual meson-level description with $M \to M_B = 5.3$ GeV, the endpoint is shifted higher by 350 MeV, precisely overlapping the charm-free window so critical to experimental identification of the decay as $b \to u\,\ell\,\nu$.

prediction of the lepton energy spectrum in this region is out of theoretical reach, dictated by the inaccessible details of the b quark motion in the meson.

Thus we are stuck between a rock and a hard place: The lepton energy cut required for experimental identification of the $b \to u\,\ell\,\nu$ signal restricts us to a region of $b \to u\,\ell\,\nu$ phase space where there can be no easy theoretical prediction or even interpretation, essentially orphaning the measurement. But how else to get $|V_{ub}|$?

The lepton energy cut is but one possible experimental technique for isolating $b \to u\,\ell\,\nu$. Cuts on other kinematic variables are possible, with different sacrifices in efficiency and theoretical adequacy. But the interplay between theory and experiment that this example illustrates is typical of almost any analysis of $b \to u\,\ell\,\nu$ decay.

Chapter 3

Theoretical Framework

We attempt to give the flavor of the rich phenomenology of inclusive $b \to u\,\ell\,\nu$ decays. The focus is on the few remaining hurdles on the theoretical side that continue to stall efforts to deliver a well-defined, rigorous, and experimentally viable recipe for a precision extraction of the CKM element $|V_{ub}|$. The story begins with a description of heavy quark effective theory and the analysis of inclusive decays, and then moves deeper into the territory of $b \to u\,\ell\,\nu$ phenomenology to review recent proposals for measuring $|V_{ub}|$, concluding with a survey of lingering difficulties.

3.1 Motivation

Current best-knowledge [2] for $|V_{ub}|$ from inclusive measurements is

$$|V_{ub}| = (4.63 \pm 0.72) \times 10^{-3}. \tag{3.1}$$

Despite almost two decades of experimental and theoretical work on inclusive $b \to u\,\ell\,\nu$, the total precision is still at the 18% level, and remains dominated by theory errors. This simple observation prompts several immediate questions:

- What's limiting the progress? Why can't we do better at $|V_{ub}|$?
- What's been happening for the past 15 years?
- How does the current study of weak annihilation fit into the larger picture of measuring $|V_{ub}|$ to higher accuracy?
- Will the situation ever improve?

To explore these issues and the story that threads them together, we need to first review some of the theoretical machinery commonly used in the study of $b \to u\,\ell\,\nu$ decays. The operator production expansion (OPE) organizes the separation of short- and long-distance contributions to weak amplitudes, and heavy quark effective theory (HQET) provides a rigorous formalism for handling the non-perturbative physics that continually threatens to engulf any calculation. When applied to the computation of inclusive quantities such as (differential) decay rates, these tools offer a systematic framework for calculation and evaluation of correction terms and theoretical errors, especially when working in restricted regions of the full $b \to u\,\ell\,\nu$ phase space.

Once this footing is secure, we will advance to some of the proposals for extracting $|V_{ub}|$ from experimental measurements of semileptonic B decay. The endpoint of the lepton energy spectrum seems simple and very enticing: Historically, it's where the first observation of charmless b decay was made, and in this region, $b \to u\,\ell\,\nu$ has a very simple experimental signature. But as we'll see, unusual efforts are required to make any theoretical sense of decays in this part of the spectrum, and ultimately other experimental measurements (such as the photon energy spectrum in $b \to s\gamma$) are needed to extract $|V_{ub}|$ with any hope of well-understood errors. This example will expose more fully the tension between theoretical and experimental efforts to pursue $|V_{ub}|$. Two other possible approaches to $|V_{ub}|$ will further illustrate the depth and complexity of this convoluted relationship: a cut on the mass of the hadronic system (X_u), or a cut on the mass of the dilepton $(\ell\,\nu)$ pair. These examples will serve as a springboard for discussion of the remaining challenges that still make the path to a successful, unqualified measurement of $|V_{ub}|$ a rocky, uphill climb in wind-searing cold.

3.2 Infrastructure

Even in the weak decay $b \to uW$, the effects of the strong interaction cannot be ignored. Due to a property of QCD called confinement, the b quark is always embedded in a strongly-interacting environment of other quarks and gluons, distracting from the essential weak physics of the decay. The consequences for calculation are dire: no meaningful prediction about hadrons in the real world can be made without accounting for the confinement of the quarks within them, but the physics of this binding is fundamentally non-perturbative, arising from a theory that is strongly-coupled in this regime. Using the tools of the OPE, renormalization, and HQET, however, the problem can be broken into well-constrained, manageable pieces. The key insight is the identification of several separations of scale in the weak physics describing B meson decay [23–31, 33].

3.2.1 Operator Product Expansion

There is a natural separation in scales between the typical hadronic energy scale $\mathcal{O}(1\ \text{GeV})$ that characterizes the binding of quarks into hadrons and the scale $\mathcal{O}(M_{W,Z})$ of weak interactions. The operator product expansion (OPE) is a computational and conceptual tool for taking advantage of this separation in a rigorous way, particularly when calculating the amplitude for weak quark decay. Very generally, the OPE isolates the short-distance physics that can be readily calculated with the tools of perturbation theory and the renormalization group, and replaces the long-distance physics with a set of hadronic matrix elements that wrap up the non-perturbative subtleties.

As a rather trivial but instructive example of this technique, consider the four-quark interaction $c \to su\bar{d}$,[1] shown at tree-level in Fig 3.1. The amplitude A for the process, constructed according to the usual Feynman rules, is

$$A = -\frac{G_F}{\sqrt{2}} V_{cs}^* V_{ud} \, \frac{M_W^2}{k^2 - M_W^2} \, (\bar{s}\gamma^\mu P_L c) \, (\bar{u}\gamma_\mu P_L d), \tag{3.2}$$

where P_L is the projection operator for the left-handed component of weak isospin, and k is the momentum transfer through the W propagator. Noting that k^2 is small compared to M_W^2, we expand the propagator and find

$$A = -\frac{G_F}{\sqrt{2}} V_{cs}^* V_{ud} \, (\bar{s}\gamma^\mu P_L c) \, (\bar{u}\gamma_\mu P_L d) + \mathcal{O}(\frac{k^2}{M_W^2}). \tag{3.3}$$

In the limit of small k^2, we discard the second term and approximate the full amplitude A by the first, creating a (very) simple effective theory for this charged-current weak interaction. Pictured on the right of Fig 3.1 is the new vertex that results, describing an effective four-quark interaction with a new coupling. By including the next few higher-order terms from the expansion of the propagator, we can introduce additional new local interactions and thus expand the range of the effective theory.

The discarded terms in Eqn 3.3 can alternatively be considered as higher-dimension local operators inherited from a larger theory that have been lost (or have become insignificant) in the low-energy limit. They represent *corrections* to the low-energy effective four-quark interaction we are left with at scales $k \ll M_W$, which we view as a legitimate, comprehensive theory in its own right, in this regime.

The notion of an operator product expansion fully generalizes this kind of identification in a completely rigorous and systematic way.

Note the OPE series is equivalent to the original theory when considered to all orders. The truncation of the series yields a systematic approximation scheme for describing low-energy processes, and only neglects contributions suppressed by higher powers of $1/M_W$.

More generally, for any weak process, the OPE prescription tells us how to write an *effective* weak Hamiltonian as a collection of local operators with (running) coefficients:

$$\mathcal{H}_{\text{eff}} = \frac{G_F}{\sqrt{2}} \sum_i V_{\text{CKM}}^i C_i(\mu) Q_i \tag{3.4}$$

The operators Q_i are treated as new interaction terms, and the (Wilson) coefficients $C_i(\mu)$ are viewed as scale-dependent couplings resulting from the short-distance, calculable physics, evaluated with the tools of perturbation theory and

[1] This example is borrowed from a nice presentation by Buras [15].

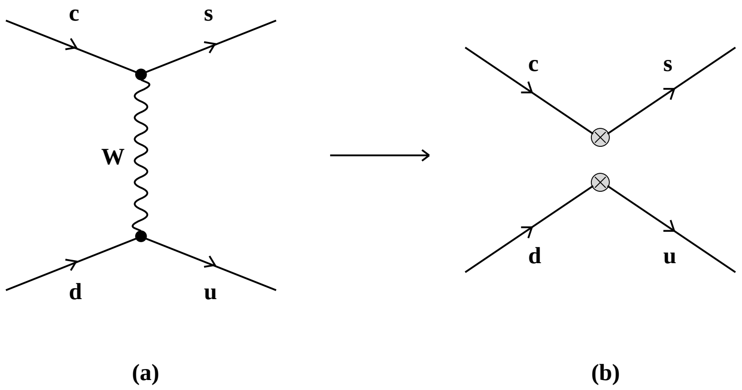


Figure 3.1: Tree-level diagram for the quark-level transition $c \to su\bar{d}$. The diagram in the full theory is shown on the left, and after "integrating out" the W propagator, we are left with the effective interaction shown on the right.

renormalization group methods. Roughly speaking, the scale μ marks the boundary between the high-energy contributions captured in the C_i, and the low-energy contributions pieces left in the Q_i.

Clearly, in order to compute the amplitude for some weak process $i \to f$, the hadronic matrix elements $\langle f|Q_j|i\rangle$ must be evaluated. However, these quantities contain the non-perturbative physics of the strong interaction and a different approach is required. To meet these needs, tools such as lattice calculations, the $1/N$ expansion (where N is the number of colors), QCD sum rules, hadronic sum rules, and chiral perturbation theory have been developed. For the B, heavy quark effective theory (HQET) (next up; see Ch 3.2.2) and the heavy quark expansion (HQE) address these issues. Despite the myriad techniques available, however, none is without limitation: In the end, the dominant theoretical uncertainties on these calculations still come from the evaluation of the matrix elements $\langle Q_j\rangle$.[2]

It is unfortunate that in many cases, the hadronic matrix elements do not yield to our current theoretical tools and cannot be readily evaluated. In a few instances, it is possible to fall back on experimental measurements of several key matrix elements, or to employ suitable ratios and combinations of amplitudes in ways such that the unknown terms cancel. Finally, it is sometimes possible to derive constraints or other bounds on the matrix elements using flavor or other approximate symmetries, provided the symmetry-breaking effects can be reliably estimated.

We conclude with a short remark about renormalization schemes (RS) and scale dependence. It is a a fundamental theorem of renormalization that physical observables such as cross-sections and decay rates do not depend on the choice of

[2]Here we include the additional uncertainty arising from the non-perturbative parameters introduced in the HQE.

RS, when computed to all orders in perturbation theory [3]. However, a truncated perturbation series *will* exhibit RS dependence, and quantities calculated to finite order will thus display some residual sensitivity to the choice of RS. At leading and next-to-leading order (NLO) in perturbation theory, the RS dependence is fully specified by the renormalization scale μ, but at next-to-next-to-leading order (NNLO),[3] this is no longer sufficient and the specification of the scale μ must be augmented by the specification of the scheme as well. It is generally agreed that appropriate choices for μ can lead to reduced RS- and scale-dependence, but there are no clear criteria for determining the "best" choice, even within a given scheme.

3.2.2 Heavy Quark Effective Theory[4]

Since the development of the OPE and renormalization methods for evolving physics between scales, particle physicists have learned to view the world as a series of effective theories, to which perturbative corrections are made in a well-defined and consistent way. Indeed, the success of the Standard Model is attributed to the fact that, in some sense, it is the unique low-energy limit of whatever physics really prevails at high energies.[5] The program set forth by Wilson—the operator product expansion just discussed—provides a rigorous framework for making calculations in a theory that seems initially plagued with infinities and unresolvable ambiguities. The trouble is that the coupling constant of QCD does not lend itself to a perturbative expansion at large distance scales: *confinement* leads to a coupling that increases in the low-energy limit, preventing bare quarks from existing individually and in isolation. Hence, even in the user-friendly Standard Model, there are calculations that are rendered impossible from first principles, since they involve fundamentally non-perturbative effects that cannot be handled with the usual assortment of theoretical tools.

A Physical Picture

Enter the heavy quark limit, an insight into QCD that arises from another separation of scales, this time springing from the (happen-chance) separation between the masses of the so-called heavy quarks (b, c) and the scale $\Lambda_{\rm QCD}$ where the strong

[3]Leading-order calculations typically do not offer enough precision to be useful for careful tests of the Standard Model, so the more tedious and difficult NLO and NNLO calculations are almost always required for useful theoretical prediction. Yup, life as a theorist is hard.

[4]Some of this material was prepared previously [33], but why skip the chance to have it preserved forever on microfilm by including it here?

[5]Such physics, of course, may not be described by something as straightforward as a quantum field theory. But it will give rise to a field theory in the limit of the low energies within the range of current experiment.

coupling α_s becomes of order unity.[6] First articulated in the early 1990's, this idea essentially exploits the observation that in a hadron H_Q containing a heavy quark Q, the light degrees of freedom (ldf) are insensitive to the actual values of the mass m_Q and spin orientation s_Q of the heavy quark. A simple physical argument suffices to motivate this conclusion. Picture a heavy hadron as composed of a single heavy quark surrounded by a complicated and strongly-interacting cloud of light quarks, antiquarks, and gluons. The Compton wavelength of the heavy quark scales as the inverse of its mass, $\lambda_Q \sim 1/m_Q$. The ldf, in contrast, have momenta characterized by the scale $\Lambda_{\rm QCD}$, corresponding to wavelengths of size $\lambda_{ldf} \sim 1/\Lambda_{\rm QCD}$. Since $\Lambda_{\rm QCD} \ll m_Q$, implying $\lambda_{ldf} \gg \lambda_Q$, the ldf cannot resolve any detailed features of the heavy quark beyond its conserved quantum numbers.

A similar conclusion follows from considering the situation in momentum space. Clearly, the structure of the hadron H_Q is determined by non-perturbative strong interactions. Recall that the asymptotic freedom of QCD implies that when quarks and gluons exchange momenta p much larger than the scale $\Lambda_{\rm QCD}$, the process can be treated perturbatively in the strong coupling constant $\alpha_s(p)$. The typical momenta exchanged by the ldf with each other and with the heavy quark are of order $\Lambda_{\rm QCD}$, for which a perturbative expansion is of no use. But in an exchange of momentum $p \ll m_Q$ between the ldf and the heavy quark, the heavy quark will essentially not recoil: it remains at rest in the frame of the hadron. In this limit, Q acts as a *static* source of electric and chromoelectric field. This latter field is what binds the constituents of H_Q together, but the field is actually independent of m_Q in this limit. Again, though the field is responsible for the clearly non-perturbative coupling that glues Q and the light degrees of freedom together, it doesn't depend in detail on the *mass* that produces it. The generalized result is that the properties of the light degrees of freedom depend only on the presence of the static gauge field, independent of the flavor and mass of the heavy quark carrying the gauge charge.

Further consideration along these lines illustrates that the light degrees of freedom are also blind to the spin orientation of the heavy quark. (Essentially, only the color electric field is relevant, since the color magnetic interaction vanishes in the $m_Q \to \infty$ limit.) These observations immediately lead to relations between the masses of various of the heavy mesons B, D, B^*, and D^*. These relations can be systematically improved by including corrections to the heavy-quark limit, as we discuss below. Excellent agreement with the data results, suggesting that this "heavy quark symmetry" is, in fact, a reasonably "good" symmetry, broken only to a modest extent.

This separation in scales is of incredible utility when making theoretical calculations. And, of course, it is satisfying to be able to make (some) predictions based

[6] For completeness, we note that current estimates place $m_b \sim 4.7$ GeV and $\Lambda_{\rm QCD} \sim$ 200-500 MeV, although both of these parameters are subject to some ambiguity determined by the scheme within which they are applied.

on properties of QCD itself (in a certain limit). Much of the relevant physics—the short-distance effects—can be computed perturbatively using renormalization group methods; the remaining long-distance effects may simplify due to the realization of approximate symmetries which connect a large number of long-range properties to only a small number of hadronic matrix elements. In this manner, a calculation can actually make statements beyond the range of applicability of ordinary perturbation theory.

Heavy Quark Expansion

The heavy quark expansion provides a systematic treatment of this limit, organizing it into an expansion in powers of Λ/m_Q,[7] the ratio of characteristic scales across which the physics of the problem separates. The intuitive appeal of the picture painted above is translated into a powerful formalism for handling non-perturbative QCD in an organized fashion. In particular, this allows symmetry-breaking corrections to be computed in a controlled and model-independent fashion, and provides for the estimation of theoretical uncertainties. This latter feature of HQET represents a substantial improvement over the previous generation of quark and potential models that have been used to calculate the effects of the strong interaction. Despite the predictive power of such models, they are crippled by an inability to assess the validity of the assumptions used in the model-building phase, and it is typically not even possible to confidently assign errors within the context of a particular model. At the sacrifice of some predictive power, heavy quark symmetry grounds a computation firmly in QCD and yields useful phenomenological results.

Our presentation here is essentially a reproduction of treatments given by Neubert ([23, 26, 28]) and Falk [25]. The idea is that by explicitly separating the scales in the heavy quark field, one can derive the Lagrangian for HQET directly from QCD in a model-insensitive way. Further, one can include correction terms to the $m_Q \to \infty$ limit simply by keeping terms of order Λ/m_Q and higher during the initial calculation.

The typical starting point for this argument is to note that the heavy quark essentially carries all of the momentum of the hadron, and so it is useful to decompose the heavy quark momentum as follows

$$p_Q^\mu = m_Q v^\mu + k^\mu, \tag{3.5}$$

where v^μ is the four-velocity of the hadron, and $k^\mu \sim \Lambda_{\mathrm{QCD}}$ is the small, fluctuating "residual momentum" contributed by the light degrees of freedom. The

[7] Here and elsewhere, it is to be assumed that, if otherwise unadorned or unlabeled, the symbol Λ refers to the scale Λ_{QCD}.

heavy quark limit suggests Q is never far from its mass shell, $p_Q^2 = M_Q^2$; explicitly requiring this leads to the constraint

$$m_Q^2 = (m_q v_\mu + k_\mu)^2 = m_Q^2 + 2m_q v \cdot k + k^2, \tag{3.6}$$

which forces us to conclude that

$$v \cdot k = 0, \tag{3.7}$$

since the term k^2 can be neglected in the heavy quark limit. Eqn 3.7 is sufficient to identify the operators

$$P_\pm = \frac{1 \pm \not{v}}{2} \tag{3.8}$$

as projection operators: They project onto the positive (P_+) and negative (P_-) frequency parts of the Dirac field of the heavy quark, $Q(x)$. From the Dirac representation in the rest frame of the heavy quark, it can be seen that P_+ and P_- project, respectively, onto the upper two and lower two components of the heavy quark spinor. In the limit $m_Q \to \infty$, in which Q is essentially on-shell, only the "large" upper components of the field Q propagate, and the "small" lower components become negligible; mixing of the upper components with the lower is suppressed by a factor of $1/2m_Q$. We can see this explicitly in the action of the projection operators on the field:

$$P_+ Q(x) = Q(x) + \mathcal{O}(1/m_Q), \qquad P_- Q(x) = 0 + \mathcal{O}(1/m_Q). \tag{3.9}$$

Thus one is led naturally to introduce "large" and "small" component fields h_v and H_v given by

$$h_v(x) = e^{im_Q v\cdot x} P_+ Q(x), \qquad H_v(x) = e^{im_Q v\cdot x} P_- Q(x) \tag{3.10}$$

so that

$$Q(x) = e^{-im_Q v\cdot x}\left[h_v(x) + H_v(x)\right]. \tag{3.11}$$

Note that the effective field $h_v(x)$ is independent of the heavy quark mass m_Q; it carries only a velocity label v and is now a two-component object. The modifications to the full field $Q(x)$ basically project out the positive frequency part and ensure that the states annihilated by $h_v(x)$ have no dependence on m_Q. Similarly, the field $H_v(x)$ carries the effects of order $1/m_Q$ and vanishes in the heavy quark limit. In this same limit, we can write

$$Q(x) \sim e^{-im_Q v\cdot x} h_v(x) + \cdots \tag{3.12}$$

where we mean that terms of $\mathcal{O}(1/m_Q)$ are to be dropped.

Recall now that the QCD Lagrangian for a heavy quark of mass m_Q is

$$\mathcal{L}_{QCD} = \overline{Q}(x)(i\not{D} - m_Q)Q(x), \tag{3.13}$$

where $D_\mu = \partial_\mu - igA^a_\mu T^a$ is the usual gauge-covariant derivative for the strong interaction. To explore the form of QCD in the heavy quark limit, we substitute the expression (3.11) for $Q(x)$ into the Lagrangian (3.13):

$$\mathcal{L} = \overline{h}_v\, iv\cdot D\, h_v - \overline{H}_v(iv\cdot D + 2m_Q)H_v + \overline{h}_v\, i\, \not{D}_\perp H_v + \overline{H}_v i\, \not{D}_\perp h_v. \tag{3.14}$$

Here we've used the popular notation $D^\mu_\perp = D^\mu - v^\mu v\cdot D$, clearly orthogonal to the heavy quark velocity. It is clear now that h_v describes massless degrees of freedom, while H_v corresponds to fluctuations with twice the heavy quark mass m_Q. By expanding according to the leading-order prescription for $Q(x)$ in Eqn 3.12, we finally obtain the Lagrangian for HQET:

$$\begin{aligned} \mathcal{L}^0_{HQET} &= \overline{h}_v\, iv\cdot D\, h & (3.15)\\ &= \overline{h}_v\left(iv^\mu\partial_\mu + gT^a v^\mu A^a_\mu\right) h_v\, . & (3.16) \end{aligned}$$

The Feynman rules for this theory are simple. The propagator for the heavy quark is simply

$$\frac{i}{v\cdot k + i\epsilon}, \tag{3.17}$$

and there is a simple quark-gluon vertex with factor

$$igT^a v^\mu A^a_\mu. \tag{3.18}$$

Of course, both propagator and vertex factor are independent of the value of the heavy quark mass, reflecting the original symmetry that we sought to expose in this limit. They also have no Dirac structure, reflecting the additional spin symmetry that arises in this limit.

It is straightforward to include power corrections to the Lagrangian $\mathcal{L}_{HQET}$. Starting again from the exact QCD Lagrangian in Eqn 3.14, we apply the classical equation of motion

$$(iv\cdot D + 2m_Q)H_v = i\, \not{D}_\perp h_v \tag{3.19}$$

(which can be obtained by taking the variation of the Lagrangian with respect to the field $\overline{H}_v$), and find

$$\mathcal{L} = \overline{h}_v iv\cdot Dh_v + \overline{h}_v i\, \not{D}_\perp \frac{1}{iv\cdot D + 2m_Q} i\, \not{D}_\perp h_v. \tag{3.20}$$

Finally, expanding the non-local operator on the right in powers of $1/m_Q$, we arrive at the final form

$$\mathcal{L}_{HQET} = \overline{h}_v iv\cdot Dh_v + \frac{1}{2m_Q}\left[\overline{h}_v(iD_\perp)^2 h_v + \frac{g}{2}\overline{h}_v\sigma^{\alpha\beta}G_{\alpha\beta}h_v\right] + ... \tag{3.21}$$

where the first-order $1/m_Q$ corrections to the heavy quark limit are now apparent. The infinite series of local operators that has been neglected is formally $\mathcal{O}(1/m_Q^2)$; the combination as a whole forms the basis for the OPE of HQET.

These leading corrections have a simple interpretation, which becomes clearer in the rest frame where $v^\mu = (1, 0, 0, 0)$. The first term, spin-independent, becomes

$$\frac{1}{2m_Q}\mathcal{O}_{kin} \equiv \frac{1}{2m_Q}\overline{h}_v(iD_\perp)^2 h_v \to -\frac{1}{2m_Q}\overline{h}_v(i\vec{D})^2 h_v \tag{3.22}$$

which is readily identified as the gauge-covariant extension of the kinetic energy arising from the off-shell residual motion of the heavy quark. (In the rest frame, $(iD_\perp)^2$ is the operator for $-\vec{k}^2$.) The second, spin-dependent part becomes

$$\frac{1}{2m_Q}\mathcal{O}_{mag} \equiv \frac{1}{2m_Q}\frac{g}{2}\overline{h}_v\sigma^{\alpha\beta}G_{\alpha\beta}h_v \to \frac{1}{4m_Q}\overline{h}_v\sigma^{ij}T^a h_v \times gG^a_{ij} \equiv g\vec{\mu}^a_Q \cdot \vec{B}^a \tag{3.23}$$

Here we've explicitly identified the chromomagnetic moment of the heavy quark, and $B^i = -\frac{1}{2}\epsilon^{ijk}G^{jk}$ are the components of the chromomagnetic gluon field. This term represents the coupling of the spin of the heavy quark to the chromomagnetic field, akin to the familiar hyperfine interaction in atomic physics. The coupling breaks both the heavy quark flavor (mass) and spin symmetry, and is responsible, for instance, for the mass-splittings between the D–D^* and B–B^* states. In heavy quark effective theory, both of these correction terms are treated as part of an *interaction* Lagrangian, even though $\mathcal{O}_{kin}$ has a piece which is a pure bilinear in the quark field.

Our discussion so far has dealt with the construction of the HQET Lagrangian, but it is not yet complete. In order to use the theory, short-distance physics needs also to be properly included. The typical program for doing so is to match the full Standard Model onto the four-fermion theory at the scale M_W, and then run it down to the scale relevant for the physics of interest, *e.g.* m_b. It is then that rates, differential distributions, etc. can be calculated in HQET in an expansion in powers of $1/m_Q$. The result of this "matching" is a renormalization of the coefficients of the operators in the heavy quark expansion. We jump directly to the end result: In the context of the operator product expansion, the Lagrangian for HQET is often recast as [28]

$$\mathcal{L}_{HQET} = \overline{h}_v iv\cdot Dh_v + \frac{C_{kin}}{2m_Q}\overline{h}_v(iD_\perp)^2 h_v + \frac{gC_{mag}}{4m_Q}\overline{h}_v\sigma^{\alpha\beta}G_{\alpha\beta}h_v + \mathcal{O}(1/m_Q^2), \tag{3.24}$$

where the Wilson coefficients C_{kin} and C_{mag} result from short-distance effects and generally depend on the scale at which the operators are renormalized. The remaining HQET operators capture the non-perturbative physics, and present the chief computational challenge, if one can guarantee that the perturbative effects are under control.

The matrix elements of the kinetic energy and hyperfine interaction terms appear in many applications of HQET, and it is conventional to define for the B system:

$$-\lambda_1 \;=\; \frac{1}{2M_B}\langle B|\mathcal{O}_{kin}|B\rangle \tag{3.25}$$

$$3\lambda_2 \;=\; \frac{1}{2M_B}\langle B|\mathcal{O}_{mag}|B\rangle \tag{3.26}$$

These parameters characterize the order $1/m_b^2$ corrections in HQET; at $\mathcal{O}(1/m_b^3)$ seven additional operators arise in the OPE, but they are not as well constrained theoretically. Some workers in the field use the parameters μ_π^2 and μ_G^2, which are related to the parameters above in a simple way, but receive short-distance corrections in a different fashion [31]. (Note that for the case where the meson H_Q is actually a vector particle, the definition for λ_1 is unchanged, but that for λ_2 acquires a factor of -3 on the RHS [39].)

As mentioned above, the parameter λ_2 is related to the mass-splitting between vector and pseudoscalar meson states through

$$M_{B^*}^2 - M_B^2 = 4\lambda_2 + \mathcal{O}(1/m_b) \tag{3.27}$$

which, by using the experimentally determined values of the meson masses, evaluates numerically to

$$\lambda_2 = 0.12 \;\; \mathrm{GeV}^2. \tag{3.28}$$

On the other hand, the parameter λ_1 cannot be determined from hadron spectroscopy, although it has been shown to be related to the difference of the pole masses for charm and bottom quarks through [28]

$$m_b - m_c = (\overline{M}_B - \overline{M}_D) + \lambda_1(\frac{1}{2\overline{M}_B} - \frac{1}{2\overline{M}_D}) + \cdots \tag{3.29}$$

Here $\overline{M}_B = \frac{1}{4}(M_B + 3M_{B^*})$ (similarly for $\overline{M}_D$) denotes the spin-averaged meson mass, defined so that there is no contribution from the chromomagnetic interaction. Theoretical estimates [26] for λ_1 are in the range $-(0.3 \pm 0.2)$ GeV^2. The current opinion is that this parameter is not physically meaningful and suffers from an intrinsic ambiguity of order Λ_{QCD}^2. For further discussion of the status of these parameters, see Ref [28] and the references therein.

A third HQET parameter arises in the analysis of the heavy quark systems at this order,

$$\overline{\Lambda} = \lim_{m_b\to\infty}(M_B - m_b), \tag{3.30}$$

which measures the "mass" carried by the light degrees of freedom. It is connected to a residual mass term that can appear in the HQET Lagrangian; this presence leads to an ambiguity of order Λ_{QCD} in the heavy quark mass. With a particular choice of the expansion parameter m_Q, this residual mass term can be made to vanish, and then m_Q coincides with the pole mass used in perturbation theory. Other choices for the expansion parameter will of course adjust the value of $\overline{\Lambda}$, but this arbitrariness disappears in the calculation of actual physical quantities. This parameter is often used to re-express predictions in terms of meson masses through relations of the form [24]

$$M_Q = m_Q + \overline{\Lambda} - \frac{\lambda_1 + 3\lambda_2}{2m_Q} + \cdots. \tag{3.31}$$

This introduction to the formalism of heavy-quark symmetry is readily extended to make model-independent calculations on many phenomenological fronts. Correctly handling the perturbative effects and understanding the dependence on the HQET parameters are generally difficult issues, but the essential non-perturbative physics has been isolated and, to some extent, rendered tractable. We will not go into the details of these issues here; for more information, see the references mentioned above.

3.2.3 Theory of Inclusive Decays

The calculation of inclusive quantities such as (differential) decay rates is one area that has benefited from the equipment of HQET. Recall that in these calculations, we sum over all possible hadronic final states allowed by the kinematics or some set of global quantum numbers. Although the key theme is again the separation of long- and short-distance physics, there is a new ingredient: the notion of so-called "quark-hadron" duality which also relies on the heavy quark limit. From the theoretical point of view, such decays have two advantages: First, bound-state effects related to the initial state—such as the "Fermi motion" of the heavy quark within the confines of the hadron—are accounted for in a systematic way with the heavy quark expansion. Secondly, the fact that the final state is a sum over many hadronic channels eliminates bound-state effects in the final state hadrons. Essentially, the final state quarks are guaranteed with unit probability to hadronize, and so the particular details of how this comes about are irrelevant. That this process does not "interfere" with the heavy quark decay can be seen by noting that the timescale Δt for the b quark decay is of size $1/m_b$, while the timescale associated with hadronization is again set by QCD: $\Delta t \sim 1/\Lambda_{\rm QCD}$, *i.e.* the hadronization occurs at a much later time than the initial b decay, so the physics decouples. It is typical of the success of HQET that—at leading order—its predictions match the more naïve parton-model calculations carried out in the early days, and further indicative of its power that it also includes a prescription for computing sub-leading corrections to these rates—which subsequently compare well with experiment.

It is instructive to examine the outline for an inclusive calculation to see just how the machinery of HQET is applied, and to gain some understanding of how the assumption of duality arises as a necessary part of the computation. To keep the analysis general, we consider the rate $\Gamma(B \to X)$, where X represents the set of final states consistent with whatever inclusive quantities have been specified. The discussion here once again closely follows the one presented by Falk [25] and Neubert [28]. The inclusive decay is a sum over all possible final states, which is really a sum over exclusive modes, followed by an integral over the phase space for each mode. Hence we begin by writing

$$\Gamma(B \to X) = \sum_X \int \mathrm{d}[\mathrm{P.S.}] \; |\langle X|\mathcal{O}_{bX}|B\rangle|^2 \tag{3.32}$$

where the operator $\mathcal{O}_{bX}$ is a convenient shorthand for whatever effective operators are relevant for the process under consideration. An analog of the optical theorem for QCD allows us to express the transition rate instead as the imaginary part of a forward scattering amplitude:

$$\Gamma = -2\,\mathrm{Im}\, i \int \mathrm{d}x\, e^{-ik\cdot x} \langle B|T\{\mathcal{O}^\dagger_{bX}(x), \mathcal{O}_{bX}(0)\}|B\rangle \equiv 2\,\mathrm{Im}\,\mathbf{T} \tag{3.33}$$

The next step is to expand the time-ordered product $T\{\mathcal{O}^\dagger_{bX}(x), \mathcal{O}_{bX}(0)\}$ in an operator product expansion, in which the transition operator $\mathbf{T}$ is represented as a series of local operators containing the heavy-quark fields. The applicability of this expansion, and its computation in perturbation theory, rests on the heavy quark limit $m_b \gg \Lambda_{\mathrm{QCD}}$.

For concreteness, we now restrict the discussion to the case where the final state X is semileptonic. In this case, the transition operator $\mathcal{O}_{bX}$ factors into separate leptonic and hadronic currents. The core of the calculation is then the evaluation of a hadronic tensor of the form

$$T^{\mu\nu} = -i \int \mathrm{d}x\, e^{-iq\cdot x} \langle B|T\{J^{\mu\dagger}_{bX}(x), J^{\nu}_{bX}(0)\}|B\rangle \tag{3.34}$$

where we've introduced the total momentum q^μ transferred to the lepton system. It is standard to express this general tensor as a sum of five individual terms, each with definite Lorentz structure. When considering semileptonic B decay to massless u-quarks, only three of these are non-vanishing. However, such details take us beyond the scope of our current discussion.

A standard property of quantum field theory is that the imaginary part of a hadronic matrix element such as $T^{\mu\nu}$ is non-vanishing whenever there is a real intermediate state, that is, when the intermediate particles are all on mass-shell. Careful consideration of the possible avenues for the production of such physical intermediate states leads to the identification of poles and cuts in the complex $v \cdot q$ plane which neighbor the usual path of integration. Much can be said about the analyticity of the transition matrix, but the summary is that the location of the singularities along the real axis depends intimately on the details of QCD at long distances, since these physical states represent non-trivial bound states.[8] Hence the evaluation of the integral depends in an essential way on physics that is incalculable. The perturbative treatment we have been pursuing suddenly seems to have broken down.

The solution is to deform the contour of integration away from the difficulties on the real axis, into complex $v \cdot q$ space. As Falk [25] argues, the momentum scale in the problem is set by m_b, so that the new contour is now predominantly a distance $m_b \gg \Lambda_{\mathrm{QCD}}$ away from the difficult resonances, and a perturbative treatment

[8] Remember, perturbative QCD doesn't even predict these!

is again applicable, since on the new contour, $|q^2|$ is once again large. And in the small region where the contour is still close to the cut, we know the allowed phase space vanishes, so the calculation is once again on familiar perturbative ground. This claim, that the average value of a hadronic quantity can be calculated perturbatively when at each point the quantity depends on the details of non-perturbative physics, is the notion of (global) *parton-hadron* duality. This duality ranks as more than just a convenient theoretical assumption, since it is known to hold in the $m_b \to \infty$ limit, but it does not have the status of a standard approximation because there is no framework for calculating the leading corrections to this limit. For instance, it is known to hold only qualitatively [29] in $b \to c\bar{c}s$ processes. In any case, its use is key to the completion of the inclusive calculation. We will have more to say on this topic later (see Sec 3.4.1).

What remains is the construction of an operator product expansion for the transition operator to handle the short-distance (and therefore perturbative) physics properly. One can list the operators that can appear in this expansion quite generally. The lowest dimension operator that can contribute is $\bar{b}b$, with dimension $d = 3$. Likewise, there is only one possible gauge-invariant dimension 4 operator, $\bar{b}i\not{D}b$, but this one can be replaced by $m_b\bar{b}b$ with the application of the equations of motion. The first operator that is different from the simple $\bar{b}b$ has dimension 5 and includes the gluon field; it is the familiar $\bar{b}g\sigma_{\mu\nu}G^{\mu\nu}b$. (From dimensional considerations, one can see that the matrix elements of a dimension d operator are suppressed by inverse powers of the heavy quark mass.) Once these operators have been derived, the tools of HQET are used to replace the operators b and $\bar{b}$ with the HQET fields h_v and $\bar{h}_v$ according to the usual prescription. The Wilson coefficients associated with each operator are then computed to some order in perturbation theory. The result is that any inclusive decay rate can be written in the form [28]:

$$\Gamma(B \to X_f) = \frac{G_f^2 m_b^5}{192\pi^3}\left\{c_3^f\left(1 + \frac{\lambda_1 + 3\lambda_2}{2m_b^2}\right) + c_5^f\frac{\lambda_2}{m_b^2} + \mathcal{O}(1/m_b^3)\right\}. \tag{3.35}$$

The prefactor is the usual result of the loop and phase space integrations, and the c_n^f are coefficient functions depending on the quantum numbers of the final state X_f; naturally, they also contain the relevant CKM elements for the process. The leading contribution is just the result obtained from a naïve parton-level calculation. This is an instance of the KLN theorem [29] which asserts that the cancelation of the infrared singularities present in the exclusive rates gives a total inclusive width that is insensitive to the details of hadronization. It is worth noting that the kinetic energy contribution $\lambda_1/2m_b^2$ is nothing but the field-theory generalization of the expected Lorentz factor $(1 - \vec{v}_b^2)^{1/2} \simeq 1 - \vec{k}^2/2m_b^2$; this is the usual time-dilation factor that increases the lifetime $\tau = 1/\Gamma$ of a moving particle. The absence of first-order corrections of order $1/m_b$ in the above expression is also noteworthy, and is a natural consequence of the equations of motion in HQET. In this case, the terms disappear due to the choice of the residual mass term mentioned earlier [37]. In conclusion, we emphasize that the utility of HQET lies in the model-independent prediction of quantitative *corrections* to these basic results.

We have only grazed the surface of the issues (both technical and physical) associated with the calculation of inclusive quantities. Observe that within the setting of HQET, inclusive rates depend only (up to order $1/m_b^2$) on the three HQET parameters λ_1, λ_2, and $\overline{\Lambda}$, and calculable coefficient functions. It should be clear that the procedure is well-established, grounded in a clear physical picture, and depends indispensably on the heavy quark limit of QCD.

3.3 Phenomenological Road to $|V_{ub}|$[9]

Our short review of some of the theoretical tools is now complete, and we turn to the advertised program: extraction of $|V_{ub}|$ from semileptonic B decays.

Falk [25] presents the results of applying this technology to the calculation of the total charmless semileptonic rate. Taking $m_u = 0$, he reports

$$\begin{aligned}\Gamma(B \to X_u l\nu) = \frac{G_F^2 |V_{ub}|^2}{192\pi^3} m_b^5 \Big[1 \quad &+ \quad \left(\frac{25}{6} - \frac{2\pi^2}{2}\right) \frac{\alpha_s(m_b)}{\pi} \\ &- \quad (2.98\beta_0 + C_u) \left(\frac{\alpha_s(m_b)}{\pi}\right)^2 + \ldots \\ &+ \quad \frac{\lambda_1 - 9\lambda_2}{2m_b^2} + \ldots \Big].\end{aligned} \tag{3.36}$$

The occurrence of the heavy quark mass m_b in the prefactor is troublesome since, as an unphysical parameter, it is subject to considerable theoretical ambiguity; in fact, this uncertainty dominates subsequent attempts to extract $|V_{ub}|$. The expression for the decay rate can be recast in several ways to reduce this sensitivity by exchanging some portion of it for additional input from experiment. For instance, one can substitute reference to the charm mass, employ an alternative kinematic mass for the b quark [14], or use a so-called “Upsilon expansion” to eliminate the quark mass in favor of the mass $m_{\Upsilon(1S)}$ of the $\Upsilon(1S)$ resonance [32].

The formula for the total decay rate can be inverted to provide for the extraction of $|V_{ub}|$ from a direct experimental measurement of the charmless semileptonic branching fraction. A recent assessment by the LEP VUB working group gave [12]:

$$|V_{ub}| = 0.00445 \left(\frac{\mathcal{B}(B \to X_u \ell\nu)}{0.002}\right)^{\frac{1}{2}} \left(\frac{1.55\text{ps}}{\tau_B}\right)^{\frac{1}{2}} (1 \pm 0.020 \pm 0.052). \tag{3.37}$$

[9] We have been less than meticulous about notation and ν versus $\overline{\nu}$ conventions in preparing the following material. For a consistent and more accurate discussion of any of these topics, please consult any of the references cited at the start of this chapter.

The quoted errors arise from the OPE (first) and dependence on the mass m_b (second). This approach is theoretically quite straightforward, and results in a determination that has total theory error at about the 5% percent level. (Similar but not identical formulae are quoted in papers by Bigi [31], Ligeti [32], and Uraltsev [63].) As we shall see later, however, there are additional theoretical uncertainties not treated in this evaluation. More importantly, also overlooked is the significant impact that experimental practice can have on this rather naïve agenda.

Experimentally, we cannot neglect the need to separate the charmed semileptonic decays of the B from the charm-less ones. An assortment of kinematic variables have been considered to provide such discrimination, all of which rely on the basic fact that $m_u \ll m_c$. It is at kinematic endpoints that this difference can become significant and useful. The highest lepton energy E_ℓ, the smallest possible hadronic mass M_X, or the largest possible momentum transfer q^2 to the lepton system—for each of these, there is still some allowed phase space for $b \to u$ decays beyond the endpoint for charm. In short, $b \to u$ has more kinematic reach. The choice of discriminator and the location of the experimental cut together determine the fraction of $b \to u\,\ell\,\nu$ decays that are accessible with each method.

Working near the boundaries of phase space has its disadvantages, both experimental and theoretical. Briefly, the imposition of a cut changes the theoretical expansion parameters from α_s for perturbative QCD and Λ/m_b for the non-perturbative physics (*i.e.* HQET) to ones of the form $\alpha_s \log\rho$ and $\Lambda/(m_b\rho)$, where ρ is some measure of the area of phase space to which the analysis has been restricted. It may be the case that one or both of these new effective parameters will be of order unity, necessitating a reorganization of the entire calculation. Essentially, the heavy quark expansion must be generalized to account for the "Fermi motion" of the b quark within the B-meson; this is critical in the endpoint region because it is this residual momentum that accounts for the shift from the free-quark endpoint $m_b/2$ up to the physical endpoint $M_B/2$.

Due to the unfortunate reality of the ever-present $b \to c\,\ell\,\nu$ decays, theorists are forced to produce a calculation appropriate for the hard experimental cuts that must be made to isolate the signal events in a background-free region. Once again, various models have been employed to extrapolate from the rate measured in some kinematic window to the full charmless semileptonic rate, but these are hampered by questionable (or at least unverifiable) assumptions. In recent years, new techniques have been developed to deal with these problems; they rely heavily on the formalism of the heavy quark expansion but also owe their success to equally clever ideas for tackling both the experimental and theoretical analysis.

We will briefly review three such methods here, in rather broad strokes. The intent is not to capture the state-of-the-art, but to illustrate the common and disparate problems and triumphs that complicate the inclusive measurement of $|V_{ub}|$. For more detailed (and current and accurate) discussions of these and other techniques, see the references, *e.g.* Ref [2, 14, 72, 73]. In outline, the three methods are:

- $E_\ell > (M_B^2 - M_D^2)/2M_B$
  An analysis of the lepton energy spectrum in the endpoint region above the kinematic reach of $b \to c\,\ell\,\nu$, with non-perturbative effects described by a shape function that can be constrained from measurements of the photon spectrum in inclusive $b \to s\gamma$ decays [35].

- $M_X < M_D$
  An analysis of the spectrum of the invariant mass M_X^2 of the recoiling hadronic system [44–46]. This offers the experimental advantage that charm background can be excluded while retaining a much larger fraction of $b \to u$ decays. On the other hand, the calculation is extremely sensitive to the precise value of the cut, and the shape function mentioned above is still relevant. Experimentally, this approach demands excellent resolution on the hadronic mass.

- $q^2 > (M_B - M_D)^2$
  An analysis of the spectrum of the lepton invariant mass, where again charm contamination can be excluded with a wise cut on q^2, leaving behind a reasonable fraction of the $b \to u$ signal [51]. In the region of phase space selected with this method, the theoretical calculation seems to be on excellent footing, but the scale of the intermediate physics is m_c instead of m_b, so $1/m_b$ correction terms blow up to $1/m_c$. Success on the experimental side depends on the resolution with which the lepton and neutrino kinematics can be reconstructed.

The issue in each case is addressing the perturbative and non-perturbative contributions to the decay rate in the selected region of phase space, and then estimating consistent errors for various correction terms. The general consensus [1] is that, due to the complex—and in some cases, correlated—issues that trouble these approaches, the best course is to pursue them in parallel, applying improvements as our knowledge of $b \to u\,\ell\,\nu$ grows. A well-constrained value for $|V_{ub}|$ will only be possible if our picture of $b \to u\,\ell\,\nu$ has been firmly established and tested in several independent ways.

3.3.1 Lepton Energy E_ℓ

With a precise prediction of the lepton energy spectrum for charmless semileptonic B decays, experiments could readily extrapolate from the partial rate measured in the endpoint region[10] to the total inclusive rate and then extract $|V_{ub}|$

[10]The "endpoint region" as it is used here and in what follows refers to the region of phase space near the upper kinematic endpoint for the lepton energy E_ℓ in inclusive

according to Eqn 3.37 or similar. Typical values of such a cut accept only about 10% of the total rate. A clean theoretical analysis, however, is limited by the breakdown of the usual heavy quark expansion in the endpoint region, where an additional expansion parameter $\Lambda/(m_b - 2E_\ell)$ becomes important, making it hard to evaluate what fraction of the total rate has been sampled by a given cut. In what follows, we first briefly describe the nature of the problem, and then outline the proposal of Neubert [35], followed by others, to collapse the non-perturbative physics into a shape function that can be used to smear the quark-level decay into the real hadronic process. This shape function turns out to be universal in B-decays (up to higher-order corrections), and arises naturally in the discussion of the radiative decays $b \to s\gamma$, where it can be experimentally determined.

Difficulties in the Endpoint Region

An understanding of the lepton energy spectrum in the endpoint region begins with a calculation of the differential rate $d\Gamma/dE_\ell$. To leading order in the non-perturbative corrections, the standard theoretical result, ignoring all QCD perturbative corrections, is ([35, 38])

$$\frac{1}{2\Gamma_0}\frac{d\Gamma}{dy} = yF(y)\Theta(1-y) - \frac{\lambda_1 + 33\lambda_2}{6m_b^2}\delta(1-y) - \frac{\lambda_1}{6m_b^2}\delta'(1-y), \tag{3.38}$$

where the new variable

$$y = \frac{2E_\ell}{m_b} \tag{3.39}$$

parameterizes the differential rate in terms of the excursion from the free-quark kinematic limit of $y = 1$. The factor Γ_0 is similarly the free-quark decay width

$$\Gamma_0 = \frac{G_F^2|V_{ub}|^2}{192\pi^3}m_b^5. \tag{3.40}$$

The symbol $\delta'(\cdot)$ denotes the derivative of the usual Dirac Delta function $\delta(\cdot)$.[11] The function $F(y)$ is a slowly-varying function of y given by

$$F(y) = (3-2y)y + \frac{5y^2}{3}\frac{\lambda_1}{m_b^2} + (6+5y)y\frac{\lambda_2}{m_b^2}. \tag{3.41}$$

$b \to u\,\ell\,\nu$ decays; this is the region free of charm and has size Λ/m_b. Here Λ is again taken as the typical low-energy scale of QCD.

[11]Recall that the δ-function is rigorously defined in the theory of generalized functions (or distributions) as a linear functional L that acts on the space of more ordinary functions f (usually square-integrable), with an appropriate integral definition of the action $L(f)$. The derivative δ' of the δ-function can be defined in an analogous way, where a formal integration by parts makes it clear that while $\delta(f) = f(0)$, $\delta'(f) = -f'(0)$.

Once again, we note that there are no corrections of order $1/m_b$ due to the natural requirement that there be no residual mass term for the heavy quark field in the effective theory.

The step function is present to enforce the turn-off of the decay distribution above the tree level endpoint in the parton model. The more singular terms (the δ-function and its derivatives) arise from higher-order terms in the $1/m_b$ expansion that have y-derivatives of lower-order terms. All together, the appearance of this infinite series of singularities signals the non-perturbative effects that shift the endpoint from the free-quark endpoint to the the physical one $E_{\ell,\max}^{\rm phys} = M_B/2$, and illustrates the difficulty in analyzing the shape of the theoretical distribution in this region; see Fig 3.2. The $1/m_b$ expansion breaks down in this region where $1 - y \sim \Lambda/m_b$, or equivalently, where $(m_b - 2E_\ell) \sim \Lambda$. The program for determining $|V_{ub}|$ is thus derailed by the fact that the experimental cuts made to suppress charm background force us directly into the regime where the terms in the expansion (3.38) are formally $\mathcal{O}(1)$ and must be resummed to all orders.

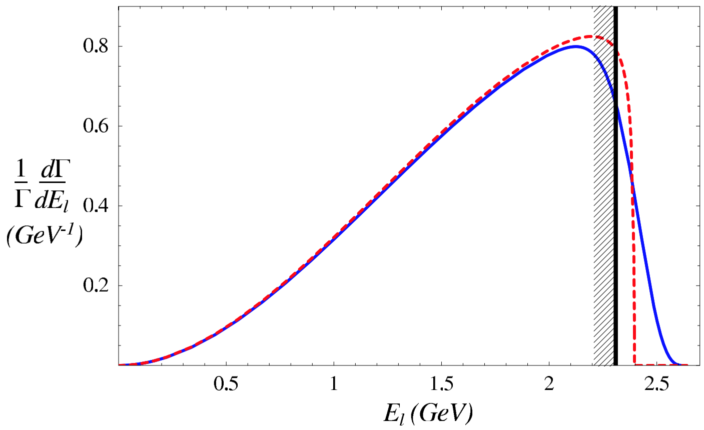


Figure 3.2: An illustration of the importance of the non-perturbative physics governing the binding of the b quark in the B meson. The red curve (dashed) shows the parton-level result for the lepton energy spectrum in $b \to u\,\ell\,\nu$, with endpoint at $m_b/2$. The blue curve (solid) shows the physics spectrum obtained by implementing the non-perturbative physics according to a model for the Fermi motion of the b quark, with endpoint higher at $M_B/2$. As is clear, the details of the hadronization significantly affect the decay rate in the region above the endpoint for semileptonic decays to charm at 2.3 GeV, indicated by the black vertical line.

To further elucidate this problem and understand the formal solution (rather, the rigorous parameterization of our ignorance), we separate the expression for the differential rate into two pieces according to the following prescription (first suggested by Neubert) [35]:

$$\frac{1}{2\Gamma_0}\frac{d\Gamma}{dy} = F(y)\Theta(1-y) + S(y) \tag{3.42}$$

where now the function $S(y)$ captures all of the non-perturbative physics. When constructing the operator product expansion for $S(y)$, an additional expansion parameter $\Lambda/[m_b(1-y)]$ arises in addition to the usual one, Λ/m_b. Over most of the kinematically allowed region, this new parameter is small, until one approaches the endpoint region, where $(1-y)$ is of order Λ/m_b and the expansion becomes ill-behaved. From our discussion above, it should be clear that the support for the function $S(y)$ is restricted to a small interval of size 2Δ around $y = 1$, where Δ is of size Λ/m_b.

Introduction of a Shape Function

In the original experimental analyses of the lepton energy spectrum [124, 128], various models were employed to estimate the fraction of the total charmless semileptonic rate in the endpoint region, which could then be used to translate a partial rate measured in the endpoint region into an absolute total rate. Then a formula like Eqn 3.37 could be applied to determine $|V_{ub}|$. However, the value for $|V_{ub}|$ extracted in this way is heavily dependent on the model that is used, and there is also no clear prescription for evaluating the effective systematic error inherent in the use of one model over another.[12] Although these initial observations indicated quite clearly that $|V_{ub}| \neq 0$, there was little that could be done to either properly assess or improve the associated theoretical uncertainties.

The first theoretical inroads on this problem were made by Neubert [35], who was able to resum the *leading* singular contributions to the function $S(y)$ to all orders in $\Lambda/[m_b(1-y)]$.[13] The result is a smooth—but rapidly varying—shape function or form factor that describes the fall-off of the lepton spectrum in the endpoint region in a genuinely non-perturbative way.

A formal definition of the shape function $S(y)$ can be obtained from an operator product expansion of the differential decay rate, in conjunction with an expansion of the hadronic matrix elements in powers of $1/m_b$, as provided by HQET. The result yields a formal expression for $S(y)$ as a sum of forward matrix elements of

[12] The original CLEO paper [128] considered a range of models that gave a spread of more than 50% in $|V_{ub}|$, but could find no defensible recipe for translating this spread into some sort of confidence interval around the central value.

[13] Mannel and Neubert extend this in Ref [37] to include the case of final-state quarks with finite mass $m_q > 0$, which introduces a parameter $\rho = m_q^2/m_b^2$ in the new shape function.

local operators:

$$S(y) = \sum_{n=1}^{\infty} \frac{1}{n!}\frac{A_n}{m_b^n}\delta^{(n-1)}(1-y) + \text{less singular terms} \tag{3.43}$$

where the coefficients A_n are as defined in Ref [35] and contain the aforementioned hadronic matrix elements. The conclusion of this analysis is that, by comparing to the initial expression for the rate in Eqn 3.38, the first few moments of the shape function $S(y)$ can be deduced, leading to the (non-trivial) identification

$$A_0 = 1, \quad A_1 = 0, \quad \text{and} \quad A_2 = -\lambda_1/3. \tag{3.44}$$

These first few moments already significantly constrain the broad features of the shape function; while higher-order moments have a small impact on integrated quantities such as the total decay rate, they can still substantially alter the resulting spectral shape in the endpoint region. It is important to note that in this region, the theoretical spectrum can be drastically unfaithful to the real one, and yet—frustratingly so—this area is the same place where experimentalists must focus to fight contamination from leptons produced in $b \to c$ decays.

Later treatments of this same approach develop a function $f(k_+)$ that determines (to leading order in $1/m_b$) the probability of finding a b quark with residual light-cone momentum k_+ inside the B-meson, *i.e.*

$$b_B(x)dx = \{f(k_+) + \mathcal{O}(1/m_b)\}dk_+ \tag{3.45}$$

where k_+ and x are related through $k_+ = M_B x - m_b$. In terms of the formalism just developed, this light-cone distribution function is defined by

$$f(k_+) = \langle B(v)|\overline{h}_v \delta(k_+ - iD_+)h_v|B(v)\rangle \tag{3.46}$$

and is naturally related to the shape function discussed previously:

$$\begin{aligned}\Theta(1-y) + S(y) &= \left\langle \Theta\left(1 - y + \frac{in\cdot D}{m_b}\right)\right\rangle + \text{less singular terms} \quad (3.47)\\ &= \int dk_+ \Theta\left(1 - y + \frac{k_+}{m_b}\right)\{f(k_+) + \mathcal{O}(1/m_b)\}\end{aligned}$$

Once again, it is the moments of this new shape function that are related directly to matrix elements in QCD; one can write [36]

$$f(k_+) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} A_n \delta^{(n)}(k_+) = \delta(k_+) - \frac{\lambda_1}{6}\delta''(k_+) - \frac{A_3}{6}\delta'''(k_+) + \cdots \tag{3.48}$$

where

$$A_n = \int dk_+ k_+^n f(k_+) = i^n \tilde{f}^{(n)}(0) = \langle B(v)|\overline{h}_v (iD_+)^n h_v|B(v)\rangle, \tag{3.49}$$

and $\tilde{f}^{(n)}(0)$ stands for the evaluation of the n-th derivative of the Fourier transform $\tilde{f}(t)$ at the point $t = 0$. Ref [37] presents an analysis of the first few of these moments, and concludes that the light-cone structure function is centered around $k_+ = 0$ with a width of order 200-300 MeV determined by the average kinetic energy of the b quark within the B-meson. It also follows that in the heavy quark limit, the range of k_+ is given by $-\infty < k_+ \leq \overline{\Lambda}$, and that $f(k_+)$ should be exponentially small for $k_+ \ll \overline{\Lambda}$.

The decay rate can now be expressed more clearly as a convolution of the simpler parton-level result with this function $f(k_+)$, with the requirement that in the partonic result, the b quark mass m_b is replaced by the mass $m_b^* = m_b + k_+$ [37].

$$\frac{d\Gamma}{dE_\ell} = \int dk_+ f(k_+) \frac{d\Gamma_{\text{parton}}}{dE_\ell}\bigg|_{m_b \to m_b^*} \tag{3.50}$$

We thus discover that the leading bound-state corrections amount to an averaging of the parton-model rate for the decay of a quark with mass $m_b^*(k_+)$ over the distribution $f(k_+)$. The free-quark result is recovered in the obvious limit $f(k_+) \to \delta(k_+)$.

The analysis of the endpoint region is thus reduced to the determination of a smearing function that captures the non-perturbative physics of the b quark's Fermi motion within the B meson. Further, it is faithful to the underlying theory of QCD in a clear, definite way. However, the theoretical considerations do not deliver the smearing function in explicit form; rather, it is loosely determined by the values of its first few moments, which are related to quantities like m_b (or $\overline{\Lambda}$) and λ_1.

Further work with the shape function resorts to the building of toy models that mock up the effects of the Fermi smearing. Functional forms are chosen to conform to the general limiting behavior described above, and free parameters are tuned to comply with the few constraints on the moments (*e.g.* as listed in Eqn 3.44). Three models as found in the recent literature are compared in Fig 3.3 [99, 100].

It is worth emphasizing the physical picture that has resulted from this analysis. The shape function $S(y)$, or equivalently the light-cone distribution function $f(k_+)$, was introduced to capture the non-perturbative subtleties in the lepton-energy spectrum near its endpoint; effectively, the infinite number of (leading) singularities in the perturbative expansion are summed into a single, unknown function. A surprisingly intuitive interpretation results: the the parton-level result is averaged or smeared out by a weighting function that encodes the non-perturbative bound-state effects present within the environment of the B-meson. Further, as we might expect, this smearing function contributes only over a range $-\overline{\Lambda} < k_+ < \overline{\Lambda}$, and pushes the parton-level endpoint (set by m_b) up to the physical endpoint (set by M_B). As Falk [38] notes, it is easy to see how this change comes about. If the energy of the b quark is allowed to fluctuate from its on-shell value, then it will occasionally have an energy larger than its free value m_b. This fluctuation can

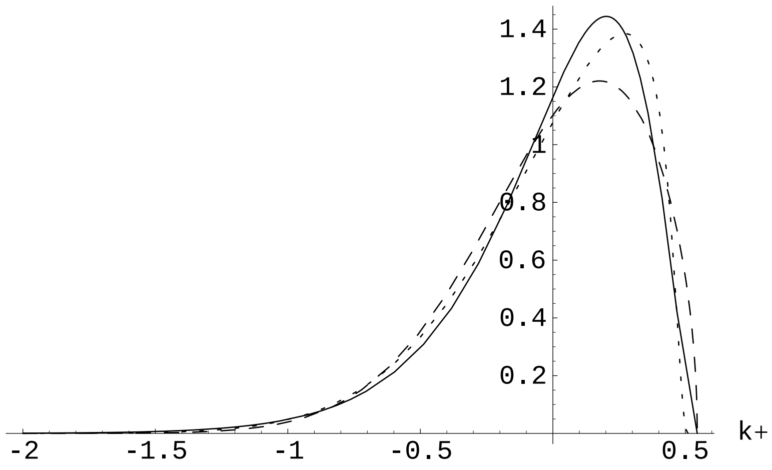


Figure 3.3: Comparison of three models for the non-perturbative shape function $f(k_+)$. Parameters are chosen to meet the requirements $\overline{\Lambda} = 0.545$ GeV and $\lambda_1 = -0.33$ GeV^2. The solid line is the exponential form; dashed, Gaussian; dotted, Roman [100].

be considered a result of a situation in which the quark has temporarily absorbed some energy (of order Λ) from the light degrees of freedom in the B-meson. If the quark decays at this moment, the resulting lepton is capable of having an energy greater than the limit $E_\ell = m_b/2$ set in the parton model.

The shape function describes the Fermi motion inside the B meson, and as such, is a universal property of B decays, independent of the particular decay channel. Consequently, it can in principle be extracted from a different heavy $\rightarrow$ light process and then employed in inclusive $B \rightarrow X_u\,\ell\,\nu$. Such additional input from experiment further reduces the dependence of this method on shape-function modeling, as we shall see next.

Connection to $b \rightarrow s\gamma$

Much of the work outlined above is not specific to the case of inclusive $b \rightarrow u\,\ell\,\nu$ decays, suggesting that there may be some universal features which will help resolve the unknown details of the non-perturbative shape function $f(k_+)$. An analysis addressing the possibility of using the photon spectrum from radiative $b \rightarrow s\gamma$ decays was first presented by Neubert [36] and subsequently further explored and expanded by Leibovich *et al.* and others [40–43]. (These radiative decays are interesting in their own right because they are induced by penguin diagrams with virtual top or charm quark exchange. Interestingly, the experimental analysis also requires a hard experimental cut, this time on the photon energy E_γ to limit the systematic error from the $B\overline{B}$ background subtraction.)

In a manner similar to the one applied in the semileptonic case, the leading non-perturbative corrections to the photon spectrum can be summed to all orders in $1/m_b$. The resulting spectrum is determined by an analogous fundamental structure function describing the distribution of the light-cone momentum of the b quark within the hadron.[14] A connection between the radiative and semileptonic decays is readily made at this point.

We omit most of the steps taken in arriving at the $b \to s\gamma$ spectrum; they are essentially an application of the inclusive formalism described in Section 3.2.3. A spectral function $s(y,\rho)$ is introduced to describe the inclusive spectrum according to

$$\frac{d\Gamma}{dy} = \frac{G_F^2 \alpha m_b^5}{32\pi^4} |V_{tb}V_{ts}^*|^2 |c_7(m_b)|^2 (1+\rho)(1-\rho)^3 \eta_b s(y,\rho) \tag{3.51}$$

where here $y = 2E_\gamma/m_b$, $\rho = m_s^2/m_b^2$, and the parameter η_b captures the leading-order non-perturbative effects. Once again the leading singular terms in $s(y,\rho)$ are summed to all orders to arrive at [36]

$$\begin{aligned} s(y) &= \left\langle \delta\left(1 - y + \frac{iD_+}{m_b}\right)\right\rangle + \text{less singular terms} \qquad (3.52)\\ &= \int dk_+ \delta(1 - y + \frac{k_+}{m_b})\{f(k_+) + \mathcal{O}(1/m_b)\}, \end{aligned}$$

which is reminiscent of the corresponding expression arrived at in the analysis of $b \to u\,\ell\,\nu$, Eqn 3.47. This formal resemblance is a demonstration of the common role the shape function plays (to leading order) in these two different decay processes.

The experimental approach to date has pursued the use of information from the observed $b \to s\gamma$ photon spectrum to further constrain the unknown attributes of the shape function [128]: a candidate form for $f(k_+)$ is chosen, and its parameters are determined by a fit to the experimental data.

Alternatively, the parallels between the two analyses can be exploited directly at the phenomenological level, leading to the remarkable relation [36]

$$\left|\frac{V_{ub}}{V_{cb}}\right|^2 \simeq \left|\frac{V_{ub}}{V_{tb}V_{ts}^*}\right|^2 = \frac{3\alpha}{\pi}|c_7(m_b)|^2 \eta_{QCD} \frac{\hat{\Gamma}_u(E_0)}{\hat{\Gamma}_s(E_0)} + \mathcal{O}\left(\frac{\Lambda_{\text{QCD}}}{m_b}\right). \tag{3.53}$$

Here the η_{QCD} factor contains radiative corrections, and the two integrals over the endpoint region are defined via

$$\hat{\Gamma}_u(E_0) \equiv \int_{E_0}^{\infty} dE_\ell \frac{d\Gamma(B \to X_u \ell\nu)}{dE_\ell} \tag{3.54}$$

$$\hat{\Gamma}_s(E_0) \equiv \frac{2}{M_B}\int_{E_0}^{\infty} dE_\gamma (E_\gamma - E_0) \frac{d\Gamma(B \to X_s\gamma)}{dE_\gamma}, \tag{3.55}$$

[14] The appearance of this function is the way that QCD naturally accounts for the "Fermi motion" of the initial state quark within the meson; in the past, various models have introduced this artifact by hand in a rather *ad hoc* fashion.

where E_0 is the experimental cut applied in the lepton endpoint region. The conclusion is that a measurement of the integrated quantities $\hat{\Gamma}_u(E_0)$ and $\hat{\Gamma}_s(E_0)$ yields a direct determination of the ratio $|V_{ub}/V_{cb}|$. The value of this technique is that intermediate extraction of the non-perturbative structure function is sidestepped completely. In addition, the use of the ratio helps reduce various hadronic uncertainties to the level of smaller $1/m_b$ power corrections.

Prospects

The basic strategy for determining $|V_{ub}|$ by looking at high momentum leptons dates back to the original observation of charmless semileptonic B decays, and has greatly evolved in experimental and theoretical sophistication and precision since then. However, even with the parameterization of our non-perturbative ignorance with a QCD-based shape function, whose details can be extracted from experimental features of $b \to s\gamma$, the approach still suffers from three as-yet unavoidable weaknesses [14]:

1. The identification between the shape function in $B \to X_u \,\ell\, \nu$ and $B \to X_s\, \gamma$ decay only holds to leading order in the $1/m_b$ and α_s expansions [14]. Perturbative QCD and "higher-twist" corrections inevitably confound the "universal" nature of the shape function. These complications can be ameliorated slightly by the use of relations like that in Eqn 3.53 [42], which avoids the extraction of the shape function as an intermediate step, but its application is limited on the experimental side by the fact that current lepton spectrum measurements are in the $\Upsilon(4S)$ rest frame rather than the B frame.[15]

2. There are contributions specific to $b \to u\,\ell\,\nu$ that are still not calculated or not calculable in principle, any of which could be quite sizeable. These include contributions from dimension-6 operators such as the one describing weak annihilation and other sub-leading corrections to the shape function. (See later Sec 3.4.2 for more discussion.)

3. The endpoint region contains such a small part of the full $b \to u\,\ell\,\nu$ phase space that it may be vulnerable to violations of (local) quark-hadron duality. (See later Sec 3.4.1 for more discussion.)

Experimental interest in this approach has not wavered, and the analysis techniques have seen considerable sophistication as well. Recently, CLEO and the B factories [128–130] have released a new round of results that combine the latest information from inclusive $b \to s\gamma$ with measurements of the semileptonic rate above

[15]With the advent of B-tagging and other-B reconstruction methods at the B-factories, the application of these relationships or improved ones will only become more practical.

$\sim$2 GeV to extract $|V_{ub}|$ with much better-controlled uncertainties than ever before. They deliver values for $|V_{ub}|$ with a properly-estimated precision at better than the 20% level.

3.3.2 Hadronic Mass M_X^2

One of the disadvantages of working in the endpoint region of the lepton energy spectrum is that only about 10% of $b \to u\,\ell\,\nu$ decays lie above the kinematic limit for charm. A different and kinematically more direct approach was first suggested by Barger *et al.* in 1990 [44]. Recall that the essential idea behind the lepton energy cut $E_\ell > (M_B^2 - M_D^2)/2M_B$ is to exclude semileptonic decays where a heavy, charm-containing meson is produced. But since there is no unique, single-valued mapping between M_X and E_ℓ (due to the presence of the neutrino), the cut on E_ℓ is not particularly efficient at implementing the implicit bound on M_X. Indeed, a more efficient method is to reconstruct the hadronic system and apply the intended cut directly, requiring $M_X < M_D$. The gain in efficiency is significant: a simple cut $M_X < M_D$ suffices to eliminate the lightest possible charm-containing hadronic system, yet retains as much as 90% of the $b \to u$ events. On the experimental side, one needs to somehow infer the invariant mass of the recoiling hadronic system, a more demanding task than identifying high-momentum leptons. Resolution issues will force the actual cut on the reconstructed mass to be somewhat lower than M_D to reduce leakage from charm down into the signal region for $b \to u\,\ell\,\nu$ thus defined.

The extraction of $|V_{ub}|$ proceeds directly from a calculation of the differential rate $d\Gamma/dM_X$ (or equivalently, $d\Gamma/dM_X^2$). An experiment essentially counts the number of $b \to u$ decays with $M_X < M_X^{\max}$, which translates to a measurement of the partial rate for that mass window. If the fraction of the total rate that falls into the given window is well-predicted theoretically, an extrapolation to the total rate can be done readily, and then Eqn 3.37 once again provides access to $|V_{ub}|$. The clear obligation on the theoretical side is thus to provide this fraction

$$\Phi(M_X^{\max}) = \frac{1}{\Gamma(B \to X_u l \nu)} \int_0^{M_X^{\max}} dM_X \frac{d\Gamma}{dM_X}, \tag{3.56}$$

with reasonable uncertainties.

The advantage in the theoretical evaluation of the spectrum $d\Gamma/dM_X^2$ compared to $d\Gamma/dE_\ell$ is that a cut on M_X^2 does not preferentially weight the light hadronic states: all states satisfying $M_X < M_D$ are accepted on equal footing. This selection contrasts with the case of a cut on the lepton energy, where the endpoint is dominated by contributions from the lightest states, such as π and ρ. Consequently, it is much more likely that the first few terms of the OPE will be an adequate description of the decay rate in the region $M_X^2 < M_D^2$ than in the region

$E_\ell > (M_B^2 - M_D^2)/2M_B$. This observation also leads to the expectation that the (restricted) mass spectrum will be less sensitive to duality violations.

Standard application of the OPE results in a differential rate $d^2\Gamma/ds_h dE_h$ calculated in perturbation theory to some order, in terms of the *partonic* variables $s_h \equiv m_h^2$ and E_h, the invariant mass and energy of the partons arising from the b quark decay. By integrating over the allowed range for the partonic energy E_h, one then naturally obtains the spectrum in the partonic invariant mass variable s_h. However, in order to apply a cut on the physical hadronic mass, one needs to change to the hadronic variables

$$\frac{s_H}{m_b^2} = s_h + \epsilon E_h + \epsilon^2 \tag{3.57}$$

$$\frac{E_H}{m_b} = E_h + \epsilon, \tag{3.58}$$

where $\epsilon = \overline{\Lambda}/m_b$ captures the difference between the partonic and hadronic variables.

Unfortunately, after this transformation, a double logarithmic singularity appears at $s_H = \overline{\Lambda} M_B$, and at higher orders, more logarithmic singularities appear [45–50]. These problems make the differential rate $d\Gamma/ds_H$ hard to predict reliably even in perturbation theory. The behavior of the spectrum near this singularity is less important for observables that average over larger regions of the spectrum significantly beyond $\overline{\Lambda} M_B$. Thus keeping the experimentally-imposed cut as high as possible is crucial to control of the perturbative expansion. The problem is that for reasons of finite experimental resolution, the ideal cut at $M_D^2 \approx 3.5\ \mathrm{GeV}^2$ will need to be lowered, bringing it very close to $\overline{\Lambda} M_B$, which is numerically already at the same scale.

Non-perturbative effects—which are significant across the entire low mass region $s_H \leq \overline{\Lambda} M_B$—are applied using the usual smearing procedures developed in the analysis of the lepton endpoint region. Since the shape function is essentially unknown, one can only study these effects only in the context of toy models. As in the lepton endpoint region, the smearing has a significant effect on the spectrum, re-distributing the rate across any cut at the scale of M_D^2. Fig 3.4 illustrates the effects of this non-perturbative physics with a model implementation of the shape function.

Alternatively, as suggested by Leibovich *et al* [48], one can again consider the use of radiative $b \to s\gamma$ decays to determine the non-perturbative structure function and instead extract the ratio $|V_{ub}|/|V_{ts}|$. Much like the suggestion in the endpoint analysis, the idea is to eliminate ignorance of the Fermi motion of the heavy quark by incorporating measurements from another B decay that involves the same fundamental distribution function. But this procedure only mitigates the crippling ignorance of the non-perturbative effects, and cannot resolve the fundamental problems in the perturbative expansion.

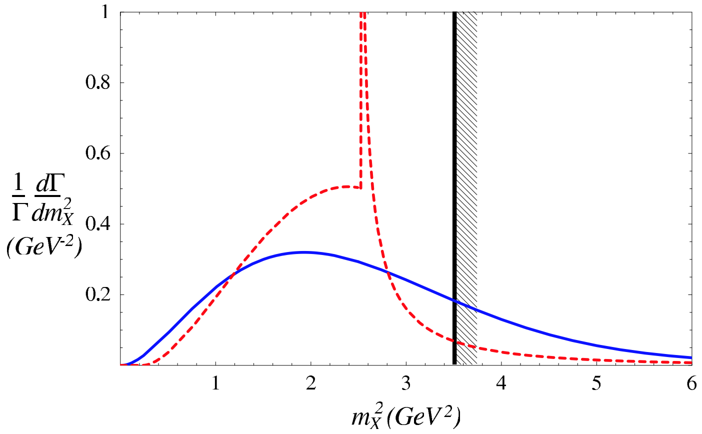


Figure 3.4: Smoothing effect of non-perturbative shape function on the hadronic mass spectrum. The red (dashed) curve shows the first-order perturbative result, after translation to hadronic variables, giving rise to a singularity at $s_H \sim \overline{\Lambda} M_B$. The blue (solid) curves shows the physical spectrum after convolution with the shape function that effectively implements the Fermi motion of the b quark in the initial-state B meson.

In summary, a straightforward cut $M_X < M_D$ is quite efficient at singling out $b \to u$ decays from $b \to c$ and accepts a large patch of phase space. This last fact leads one to hope that the analysis will be less sensitive to local duality violations. However, as we have seen, the ability to even perform the theoretical calculations is extremely sensitive to the placement of the experimental cut. The numerical coincidence that $M_D^2 \simeq \overline{\Lambda} M_B$ gives rise to significant non-perturbative corrections in the region retained by the cut, and perturbative effects lead to an unphysical singularity at the same point. The result is a direct conflict between the experimental need to apply a more restrictive cut $M_X^{\rm cut} < M_D$ and the theoretical desire to save the calculation by raising the cut as high as possible.

3.3.3 Dilepton Mass q^2

Bauer, Ligeti, and Luke [51, 52] have proposed an inclusive determination of $|V_{ub}|$ from the dilepton invariant mass spectrum. With a cut $q^2 > (M_B - M_D)^2$, manifestly above the charm endpoint, $b \to c$ decays can be entirely eliminated while retaining close to 20% of the $b \to u$ rate. As just discussed, the cuts on lepton energy or hadronic mass required to isolate $b \to u\,\ell\,\nu$ from $b \to c\,\ell\,\nu$ spoil the convergence of the OPE, and hence require an infinite set of non-perturbative contributions to be resummed into an unknown structure function. In contrast, this proposal uses an observable for which the OPE does not break down in the region that is free from the charm background.

The key benefit is that in the restricted region of phase space singled out by a cut on q^2, the influence of the non-perturbative shape function is reduced and the OPE expansion is much better-behaved. The non-perturbative smearing of the q^2-spectrum is demoted to a sub-leading effect; indeed, the nature of the cut on q^2 is to forbid the hadronic final state from moving fast in the B rest frame, so the light-cone expansion which gives rise to the shape function is rendered irrelevant. Non-perturbative effects are only significant at large values of q^2, where the difference between the hadronic and partonic endpoints is roughly $M_B^2 - m_b^2 \simeq 2\overline{\Lambda} M_B$, and the contribution of these additional terms will only be important if the spectrum is integrated over a region of similarly small width. Fig 3.5 illustrates these claims with a a comparison of the parton-level q^2 spectrum before and after the application of non-perturbative smearing via a model shape function.

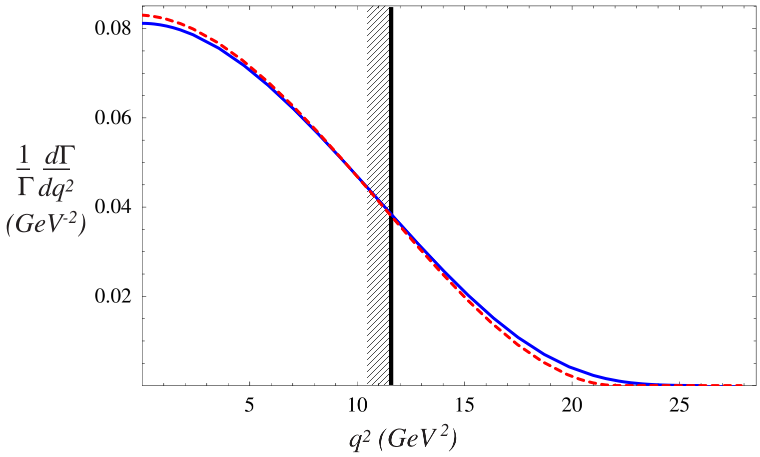


Figure 3.5: Impact of non-perturbative "smearing" on the parton-level q^2 spectrum. The interpretation of colors and line styles is the same as in the previous two figures. The message here is that the non-perturbative corrections are parametrically suppressed with this approach, so that shape function effects are far less important.

In the initial proposal, Bauer *et al.* calculate the fraction $F(q_0^2)$ of all $b \to u\,\ell\,\nu$ events that lie above a cut $q^2 > q_0^2$ by integrating the differential rate $d\Gamma/dq^2$, including perturbative and non-perturbative corrections to $\mathcal{O}(1/m_b^2)$ and $\mathcal{O}(\alpha_s^2\beta_0)$.[16] Normalizing to the well-known full (charmless) semileptonic width, they use (a variation of) the result in Eqn 3.37 to extract $|V_{ub}|$. Re-expressing the dependence on the b quark mass with the use of the upsilon expansion, they find

$$|V_{ub}| = (3.04 \pm 0.06 \pm 0.08) \times 10^{-3} \left[\frac{\mathcal{B}(B \to X\ell\nu)|_{q^2 > q_0^2}}{10^{-3} F(q_0^2)} \frac{1.6 \text{ ps}}{\tau_B} \right]^{1/2}, \qquad (3.59)$$

[16] The function $F(q_0^2)$ is the analog for the dilepton mass analysis of $\Phi(M_x^{\max})$ for the hadronic mass analysis.

where the errors are estimates of the perturbative and non-perturbative uncertainties, respectively, in the upsilon expansion. The dominant uncertainty remains the error on the short-distance b quark mass, whichever way it is defined.

For the nominal cut at $q_0^2 = (M_B - M_D)^2 \approx 11.6\ \mathrm{GeV}^2$, they find 18.6% of the full rate is accepted (using $\bar{\Lambda} = 0.4$ GeV, $\lambda_1 = -0.2\ \mathrm{GeV}^2$, and $\alpha_s(m_b) = 0.22$). As the cut is raised, the fraction of the rate that is sampled falls rather steeply; for instance, with a cut $q_0^2 = 15\ \mathrm{GeV}^2$, the fraction is reduced to 6.7%. Thus, the advantage of not having to employ information from $b \to s\gamma$ and the absence of $1/m_b$ power corrections is offset by the fact that the cut eliminates a large fraction of the rate. Worries about duality violations begin to lurk once more. In addition, issues of finite experimental resolution may require the cut to be significantly higher than the charm endpoint value; since the theoretical error on $|V_{ub}|$ in this scheme grows rapidly as the cut is increased, the analysis may be pushed into a regime of low acceptance and large errors.

In a subsequent analysis, Neubert [53, 54] has investigated in more detail the structure of the heavy quark expansion as it applies to the computation of $F(q_0^2)$. He finds that the relevant mass scale μ_c controlling the size of the corrections is really of order the charm quark mass m_c rather than the heavier b mass. This conclusion follows from the observation that the largest values of the hadronic mass and energy accessible in the presence of this cut are only of order m_c.[17] The corrections must then be analyzed using a two-step, hybrid expansion in terms of the parameters μ_c/m_b and Λ/μ_c. The scale μ_c is the typical hadronic momentum scale, of order m_c or less, and depends on the value q_0^2 of the cut. A modified version of HQET can then be used to disentangle the physics associated with the scales m_b and μ_c, while the OPE is sufficient for separating the physics at the scale μ_c from the truly long-distance physics.

Using a renormalization group-improved (RG) result and a potential-subtracted mass m_b^{PS} that improves aspects of the perturbative calculation, Neubert revises the estimates of the various uncertainties considered by Bauer *et al.* in their analysis. In addition to a heightened sensitivity to m_b, he finds that the higher-order power corrections now at the lower scale of $(\Lambda/\mu_c)^3$ are of increased signigicance. As the cut on q^2 is raised from the efficient value of $(M_B - M_D)^2$, the size of these latter terms changes rapidly from sub-leading to dominant, worsening from an 8% projected error on $|V_{ub}|^2$ to 25% for $q_0^2 = 15\ \mathrm{GeV}^2$; the total estimated theoretical uncertainty on $|V_{ub}|$ grows from 10% to almost 20%. He concludes that the q^2 proposal is sound, but the initial promise of 5–10% resolution on $|V_{ub}|$ is too optimistic. With a reasonable cut in the vicinity of $q_0^2 = 12.5\ \mathrm{GeV}^2$, a precision

[17] Starting from Eqn 2.5 defining q^2 in terms of the hadronic variables, we can apply the inequality $q^2 > (M_B - M_D)^2$ and immediately arrive at the bound $E_X \leq M_D + (M_X^2 - M_D^2)/2M_B$, showing that indeed the hadronic energy is of the order of the charm mass. Since the invariant mass of the hadronic system cannot exceed its energy, the same bound applies to the mass M_X.

better than 15% is still theoretically possible.

Recently, a strategy relying on combined cuts on the dilepton and hadronic masses has been proposed [55]. The basis for the suggestion is the observation that a twist expansion is required if the phase space surviving a cut is dominated by hadronic states with energy much larger than their invariant mass. This problem arises for the cases of a cut on E_ℓ or a cut on M_X^2. However, a cut on the dilepton mass rejects the low q^2 region, and restricts the hadronic energy to $E_X < M_B - \sqrt{q_0^2} < M_D$, which eliminates the "dangerous" region of phase space. In essence, the strategy employs a hadronic mass cut to reduce the charm background while keeping a large fraction of the $b \to u$ rate, and simultaneously applies a lower cut on q^2 to avoid the shape function region where non-perturbative effects can erode the theoretical calculation. Fig 3.6 illustrates these considerations. Compared to a pure q^2 cut, the uncertainty from unknown $1/m_b^3$ terms in the OPE is significantly reduced and the fraction of $b \to u$ events retained is roughly doubled; compared to a pure M_X^2 cut, the uncertainty attributed to ignorance of the shape function is significantly reduced as well. The originators of this new mixed strategy estimate that determining $|V_{ub}|$ with a theoretical error of 5–10% should be possible.

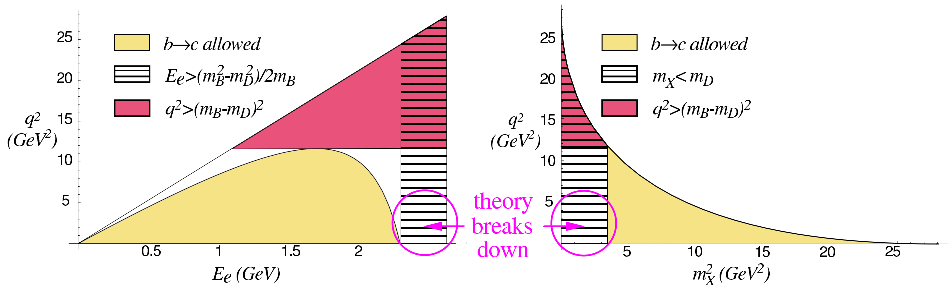


Figure 3.6: A Dalitz plot for $b \to u\,\ell\,\nu$ in the q^2 *vs.* M_X^2, E_ℓ planes. The regions free from charm are highlighted, as well as the areas passing various experimental cuts. The circled region in the corner of each phase space plot indicates where the local OPE breaks down and a twist expansion is required. From Ref [55].

3.3.4 Summary

The theoretical challenge in extracting $|V_{ub}|$ is kindled by the experimental requirement that the measurement be made in some restricted region of phase space, where background $b \to c$ decays are kinematically forbidden. This restriction increases the relevance of non-perturbative corrections, which can be large enough to cripple the usual OPE in these extreme regions. Table 3.1 summarizes this tension by listing the size of the region over which non-perturbative effects are important along with the size of the region remaining after the phase space cut to capture the $b \to u$ decays.

Table 3.1: A comparison between the three methods for extracting $|V_{ub}|$ discussed in this chapter. Each strategy relies on a cut on a different kinematic distribution to select a region free from contamination from $b \to c$ decays; the size of this region is appears in the second column. But for each spectrum, there is also a natural scale at which the contribution of non-perturbative (or hadronic) effects becomes important, indicated in the third column. Since $M_D^2 \sim \Lambda M_B$, the width of this region for the first two strategies is in fact comparable to the charm-free region selected by the cut, causing calculations in this experimentally-accessible region to break down. The final column lists the estimated efficiency for each strategy. Experimental issues are not addressed in this table. (Modeled after a similar table presented in Ref. [51].)

Decay distribution	**Region free of charm**	**Region where had. effects signif.**	$b \to u$ **Eff.**
$d\Gamma/dE_\ell$	$\Delta E_\ell = M_D^2/2M_B$	$\Delta E_\ell \sim \Lambda$	$\sim 10\%$
$d\Gamma/dM_X^2$	$\Delta M_X^2 = M_D^2$	$\Delta M_X^2 \sim \Lambda m_b$	$\sim 80\%$
$d\Gamma/dq^2$	$\Delta q^2 = 2M_B M_D - M_D^2$	$\Delta q^2 \sim \Lambda m_b$	$\sim 20\%$

The general feature that, numerically, M_D^2 is the same size as ΛM_B is an unfortunate coincidence,[18] since it leads to (large) non-perturbative effects in both the endpoint of lepton energy spectrum, where $E_\ell \sim m_b/2 + \Lambda$, and in the region $M_X \sim M_D$, which suffers equally from the same non-perturbative physics. (Neglecting the neutrino, it is easy to see that $M_X \sim \Lambda M_B \Rightarrow E_\ell \sim M_B/2 \pm \Lambda$.) As we have seen, the non-perturbative effects can be packaged into a universal shape function that describes the Fermi motion of the b quark within the initial-state meson, but this function is essentially unknown beyond its first few moments. The size of the region over which this shape function is smeared controls the size of the contributions from the higher-dimension operators it conceals. The lepton endpoint and the region $M_X < M_D$ in hadronic mass are both inadequate for diluting these contributions, and so the shape of these spectra depend heavily on the details of this function. Both of these methods must rely on input data from *other* inclusive B decays to proceed. The q^2 approach improves on these earlier proposals slightly, but the expansion parameter in the OPE for a q^2 analysis turns out to be of size Λ/m_c, which tends to inflate higher-order contributions, and a cut on q^2 much beyond the charm endpoint is rather inefficient. A combined q^2-M_X^2 analysis may prove more effective than any of these single-variable analyses.

[18] I place the blame for this unlucky state of affairs on the Higgs boson. It's responsible for so much of the trouble in the (particle physics) world today.

The question of which measurement technique to use to determine $|V_{ub}|$ has no single answer. While none of the preceding proposals for extracting $|V_{ub}|$ is markedly superior to the others, the combined q^2-M_X^2 approach has been well-received as the first step to constructing a strategy that is effective on multiple theoretical fronts while also respecting experimental limitations [2]. The approach has reduced dependence on the shape function yet still includes a sufficient fraction of the full rate to soften concerns about weak annihilation contributions and violations of duality. On the experimental side, the large acceptance and excellent discrimination offered by the hadronic mass cut is valuable, as is the utility of the q^2 cut in rejecting non-$B\overline{B}$ backgrounds.

It is likely that the optimal measurement strategy may only be realized after the fact, when a final analysis of $b \to u\,\ell\,\nu$ is complete [56]. Meeting the challenges posed by $b \to u\,\ell\,\nu$ requires an investigation of theoretical errors in conjunction with experimental uncertainties, for they are deeply entwined in a fundamental and unavoidable way, a feature made startlingly clear in the study of the endpoint of the lepton energy spectrum. Depending on the details of experimental efficiency and resolution, some combination of cuts on all three variables could turn out to be the most effective way to minimize the overall uncertainty on $|V_{ub}|$.[19]

The first step toward identifying an optimal strategy is a dedicated and careful analysis of *all* sources of theoretical uncertainty,[20] which includes a realistic structure function well-constrained by data, and exhaustive, conservative errors that assess contributions from all expected—and neglected—correction terms.[21]

It would be remiss not to include mention of recent work by Bosch, Lange, Neubert, and Paz that uses the tools of effective field theory to compute, for the first time, a complete next-to-leading order prediction for decay rates and spectra in the shape function region [57]. They employ a systematic treatment that first matches QCD onto soft-collinear effective theory (SCET), integrating out hard quantum fluctuations; then they match SCET onto HQET, integrating out

[19]Any experimental analysis will, practically speaking, be forced to cut on all three variables, if only indirectly due to acceptance or other detector effects.

[20]One could argue that a large part of the progress on $b \to u\,\ell\,\nu$ in the past 20 years has been in recognizing new sources of theoretical uncertainty. It is time to put this program into practice and start delivering.

[21]Such an effort is likely to be tough work, and it's not entirely clear where to start. An old Sufi mondo comes to mind:
Once some pilgrims came upon an old Sufi digging through the dirt on the side of a road. When asked what he was looking for, he replied: "A treasure I have lost." So all the pilgrims joined in and searched until they were hungry and tired. Finally, one of them asked the Sufi if he was sure this was where he lost his treasure. "Oh my, no! I lost it over there, on the other side of those mountains." To which they replied: "Then why in the name of Allah are you looking for it here?" He answered, "Because here there is more light."

the hard-collinear fluctuations. The final theory provides new model-independent insights into the nature of the shape function, significantly reduces the model dependence usually incurred in relating $b \to s\gamma$ to inclusive $b \to u\,\ell\,\nu$ via the shape function, and suggests a new method for a high-precision determination of $|V_{ub}|$ using the hadronic light-cone momentum variable $P_+ = E_X - |\vec{p}_X|$. Further discussion of this important progress[22] is beyond the scope of this work.

3.4 Untamed Uncertainties

The phenomenological understanding of $b \to u\,\ell\,\nu$ has evolved considerably since the first observation in 1990 of high-momentum leptons from B decay beyond the endpoint for $b \to c\,\ell\,\nu$ [123, 124]. Somewhat naïve, model-based extractions of $|V_{ub}|$ have been replaced by a rigorous treatment using the tools of HQET and the OPE, and this framework has allowed for a significant increase in confidence and precision in efforts to extract $|V_{ub}|$ from semileptonic B decay. The breadth and depth of this program has hopefully been amply demonstrated in the preceding discussion of approaches for extraction of $|V_{ub}|$. Nonetheless, there remain some theoretical blindspots that only recently have been clearly identified. Unfortunately, these uncertainties are not yet fully quantified, partly because they are non-parametric by nature, and constraints of such a sort are difficult to come by. The growing consensus in the theoretical community [14], however, is that these lingering issues need to be resolved to meet the hopes for a precision value of $|V_{ub}|$ with well-constrained errors.

The concerns are tied to basic assumptions that are implicit in almost any systematic treatment of inclusive $b \to u\,\ell\,\nu$ decay, and thus are pervasive, ingrained, and correspondingly difficult to isolate and quantify. Although the categorization below is not unique, and could be re-grouped in multiple ways under different headings, we follow [2] and list the most important outstanding theoretical issues as follows:

- **Quark-hadron duality.** Within the framework described in Sec 3.2.3, the assumption of quark-hadron duality underpins the calculation of any inclusive quantity, and in particular, the notion of *local* duality is vital to calculation of differential spectra or partial rates in reduced regions of phase space. Yet it is clear that this duality is by no means guaranteed or exact.[23]

[22] Arguably, these calculations represent a "quantum leap" forward in theoretical understanding of these decays in the shape function regime. . . . !

[23] For instance, high-momentum leptons are clearly seen above the parton-level limit of $m_b/2$ for $b \to u\,\ell\,\nu$ decays, a region where the parton-level rate vanishes. That is, with too tight a lepton energy cut, local duality simply cannot hold—the existence of such leptons would be ruled out!

- **Neglect of weak annihilation.** One contribution to the $B \to X_u \ell \nu$ decay rate is a series of annihilation-like diagrams similar in part to valence quark annihilation in the purely leptonic decay of a charged B. The magnitude of this process is expected to be small compared to the usual $b \to u \ell \nu$ rate computed at lower order, but it will be concentrated in a small region of phase space, where its neglect could have significant impact on the extraction of $|V_{ub}|$. *Constraining this uncertainty is the subject of this thesis.*

- **Unknown corrections to the shape function.** Sub-leading corrections to the non-perturbative shape function are still poorly estimated and remain somewhat controversial, but formalisms for estimating these corrections are now emerging.

We elaborate on these issues briefly, trying to place them appropriately in the context already developed.

3.4.1 Quark-Hadron Duality

The notion of quark-hadron duality (or simply "duality") is rooted in the idea that a quark-level calculation should, in some real sense, approximate physical hadronic observables. It is implicitly invoked repeatedly to connect quantities evaluated at the quark-gluon level to the observable world of hadrons, but, particularly as it applies to the study of $b \to u \ell \nu$, it stands really only as an approximation, for which the attendant uncertainty still needs to be properly assessed.[24]

In order to understand the limitations of duality, it must be clear to what exactly the concept refers. Several versions have been described in the literature, and there is now a categorical distinction[25] between at least two types: *global* duality and *local* duality. (The discussion followed here is based on that in Ref [14].)

The classic example of the explicit application of duality arises in the discussion of the cross section for the scattering process $e^+e^- \to$ hadrons. Naively, for large center-of-mass energies $\sqrt{s} \gg \Lambda_{\mathrm{QCD}}$ it would seem that the calculation could be handled perturbatively in terms of quarks and gluons, due to the asymptotic freedom of QCD. However, the presence of production thresholds, such as those for charm, corrupts the straightforward computation of $\sigma(e^+e^- \to \text{hadrons})$, by introducing relevant but intrinsically non-perturbative physics. The resolution is

[24]Historically, concerns about the adequacy of duality assumptions were raised essentially at the same time that is was first used in making inclusive calculations at the dawn of the OPE and the recognition of the heavy quark symmetry, specifically in the context of deep inelastic scattering. Its role in inclusive B decays and the potential impact of its violations for values of $|V_{ub}|$ extracted from $b \to u \ell \nu$, however, remains still incompletely understood.

[25]The distinction is not without its dissenters; see *e.g.* Shifman's critique in Ref [59].

to posit that equality between the quark-level result and the physical world holds after an averaging or "smearing" over an appropriate energy interval is applied. In general, for an observable T, the claim is that

$$\langle T^{\text{hadronic}} \rangle_w \simeq \langle T^{\text{partonic}} \rangle_w \tag{3.60}$$

where $\langle \ldots \rangle_w$ denotes the smearing using a smooth weight function $w(s)$:

$$\langle \ldots \rangle_w = \int ds \ldots w(s). \tag{3.61}$$

The extent to which we can trust that $\langle T^{\text{partonic}} \rangle_w$ actually represents the physical quantity $\langle T^{\text{hadronic}} \rangle_w$ depends on the details of the weight function—namely, its width. It might be broad compared to the structures that appear in the hadronic spectrum, or it could be quite narrow; in the extreme, we could have $w(s) \sim \delta(s - s_0)$. The convention is to refer to the first case (averaging over a region large compared to hadronic structure) as relying on *global parton-hadron duality*, and to the second case (point-by-point equality of the hadronic and partonic quantities) as an instance for *local parton-hadron duality*. The nomenclature "parton-hadron" emphasizes that the duality is between the physical, hadronic quantity and the partonic result which is computed from quark *and* gluon contributions. (However, we will still adopt the shorter term "quark-hadron.") Clearly quantities like total widths will rely on global duality, since they involve integration over all available phase space, while a differential rate that integrates out only some degrees of freedom will rely on the validity of local duality.

The features of duality violation have been clarified by theoretical consideration over the past several years [58–61]. A precise definition of duality requires a more technical treatment than we can afford space and patience for here, but the discussion above in Sec 3.2.3 introduced some of the essential ingredients. The essence of the OPE is an expansion of a combination of Green's functions as a series of local operators with short-distance coefficients. The entire analysis is conducted in the Euclidean domain, far from any singularities induced by hadronic thresholds, and has to be continued analytically into Minkowski space to connect to observable hadronic quantities. The assumption of duality is essentially the statement that the truncated series in Euclidean space corresponds in a natural way to—or is dual to—the truncated physical calculation in Minkowski space. The extent to which this assumption is flawed measures the extent to which duality is violated. Typically, neglected terms in the Euclidean domain are exponentially small, but under analytic continuation, they are mapped to oscillating functions of momentum transfer which are only power-suppressed to an extent that depends on the details of the process under consideration. These neglected terms embody the local violations of duality.

Duality then is not an additional assumption applied in the construction of the OPE, but a natural result of translating the OPE back to physical space-time coordinates. Violations of duality hinge on our ignorance of the exact analytic solution to QCD in the Euclidean regime, and the subsequent uncertainty that arises upon the analytic continuation of a *finite* number of terms from the OPE result back into physical Minkowski space.

Bigi and Uraltsev [58] also emphasize that the effects of local duality violation must oscillate as a function of energy scale and have vanishing averages or be exponentially small. So it cannot be blamed for the systematic excess or deficit of decay rates.[26]

Current progress on understanding duality violations is along two fronts. Theorists are exploring toy models (far simpler than full QCD) for which exact solutions are known and the extent of duality violations can be explicitly determined. Mapping these conclusions back onto QCD is still a challenging enterprise. The second approach is the oft-quoted experimental dictum: determine the same quantity in several independent ways and compare the results. Heavy quark physics lends itself well to such tests, since it can predict numerous decay rates and other spectral properties from only a small handful of basic input parameters (specifically: quark masses and hadronic expectation values). One test of duality in this regard is a comparison of the inclusive and exclusive determinations of the CKM elements $|V_{cb}|$ and $|V_{ub}|$. For both, the inclusive and exclusive values are consistent [2]. Further support for the operator product expansion used in these analyses comes from the general agreement seen for HQET parameters extracted from numerous moments of various semileptonic spectra. However, other uncertainties are also present in these measurements, making it hard to accept this comparison as a yardstick for measuring the size of duality violations, rather than simply as a (limited) test of general theoretical control over the program as a whole.

Violations of duality will have different impact depending on the experimental route taken for the identification of $b \to u\,\ell\,\nu$ and the subsequent extraction of $|V_{ub}|$. It is an adequate but not necessary rule that as a measurement becomes more inclusive, covering more of phase space (uniformly), the impact of duality violations will diminish. In particular, the endpoint analysis may be subject to violations of local duality since it considers such a narrow slice of phase space (only $\sim$10% of the total $b \to u\,\ell\,\nu$ rate), but the hadronic mass analysis should be much less sensitive, since it accepts 80% of the available rate.

In general, recent attempts to quantify this particular uncertainty suggest that it will be sub-dominant compared to other theoretical uncertainties that currently limit our knowledge of $|V_{ub}|$. It is likely that future understanding of duality

[26]Before theoretical scrutiny was brought to bear on the question of duality, it was proposed that the discrepancy between the measured and expected B semileptonic widths could perhaps be attributed to duality violation. The current understanding rules out this possible escape hatch, since the extent attributable to duality violations is constrained to be less than a fraction of a percent [14].

violations will keep pace with attempts to reduce these other theoretical errors, suggesting that uncertainty about quark-hadron duality is unlikely to ever be the limiting factor in efforts to measure $|V_{ub}|$.

3.4.2 Weak Annihilation

When the OPE for the charmless semileptonic width is carried to higher order in Λ/m_b, new terms naturally emerge. A few of these can be broadly characterized as "spectator-dependent" corrections to the lower-order terms since, for the first time, the light spectator quark in the B meson becomes involved. Some of these corrections are simply the $1/m_b$-suppressed pieces of the kinetic and chromomagnetic operators, and are expected to be rather small compared to their leading-order contributions [63]. But two new operators arise as well: One is the so-called Darwin term,

$$\rho_D^3 = -\frac{1}{2M_B}\langle B|\frac{g_s^2}{2}\bar{b}\gamma_\alpha t^a\, b\sum_q \bar{q}\gamma_\alpha t^a q|B\rangle, \tag{3.62}$$

and is estimated to be a small (downward) correction to the total width of about $-(1$–$2)\%$ [63]. The second operator is also of a four-fermion form, but includes only the u quark and has a different color and Lorentz structure:

$$\frac{1}{2M_B}\langle B|\bar{b}_L\gamma_\alpha u_L\ \bar{u}_L\gamma_\beta b_L|B\rangle\,(\delta_{\alpha\beta} - v_\alpha v_\beta), \tag{3.63}$$

where the new quantity $v_\mu = p_\mu^B/M_B$ is the four-velocity of the B. This term describes the process of *weak annihilation* in the B meson, similar to the leptonic decay of the $B^\pm$ where the constituent b and u quarks annihilate. In general, it affects charged and neutral B mesons differently, and so can lead to a decay rate difference between $B^\pm$ and B^0 mesons. The magnitude of this contribution cannot be formally calculated, but some initial qualitative observations can be made. Because there are only two particles in the final state, the term is enhanced by a factor of $16\pi^2$ relative to the usual terms in the OPE for $b \to u\,\ell\,\nu$. Further, the weak annihilation channel contributes formally at the endpoint of the q^2 spectrum,

$$\frac{d\Gamma_{\rm WA}}{dq^2} \sim \delta(q^2 - m_b^2). \tag{3.64}$$

The relative contribution of this effect is thus maximal for a partial rate measurement that focuses only on the q^2 endpoint, and is only diluted by the inclusion of additional phase space. In other words, the corresponding uncertainty from our ignorance of of weak annihilation actually *grows* as a q^2-like or E_ℓ-like cut is tightened (so long as the high q^2 region is included). The localization of this process to a tiny region of phase space makes it effectively independent of cuts on M_X^2 or

q^2. In particular, there is potential for a large impact on an endpoint measurement, which is typically confined to the same small region where weak annihilation contributes in its entirety.

The contribution from weak annihilation vanishes in the QCD factorization approximation, so estimates of its size can be made by exploring violations of this hypothesis [62–65]. Specifically, the contribution to the total charmless semileptonic width from WA reads as [65]

$$\delta\Gamma(B \to X_u \ell\nu) = \frac{G_F^2|V_{ub}|^2 f_B^2 m_b^2 M_B}{12\pi}(B_2 - B_1) \tag{3.65}$$

where f_B is the B meson annihilation constant, and the phenomenological "bag constants" $B_{1,2}$ parameterize the matrix elements of two four-quark operators as shown below,

$$\begin{aligned} \langle B|\bar{b}_L\gamma_\alpha u_L \; \bar{u}_L\gamma_\alpha b_L|B\rangle &= \frac{f_B^2 M_B}{8} B_1 \\ \langle B|\bar{b}_L\gamma_\alpha u_L \; \bar{u}_L\gamma_\alpha b_R|B\rangle &= \frac{f_B^2 M_B}{8} B_2 \end{aligned} \tag{3.66}$$

In the limit of naïve factorization, the product of the bilinear operators is saturated by the simple vacuum insertion ($|0\rangle\langle 0|$), and the two parameters are equal, depending only on the flavor of the light quark in the B meson. For B_u, the result is $B_1 = B_2 = 1$, while for $B_{d,s}$, $B_1 = B_2 = 0$. Clearly, in this limit, the effect on the rate expressed in Eqn 3.65 vanishes.

However, the conventional assumption in the literature is that the factorization approximation holds to no better than about 10% accuracy; that is, correctly handling non-factorizable terms[27] will amount to corrections of about this magnitude. The standard arena for estimating the validity of factorization is charmed D mesons, where violations can potentially explain differences between the expected and observed values for the D semileptonic width, and perhaps also lifetime differences between the D_s and the D.[28] Using ingredients from several other theoretical investigations, Uraltsev estimates [63] that

$$\frac{\delta(\Gamma_{\rm SL}(b \to u))_{\rm WA}}{\Gamma_{\rm SL}(b \to u)} \approx 1\%, \tag{3.67}$$

[27] Perhaps rather obviously, "non-factorizable" labels a term that is non-vanishing only in the case that factorization is violated.

[28] Some recent analyses attribute all of the "missing" 50% of the D semileptonic width to contributions from non-factorizable terms [65], but it is by no means conclusively decided that violations of factorization are the only explanation for either the problem of the D semileptonic width or the lifetime difference $\tau_{D_s} - \tau_D$. Other possibilities include flavor SU(3) symmetry-breaking, the failure or inaccuracy of the OPE at scales as low as the charm mass, or some version of duality violations.

which only sets the scale for the significance of the potential contributions; he allows for a factor of two increase in the effects. Others [72, 73] have translated a 10% violation of factorization into a possible 2–3% enhancement for the total semileptonic width $\Gamma_{\rm SL}(b \to u)$. For an endpoint cut that retains only 10% of the total $b \to u\,\ell\,\nu$ rate, the contribution from weak annihilation can thus be magnified to a 20–30% correction to the partial rate. For a q^2 cut at the critical value of $(M_B - M_D)^2$, the correction could still be $\sim$10%, increasing in significance as the q^2 cut is raised.

Bigi *et al.* [62] have explored the role of helicity suppression in the decay rate difference between charged and neutral B mesons, still considered one of the best experimental avenues for constraining the contribution from weak annihilation. They find that the width difference due to the term in Eqn 3.63 can be parameterized as

$$\Delta\Gamma_{\rm SL} \equiv \Gamma_{\rm SL}(B^-) - \Gamma_{\rm SL}(B^0) \simeq \frac{G_F^2 |V_{ub}|^2 f_B^2 M_B^3}{8\pi}(1 - \frac{m_\ell^2}{M_B^2})(v\frac{m_\ell^2}{M_B^2} + 2g) \qquad (3.68)$$

The first term in the final factor (proportional to v) vanishes in the limit $m_\ell \to 0$, and reflects the conventional chirality suppression of weak annihilation as it applies to semileptonic decays. On the other hand, it is clear that even in the limit of massless leptons, the second term still survives and the size of the contribution depends on the parameter g. This term captures the non-factorizable contributions of the underlying matrix element. Thus, while the decays $B \to \ell\nu$ are indeed subject to helicity suppression, the reduction is subtly circumvented when hadrons are present in the final state. Choosing an arbitrary value of $g \sim 1/3$, Bigi finds that the contribution from weak annihilation could be as much as six times the rate for $B \to \tau\nu$. Current estimates of the branching fraction for this purely leptonic decay are in the neighborhood of 5×10^{-5}, suggesting that weak annihilation could be as large as a 15% effect, relative to a nominal rate of $\mathcal{B}(b \to u\,\ell\,\nu) = 1.75 \times 10^{-3}$.

There is speculation [66] that the non-factorizable contributions from the dimension-6 operator highlighted in Eqn 3.63 do not vanish for neutral B mesons. Whether these contributions should be interpreted physically as manifestations of "weak annihilation" is a matter of definition (or opinion).

We note that lattice QCD also has the potential to shed light on these effects.

3.4.3 Sub-leading Corrections to the Shape Function

From a more distant perspective, almost all of the complications in the analysis of $b \to u\,\ell\,\nu$ arise from problems with approximating an infinite (asymptotic) series by only its first few terms.[29] Duality violations can be traced to the neglect of

[29]Most of the remaining trouble is caused by our ignorance about the few terms that are retained in the truncated series.

terms in Euclidean space that become relevant in the analytic continuation back to Minkowski space; weak annihilation contributions arise from a neglected higher-order term in the OPE of HQET in Λ/m_b. Similarly, in the analysis of the endpoint region we encountered a breakdown of the usual OPE because an infinite number of terms were formally all $\mathcal{O}(1)$; the resummation of the leading singular terms led to the shape or light-cone momentum distribution function $f(k_+)$. The natural next step in understanding the OPE for $b \to u\,\ell\,\nu$ is the investigation now of the *sub-leading* terms in the shape function region. Unfortunately, much less is known about the new terms that arise at next-to-leading order.

We present a lightning review of the appearance of the shape function in the now-standard revision of the OPE for the endpoint region to illustrate where and how new sub-leading contributions to the leading-order result can appear. Recall that hadronic states in the endpoint region are kinematically constrained to have high energy and low invariant mass:[30]

$$E_X \sim m_b,\ M_X^2 \sim \Lambda m_b \gg \Lambda^2, \tag{3.69}$$

describing a region where the OPE approach is still valid[31] but the expansion parameter must be modified. The most singular terms in the modified expansion may be resummed into a non-local operator[32]

$$O_0(\omega) = \bar{h}_v\ \delta(\omega + in\cdot\hat{D})\ h_v \tag{3.70}$$

where n^μ is a light-like vector in the direction of the final hadrons, h_v is the HQET heavy quark field introduced previously, and $\hat{D}^\mu \equiv D^\mu/m_b$ is the reduced gauge-covariant derivative. The matrix element of this operator in a B meson defines the light-cone structure function for the meson:

$$f(\omega) = \frac{1}{2M_B}\langle B|O_0(\omega)|B\rangle. \tag{3.71}$$

The rate in the endpoint region is determined by $f(\omega)$,

$$\frac{d\Gamma}{dE_\ell} = \frac{G_F^2|V_{ub}|^2 m_b^4}{96\pi^3}\int d\omega\ \Theta(m_b - 2E_\ell - \omega) f(\omega), \tag{3.72}$$

but since the shape function is fundamentally non-perturbative and cannot be determined analytically, the rate in this region is intrinsically model-dependent even at leading order in Λ/m_b. (We've already discussed various efforts to reduce

[30]Observe that the high-energy, low-mass four-vectors are by definition light-like. Hence the alternative nomenclature describing the shape function as the "light-cone distribution function."

[31]Note that the OPE completely breaks down in the resonance regime $M_X^2 \sim \Lambda^2$ and becomes basically unsalvageable.

[32]We switch notation here and adopt the reduced (dimensionless) variable $\omega \equiv k_+/m_b$.

this basic ignorance with constraints on the moments or input from the photon spectrum in $b \to s\gamma$.)

Theoretical considerations have now advanced to the next terms in the non-perturbative expansion, suppressed by higher powers of Λ/m_b. Since these so-called "sub-leading twist corrections" are different between $b \to u\,\ell\,\nu$ and $b \to s\gamma$, the destroy the universality of the shape function across B decays. Hence the remarkable relation in Eqn 3.53 must, for instance, be modified to account for the differing corrections to each process. Recent analyses [67–69] identify five new non-local operators that emerge at the next order in the twist expansion, which reduce to four (unknown) structure functions for B meson decays. Bauer *et al.* evaluate the effects of these news terms on an endpoint analysis, using a standard ansatz for the structure function. They find the corrections to the partial rate are as large as 15% with a lepton energy cut $E > 2.2$ GeV, and conclude that since their analysis is strongly model-dependent, the impact for $|V_{ub}|$ could be equally as large. Leibovich *et al.* have carried out a similar analysis with comparable results.

Neubert [70] has shown that the anomalously large corrections reported by Bauer *et al.* are dominated by the first moment of the sub-leading shape function, which is known in terms of the hadronic parameter λ_2, making the correction better known that initially believed. Further, while the power correction to the shape function does lead to a ~25% enhancement in the partial rate (a 13% impact for $|V_{ub}|$), it is under good control, since the correction can absorbed into a redefinition of the leading-order shape function. Residual corrections from sub-leading shape function effects estimated across a range of models are only in the few percent range, so there is no significant loss of predictive power with the neglect of these additional terms. With the first few moments of the structure function known, the correction is largely in hand; hence the sub-leading power corrections do not pose a fundamental limitation on the extraction of $|V_{ub}|$.

More recently, Burrell, Luke, and Williamson [71] have explored the impact of sub-leading power corrections on the hadronic mass spectrum and come to similar conclusions when the extraction of $|V_{ub}|$ is augmented with input from $b \to s\gamma$. They find that the theoretical uncertainty in $|V_{ub}|$ due to higher-order shape function effects is at the few percent level, substantially less than the corresponding corrections to an endpoint analysis, and sub-dominant when compared to other sources of theoretical and experimental error.

We have now completed our tour of the extensive and mature theoretical program to extract $|V_{ub}|$ from inclusive $b \to u\,\ell\,\nu$ decays.

3.5 Summary and Outlook

Several recent proposals in the phenomenological literature respond to the challenge of extracting $|V_{ub}|$ from realistically obtainable data. No method is particularly stunning, nor fully immune to experimental and theoretical complications. In partial consequence, there is good reason to prefer the pursuit of multiple, competing approaches rather than relying on a single strategy with its array of associated uncertainties. In the end, each strategy promises to deliver a determination of $|V_{ub}|$ with a theoretical error of order 10% or better, but consensus within the theoretical community on quantitative versions of these estimates has grown only slowly. The new precision arises largely from the use of the model-independent framework of HQET and a judicious choice of experimental cuts, but every case displays a subtle interplay between what is theoretically calculable and what is experimentally feasible.

One of the essential tools in the analysis of $b \to u\,\ell\,\nu$ has been the heavy quark expansion, which connects inclusive calculations to the underlying theory of QCD. Inherent in this OPE-based approach, however, are three unavoidable sources of theoretical uncertainty:

1. Unknown (perturbative) terms of higher order in α_s,
2. Unknown (non-perturbative) terms of higher order in $1/m_b$,
3. Uncertainties in the input parameters α_s, m_b, and the expectation values of various local operators in the OPE.

This list is essentially exhaustive, since it recognizes the parameterization of our ignorance in all parts of the OPE approach. For instance, weak annihilation and sub-leading corrections to the shape function fall into the second category. Violations of duality are perhaps best regarded as additional uncertainties arising from the theory at a lower level, entangled with issues of the adequacy of the OPE program, but they can also be viewed as the failure of an asymptotic expansion of sorts, where corrections of unknown size arise from neglected terms.

Reducing the errors associated with each entry in this list remains the basic theoretical mandate for phenomenological study of $b \to u\,\ell\,\nu$ and progress on $|V_{ub}|$. In a recent presentation, Ligeti [73] outlines a possible program for a response. Among his suggestions are:

- Evaluate the full $\mathcal{O}(\alpha_s^2)$ perturbative corrections to all relevant kinematic distributions (they are currently only known for the total rate and q^2 spectrum).
- Pursue a precision determination of m_b. All rate calculations include a factor of m_b^5, and sensitivity to m_b increases in the presence of cuts.

- Improve the measurement of the photon energy spectrum in $b \to s\gamma$, and employ the spectrum directly in E_ℓ and M_X^2 analyses, rather than using it to constrain some intermediate parameterization of the shape function.

- Bring the experimental cuts as close to charm threshold as feasible to include as much $b \to u\,\ell\,\nu$ phase space as possible.

- Constrain the contribution from weak annihilation.

The first few items are specific challenges on the theory front; the last ones are directives for experimental work. Most of these tasks, however, require input and expertise from both arenas, continuing the long tradition of the intertwining of theory and experiment in $b \to u\,\ell\,\nu$.

New analyses at the B factories are implementing advanced experimental techniques to: reduce the contamination from charm backgrounds (for instance, by fully reconstructing the other B decay in the event); improve the efficiency and purity of identifying $B \to X_u\,\ell\,\nu$; and reduce the systematic errors on measurements of both $b \to u\,\ell\,\nu$ and $b \to s\gamma$. A measurement of $|V_{ub}|$ with experimental errors comparable to the tantalizing promise of a 10% theory error is not far away.

Estimates of the remaining unknown theory errors have already been made. In a recent review [2], Gibbons has attempted to use existing experimental data to derive constraints on the theoretical uncertainties that are not yet fully quantified. By identifying how the unknown corrections to the partial rate vary across phase space, he is able to constrain, for the first time, the uncertainties associated with the three areas discussed in detail above: local quark-hadron duality, weak annihilation, and sub-leading corrections to the shape function. The assembled experimental data (2003) include lepton endpoint [128–130], M_X^2-only [131, 132], and q^2-M_X^2 [133] analyses at various of the B experiments—CLEO, BaBar, and Belle. His efforts result in the combined inclusive result:

$$\begin{aligned} |V_{ub}| = (4.63 &\pm 0.28_{\text{stat}} \pm 0.39_{\text{syst}} \pm 0.48_{f_{qM}} \pm 0.32_{\Gamma_{\text{thy}}} \\ &\pm 0.11_{\text{QHD}} \pm 0.27_{\text{WA}} \pm 0.31_{\text{SSF}}) \times 10^{-3} \end{aligned} \tag{3.73}$$

for a total theory error of 15% and an overall precision of 18%. (The Γ_{thy} term is the contribution to the error from uncertainty on the semileptonic width; f_{qM}, the error in the fraction of the rate above Belle's combined q^2-M_X^2 cut; QHD, from violations of local quark-hadron duality; WA, contributions from weak annihilation; and SSF, uncertainty in sub-leading corrections to the shape function.)

The relevance of the work described in this thesis should now be clear. By analyzing semileptonic B decays at CLEO, we attempt to independently bound the contribution of weak annihilation to the analysis of $b \to u\,\ell\,\nu$, and so address directly one of the critical tasks required for a precision measurement of $|V_{ub}|$. It amounts to a refinement of the same kind of data-driven bound Gibbons was able to achieve by reviewing existing analyses, but by performing the search for weak annihilation directly in data, the constraints are both more informative and more powerful.

In summary, by constraining the contribution of weak annihilation effects to traditional $b \to u\,\ell\,\nu$ decays, we help whittle away at one of many theoretical uncertainties that still limit our knowledge of the CKM element $|V_{ub}|$.

Chapter 4
Experimental Apparatus

The data analyzed in this thesis were collected over the course of a decade, using advanced electronic instruments to record hundreds of millions of collisions between electrons and positrons. In this collaborative data-taking venture, the accelerator's role is the creation, storage, and control of the particle beams, and the detector's function is to analyze the results when those beams collide, recording the events that meet certain standards of interest. In a properly orchestrated collision, the electron and positron annihilate into energy that re-materializes as a much less common form of matter: a $b\bar{b}$ quark pair. Studying the subsequent decay of these particles has been the chief focus of particle physics research at CLEO, a specialized laboratory located on the campus of Cornell University in Ithaca, NY.

The details of every collision are recorded by thousands of sensing elements in the CLEO detector, each optimized to record some aspect of the event, such as the path or speed of a charged particle, or the total energy carried by a neutral one. Reviewing the digitized clues left in the detector in each event, we try to identify what particles were initially present and reconstruct how and where each ultimately expired. These deductions require the methodical analysis of huge volumes of detector information using sophisticated software tools. Critical to interpreting the real data is the analysis of simulated data, which can teach us in detail how the detector would respond if a certain decay were to occur. Since we can only infer what happened in an event based on the aftermath left in the detector's sensors, knowledge gained from this kind of simulation is crucial to the research program.

4.1 Birth, Life, and Death of the B Meson

Precision B physics requires a plentiful source of B mesons and conditions well-suited for studying their decay products. Although B's are not ordinarily found in terrestrial environments, they can be produced copiously with modern particle accelerators under controlled conditions. The B decays studied in this thesis were obtained from the annihilation of sufficiently high-energy electrons and positrons in such a setting.

In general, the result of colliding a high-energy electron (e^-) with an equally high-energy positron (e^+) depends on the energy of the collision. At low, non-relativistic energies, the electric attraction between the two opposite charges will ensure that they meet each other and annihilate, but the result (a photon) will have insufficient energy to do anything more interesting than perhaps reconstitute the e^+e^- pair in a kind of mirror process called pair creation: Just as the photon is created from the annihilation of the pair of fermions, it can also decay back into a fermion pair. The key is that the final fermion pair need not be the same kind that was present initially—any type of fermion is possible, provided there is enough energy available. As Fig 4.1 illustrates, the entire reaction is considered to be mediated by a short-lived or *virtual* photon in this case. As the collision energy is increased, the energy of the virtual photon increases, and when it exceeds the threshold $2m_f$ for creating a fermion-antifermion pair each with mass m_f, the $f\bar{f}$ production channel becomes available. When the initial electron and positron are extremely ultra-relativistic ($E \gg m_e$), the annihilation can result in the transformation of the incident e^+e^- pair into many other kinds of particles, such as the other leptons $\mu^+\mu^-$ or $\tau^+\tau^-$ or even a quark pair $q\bar{q}$. The latter is precisely the production channel employed at Cornell to manufacture b quarks: The center-of-mass collision energy is carefully tuned to $E_{\rm cm} = 10.58$ GeV, where the production rate of $b\bar{b}$ quark pairs is particularly enhanced.

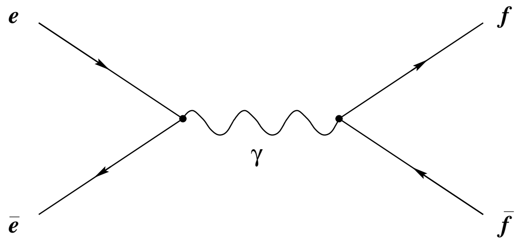


Figure 4.1: Electron-positron annihilation, resulting in the production of a new fermion pair $f\bar{f}$, where $f = \ell, q, ...$ could be a lepton, a quark, or something entirely new, as long as it is consistent with overall mass-energy conservation. As the center-of-mass collision energy E increases, more final states become available. (In this figure and similar ones, time runs from left to right, and the actual interaction is considered to occur at the point-like vertex where the particle paths intersect.)

Under the conditions at CLEO, the possible results of an interaction—if it actually occurs—are roughly distributed as follows [22]. (The listed cross sections are only approximate, but do reflect detector acceptance effects.)

- Bhabha scattering: $\sigma \sim 40$ nb
  The final state is just another electron and positron, the result of a purely elastic collision that may have occurred through actual annihilation or just an (t-channel) exchange of a virtual photon. Events of this type are so numerous that they must be identified at run-time and discarded in order to save space for the much rarer processes below.[1]

[1]Rejection of a fraction of the events from dominant processes like bhabha scattering is

- $\mu^+\mu^-$ or "mu-pair": $\sigma \approx 1.2$ nb
  Muons are relatively long-lived and have a relatively low probability of interacting with matter, so escape the experiment with little further notice. These events are more useful for calibration than as a backdrop for new physics.

- $\tau^+\tau^-$: $\sigma \approx 0.9$ nb
  The τ lepton is heavy enough to have access to a wide range of interesting decay channels with broad physics implications. The study of such τ physics is outside the scope of the present work.

- $q\bar{q}$, with $q = u, d, s, c$ "light": $\sigma \approx 3.4$ nb
  Because these quarks are light in mass, they emerge with large momenta; their subsequent hadronization leads to a characteristic "jettiness" in the shape of the event.

- $b\bar{b}$: $\sigma \approx 1.05$ nb
  Although by no means the most frequent outcome, this process is the the one central to the B physics program.

Note there is no determinism about which decay channel the virtual photon will follow.

The function of the accelerator is to provide the collision conditions, and part of the job of the detector is to identify which of the possible outcomes above was actually realized. Various features of the different processes allow for some discrimination between them, but no separation criteria is perfectly efficient with no rate for mistakes. For instance, sifting out the b quark-containing events from the larger hadronic sample of τ and light quark events requires analyzing many features of the event, and still the competition from the lighter quarks is fierce. (Successful methods for eliminating this source of contamination from the data sample are discussed in Sec 5.5.)

4.1.1 From $b\bar{b}$ to $B\overline{B}$

In the simple quark model, the B meson is a bound-state between a b quark and another light quark, with the following constitutive definitions:

$$B^0 \equiv (\bar{b}d) \qquad \overline{B}^0 \equiv (b\bar{d})$$

$$B^+ \equiv (\bar{b}u) \qquad B^- \equiv (b\bar{u})$$

called pre-scaling. At CLEO, endcap bhabhas in particular are pre-scaled. (The bhabha cross-section grows at low angles, and so these events dominate the bhabha sample so much that many of them are dropped in favor of barrel bhabhas.)

The hadronization of the $b\bar{b}$ quark pair at the $\Upsilon(4S)$ resonance amounts to the creation of an additional $u\bar{u}$ or $d\bar{d}$ pair from the energy remaining after the creation of the b quark. The pair then "divide," one light quark combining with each heavy b quark to make a meson.

The QCD property of confinement implies that a b quark will never be found "bare" in nature—it will always be cloaked by a shroud of other quarks and gluons in some form. In the circumstances described here ($e^+e^- \to q\bar{q}$), the consequence is that the quarks cannot exit the interaction as individual, isolated particles: they must somehow become bound up into hadrons to meet the rules of the strong interaction. If the quarks produced in the collision are moving relativistically with respect to each other ($E_{\rm cm} \gg m_q$), they won't have the opportunity to take common shelter in some sort of bound-state; instead they "hadronize" as they fly away from the interaction point, "popping" additional quark-antiquark pairs from the vacuum to build up a shroud of strongly-interacting companion particles to satisfy the demands of confinement, as shown in Fig 4.2. But if the collision energy is chosen appropriately so that all of the available center-of-mass energy is exhausted simply in creating the the mass of the new quarks, they will have essentially no kinetic energy and may then actually bind to each other. This "resonance" phenomenon leads to a (sometimes slight) enhancement of the cross section. A series of these peaks for the bound states of the $b\bar{b}$ system can be seen in Fig 4.3 as a function of the center-of-mass energy of the e^+e^- collision.

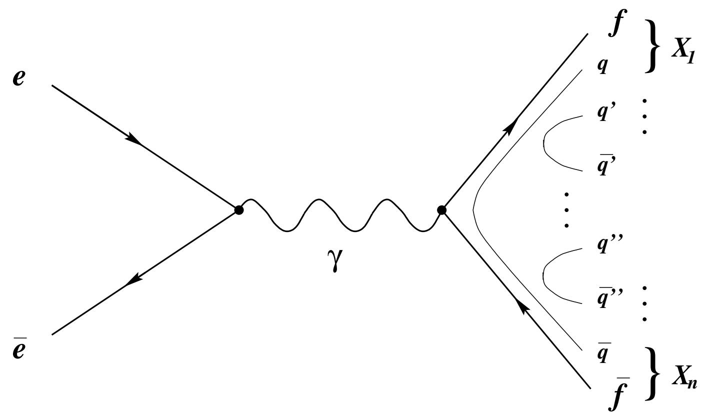


Figure 4.2: Production and subsequent hadronization of a pair of quarks $q\bar{q}$ from e^+e^- annihilation. As the quarks separate, it becomes energetically more favorable to produce additional pairs of light quarks $q'\bar{q}'$, etc. to bind with the original pair than to increase the separation between the bare quarks, a basic consequence of QCD confinement. The assortment of quarks in the final state eventually settles down into one or more hadrons $X_1 \cdots X_n$.

The three lighter resonances $\Upsilon(1S)$–$\Upsilon(3S)$ in Fig 4.3 all decay via the strong interaction in such a way that the b quarks quickly disappear. But the mass of the

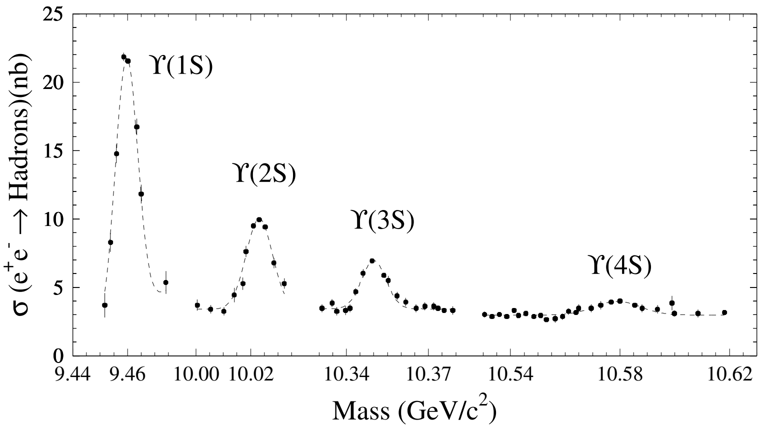
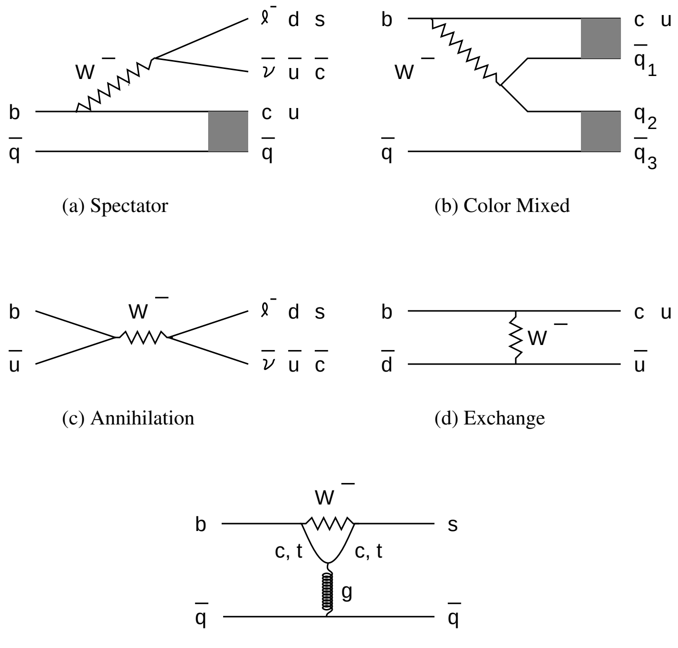


Figure 4.3: The Υ series of $b\bar{b}$ resonances lying near 10 GeV. The highest $\Upsilon(4S)$ resonance has sufficient mass to decay freely into a pair of B mesons.

fourth resonance, $\Upsilon(4S)$, is just above the threshold for decay into a pair of real B mesons. Sufficient energy is available to allow each b quark to hadronize into a (slow-moving) B in which the b quark still lives on (if but for a short time). The peak of the $\Upsilon(4S)$ resonance lies at 10.58 GeV, and with $M_B = 5.28$ GeV, only about 300 MeV is left as energy of motion to be shared between the mesons. In the center-of-mass frame, then, the B's are moving very slowly ($\gamma - 1 \sim 0.001$), and in the 1.5 ps before they decay, they will only travel $\mathcal{O}(10\ \mu\mathrm{m})$. These facts influence the design of the detector used to observe the decay products.[2]

As shown in Fig 4.4, the dominant channels for weak decays of the B can be categorized into five broad categories, depending on the diagram's topology and the location and connection of the weak interaction vertex. The decays analyzed in this work are of the "spectator" or external-W type, shown schematically in Fig 4.4 (a).

[2]We relegate the B factories to a footnote in this discussion, noting merely that by colliding electrons and positrons in a center-of-mass frame that is moving (non-relativistically) with respect to the lab, these experiments are able to stretch the B decay distance to an extent that it is actually detectable with modern equipment. (Note that the increased flight distance is not a relativistic effect, but results simply from the Newtonian motion $\beta\gamma \approx 0.6$ of the B's in the lab.) This technique opens up a whole new vista of possible B physics measurements, many of which are inaccessible to CLEO. Sniff, sniff.



(a) Spectator (b) Color Mixed

(c) Annihilation (d) Exchange

(e) Penguin

Figure 4.4: Various channels for weak decay of the B meson.

4.2 Accelerator

From the particle physicist's point of view, the role of the particle accelerator is to prepare and then deliver the electrons and positrons to the interaction region, where they are efficiently collided to produce as many physics events as possible.

Fig 4.5 shows the basic components of the CESR accelerator facility. Electrons originate in an electron gun, where they are "boiled" off a cathode heated by the passage of electrical current. After production, the electrons are accelerated to about 150 MeV by a small linear accelerator or linac, shown just below the center of the figure. Inside the linac, carefully synchronized microwave EM fields propel the electrons to relativistic speeds and also group the charge into "bunches," transforming the continuous stream of charge into a chain of discrete clumps of charge. In addition to supplying raw accelerating power, the linac also weeds out the particles that are are too fast or too slow relative to the bunch, herding the electrons into a small region of momentum space. That concerted acceleration and damping process continues when the bunches are inserted into the larger synchrotron ring, where the bunches are brought up to the final energy of $\sim$5 GeV. Positrons are produced by introducing a tungsten target into the path of the electron beam near the end of the linac; the spray of particles out the backside includes positrons that can be isolated, bunched, and accelerated in a manner very similar to that for electrons.

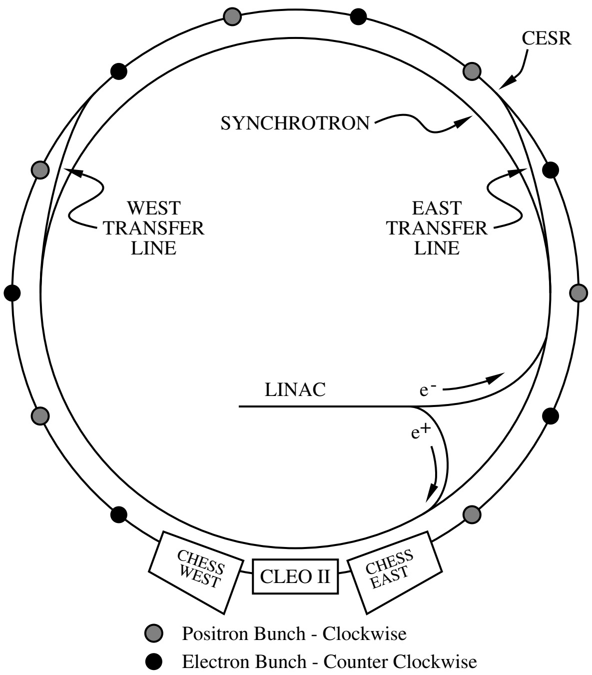


Figure 4.5: Schematic of the accelerator facility at Cornell. Electrons and positrons are separately accelerated through the linac and synchrotron, but stored simultaneously in the storage ring CESR as counter-rotating bunches of charge. The acceleration process takes the particles from essentially being at rest to 5 GeV, where the relativistic boost is $\gamma \approx 10^4$.

As the momentum of the electrons in the synchrotron increases, the magnetic fields that steer the bunches around the ring must be ramped up in a synchronous fashion to keep them traveling along the same orbit inside the machine. (Recall that charged particles will describe circular orbits in a fixed magnetic field, with radius of curvature proportional to their momentum.)

There are four accelerating sections around the mile-and-half circumference of the synchrotron; each of these must be mechanically (but automatically) reconfigured to instead accelerate positrons, since by design, electrons are accelerated in a counter-clockwise fashion, positrons, clockwise. This trick—reversing the direction of travel with the change in charge—is what allows a single ring of fixed magnets to be used to accelerate and steer two kinds of particles.

Each bunch spends less than 10 ms in the synchrotron, making 40,000 circuits in the journey from 150 MeV up to 5 GeV. At that time, the charge is smoothly transferred into the storage ring CESR (short for Cornell Electron Storage Ring). This device was added to the pre-existing synchrotron facility at the end of the 1970's as a staging area, a place to put the bunches into a stable holding pattern while even more charge is accumulated. The injection process that funnels the charge from the synchrotron into the storage ring is complex but extremely efficient. Once all of the bunches have been emptied out of the synchrotron and packed into the storage ring, the synchrotron is ready for use once more, to add more electrons to those already stored, or to begin the acceleration cycle for positrons. The injection sequence is repeated at a rate of 60 Hz to accumulate beam currents of 200 to 400 mA.

After injection of both beams into the storage ring is complete, the beams are brought into collision. The stored beams suffer particle losses due to collisions with residual gas in the vacuum chamber or with the walls of the chamber, causing a gradual decrease in the collision rate in the experimental area. After a 60–90 minute data-taking run, the machine is refilled, and data-taking resumes about 10 min later. The length of the data runs is chosen to optimize the total number of delivered collisions, given beam lifetimes and refilling times.

Steering oppositely-charged bunches around the storage ring is a non-trivial task. At turn-on in 1979, CESR had only one bunch per beam and there were two collision areas, CLEO in the south and an experiment called CUSB in the north. The two bunches in the machine naturally collided twice per revolution, once in each experimental hall. Over time, as part of a continual program to improve machine luminosity, the RF accelerating structures of the machines were modified to accommodate more complex bunch arrangements, including multiple bunches per beam. Eventually the CUSB experiment was dismantled as well. Both of these changes introduced so-called parasitic crossings between the bunches, collision points elsewhere on the ring that are both unnecessary since any charge lost to collision not inside CLEO is essentially wasted, and problematic, as these crossings provide additional opportunities for one beam to disturb the other. To eliminate these parasitic crossings, small displacements are introduced in the e^+ and e^- orbits at various points around the ring, effectively steering the bunches out of each other's way. The result is curious "pretzel-shaped" orbit, whose evolution near the interaction region is shown in Fig 4.6. The perturbations are introduced and controlled with a combination of different accelerator elements, including electrostatic separators, where a stationary electric field transverse to the beam axis deflects the positrons and electrons in opposite directions, and thus prevents a potential collision.

As the bunches coast around the storage ring, they radiate some of their energy away in a stochastic process called synchrotron radiation, associated with the accelerated motion of any charged particle. In CESR, a typical electron loses about 1 MeV per turn, meaning that without any remedial action, a bunch would last only a few thousand turns before it would dissipate all of its energy.[3] Since

[3]Long before this eventuality, the static dipole field of the storage ring would no longer steer the particles down the center of the ring; as the particle slow, they would be bent "too hard" and collide with the inner wall of the vacuum chamber.

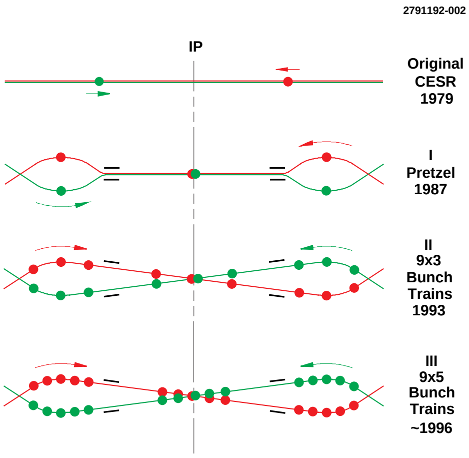


Figure 4.6: Evolution of the electron and positron orbits in CESR over time. Without special insertions (kickers, separators, and other devices) the positron and electron bunches would cross in many places around the ring. The "pretzel" orbit helps minimizes these potential intersections by taking advantage of the gaps in the bunched structure of the beam. By packing more particles into a bunch, or more bunches into a train, the delivered luminosity can be increased.

the charged bunches travel at essentially the speed of light, this means the beam lifetime would only be about 50 ms, not offering much of a collision opportunity, and requiring continuous operation of the synchrotron just to keep charge in the machine! Instead, since only a small amount of energy needs to be restored on each circuit, special radio frequency accelerating cavities are inserted into the storage ring to nudge up the energy of slower particles and slightly retard any faster particles, keeping the bunch as coherent and compact as possible. With these advanced accelerating devices in place, the beam can last for hours in the machine, circling around the ring billions of times. This repetition means that accelerator optics must be incredibly uniform—even a small orbit error can have devastating effect when encountered a billion times. The damping offered by the emission of synchrotron radiation serves as an effective counter to this resonant effect, since within a few thousand turns, the particles have lost all "memory" of earlier machine conditions.[4]

For more discussion of the design and operation of CESR, consult Ref [81] and

[4]The invaluable damping effect of synchrotron radiation reduces as the fourth power of beam energy. At the new CLEO-c energies of 1.8 GeV per beam, the damping is down by a factor of almost 100, vastly increasing the sensitivity to field and other orbit errors. The new CLEO-c wiggler magnets are designed solely for the purpose of increasing the amount of synchrotron radiation lost per turn to shorten the beam's "memory" back to the levels enjoyed at the $\Upsilon(4S)$.

the sources therein.

The performance of the accelerator over the last 25 years has been spectacular. Continual evolution in machine configuration and optics design has maintained a steady, near-exponential growth in the annually-accumulated luminosity, as seen in Fig 4.7.

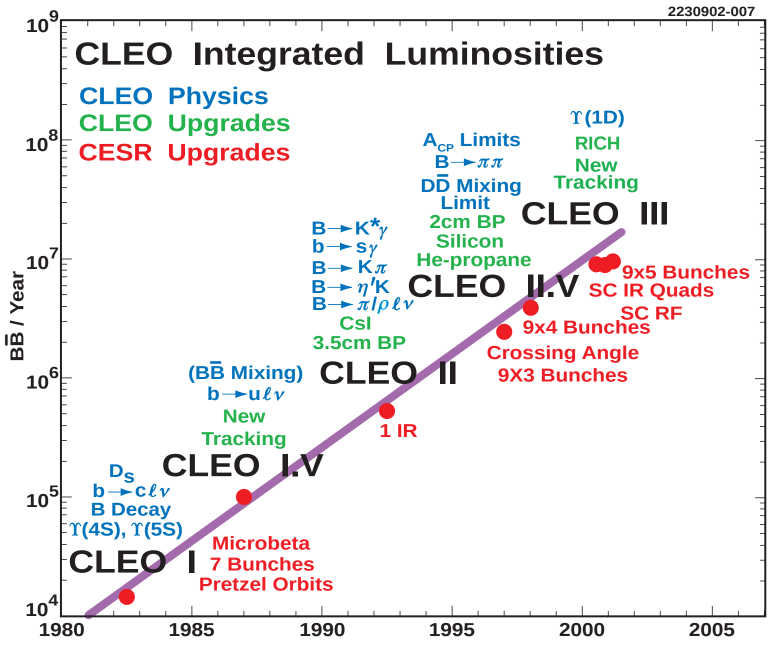


Figure 4.7: Annual integrated luminosity recorded by CLEO over the past two decades, with various advances in machine design and detector configuration marked. The data analyzed in this thesis were collected between 1990 and 1999 with the CLEO II and II.V versions of the detector.

Luminosity $\mathcal{L}$ is a measure of the collision rate per unit area at the interaction point, measured in units of inverse area per unit time. The rate of some process with cross section σ is readily computed as $dN/dt = \mathcal{L}\,\sigma$. By integrating the instantaneous luminosity record over a period of time, one can discuss the "integrated luminosity," also typically denoted $\mathcal{L}$, that when multiplied by a cross-section, gives the expected total number of events for that process in the data sample.

Some of the operating parameters and other dimensions of the accelerator are summarized in Table 4.1.

Table 4.1: Operating parameters and other trivia describing CESR.

Physics start (end) date	1979 (1998)
Maximum beam energy	6 GeV
Circumference	768 m
Typical beam size	460 μm horiz., 4 μm vert.
β^*, amplitude function at IP	1 m horiz., 15 mm vert.
Peak instantaneous luminosity	1.2×10^{33} cm^{-2} s^{-1}
Particles per bunch	1.6×10^{11}
Bunch structure	9 trains, 5 bunches each
Typical beam current	340 mA per species
Crossing angle	±2 mrad
Beam circulation frequency	396 kHz (2.5 μs per turn)
RF frequency	500 MHz
Number of RF buckets	180
Time between collisions	14 to 220 ns
Average RF power	500 kW
Luminosity lifetime	$\sim$ 3 hrs
Typical vacuum in ring	10^{-8} Torr
Ring magnets	86 dipoles, 105 quadrupoles

4.3 Detector

The CLEO detector surrounds the interaction region, where CESR has arranged the electron-positron collisions to occur. When an actual e^+e^- interaction takes place, the particles created in the collision pass through the beam pipe and into the detector volume, where layers of instrumentation attempt to record enough detail about the decay products to allow later reconstruction of the event. Rather than being optimized for a particular physics goal, CLEO is designed as a general-purpose detector, capable of recording the results of many different kinds of B, τ, and charm decays.[5] To meet these needs, it employs a variety of different detection and identification techniques to obtain as much information as possible for the reconstruction of the underlying physics in each collision event.

It is important to note that all of the primary "physics" of the interaction and initial decay takes place *inside* the accelerator vacuum chamber; it is only the secondary and later decay products that escape the beampipe that are observed in the detector. (For instance, when CESR produces B meson pairs at the $\Upsilon(4S)$ resonance, each B drifts no more than a few tens of microns before it decays.)

[5] My apologies to the fans of two photon physics for not mentioning these processes in the main text. Instead, it earns a footnote.

Reducing the diameter of the beampipe to bring the sensitive elements of the detector as close as possible to the initial decay is limited by the finite size of the interaction region, *i.e.* the volume of space where the beam envelopes intersect. Thus the ability to directly observe the primary decay is influenced by the design of the accelerator, beam backgrounds, and other machine dynamics.

4.3.1 Basic Strategy

In particle physics, a particle's identity rests fundamentally on the full specification of its quantum numbers, since they capture all of its elementary properties. But on a practical basis, particles are identified simply by their mass. This simplification is actually a sound one, since most of the symmetry properties described by the quantum numbers are actually broken to some extent in nature, making a particle's mass or self-energy a (fairly) unique function of its set of quantum numbers.

Thus the name of the game for experimental particle identification (PID) is mass determination. Special relativity connects a particle's (invariant) mass to its energy and momentum in a very simple way:[6]

$$E^2 = p^2 + m^2. \tag{4.1}$$

This relationship is the basis for most experimental methods for particle identification: By measuring some pair of kinematic quantities of a particle, such as (E and p) or (p and v), the particle's mass is implicitly determined, and so, then, its unique identity.

No single PID strategy is effective across the full range of particle momenta seen at CLEO, so different elements of the detector are designed to make complementary measurements to provide as complete PID coverage as possible. Nonetheless, even with partial redundancy at some momenta, unambiguous PID assignment is not generally possible at CLEO or elsewhere. Each PID strategy is also hampered by finite detector resolution. The result is that each method is typically only useful across a small window of particle momentum (or velocity), offering at best only limited discrimination between a few different particle species.

By following clues in the kinematics and identities of the observed final-state particles, it is often possible to "reconstruct" a decay chain, building up short-lived intermediate parent particles from their observed decay products. Once the entire decay chain of the B meson is determined, the physics program can begin. Unfortunately, such unequivocal reconstruction is rarely possible, due in part to

[6]Following standard shorthand, we take $c = 1$ (and have done so earlier in this document). Dimensional analysis of any quantity in this scheme is sufficient to determine what factors of c and $\hbar$ are required to restore more "conventional" units.

limitations of PID and experimental resolution, but predominantly because not all particles produced in the B decay are observed. Further, despite the thousands of bits of information recorded in a single event, there is still insufficient data to fully constrain what happened in the primary decay; many likely scenarios are often consistent with the same set of observations. However, a rational program of deduction and statistical analysis, matched by detailed expectations, can identify patterns that may be evident across large numbers of events, even if not manifest in any particular one. By looking at distributions or correlations even in the face of uncertainty about individual events, it is possible to extract useful information about the unobserved initial-state physics.[7]

4.3.2 Subdetectors

Respecting the azimuthal symmetry about the beam axis, the CLEO detector is a barrel-shaped arrangement of cylindrical subdetectors, each designed to measure some property or quality of the particles that pass through it, such as trajectory, energy, velocity, etc. These measurements are then combined to identify each particle and reconstruct its kinematic properties.

The basic elements of CLEO detector are shown in Fig 4.8. A second view, showing the central portion of the detector as viewed from the end, appears in Fig 4.9. The detector has evolved through several incarnations since its beginnings at the dawn of the CESR era. The version depicted in Fig 4.8 is known as CLEO II.V, commissioned in 1996 but largely unchanged from the CLEO II version constructed in 1989. Details of the subsequent upgrades to CLEO II.V and CLEO III are described in Sec 4.4. Until then, discussion will focus on the CLEO II configuration.

The cylindrical "barrel" portion of the detector is capped by two fully-instrumented "endcap" assemblies that complete the nearly 4π coverage.

[7]A pool table scenario can be manufactured to portray the experimental challenge. Supposed that the initial break in a game of pool is concealed with a cloak, and then a random number of billiard balls are removed from the table. Just before the remaining balls stop rolling, the obscuring cloak is lifted and you get to film the last few seconds of action. The problem is now to pore over the film and determine exactly what transpired, starting from when the cue ball hit the first billiard ball, down to the details of which ball contacted which, and when. A unique solution is usually impossible. But suppose that now you get to observe this relatively dull one-shot game millions of times. Patterns might emerge. For instance, you might observe a tendency for the balls to pile up on one side of the table, suggesting that the person striking the cue ball has a slight tendency to favor one side or the other when slamming the cue stick. Such a statistical analysis of the patterns that might arise is exactly the game of experimental particle physics. (Inelastic collisions, where the billiard balls are allowed to break up, are not present in this analogy, but certainly further complicate the particle physics endeavor.)

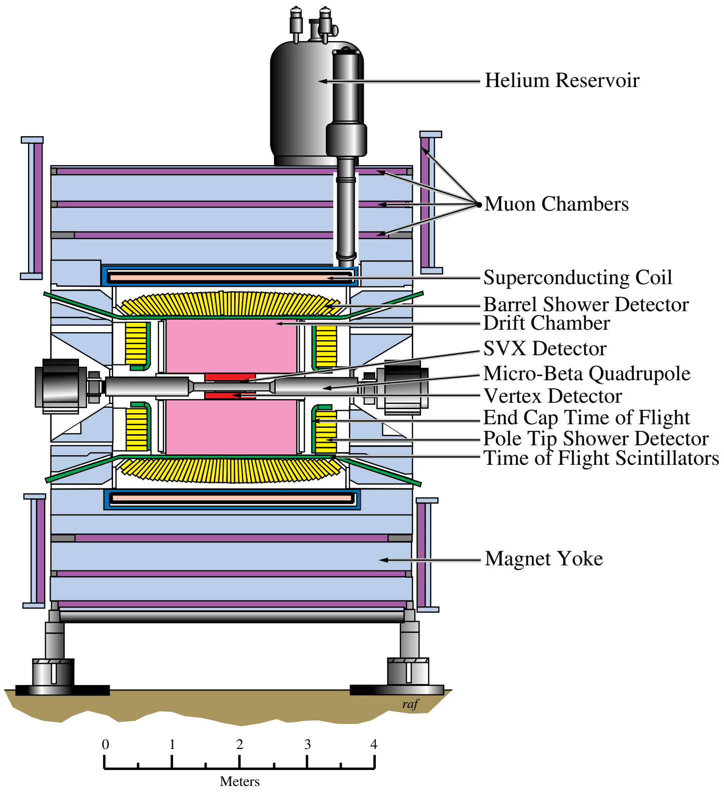


Figure 4.8: Side view of the CLEO II.V detector.

The customary coordinate system used to describe the detector and the physics that occurs within is defined by the beam axis and the plane of the accelerator ring. The $+z$ direction heads west, in the direction positrons head as they pass through the detector. The $+x$ unit direction points out of the ring in the horizontal direction, and $+y$ thus points up. The origin of the coordinate system is canonically taken as the geometric center of the detector, roughly coincident with the interaction point (IP). The polar angle θ thus measures the angle off the beam axis, the azimuthal angle ϕ describes orientation around this axis, and the cylindrical (often "radial") coordinate r is taken to be the perpendicular distance to this axis.

The detector has been documented in detail elsewhere [80]. Briefly, a series of tracking chambers form the innermost portion of the detector, providing information on the trajectories of charged particles. The tracking volume is immersed in a strong solenoidal magnetic field, causing the particles to curve in the transverse plane, which allows for momentum determination. Scintillators outside the tracking volume allow precise timing measurements that record the transit time

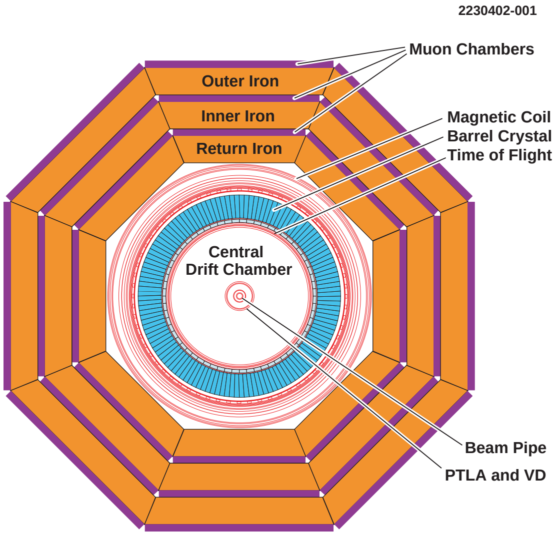


Figure 4.9: An end view of the CLEO II detector.

for particles arriving from the origin. Beyond that system is an electromagnetic calorimeter which records the energy deposited by particles traversing crystals of cesium iodide; the excellent granularity also allows for determining the position of showers from otherwise unobserved neutrals such as photons. The superconducting coil responsible for the solenoidal field lies outside the calorimeter, and a multi-layer muon tracking system forms the final and outermost detection element, located between layers of the steel flux return for the solenoid. Table 4.2 provides some more information about these different systems.

Each subdetector relies on certain properties of the interactions of relativistic particles passing through matter. The survey of detector systems will highlight the relevant features as appropriate.

Tracking Chambers

Immediately surrounding the interaction region are three nested drift chambers that provide basic tracking information about the charged particles produced in each event. These devices are refinements of the basic multi-wire proportional counter (MWPC), where measurement of the arrival time of the leading ionization electrons in each drift cell allows for a precise determination of the position of the track in the cell. A remaining left-right ambiguity is resolved with the use of a

Table 4.2: Summary jargon and numerology for the barrel subdetectors of CLEO II. Analogous systems for all but the tracking chambers are implemented in the endcaps.

System		**Outer Radius**	**Notes**
Beampipe		3.5 cm	0.5 mm beryllium
Precision Tracker	(PT)	7.5 cm	Straw-tubes; 6 layers, 384 sense wires
Vertex Detector	(VD)	17.5 cm	10 layers, 800 sense wires, read out both ends
Drift Chamber	(DR)	0.95 m	51 layers (11 stereo), 12240 sense wires
Time of Flight	(TF)	1.02 m	Plastic scintillator; 64 segments in ϕ
Crystal Calorimeter	(CC)	1.44 m	7800 CsI crystals
Solenoid Coil		1.55 m	Superconducting; 1.5 T field, uniform to 0.2%
Muon Chambers	(MU)	>1.74 m	Interleaved with iron of magnet flux return

half-cell stagger between layers of cells.

The fundamentals of drift chamber operation depend on the ionization that high-energy charged particles cause as they pass through matter. As a particle traverses the gas volume of a drift chamber, it will ionize molecules of the gas. By collecting the resulting electrons on a finely-spaced, regular array of sense wires, it can be determined in which cells ionization occurred, and hence where the particle traveled. But in fact much more precise position information is possible—the location of the track within each sense cell.

The basic unit of a drift chamber is the drift cell, depicted in Fig 4.10. The so-called field wires at the corners of the cell are held at ground, and the sense wire in the center is held at positive high voltage, setting up an electric field configuration that encourages ionization electrons to drift toward the central wire. As the electrons drift, they continually collide with other gas atoms, losing some energy in each encounter. But under the influence of the applied electric field, the electrons are re-accelerated, recovering some energy from the field. Since the electric field strength increases as $1/r$ close to the wire, there is a natural radius inside which the acceleration due to the field is sufficient to restore the electrons to ionizing capability after every collision, and a so-called "avalanche" occurs.[8] The

[8]The radius at which avalanche begins is basically set by comparing the potential gained at radius r by traversing one mean free path length to the minimum energy

electron gains enough energy between collisions to ionize every atom it encounters, liberating additional ionization electrons each time.

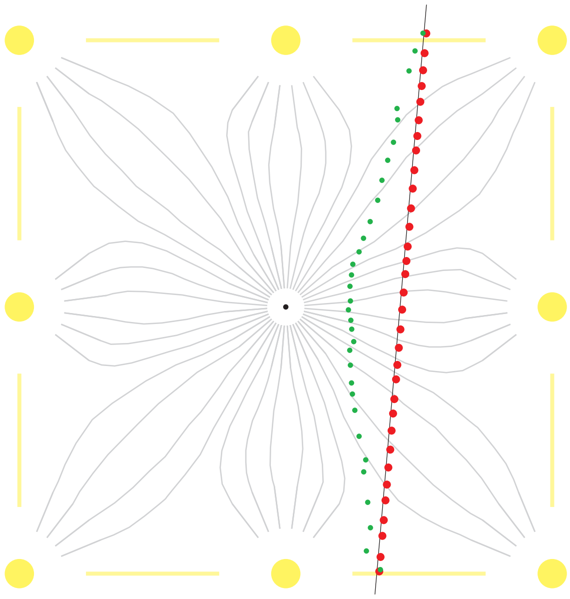


Figure 4.10: A schematic representation of a drift cell, the basic unit of a drift chamber, shown in cross-section. The yellow dots at the corners of the cell represent field wires held at ground; in the center is a sense wire held at $\sim$2 kV. These wires define field lines (in light grey) that will guide a negative test charge toward the center of the cell. A track passes through the right side of the cell, liberating ionization electrons shown in green, already starting to drift toward the sense wire; they leave behind massive, slow-moving ions in red. By timing the arrival of the first drift electrons at the wire, it is possible to deduce the position of the track in the cell. A typical cell has a half-width of several millimeters. (Many thanks to D. Peterson for providing this figure.)

By timing the arrival of the ionization charge at the wire, it is possible to determine how far from the sense wire the primary ionization occurred (given knowledge of the drift velocity for e^- in the gas), and so determine exactly where in the transverse plane the particle went through the cell. (The actual time-to-distance relationship in a real cell generalizes the concept of a single drift velocity, but the idea still carries through.) In addition, the amount of charge collected on the wire is proportional to the initial amount of ionization in the cell.

A practical drift chamber is constructed from many such drift cells assembled in a regular pattern. The arrival time and total amount of ionization charge for each cell is read out by fast electronics attached to each sense wire.

necessary for ionization of the gas in question. If this radius is larger than the sense wire, avalanche can occur.

The axial symmetry of CLEO lends itself to cylindrical chambers, with meter-long field and sense wires strung parallel to the beam axis, secured to large disc-shaped endplates that cap the chamber and seal the gas volume. Drift cell sizes range from half-widths of 5–7 mm, and applied high voltages are around 2 kV, leading to gas gains of 10^4–10^5 and drift velocities on the scale of 50 μm/ns.

The innermost drift chamber in CLEO II is called the precision tracker (PT), and implements this basic design with six layers of 64 wires each, all oriented along the beam axis. Rather than using field wires to define the edges of the drift cell, it employs aluminized Mylar tubes, making it an instance of a "straw tube tracker."

A precision silicon vertexing detector replaced the PT in 1996; this change and others made as part of the CLEO II.V upgrade are described in Sec 4.4.1.

A ten-layer intermediate drift chamber called the vertex detector (VD) lies immediately outside the PT, and provides low momentum coverage for those tracks that curl too sharply in the 1.5 T field to reach more than a few of the innermost layers of the main drift chamber. A diagram of the dimensions and wire layout for the VD and PT appears in Fig 4.11. The field cage for the innermost and outermost layers of the VD is shaped by a metallized cathode sheet that is segmented in the z and ϕ directions. The pitch in z is 5.85 mm on the inner cathodes, and 6.85 mm on the outer, while in the ϕ direction, the cathodes are divided into eight azimuthal sections to simplify association of cathode hits with sense wire hits. By reading out the induced ionization signal on these pads, it can be determined where along the z direction the avalanche occurred in the drift cell. The sense wires are read out from both ends, and by assuming that the charge signal is divided according to the relative distance in z from each endplate to the location of the avalanche, an additional constraint on the z coordinate can be obtained.

The main drift chamber follows particle trajectories from $r = 17.5$ cm out to $r = 95$ cm, providing a long lever arm for measuring the momentum vectors of charged particles at the event vertex. Additional measurement of the z coordinate of a track is provided by inner and outer cathode strips, similar to the VD. But among the 51 layers of wires are also 11 so-called "stereo" layers, where the wires are positioned to have a slight stereo angle relative to the main axis of the chamber and detector.[9] Fig 4.12 illustrates the pattern of wire groupings at one end of the drift chamber, and shows how the stereo layers are interwoven with the traditional axial layers. All together, there are 12240 sense wires in the chamber, each of

[9] The easiest way to picture a stereo angle is to imagine the result on an axial layer of taking one endcap and giving it a tiny twist about the central axis of the detector. Then the ϕ coordinate of each drift cell in this layer will vary as a function of the z coordinate along the length of the chamber. Which stereo cell a track crosses into from an adjacent axial layer will depend on the track's z coordinate. Applying this relationship in reverse, it is possible to determine the z coordinate of a hit by matching a stereo hit to an axial hit in an adjacent layer.

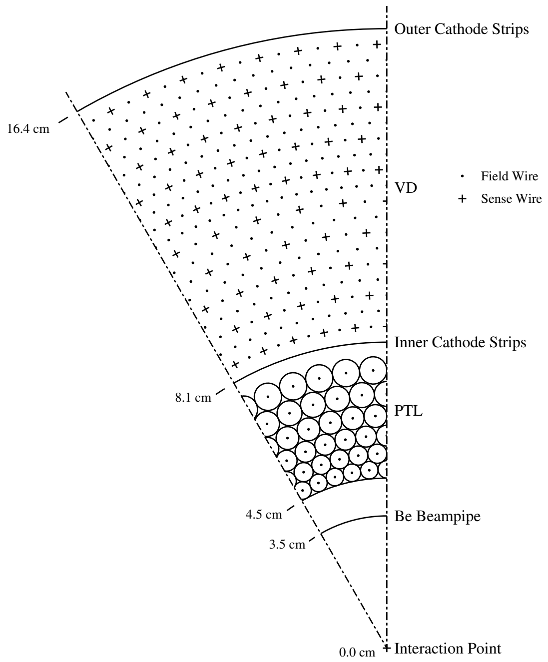


Figure 4.11: Diagram of wire layout in the innermost and intermediate drift chambers, called the precision tracker (PT[L]) and vertex detector (VD), respectively. By identifying the cells that register hits, it is possible to determine particle trajectories and momenta in this transverse plane.

which is read out with a set of dedicated electronics that measures both the time of arrival of the first ionization charge and the total amount of charge deposited in the cell. The timing information is also forwarded to the trigger system (see Sec 4.3.3) for use in ascertaining whether the activity seen in the detector for this beam crossing is indicative of an interesting physics event that should be digitized and recorded.

By examining the charge deposition in each cell through which a track passes, an estimate of the average specific ionization dE/dx can be made. For charged particles traveling through matter, this rate of energy deposition is a function only of the particle's velocity (to reasonable approximation). Coupled with the simultaneous measurement of momentum made in the tracking chambers, this leads to another strategy for particle identification, as illustrated in Fig 4.13: a measurement of the (average) charge deposition along the length of the track and its momentum separates one particle species from another. As the figure shows, however, the separation varies as a function of momentum, with significantly diminishing discrimination above 1.5 GeV.

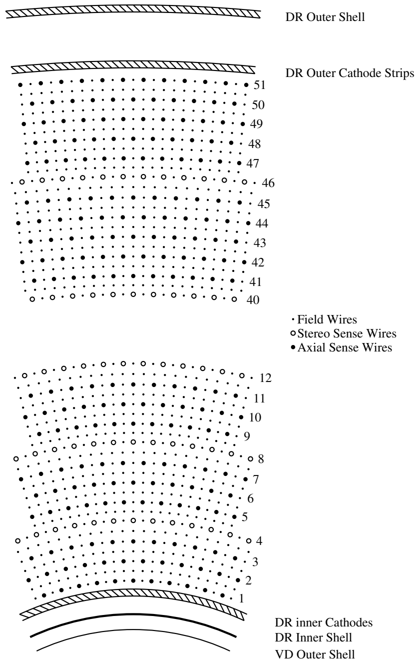


Figure 4.12: Diagram of wire layout in the outer drift chamber, called the DR.

The electronic signals recorded on each sense wire form the raw data for the tracking chambers. Reviewing the information afterwards, specially-designed pattern recognition routines assemble information from individual cell hits to build up a picture of the global trajectory of each (charged) particle through the detector volume, projected inward to the IP to estimate the initial momentum, and projected outward to match up with the responses of other detectors. An end-view of a typical hadronic event with several charged particles is shown in Fig 4.14, and makes it clear how individual hits are combined to identify the path of a particle.

As a particle travels through the different materials of the drift chamber (inner and outer skins, gas, wires), it loses energy due to ionization as just mentioned, but it also undergoes small deflections from its initial flight path due to the process of multiple scattering, the cumulative effect of many elastic collisions with the atomic electrons in the material it is traversing. Both of these effects complicate the subsequent pattern recognition and track-fitting problem, but a general approach called Kalman filtering is employed to properly take the statistical properties of these processes into account.

Additional tracking complications arise at low momentum, when tracks with small transverse momentum can describe (several) complete circles within the tracking volume due to the strong 1.5 T magnetic field in the drift chambers. These tracks are called “curlers” for obvious reasons and can quickly add so many

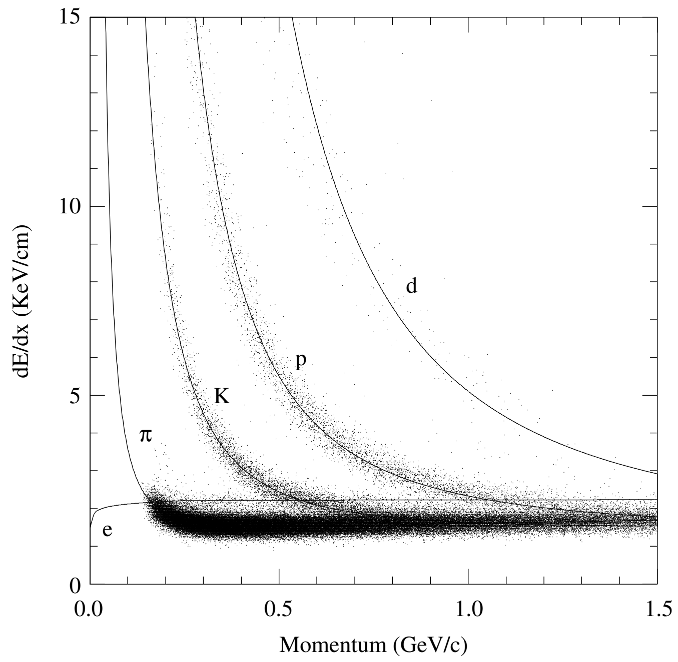


Figure 4.13: Specific ionization (dE/dx) curves for different particle species as a function of momentum.

additional hits to the event that pattern recognition for that track or others can be compromised. Alternatively, tracks may have such large z momentum that they exit the chamber "early," passing through an endplate instead of continuing out to the other barrel detectors.

Nonetheless, the combined performance of the tracking system at CLEO delivers superb momentum resolution over 96% of 4π, and covers momenta from well below 100 MeV up beyond 5 GeV.

Time of Flight

The time of flight (TF or TOF) system in CLEO II is a two-part system, implemented in both the barrel and endcap regions of the detector. The basic sensing elements are plastic scintillators that signal the passage of a high-energy charged particle with a small flash of light. Fast photomultiplier tubes at the ends of the lengths of scintillator allow precise timing of how much time elapsed from the initial e^+e^- crossing at the IP and the arrival of the particle at the known radius of the (barrel) TOF system. Combined with the measurement of the particle's momentum made in the drift chamber, the measurement of this transit time (clearly determined by particle velocity) once again delivers a combination suitable for particle identification. The separation between particle types offered by this strategy is shown in Fig 4.15.

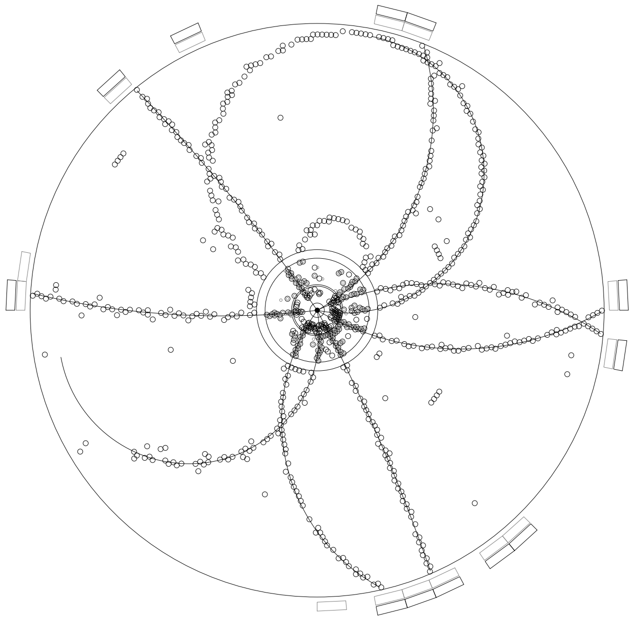


Figure 4.14: A sample hadronic event in the CLEO II drift chambers. In this transverse view, a track is readily identified by a nearly-continuous string of hits along a circular path.

The scintillators are made of a plastic material called Bicron BC-408. When molecules of the base plastic are excited by the ionization deposit from the passage of a high-energy charged particle, they can occasionally de-excite by radiating an ultraviolet photon. This UV photon is shifted to a longer, optical wavelength by the fluorescent action of other organics added to the plastic mix. Their primary purpose is to shift the valuable initial scintillation light to a wavelength where the plastic base is more transparent, significantly increasing the attenuation length in the material. In the barrel TOF system, the scintillator strips are almost 3 m long and are only read out at the ends, so achieving the 2.5 m attenuation length with the fluorescence mechanism is crucial. The strips in both endcap and barrel sections are only 5 cm thick to minimize the amount of material in front of the calorimeter while still presenting sufficient scintillator material to incident particles to guarantee the production of scintillation photons with high efficiency.

Crystal Calorimeter

The detector subsystems described so far are sensitive only to the passage of *charged* particles. The crystal calorimeter (CC) is the first (and only) detector that can register the presence of neutrals such as photons, neutrons, or K_L^0's. In

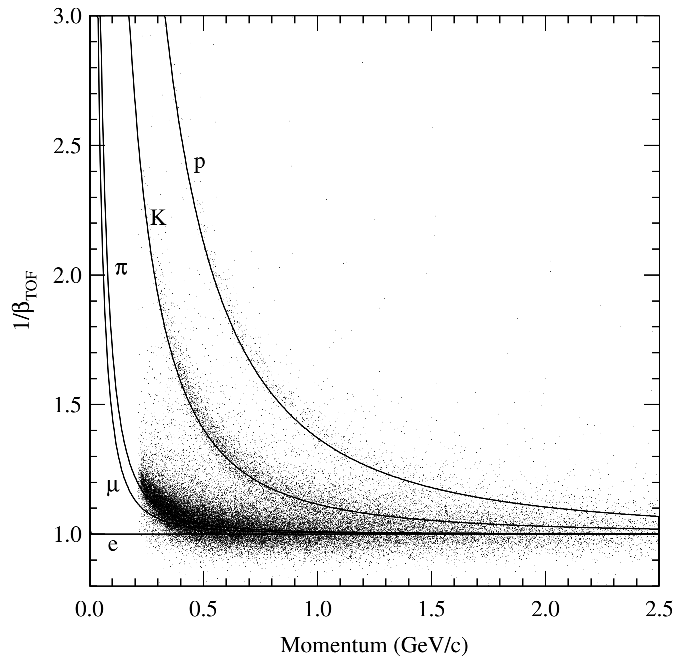


Figure 4.15: PID separation offered by the CLEO II TOF system as a function of momentum. Discrimination between species disappears once particles become sufficiently ultra-relativistic that, within the timing resolution, all species transit the inner detector in the same time.

principle, the function of a calorimeter is to absorb the full energy of any incident particle, and to be useful, it must also provide a mechanism for measuring that absorbed energy.

The CLEO calorimeter is composed of 7800 thallium-doped cesium iodide (CsI) crystals, each roughly shaped like a trapezoidal prism, with a 5×5 cm square face and a length of about 30 cm, as shown in Fig 4.16. Some of the physical characteristics of this dense inorganic scintillator material are listed in Table 4.3. The addition of thallium impurities modifies the levels of the scintillation energy diagram, and results in an increase in the probability and wavelength of the scintillation light.

In the barrel section, which covers out to $|\cos\theta| < 0.71$, the crystals are arranged in 48 azimuthal rings of 128 crystals each, with each crystal tapered to fit snugly next to its neighbor in both the ϕ and z directions. The total mass of CsI in the barrel unit is more than 27,000 kg. The crystals are arranged in a nearly projective geometry, pointing toward the center of the detector and the IP so that photons will shower primarily along the length of a crystal. A slight misalignment prevents the small gaps between crystals from lining up with a direct flight path from the origin, which would otherwise introduce a small inefficiency.

Each endcap calorimeter holds another 828 crystals, machined and stacked as

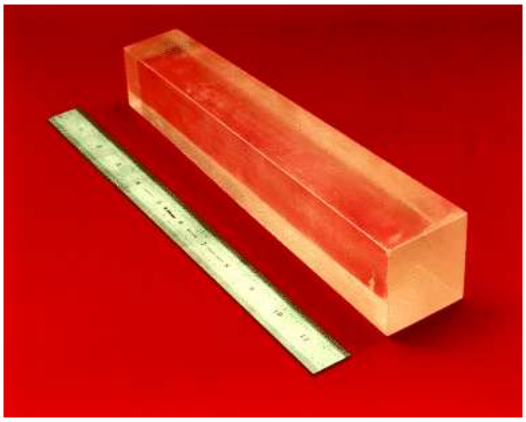


Figure 4.16: Photo of a single CsI crystal, 7800 of which are used in the crystal calorimeter.

Table 4.3: Some physical properties of crystalline CsI(Tl).

Property	**Value**
Radiation length	1.86 cm
Interaction length	37 cm
Moliere radius	3.5 cm
Light yield	50,000 γ/MeV
Emission range	325-700 nm
Emission peak	565 nm
Attenuation length	$\geq$ 1 m
Density	4.5 g/cm^3
Slightly hygroscopic	
Radiation hard	

illustrated in Fig 4.17. Each endcap container is located about 1.25 m from the IP in the $\pm z$ directions. Inside the container, there is a rough azimuthal symmetry.

The crystal size was chosen to balance the issues of position resolution, cost, shower containment, and impact on energy resolution. Smaller size of the transverse face offers finer granularity for determining shower location, but more crystals must then be summed to compute shower energy, and this can degrade energy resolution due to the increased contribution from electronic noise. The number of channels and cost of the readout electronics scales as the square of the number of crystals, providing a strong deterrent to small crystal sizes. In the radial direction, the goal is complete shower containment; as discussed below, the 30 cm length is adequate for electromagnetic showers but falls far short for hadronic showers.

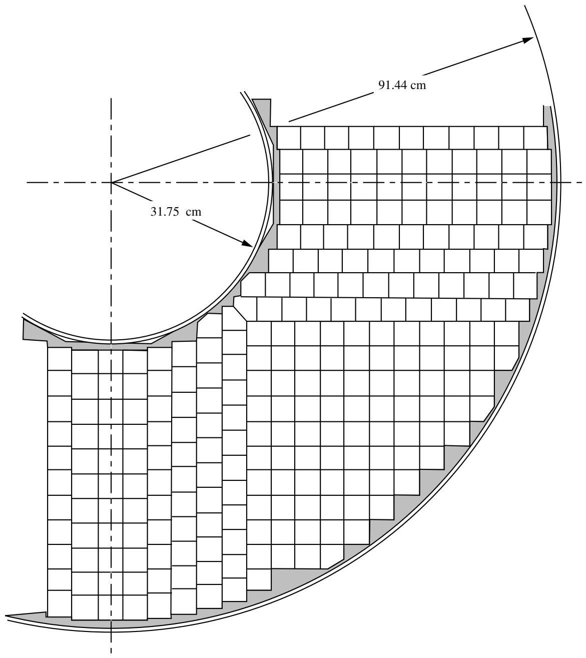


Figure 4.17: A quadrant of the CLEO II endcap calorimeter, showing the dense packing of crystals throughout the detector, limiting light leakage and offering excellent granularity for determining shower positions.

The CLEO calorimeter is designed as an *electromagnetic* calorimeter whose purpose is the full containment and measurement of the energy of electrons and photons rather than hadrons. The distinction is a consequence of the difference between the patterns of energy deposition created by e/γ and hadrons in matter. In a general scintillating substance, the energy of an incident particle is converted to scintillation light, with the quantity of light proportional to the amount of energy deposited. If the full energy of the initial particle is absorbed in the material, then the observed scintillation light is a fair measure of the primary particle's energy. Electrons and photons deposit energy in material via electromagnetic interactions, chiefly through the processes of bremsstrahlung radiation and pair production. In a thick absorber, these interactions can initiate a self-perpetuating electromagnetic cascade or shower, where secondary electrons and photons generated in the initial interaction continue to liberate even more particles of lower energy. The shower ends when the daughter energies finally fall below the threshold for further ionization, known as the critical energy. The characteristic depth scale of the shower is called the radiation length X_0 of the material, and as Table 4.3 shows, the radiation length for CsI is short, $X_0 = 1.86$ cm. The location of maximum energy deposition is only logarithmically dependent on the incident energy, making full containment feasible with reasonably-sized crystals. Linear, well-calibrated electronics are used to convert the light output to electrical signals that are directly proportional to the deposited energy.

In contrast, hadrons (neutral or charged) will lose energy via inelastic scattering and other interactions with the nuclei of the absorbing material. The cross section for these processes is much smaller, meaning much more material is required to degrade the full energy of the incident particle. The shower shape becomes highly variable, and the response of the scintillator may include both "fast" and "slow" components which can hide much of the softer hadronic energy. The nuclear interaction length λ sets the scale for the energy loss of hadrons, and for CsI, this length is $\lambda = 37$ cm, meaning that full containment requires an extraordinary length of crystal.[10]

Calorimeter performance varies considerably with polar angle because energy response is degraded by the location and quantity of material in front of the crystals, and the distribution of material is not uniform throughout the inner detectors. Material close to the crystals only degrades energy resolution through photon conversion or absorption of incident electron energy, but material farther away has a much larger impact because it disperses incident energy over a broad region. Hence, in the barrel region where the intervening material is minimized (slightly less than 20% X_0), the resolution on energy and shower angular position is best. Particles headed for the endcap calorimeters encounter significantly more material, including the 3.18 cm endplate of the main drift chamber, electronics, control and signal cables, and other mechanical supports. In addition, the endcap crystals are stacked horizontally rather than in a projective geometry, and in the outermost rings, they partially overlap the last ring of the barrel, both leading to partial shower containment and unusual patterns of energy leakage and sharing. For these reasons, the barrel and endcap detectors are almost always treated separately, whether for calibration or actual physics analysis.

The light output of each crystal is measured by an arrangement of four photodiodes mounted on the back face of the crystal; the redundancy helps prevent the loss of a crystal if one or even more components fail. The analog photodiode signals are shaped and amplified by preamplifiers also mounted inside the detector volume. Single-edged signals are carried out of the detector on twisted pair cable to analog boards that sum the four photodiode signals for each crystal and then process the signal along two paths, one for a fast input to the trigger system (see Sec 4.3.3), and the other for digitization by Fastbus ADC's. The input to the ADC is integrated with a 1 μs gate and yields a digital value proportional to the total

[10]In practice, many hadron calorimeters are actually implemented as so-called "sampling" calorimeters, where a second dense material with a much shorter interaction length is introduced in between layers of scintillators, forcing even hadrons to lose most of their energy in a short distance. Extrapolating from the measured energy deposition to the energy of the incident hadron is far more difficult than for an EM calorimeter, however, because interactions in the dense absorber are not directly observed.

energy deposited in the crystal.

The clustering of crystal energy measurements to form reconstructed showers is done in two steps. During the readout of the digital data from the detector, crystals hits are only saved if they either exceed a seed threshold or are (next-nearest) neighbors of a crystal which exceeds its seed threshold. This high-level filtering preserves essentially all hits above 5 MeV along with their neighbors, ensuring accurate shower reconstruction, while also reducing the number of extraneous hits saved in the event stream. Reconstruction software run during the data processing stage (see Sec 4.5) applies gain and pedestal corrections to each crystal, identifies clusters of nearby hits, and then sums the crystals assigned to a cluster to get an accurate determination of shower energy and position. The number of crystals included in the sum depends on the shower energy in order to minimize the relative importance of electronic noise in low-energy showers.

Muon Chambers

The key to muon identification is a basic fact about the interaction of muons with matter. Because muons are much more massive than electrons ($m_\mu/m_e \approx 200$), the energy loss through (bremsstrahlung) radiation is enormously reduced at CLEO (and typical cosmic ray) energies, leaving the dominant mechanism for energy loss as simple ionization. As leptons, muons do not interact strongly with atomic nuclei, and so penetrate matter to depths far beyond those at which other charged species, including hadrons, will have long since stopped. A basic muon detector is essentially an ordinary charged particle detector hidden behind a sufficiently thick absorber: any particle that makes it through the absorber is identified as a muon.

The CLEO muon system comprises the outermost layers of the detector, and is located behind the non-trivial amount of material presented by the crystals. In the barrel, the cold mass of the superconducting magnet and its cryostat also intervene. The sensitive elements of the detector are Iarocci tubes, also called "plastic streamer counters," operated in proportional mode. In the barrel, there are three superlayers of counters, embedded in layers of iron absorber at depths of 36, 72, and 108 cm. The innermost layer of iron in fact forms part of the flux return for the magnetic field of the solenoid. In the endcap region, there is only one superlayer of counters, behind an equivalent thickness of iron that varies from about 7.2 to 10 nuclear absorption lengths. The combined system covers 85% of the total solid angle.

A diagram of a counter and the assembly of a superlayer is shown in Fig 4.18. The left-hand figure shows an individual counter, consisting of a comb-like plastic structure that is coated with graphite, acting as the cathode. The anode wires (held at high voltage) run down the center of each cell and are ganged together to create a single counter unit, 5 m long and 8.3 cm wide. Each cell in the counter responds to the ionization caused by a passing muon in the same way as the basic drift cell discussed previously, but the detailed timing and charge information is not used; only the presence of a "hit" in the counter is registered. One coordinate of a hit is obtained by reading out the anode signal, while an orthogonal coordinate along the length of the counter is measured by copper pickup strips that run across the counter. Hence the signals in the muon chambers derive from both "wires" and "strips."

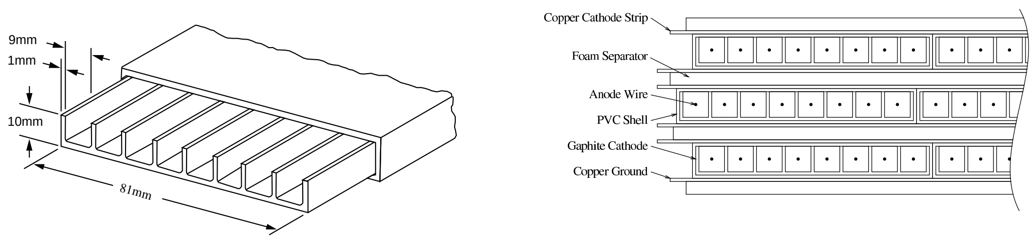


Figure 4.18: Left: Cross section of a single streamer counter in the muon detector. The exposed plastic fins form a cathode on three sides of each anode wire (not shown) running the length of the counter. Orthogonal copper strips (also not shown) are read out along with the anode signal to determine the two-dimensional location of an ionizing muon hit. Right: The assembly of multiple counters into a superlayer, embedded in the iron absorber at some depth.

Three layers of counters are stacked to form a superlayer as shown in the right-hand side of Fig 4.18; it is this package that is embedded at a given depth in the iron absorber. Azimuthally, the barrel is divided into eight octants and three superlayers ("return", "inner", and "outer); each endcap is partitioned into four quadrants, each with only a single superlayer of counters.

Electronic channel count is minimized by further connecting counters together into multiplets of 9–15 units each. The anodes from each counter in the multiplet are chained together serially with 100 Ω resistors, and the charge signal at each end is integrated and read out separately; a similar arrangement is applied to groups of strips. By looking at the pulseheight ratio $P_2/(P_1 + P_2)$ for the multiplet, it is possible to infer which counter in the multiplet actually registered a hit. Thus the small count of 576 channels in this detector disguises a much more sophisticated system of 2352 counters and 5472 cathode strips. The cost is the burden of a more complicated charge division calibration.[11]

[11] In the CLEO III readout system, the ends of the multiplet are viewed as separate electronic channels, so sparsification of these data is a more complex task than for the drift chamber. Rather than applying an independent threshold to each end, the pulse-height ratio (or sum) must be evaluated to determine whether there truly was no signal

By instrumenting at several depths in the iron, it is possible to identify muons with high purity by requiring consistent patterns of hits across multiple chambers. In addition, lower-energy muons will not penetrate to the deepest counters, so to maintain efficiency at lower momentum, the counters at shallower depths can be used. Since the innermost counter is still behind several interaction lengths of material, however, there is a minimum momentum below which muons will not reach even the first layer of counters. For this reason, it is not feasible at CLEO to identify muons below about 1.0 GeV.

4.3.3 Triggering and Readout

Not every e^+e^- crossing yields an interaction, and many of the successful collisions still do not produce events with "interesting" physics. Thus it makes no sense to record the signals registered on all channels of the detector at every beam crossing, which occurs with a frequency of 3 MHz.[12] In addition, it takes some (small) amount of time to digitize the analog signals on every channel in the detector and then transfer the results from the front end electronics to a centralized event store facility. The performance loss incurred by this time delay is known as deadtime, since during this interval the detector is not ready to acquire another event. Lastly, even in the limit of no deadtime and instantaneous readout, the sheer volume of information contained in a single event is enormous, and storage space for this data (most of it useless) would quickly shorten the operation of the experiment to something like a few weeks, ending when all available disks, tapes, and desk drawers were full.

The CLEO trigger system resolves these problems by actively determining whether an event of interest has occurred before signaling the start of full digitization and readout. It thus throttles the data acquisition rate from the potentially fatal $\sim$3 MHz down to something feasible with current hardware, discarding most uninteresting events while maintaining high efficiency for the physics of interest.

The hierarchical system consists of four levels, L0 through L3, that apply successively tighter criteria as more information becomes available during the readout process. Essentially, each level is capable of aborting the full readout sequence at a slightly later stage if its criteria are not met. The first three levels are implemented in hardware, and the last (L3) is applied at the software level to the fully-assembled event.

The first level, L0, uses early inputs from the VD, CC, and TOF systems. Although charge information is not available from the tracking chambers, timing

within the multiplet; only then can the data from both ends be safely discarded.

[12]This crossing rate has increased over time with the addition of more bunches per train. In the early days of CLEO II, it was 2.7 MHz, and at the end of CLEO II.V with 9 trains of 5 bunches each, it was closer to 3.6 MHz.

information is available as soon as a full drift time (a few hundred ns) has elapsed. The subsequent L0 decision is based on the number of loosely-identified tracks and showers observed in the event. Several different sets of criteria (often described as different “trigger lines”) are applied simultaneously, and the L0 result is the logical-OR of all lines. Event rate after the L0 stage is of order 10 kHz.

Whenever any of the L0 triggers lines are satisfied, the detector is disabled and further “interrogation” of available detector information continues. The L1 criteria can include requirements based on information from the DR, which is available at this time. The 1 μs delay required to gather the additional L1 information contributes about 2% deadtime, but reduces the rate to 10–20 Hz. The particular L2 criteria applied in the event of a successful L1 decision depend on which criteria were satisfied at the L1 stage, and can reduce the event rate by another factor of two or so. Finally, L3 criteria are applied once all event fragments are assembled in a central computer; here, although it has already been read out, an event can still be rejected if is determined not to be from an e^+e^- collision (*e.g.* beam-wall or beam-gas events). This suppression results in an additional reduction of 30–40% in the event rate.

If no trigger criteria are met at some stage, the detector and trigger electronics are reset and the detector is re-enabled to acquire additional events. The efficiency for recording $B\overline{B}$ events is better than 99.8%, and is estimated to be 100% for the particular decay chain of interest in this thesis: $B \to X_u \, \ell \, \nu$.

The trigger logic actively manages the rate at which the detector is readout and data is collected, but a separate data acquisition (“DAQ”) system actually coordinates the flow of data from the front-end crates to a central computer. Digitization and sparsification happen after an L2-accept decision is made, and occurs in parallel in the 28 front-end crates housing the digital electronics for all the subdetectors. A controller board in each crate moves the data from the data boards into local memory and then transfers the crate data into one of several link boards in another crate. The link board stores the data and waits for any of four VME-based computers to collect the data. The parallel operation at this stage allows for much higher data rates. After this transfer is complete, the front-end crates are quiet once more and the trigger is re-enabled. Meanwhile, another VME processor known as the EVBD or “EventBuilder” assembles the full event from the retrieved fragments and transfers the completed record to a Unix workstation, where L3 and other online data quality monitoring programs filter and analyze the event before it is written to tape for permanent storage.

In a typical $B\overline{B}$ event, there are perhaps 10 charged tracks and 20 showers, making the overall *occupancy* of most of the detectors rather low: only a small fraction of crystals will register significant energy deposition, and only a small number of drift cells will have seen real ionization charge. Before digitization, it is not possible to reject the information from the “empty” channels since there is no convenient reference for the analog signals. After digitization, however, it is possible to compare the digitized value to an expected "pedestal" or no-signal response, and eliminate the digital information contributed by that channel if it is determined that no signal was present. This process of eliminating channels with no useful signal us known as *sparsification* and greatly reduces the size of the event that must be stored. Although the CLEO II detector has almost 25,000 electronics channels, the average event size after sparsification is only 8 KB. For CLEO II.V, with the addition of the 26,208 channels of the silicon detector to the event stream, the event size is still only 25 KB after sparsification.

4.4 Detector Upgrades

The original CLEO II detector was installed in 1988–89 and represented a substantial improvement over an older detector since designated CLEO I. The only subdetector retained in the upgrade was the VD, first installed into CLEO I during the summer of 1984. Although data-taking with the new CLEO II detector began in October 1989, understanding and commissioning of the new hardware was not complete until mid-1990, at which time the physics program began.

CLEO II collected data on each of the four lowest-lying Υ resonances, accumulating a total integrated luminosity of more than 4.5 fb^{-1} over six full years of running. During that time, the bunch structure of the accelerator increased in sophistication; in particular, a crossing angle of 2 mrad at the IP was introduced in 1994. These improvements and others led to exceptional machine performance and record-breaking integrated luminosities. Some of the physics measurements made in this era can are summarized back in Fig 4.7.

4.4.1 CLEO II.V

In 1995–96 the PT was replaced with a new IR assembly that included a smaller-diameter beampipe and a new silicon vertex detector offering precision position information extremely close to the IP. At the same time, the drift gas in the main drift chamber (DR) was changed from argon-ethane to helium-propane.

The silicon vertex detector (SV) is shown in Fig 4.19. It consists of three concentric cylinders of thin, lightly-doped silicon wafers that register the ionization caused by the passage of charged particles. Each wafer is covered with high-density, micron-scale *p*-*n* junctions operated under reverse bias. The depletion layer in each junction forms the sensitive elements of the detector. Any charge liberated by ionization in this region is automatically swept by the depletion field to the electrodes and collected. Special low-noise electronics gather and amplify these signals from thousands of such junctions spaced with a pitch of 28 (100) μm on the inner (outer) side of the wafers.

The 96 wafers are double-sided: strips on the inner side provide r-ϕ information, and on the outer, r-z. The new gold-plated beryllium beampipe has a radius of 2 cm, bringing the first layer of silicon to within 2.5 cm of the beam. The enhanced vertexing offered by this device allows for the separation of long-lived D and τ decay vertices from the primary event vertex, but this capability is not significant for the analysis described here.

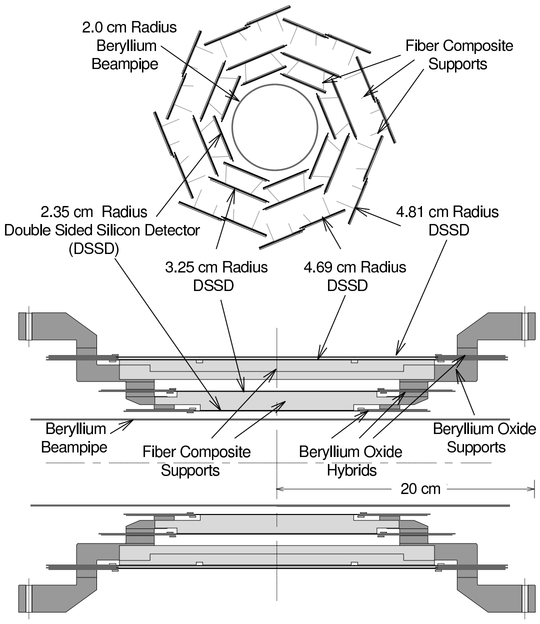


Figure 4.19: The silicon vertex detector in CLEO II.V; it replaced the inner drift chamber known as the PT.

The adoption of helium-propane in the DR increased the average ionization per cell by almost 50%, improving the hit efficiency and spatial resolution of the device, and also demonstrated a more linear drift function. The lower density gas also reduced the contribution of multiple scattering, enhancing the momentum resolution of the detector.

The CLEO II.V configuration collected an impressive 9 fb^{-1} of integrated luminosity from 1996 until early 1998, when the experiment was halted for another extensive upgrade.

4.4.2 CLEO III

The CLEO III detector upgrade was timed to coincide with further enhancements to CESR. An essential ingredient of the simultaneous accelerator upgrade was the introduction of large superconducting final-focusing quadrupoles that intruded very close to the IP. This feature required the re-design and replacement of the innermost CLEO detectors, and also provided an opportunity to upgrade the readout and data acquisition electronics across the entire detector in preparation for expected luminosity increases that could deliver physics at rates of up to 1 kHz. Another key component of the detector upgrade was the introduction of a Ring-Imaging Chêrenokov detector (RICH) to replace the TOF system and deliver well-separated PID for charged particles beyond momenta of 2 GeV for the study of the increasingly interesting but rare 2-body decays of the B meson. The superconducting quads and the inner radial constraint imposed by the RICH also compelled the design of a new main drift chamber and endcap calorimeter sections.

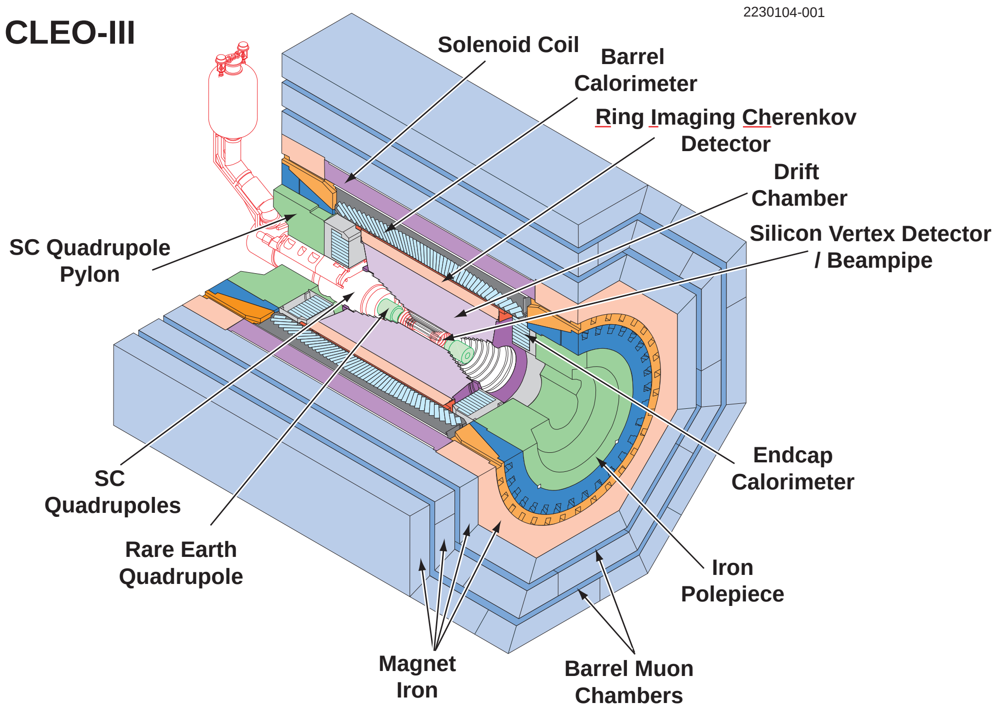


Figure 4.20: A cutaway view of the CLEO III detector.

The 5-year design effort resulted in a state-of-the-art and extremely flexible detector, shown in Fig 4.20. Inside the detector, features include a completely new beampipe assembly, a new four-layer high-precision silicon detector, a 47-layer spectacularly-performing drift chamber with a novel stepped endplate, and the new RICH detector. Outside, the digitization electronics for the calorimeter crystals and muon chambers were upgraded, and new front-end electronics were developed for each of the other detectors. The trigger system was also replaced. Finally, a new data acquisition system designed around the goals of modularity, encapsulation, flexibility, and standardization was implemented, controlling all elements of the detector and the data-taking process in a uniform way.

More information about the design and implementation of the CLEO III detector is available elsewhere; see Ref [82]. It is not described in further detail here as the data collected with this configuration are not used in this study.

4.5 Event Reconstruction

The raw data for an event is nothing more than a list of the hits registered in each detector, stored in a compact, binary format. Only after actual data collection is the low-level detector information translated into higher-level "physics" quantities like tracks, showers, and PID likelihoods. This conversion requires a large set of calibration constants specifying, for instance, the relationship between the raw ADC counts measured for a particular channel in the calorimeter and the energy actually deposited in the associated crystal. The constants are determined from a combination of ongoing electronics and physics-based calibrations, and must be updated regularly to reflect changes in detector hardware, readout thresholds, or other operating conditions.

The actual reconstruction pass through the raw data is carried out with a large software program called `pass2`.[13] In addition to track-finding and crystal-clustering, it also:

- Matches tracks to showers, hits in the muon chambers, and hits in the time of flight counters
- Computes particle identification quantities based on DR and TOF information
- Attempts to identify possible $K_S^0 \to \pi^+\pi^-$ and $\pi^0 \to \gamma\gamma$ decays
- Classifies events according to whether they appear to be from bhabha scattering, μ or τ pair production, two photon processes, cosmic rays, hadron production, or other junk

along with many other tasks. The output from this lengthy analysis sequence is written to permanent storage and made available for general use by all collaborators.

This shared output is the starting point for all subsequent physics analyses, which "skim" this enormous sample for the events that fall into their particular category of interest. Because many analyses investigate the physics of hadronic events containing B's or charm, a standardized skim known as `CCHAD` has been

[13] The name `pass1` is reserved for a smaller program that examines the data online just before it is stored, providing a first peek at the events for data quality monitoring purposes.

prepared. These files contain only the hadronic events from the pass2 output, and discard some of the detailed detector information not ordinarily required in a typical physics analysis. For certain physics investigations, it is necessary to retreat to the unfiltered pass2 output, but rarely is access of the original raw data ever required.[14] The starting point for the analysis described here is the CCHAD skim collection.

The wealth of raw data collected by the collaboration is divided into more manageable chunks called "datasets." Each dataset is collected under conditions during which the detector and accelerator were similarly stable, making all of the events within a sample more "alike" than perhaps between two different datasets. After the reconstruction pass, these differences are assumed to be reduced or eliminated, and in any case, are taken into account by careful modeling with simulated data; see Sec 4.6 below.

Datasets 4s2–4sG were taken in the CLEO II detector configuration, and datasets 4sH–4sT are from the CLEO II.V configuration. The naming scheme correctly indicates that all of these data were collected on or near the $\Upsilon(4S)$ resonance. As will be discussed in more detail in Sec 5.5, data collected slightly below the resonance are useful in understanding the contribution of non-$B\overline{B}$ processes still present underneath the $\Upsilon(4S)$ resonance. Table 4.4 summarizes the $\Upsilon(4S)$ data collected at CLEO. All listed datasets are used. The small 4sH dataset collected at the start of the CLEO II.V era is generally considered to be engineering data, but is still included in this analysis.

4.6 Event Simulation

The experiment as described so far amounts to the sampling of millions of independent, non-deterministic trials in which the primary cause (the underlying physics) is essentially unknown. The notion of a control or reference version of the experiment doesn't seem to have an analog in this framework, since there is no way of controlling the outcomes of e^+e^- collisions or B decays. Inferences and hypotheses can be checked for self-consistency in the data, which is quite a stringent

[14]Practically speaking, as the understanding of detector response improves, better reconstruction algorithms are designed. A new pass2 executable incorporating the revised code can be then re-run over the raw data to take advantage of the improvements in pattern recognition, track-fitting, PID, or whatever the case may be. At CLEO, there have been several "re-pass2's" of the data, but the evolution in CLEO II reconstruction software faded with the success of the CLEO III detector; the data has not been re-processed since about 1999, and it is this latest version that we use here. Note the careful use of language making it unclear whether this version of pass2 output could still be made even "better." As some have said in this regard and more generally, "Better is the enemy of good."

Table 4.4: Summary of CLEO II and II.V datasets. All of the listed datasets are used in this analysis. Luminosity figures are in units of pb^{-1}.

Dataset	Run Range	Collection Dates	Luminosity
CLEO II			
4s2	36233–40095	11 NOV 90–04 JUN 91	672.0
4s3	41386–45969	18 SEP 91–17 FEB 92	680.2
4s4	46035–48057	10 APR 92–26 MAY 92	317.5
4s5	48789–51406	09 JUL 92–05 OCT 92	342.7
4s6	51431–53702	03 NOV 92–19 JAN 93	316.5
4s7	53741–56976	16 MAR 93–06 JUL 93	461.3
4s8	57038–58813	01 AUG 93–27 SEP 93	274.4
4s9	58836–60576	22 NOV 93–10 JAN 94	340.1
4sA	60770–61778	20 JAN 94–28 FEB 94	190.8
4sB	61826–62855	19 MAR 94–16 MAY 94	140.9
4sC	63086–64026	16 JUN 94–15 AUG 94	141.5
4sD	64085–64674	15 SEP 94–09 OCT 94	98.0
4sE	64675–65238	09 OCT 94–01 NOV 94	128.8
4sF	65252–65818	03 NOV 94–28 NOV 94	145.9
4sG	66539–68027	19 JAN 95–09 APR 95	456.6
CLEO II.V			
4sH	78621–80924	16 NOV 95–11 MAR 96	521.0
4sJ	80925–82975	03 APR 96–08 JUL 96	677.0
4sK	82976–84602	19 AUG 96–28 OCT 96	595.0
4sL	84681–86980	17 NOV 96–10 MAR 97	1080.6
4sM	87049–88467	05 APR 97–17 JUN 97	531.8
4sN	88482–90043	19 JUN 97–17 AUG 97	714.3
4sP	90494–92806	15 OCT 97–09 FEB 98	1103.6
4sQ	92908–94096	27 FEB 98–21 APR 98	559.9
4sR	94098–94936	21 APR 98–08 JUN 98	499.1
4sS	94976–96510	27 JUN 98–05 OCT 98	1214.3
4sT	96544–98337	10 NOV 98–15 FEB 98	1583.3

test when the data sample is so large, but individual events are so complex, that it is hard to rule out or falsify a theory based on a single event. Additional scientific progress must rely on the careful and judicious use of simulation.

In particle physics, entire events are simulated using Monte Carlo techniques to implement the probabilistic features of the physical world. Essentially, the goal is not to create a sample of artificial events that are identical to a real sample on a detailed, event-by-event basis, but rather to create a sample that has the same *statistical* properties as the real one. In short, Monte Carlo simulated events are identical to real data *in distribution*. The technical details of how such simulations are implemented are beyond the scope of this discussion, but we note here that the Monte Carlo technique allows for both a physics-level simulation, where an attempt is made to actually simulate the same processes as occur in the real world, or an empirical simulation where some property is generated according to a distribution that has been previously observed in the real world, but there is no insight or commitment to how the observed pattern actually arises.

Given probability distributions and analytic constraints, the standard Monte Carlo method is used to quickly generate a set of random events that are distributed according to the input probability density functions. The limitation of this approach is that the simulated data can only exhibit patterns or relationships inasmuch as those correlations have been parameterized and modeled in the inputs. To adopt a popular aphorism: The output is only as realistic as the input.[15]

Monte Carlo (MC) simulation fills several important needs in experimental particle physics research. It provides:

- An alternative dataset of (nearly) unlimited size that can serve as a testbed for developing and tuning general analysis strategies, without fear of overusing the real data
- A data sample where the "truth" of the matter can be revealed, circumventing the ignorance encountered in the real world that prevents complete determination of what "really happened" in an event. Since the original simulands are stored as part of the simulation, efficiencies, fakes rates, and backgrounds can be studied and explored in great detail. Given an experimental signature, it becomes possible to ask "What (could have) caused that?"
- An exploratory facility for studying the consequences of theories that may not actually hold in the real world. The detailed experimental signatures of a theory such as supersymmetry can be studied, and techniques can be developed to test its validity, even if ultimately it proves "wrong."[16]

[15] So-called "bootstrap" methods and modeling using historical data are two other simulation techniques that do not suffer from these limitations (namely: missing correlations) as they work directly with real-world data. However, other issues trouble these methods, and Monte Carlo modeling is by far the method of choice for particle physics simulation.

[16] Such investigations are practical instances of attempts to evaluate the truth of what are called counterfactual conditionals in philosophy. A counterfactual claim is of the form "If it were the case that X, then it would be the case that Y" and defines a new kind of implication between two propositions X and Y. Classically, if the supposition X does not adhere in the real world, the deductive statement $X \to Y$ is always logically valid,

Of course, the final analysis uses the observations made in the real world, but often relies on simulation for the interpretation of those measurements. In this analysis, we rely on simulation to describe the shape of certain processes, but use the data to measure the relative contribution of each process and to tease out the correlations between them. In an important systematic error assessment stage, we attempt to gauge how sensitive our final result is to the assumptions that went into the Monte Carlo modeling for each process.

Simulated data at CLEO is created in two stages. In the first part, the physics of the e^+e^- interaction and the decay of all short-lived products is simulated according to input branching fractions, form factors, and other previously determined parameters. This step is carried out by a simulation package known generically as QQ which effectively creates physics at the level of momentum and position four-vectors from a basic description of the collision environment (beam energy, energy spread, crossing angle, location of the IP, etc.). The decay modes available for each particle in the simulation must be specified, and only these channels are selected from during the creation of a simulated event. Thus, in order to keep the simulation faithful to the actual world, it is important to correctly tune the input parameters to properly reproduce observed decay distributions. In several cases, such as the decay of the B, the complete list of exclusive decay channels is not known: the known channels account for less than half of the total width of the B. Hence, in the simulation, the remainder of the B width is dedicated to so-called "inclusive hadronization," where quark pairs and other particles are added to create a non-resonant final state that mocks up the contribution of unmeasured channels. Non-$B\overline{B}$ processes are simulated using other standard event generators that have been interfaced to QQ; for instance, light-quark $q\bar{q}$ processes are implemented with the JETSET package [84].

The second part of the simulation inserts these decays into the detector environment and determines both the subsequent flight paths of the particles within the detector and the likely response of the various detector elements the particles encounter. The CERN detector-modeling software known as GEANT [83] forms the heart of this part of the simulation, called `cleog`. A detailed description of the geometry and material of the detector is used to track the virtual trajectories of all particles that live long enough to exit the beampipe. Material interactions such as bremsstrahlung, multiple scattering, nuclear interactions, photon conversions, and ionization are all incorporated. Detector noise is also introduced by merging

and so becomes largely irrelevant; the truth value is tied too much to the validity of the antecedent and ignores the question of the validity of the implication itself. Scientifically, it is of incredible importance to be able to evaluate such statements, since they are the natural language in which most theories are described. By creating simulated worlds where X is true, it is possible to explore the physics implications of as-yet unverified scenarios, and perhaps determine ways of falsifying or confirming the truth of X in the real world.

in real detector data measured in beam crossings independent of trigger criteria. In addition, detector inefficiencies, mis-cablings, and many other resolution effects are included according to a detailed historical record of the state of the detector at any point in time. While the input physics is independent of the state of the detector, the output simulated events are formally rooted in a time, tied to a particular dataset corresponding to the state of the real detector. A thorough and accurate simulation must then cover all states of the detector (or all datasets) in approximately the same ratios as in the actual data.

The final output of the simulation process is thus a complete description of the detector response formatted identically to real detector raw data, but supplemented with a "truth table" describing exactly what happened in each simulated event. By then running the usual reconstruction code over the simulated data, it is possible to determine how various physics processes might appear in the CLEO detector.

The physics and detector simulations have been tuned extensively using the two decades of experience gained by operating successive generations of the detector and measuring a wide variety of physics processes over that time. In general, quantitative comparisons between the simulation and observations made in data show only small differences, at the few percent level or better.

Several kinds of simulated data are used in the analysis described in this thesis. Because of the wide interest in the physics of B decay, the collaboration has generated a so-called production sample of "generic" B decay that is roughly equivalent to five times the number of $B\overline{B}$ pairs expected in the full dataset. In this sample, every event is of the general form $e^+e^- \to \gamma^* \to B\overline{B}$, and the B's are decayed randomly according to a description that combines knowledge of known channels and the inclusive non-resonant states described above. Neutral B mixing is implemented, but the effects of final state radiation are not. This sample is labeled the *generic* $b \to c\,\ell\,\nu$ *sample* in this thesis, since it is used to study the contribution from this background to the analysis. Note that $b \to u\,\ell\,\nu$ channels are not included in the standard simulation as they represent a small contribution about which little is known.

The simulation of the signal $b \to u\,\ell\,\nu$ decays is carried out according to the same prescription as used for the generic Monte Carlo data, but using a modified version of the QQ program called InclGen, described in Appendix A. In this sample, at least one of the B mesons in every event is decayed along $b \to u\,\ell\,\nu$, with the details of the hadronic state chosen according to a combination of inputs from the inclusive spectrum expected from theory and the exclusive resonances that have been experimentally observed. The statistical size of this sample corresponds to the number of $b \to u\,\ell\,\nu$ events expected in a real data set more than 20 times larger than the actual one.

Although little is known about the actual physics of weak annihilation (WA) (recall Sec 3.4.2), a modification of the InclGen algorithm is used to create various possible manifestations of WA in the data, and then the standard detector simulation is applied. The details of this modeling will be discussed in Ch 6. The statistics of these samples is of the order 50–100 times larger than what is expected (naïvely) in the data.

4.7 Summary

This chapter has surveyed the data-taking process in some detail, describing the lengthy and complex sequence by which $B\overline{B}$ events are created, observed, and finally analyzed at the CLEO experiment. The description of that journey involved the elaboration of some basic physics principles—chiefly, the non-trivial effects particles can create when they traverse matter, and the realization and acceptance of some basic facts about the experiment—namely, that our genuine ignorance of precisely what happened in a given event can be combatted with the appropriate use of the tools of simulation and statistics.

The points material to the second half of this thesis are few. To summarize:

1. When electrons and positrons collide, the interaction can result in the production of a pair of B mesons, among many other possibilities.

2. Such collisions are deliberately arranged by CESR to take place in the middle of the CLEO detector.

3. CLEO records thousands of measurements about the particles produced in the aftermath of the e^+e^- collision. Ultimately what is observed are only the remnants after all of the interesting weak physics has taken place.

4. Among the detector elements are tracking chambers that cover more than 95% of the total solid angle, and a crystal calorimeter that is essentially hermetic, covering 98% of 4π.

5. Enormous amounts of precision data have been collected with the well-understood CLEO detector over the course of a decade of successful running.

6. Simulated data aids in the interpretation of the observed events. Monte Carlo simulation at CLEO has benefited from the combined input and experience from hundreds of physicists during more than 20 years of extremely successful and high-precision physics research.

We are now ready to begin describing the analysis itself.

Chapter 5

Inclusive Reconstruction of $B \to X\,\ell\,\nu$

Our starting point is a data sample of almost ten million $B\overline{B}$ events. The inclusive reconstruction of semileptonic B decay begins with the identification of a high-momentum lepton, either an electron or a muon, with momentum $|p_\ell| > 1.5$ GeV. We deduce the properties of the anticipated neutrino under the assumption that the deficit between what is observed and what is expected from energy conservation has been carried away by this single undetected particle. We require the kinematics of the reconstructed neutrino—indeed, the characteristics of the entire event—to be consistent with this assumption. These requirements are slightly biased against the overwhelming background from $b \to c\,\ell\,\nu$ decays, but a significant contamination still remains after all selection criteria have been applied. Other surviving backgrounds include events from non-$B\overline{B}$ processes and events where a hadron has been mistakenly identified as a lepton; both are important ingredients that contribute to the distribution observed in data. The contributions from $b \to c\,\ell\,\nu$, $b \to u\,\ell\,\nu$, and weak annihilation are ultimately disentangled with a fit described in the next chapter.

5.1 Basic Strategy

The goal of this analysis is to derive a data-driven limit on the contribution of weak annihilation (WA) to measurements of the total inclusive $b \to u\,\ell\,\nu$ decay rate, and thus constrain theoretical uncertainties on $|V_{ub}|$. One possible avenue for obtaining such a bound is to compare the total semileptonic decay rate $\Gamma(B^- \to X^0\ell^-\,\overline{\nu})$ for charged B mesons to the corresponding rate $\Gamma(\overline{B^0} \to X^+\ell^-\,\overline{\nu})$ for neutral B's: any difference can be attributed to weak annihilation effects, as discussed in Sec 3.4.2. However, determination of the charge of the B requires reconstruction of the hadronic system X or at least flavor-tagging the other B decay in the event. Large data samples that can afford the low efficiency of these techniques are required. Moreover, this method does not directly assess the impact of WA on a measurement of the much rarer $B \to X_u\,\ell\,\nu$ decay. As we've seen previously, $b \to u$ analyses are generally forced to work in restricted regions of phase space, and can thus have differing sensitivities to whatever amount of WA is actually present. In this work, we adopt an approach that attempts to address these issues directly. We effectively measure the change in a traditional inclusive measurement of $b \to u\,\ell\,\nu$ when a weak annihilation component is included in the analysis.

The investigation resembles a standard inclusive $b \to u\,\ell\,\nu$ measurement in many aspects, but the result is not a precision value for $\Gamma(b \to u\,\ell\,\nu)$ or $|V_{ub}|$. Instead, the objective is to measure the effect of allowing for a new component in the rate extraction, one that represents the possible WA contribution. The idea is to allow the data itself to dictate to what extent the process is present, and what the potential impact on the measured rate would be. An important feature of this design is that the rate measurement itself need not be particularly optimized to assess its susceptibility to contamination by WA effects.

The success of this strategy depends critically on the technology of neutrino reconstruction and the fidelity of the Monte Carlo simulations used to understand and describe the data. Significant attention will be paid to each of these topics in this chapter as we outline the general event selection and reconstruction procedure. The chapter following this one will describe the assembly of these pieces into a coherent description of the data, guided by a robust fitting procedure. But first we turn to the raw material on which the analysis machinery is brought to bear.

5.2 Data Samples

The data analyzed in this study are 13.8 fb^{-1} of e^+e^- collision data[1] collected with the CLEO detector at the CESR facility, described in the preceding chapter. Two different configurations of the detector were used during the course of the experiment: The first third of the sample was collected with the CLEO II version, and the remainder was obtained with CLEO II.V. However, the differences between the two detectors are not significant for this analysis, and every attempt is made to analyze the combined data sample in a uniform and consistent way.

The analysis also makes use of simulated "Monte Carlo" data as a tool for testing the design of the analysis and, later, at fitting time, as templates for several of the sources that are expected to be present in the data. The general philosophy behind the production of Monte Carlo data was discussed previously in Sec 4.6, but there are details particular to this analysis that require further clarification.

5.2.1 Hadronic Data

The actual input to the analysis is a processed version of the raw data called the `CCHAD` skim. The initial Level3 filtering and processing by `pass2` described in Sec 4.5 classifies events into broad physics categories and also acts to eliminate events that are unlikely to have been caused by actual e^+e^- annihilations. Rejected events include those from beam-gas or beam-wall interactions, where the particles in the beam's periphery interact with stray gas molecules or even the wall of

[1]Advance note for the numerically astute: This luminosity total includes both on- and off-resonance running at the $\Upsilon(4S)$ peak.

the vacuum chamber. The CCHAD skim contains those events that are loosely consistent with the production of hadrons, which are identified by a characteristic signature of high track and shower multiplicity and a significant amount of "visible" energy (charged and neutral) in the detector [87]. (Most of the energy in low-angle scattering events tends to be lost "down the beampipe," *i.e.* outside the detector's geometric acceptance.) The event classification algorithm is quite efficient for hadrons in general, and exceeds 99% for $B\overline{B}$ events in particular.

The surviving hadronic sample is comprised of events in three basic categories: $B\overline{B}$, light quark $q\bar{q}$, and $\tau^+\tau^-$. The first category is the one of interest, and the latter two are standard backgrounds for most B physics analyses. Selection cuts applied as part of the technique of neutrino reconstruction tend to eliminate events with τ leptons, so the τ pair background is not a particular concern for this analysis.

The more important non-$B\overline{B}$ background comes from non-resonant production of the lighter quark pairs $q\bar{q}$, where $q = u, c, s, d$ but not b; in fact, the estimates in Sec 4.1 suggest that these events outnumber $B\overline{B}$ events by a 3:1 ratio (before any filtering). Discrimination against these events relies on two distinct strategies. First, $q\bar{q}$ and $B\overline{B}$ events differ in characteristics having to do with the overall "shape" of the event. Essentially, topological differences arise from the fact that the B mesons are produced at threshold with almost no energy of motion, so their subsequent decays are isotropic in the lab frame; in contrast, lighter quarks are produced with significant additional energy that hadronizes into "jets" of particles collimated around the initial back-to-back pair of quarks. Selection criteria that combine various measures of event topology can be designed to suppress these continuum contributions quite effectively.

The second strategy relies on the fact that the $q\bar{q}$ backgrounds are not specific to the $\Upsilon(4S)$ resonance. In fact, their cross section is a slowly-varying function of energy across the entire region of Υ resonances. Non-resonant $q\bar{q}$ production is often called "continuum background" since this process has a spectrum that is essentially flat and smooth in the neighborhood of the $\Upsilon(4S)$. By running slightly *below* the $\Upsilon(4S)$ peak to collect a pure sample of continuum events, data are obtained that are essentially statistically identical to the events that contaminate the $B\overline{B}$ sample collected *on* the $\Upsilon(4S)$ peak. With the proper normalization, this non-resonant background can then simply be *subtracted* from the on-resonance data: what remains is the resonant $B\overline{B}$ physics.

(Such background subtraction is a common technique applied in particle physics analysis, and relies again on the fact that the data is studied on a *statistical* basis. While it isn't usually possible to unequivocally classify each on-resonance event as being from $B\overline{B}$ or the continuum, it is possible to state very precisely that, say, 45 in the total of 100 events must be from continuum processes, based on a knowledge of cross sections and the relative luminosity of the two samples. Then, when some feature of the sample is studied in distribution, the continuum contribution can be accounted for statistically by taking the analogous distribution measured in the off-resonant data, scaling it to 45 events, and then subtracting it from the distribution seen in on-resonant data. The resulting distribution will display the distribution of the selected property for the 55 $B\overline{B}$ events, even though those events have not been explicitly identified. Naturally, since subtraction leads to a loss of statistical precision, it is usually preferable to first reduce the continuum contribution in a sample with selection cuts, such as the topological ones mentioned previously, and then use the subtraction to account for the remaining continuum contribution.)

Before continuum suppression and other analysis cuts are applied, the full hadronic data sample consists of 80 428 106 events.[2] The on-resonance to off-resonance luminosity ratio is almost exactly 2:1. Using cross sections determined from the data and beam energy measurements recorded for each run, the total number of $B\overline{B}$ pairs in the data sample we analyze is estimated to be 9.70 M.[3]

5.2.2 Simulated Data

In addition to the hadronic event sample just described, we employ several samples of simulated data to tune analysis strategies and to model background and signal processes. Here we describe how these samples were obtained and subsequently tailored for the needs of the current analysis.

$b \to c\,\ell\,\nu$ Monte Carlo Sample

The generic $b \to c$ sample was introduced in Sec 4.6. It consists of approximately 39.7 M $B\overline{B}$ events, with each B decayed "generically" along any of literally hundreds of channels chosen from a master list of possible decay modes.[4] The

[2] The combined CLEO II and II.V datasets contain an impressive total of 823 014 925 events!

[3] For the CLEO-aware reader: The total event figure is simply the number of events seen when running over the entire `CCHAD` skim. The number of $B\overline{B}$ events is from the `RPUSER` package, as interfaced to `suez`. Recall this routine counts $B\overline{B}$ events by counting on- and off-energy events and multiplying by a dataset-dependent cross section previously determined by D. Besson [92]. The cross sections are determined by simply counting the *total* number of $B\overline{B}$ pairs in each dataset and dividing by the luminosity. The use of an effective cross section for each dataset allows for the ready calculation of how many $B\overline{B}$ pairs are included in an analysis that only samples some subset of the full dataset. Since we employ each hadronic dataset in its entirety, we are effectively using the exact tallies originally computed by Besson.

[4] Perhaps somewhat paradoxically, the list of decay channels available to the B actually includes unknown modes. Some of these are simply modes that have not actually been observed, but others are inclusive, non-deterministic directives to hadronize a quark

effects of neutral B mixing ($B^0 \leftrightarrow \overline{B^0}$) are included with an effective mixing parameter of $x_d = 0.719$.[5]

The Monte Carlo generally tracks the real detector's performance quite well, reproducing tracking and energy resolution both precisely and historically.[6] This stunning achievement is the result of careful understanding of detector effects and the painstaking calibration of the Monte Carlo simulation for each aspect of the event [91]. However, despite the original intent to have Monte Carlo samples in 1:1 correspondence with every recorded dataset, mimicking the real data in perfect, proportional detail, there are some gaps. In particular, there is no $b \to c$ Monte Carlo for the datasets `4sH`–`4sL`, which make up a little more than 15% of the recorded luminosity (and number of $B\overline{B}$ pairs). Nonetheless, the detector configuration is known to have been relatively stable during this data-taking period and it is assumed that the data in these datasets do not require custom modeling. Fig 5.1 elaborates on the issue of uniform coverage a bit further, showing the ratio of Monte Carlo to data statistics, broken down by dataset.

In order to compare the Monte Carlo sample directly with data, it is useful to normalize the simulated sample to the equivalent number of $B\overline{B}$ events expected in the data. Adopting the relatively safe assumption of "interchangeability of events," we compute a single scale factor for the entire CLEO II sample (0.2167), and similarly one for CLEO II.V (0.2767). The reciprocals of these numbers contain basically the same information as the weighted average of the ratios plotted in Fig 5.1.

The target of this analysis is charmless semileptonic B decay, $b \to u\,\ell\,\nu$. Formally, all other decay channels of the B constitute background to this mode, but if

---

state into as many realizable mesons and baryons as is feasible. The full specification of generic B decay represents a skillful blend of the measurements currently available, present estimates for as-yet undiscovered modes, and reasonable hypotheses about entire sectors about which very little is known.

[5]As the phrasing in the main text suggests, mixing is not simulated identically across all parts of the CLEO II and II.V Monte Carlo samples. The details are documented in internal CLEO literature, but essentially the value of the mixing parameter x_d varies slightly across the sample; the value quoted above is the weighted average. For comparison, note that the world average is $x_d = 0.771 \pm 0.012$ [1]. Event (dataset) re-weighting could be used to correct each dataset to the world average, but since the analysis does not explicitly depend on B flavor, this level of attention has not been brought to bear on these differences.

[6]Operationally, this success is achieved by identifying a handful of runs in each dataset that are deemed to be representative of the entire dataset and for which beam and detector conditions are well-understood. Simulated events are then generated according to the conditions of these runs. The number of events per run is taken in the same ratio as the estimated $B\overline{B}$ content of the real data. In creating a new Monte Carlo sample, an important consideration is the choice of how many Monte Carlo events to generate per effective (or representative) data event, or more exactly, per unit data luminosity.

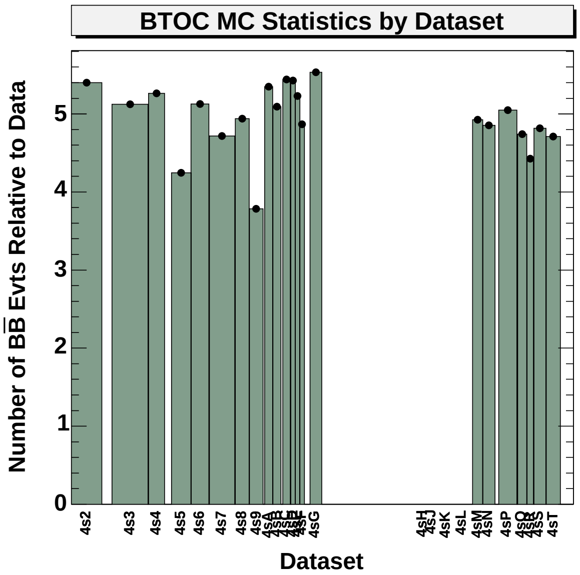


Figure 5.1: Ratio of $b \to c$ Monte Carlo statistics to number of $B\overline{B}$'s in data, broken down by dataset. The plotted points are simply the number of $B\overline{B}$ events generated in each dataset divided by the number estimated to be present in the data, based on luminosity and cross section measurements. Note the hole for the `4sH`–`4sL` datasets.

we presume perfect lepton identification, it is clear that the only significant background can come from other non-signal B decays that include leptons. Looking forward, we also know that part of the final event selection criteria for the analysis will be the requirement of a lepton with momentum above 1.5 GeV. Fig 2.8 at the end of Ch 2 then makes it clear that the dominant source of leptons above such a cut will be semileptonic B decays, *i.e.* $b \to c\,\ell\,\nu$. Thus the real use of the generic $b \to c$ sample is as a model of the dominant $b \to c\,\ell\,\nu$ background.

As described later in Sec 5.8, we account for the contribution from events with incorrectly identified leptons with the use of non-leptonic data. The chief aim of such a strategy is to remove dependence on the Monte Carlo for modeling such "fake leptons." Thus the Monte Carlo sample is used only as a model for events with real leptons.

In light of these conclusions, and to ease the data-processing burden, we have prepared a generator-level skim of the full $b \to c$ sample, selecting those events that have at least one true lepton present in the decay tree with momentum $|p_\ell| > 1.0$ GeV.[7] The selection efficiency of the skim is about 32.5%, providing

[7]Typical lepton momentum resolutions are at the 10 MeV level or better, so smearing of the lepton momentum is not a particular concern: if a track is reconstructed with a momentum of 1.5 GeV, the underlying particle most likely had that same momentum

an initial $b \to c\,\ell\,\nu$ sample of 12.5 M events.[8] Inevitably, as discussion of the analysis progresses, this sample will be referred to as both the (generic) $b \to c$ sample and the (generic) $b \to c\,\ell\,\nu$ sample. This confusion is in fact only superficial: as just argued, there is no loss of generality in replacing the fully generic sample with this reduced one, since the omitted events are not only of no consequence: the design of the analysis has rendered them entirely irrelevant. Furthermore, the skim predominantly selects real $b \to c\,\ell\,\nu$ decays; contributions from secondary leptons and other sources of hard leptons are in fact quite small.

It is this far more manageable event sample that is taken to represent the generic $b \to c$ background. From this point forward, it is "treated just like data," subjected to an analysis procedure that is (almost) identical to the one applied to the events in the `CCHAD` skim.

$b \to u\,\ell\,\nu$ Monte Carlo Sample

Simulated $b \to u\,\ell\,\nu$ decays are produced with a modified version of the standard event generation program QQ. The specialization is called InclGen, and it is designed to provide a reasonable "best-guess" full-spectrum simulation of inclusive $b \to u\,\ell\,\nu$ decay. The crux of the problem is that although the inclusive theory is on reasonably sound footing (recall Ch 3), the simulation of the inclusive prediction has to be practically implemented as a sum of various exclusive modes. Unfortunately, unlike the case for $B \to X_c\,\ell\,\nu$ where the rate is dominated by a small number of well-measured exclusive channels (chiefly D and D^*), numerous X_u resonances contribute to the total $B \to X_u\,\ell\,\nu$ rate, and the relevant form factors, spin quantum numbers, and even a few masses are poorly known for many of these channels. The decomposition of $b \to u\,\ell\,\nu$ rate among these channels is also not well-constrained, and it is not even clear what fraction of the total rate is devoted to the production of non-resonant, multi-body hadronic final states.

The InclGen algorithm is an instance of a "hybrid" approach, quite similar in principle to inclusive models developed at other experiments over the past few years.[9] Essentially, a model-independent theoretical prediction for the triply differential fully inclusive rate $d^3\Gamma/dE_\ell\,dM_X\,dq^2$ [98] is combined with a list of the X_u resonances predicted in the ISGW2 quark model [117] to yield a physically realizable decay spectrum, composed of known resonances and a broader non-resonant

(to within 10 MeV). The real reason for adopting such a low skim cut is to allow room for systematic studies and refinement of other aspects of the analysis.

[8] Various technical difficulties prevent access to the full generic Monte Carlo sample, but the loss in statistics is less than 4% of the total. The quoted number is for the sample actually used.

[9] For an early effort by OPAL that combined the ACCMM inclusive and ISGW2 exclusive models for the mass spectrum of $b \to u\,\ell\,\nu$ see Ref [125]. An approach used more recently at BaBar is described nicely in Ref [126].

background loosely consistent with the constraints of the inclusive theory. A brief outline of the procedure suffices to clarify how the information from each input is used:

1. The kinematics of the initial $B\overline{B}$ pair are generated according to the input beam conditions and other standard information.

2. Each B is considered in turn for $b \to u\,\ell\,\nu$ decay. The decision is made non-deterministically and independently for each B, but only events where at least one B has decayed along the $b \to u\,\ell\,\nu$ set of channels are actually stored in the output stream.

3. For a particular $b \to u\,\ell\,\nu$ decay, a kinematically feasible point (E_ℓ, M_X, q^2) in inclusive phase space is selected randomly, according to the distribution predicted by the inclusive model.

4. Using only the hadronic mass information generated as part of the kinematic triple in the previous step, a decision is made to either decay the hypothetical B into a nearby ISGW2 resonance, or to implement the decay non-resonantly. Briefly, the decision is made by comparing the inclusive rate $d\Gamma/dM_X$ at the generated value of M_X to the sum of the partial rates assigned to nearby exclusive resonances, each smeared out slightly from their physical width by an inclusive "smearing scale." The decision is made in a random probabilistic fashion.

   (a) If the decay is to be non-resonant, it is handled by standard Lund-like [84] CLEO software routines for hadronizing a $u\bar{q}$ system of mass M_X. The previously generated lepton energy E_ℓ and lepton-neutrino invariant mass q^2 are sufficient to determine the remaining kinematics of the system.

   (b) If the decay is to be along a particular exclusive channel, selected randomly from those "near" the initial M_X value, the generated kinematic triple is discarded and the decay is carried out according to the ISGW2 model and the relevant form factors for that resonance. The actual implementation is handled by the EvtGen package [86].

We note that unlike some past models for $b \to u\,\ell\,\nu$, this approach leads to a final hadronic mass spectrum $d\Gamma/dM_X^2$ that is not identical in detail to the inclusive prediction alone, due to the explicit inclusion of low-mass hadronic resonances. Indeed, the combination implemented in the InclGen generator is expected to describe the physical spectrum better than each component would alone. A more complete description of the simulation algorithm is available in Appendix A.

The InclGen sample is created by simulating typical $B\overline{B}$ events in the usual way, but with a new inclusive $b \to u\,\ell\,\nu$ decay mode inserted into the master decay table. The special mode is assigned a branching fraction of 0.5% for each lepton flavor (e and μ). The InclGen algorithm is triggered by the selection of this decay channel by the standard B decay machinery. Otherwise, the B is decayed fully "generically" according to the same table as used in the $b \to c$ sample. As already mentioned, only events where at least one B decay is delegated to InclGen are ultimately saved. B mixing is implemented with a fixed value of $x_d = 0.723$ across all datasets.

The probability distribution for the number of InclGen decays in an event is thus binomial, with a "success" rate of $p = 0.010$, summing across lepton flavors, and $N = 2$ trials. The coincidence rate—the chance that both B's in the event are actually decayed via InclGen—is given by $p^2/(p^2 + 2p(1-p)) = 0.005$, the probability of obtaining two InclGen decays, conditioned on the presence of at least one such decay.

As should be clear, a chief advantage of the hybrid technique is that the non-resonant background shape respects the existence (and location) of known exclusive resonances, but it is also to some extent dependent on the shape predicted by theory in the inclusive limit of sufficiently broad averaging. Rather than being finely tuned for a few of the best-known resonances (*e.g.* π and ρ), this mixed approach creates a cocktail of many resonant and non-resonant channels that respects input from both experiment and theory.

Fig 5.2 shows distributions of various kinematic quantities in a sample of 0.6 M B decays generated with InclGen. In each plot, the total generated spectrum is broken into non-resonant and resonant portions, and the sum is compared to the inclusive spectrum predicted by HQET.

The inclusive prediction used in the generator is the triple differential rate provided by DeFazio and Neubert in Ref [98], evaluated to order $\mathcal{O}(\alpha_s)$ and leading order in HQET, implemented at CLEO by Lyon *et al.* [99]. The non-perturbative smearing function is of the exponential form (see Appendix A), with parameters tuned from a recent CLEO analysis [127] of the $b \to s\gamma$ photon energy spectrum: $\lambda_1^{\mathrm{SF}} = -0.342\ \mathrm{GeV}^2$ and $\overline{\Lambda}^{\mathrm{SF}} = 0.545$ GeV. Together with the choice of $\alpha_s = 0.22$, these values fully determine the leading-order inclusive prediction for all spectra, as well as their various integrals (and so thus the total inclusive rate). For historical reasons, the nominal $b \to u\,\ell\,\nu$ event sample generated with these parameters is referred to as the InclGen `newbsg` sample.

Several systematic variations of the nominal `newbsg` sample are used to assess the sensitivity of the analysis to the modeling of $b \to u\,\ell\,\nu$ decay. Two natural variations simply employ the extreme variations of the shape function parameters consistent with the $b \to s\gamma$ measurement; these are known as the `bsglo` and `bsghi` samples. Finally, a series of "purely inclusive" samples are also available, where no resonances are allowed and the generated mass spectrum is identical to the input from the inclusive theory. Using the central and extreme values from the $b \to s\gamma$

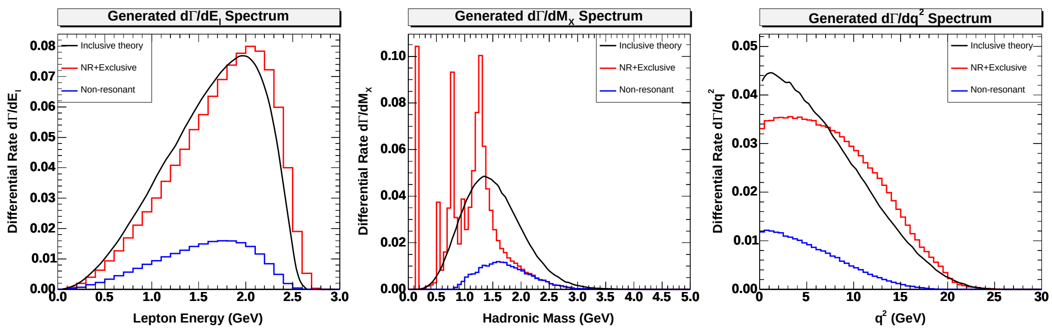


Figure 5.2: One-dimensional spectra for events generated with the InclGen inclusive generator, for a sample of 666 K B decays. The theoretical spectrum as predicted by HQET is indicated by the smoothed curve, and the non-resonant and ISGW2 resonant components as actually generated are represented by the two (stacked) histograms; the lower histogram is the non-resonant contribution. It is clear that the theoretical spectrum is ignorant of the resonant structure at low hadronic mass M_X, but the hybrid approach provides a systematic method for "borrowing" rate from the lower half of the inclusive spectrum that is then assigned to these resonances. The remaining inclusive rate is dedicated to a relatively smooth non-resonant contribution at higher hadronic mass.

analysis, this modification leads to the creation of three new samples called `x0`, `x0lo`, and `x0hi`, where the nomenclature "`x0`" is chosen to reflect that the fact the contributions from the *ex*clusive modes have been "zeroed out." Each of these five additional samples matches the size of the `newbsg` sample with approximately 1 M events. Table 5.1 gathers together the basic characteristics of these samples.

The dataset-by-dataset coverage mirrors that described for the generic $b \to c$ Monte Carlo: there is no simulated data corresponding to the `4sH`–`4sL` datasets. Fig 5.3 illustrates the statistical size of this sample relative to the data in more detail.

Weak Annihilation Monte Carlo Samples

Even less is known with any real certainty about the process of weak annihilation. Yet a model for this signal process is invaluable for understanding where to look for WA in the data and for gauging the sensitivity of the analysis to its presence. To meet these needs, a heuristic model for the kinematics of weak annihilation was created to prepare a sample of how WA might be realized in the data. By changing various properties of the toy model, different scenarios for WA can be generated.

The details of the model used to simulate the kinematics of WA decays are

Table 5.1: Basic characteristics of the various InclGen samples. One set of samples invokes the full hybrid approach to blend ISGW2 resonances with the inclusive spectrum from theory, and the other hadronizes the $u\bar{q}$ system non-resonantly (NR) with a distribution identical to the inclusive prediction. Within each sample category, the implementation of the non-perturbative smearing is varied in the range provided by the recent CLEO analysis [127] of the $b \to s\gamma$ photon spectrum.

Sample	λ_1 (GeV^2)	$\overline{\Lambda}$ (GeV)	**Notes**
`newbsg`	-0.342	0.545	NR + ISGW2 blend; $b \to s\gamma$ central values
`bsglo`	-0.130	0.475	NR + ISGW2 blend; $b \to s\gamma$ low extreme
`bsghi`	-1.130	0.781	NR + ISGW2 blend; $b \to s\gamma$ high extreme
`x0`	-0.342	0.545	All NR; $b \to s\gamma$ central values
`x0lo`	-0.130	0.475	All NR; $b \to s\gamma$ low extreme
`x0hi`	-1.130	0.781	All NR; $b \to s\gamma$ high extreme

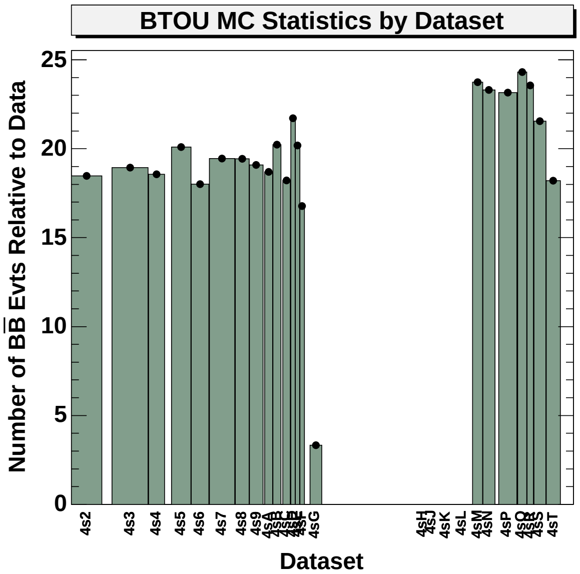


Figure 5.3: Ratio of $b \to u\,\ell\,\nu$ Monte Carlo statistics to number of $B\overline{B}$'s in data, broken down by dataset. The Monte Carlo is normalized assuming a branching fraction $\mathcal{B}(b \to u\,\ell\,\nu) = 1.5 \times 10^{-3}$, consistent with what is seen in this analysis. Note the hole for the `4sH`–`4sL` datasets.

described in Sec 6.1. Thirty different versions of WA are created with different settings for the input parameters to the model. For convenience, we identify these samples simply as models WA01–WA30. The next chapter provides translation tables (Table 6.1–6.2) that map sample index to the physical properties of each sample.

The samples are created using a natural flexibility intentionally built into the basic InclGen design. The package can accommodate different "inclusive models" to provide the kinematic triples (E_ℓ, M_X, q^2) and the various spectra required in the basic decision algorithm outlined above. Simulating charmless weak annihilation amounts to configuring InclGen to use a new model that replaces the default HQET implementation with one that returns the kinematics for WA decays. By zeroing out all of the exclusive $b \to u\,\ell\,\nu$ channels that usually compete with the inclusive model, every B that InclGen decays will automatically follow the inclusive or non-resonant path, *i.e.* weak annihilation in this configuration.

The special InclGen channel in the master B decay table, when chosen, thus leads to weak annihilation of the B meson 100% of the time. The resulting WA samples are analogous to the other InclGen samples in that each consists of $B\overline{B}$ events in which at least one B decays via the new WA channel, and the other B decays generically. Note there are no other $b \to u\,\ell\,\nu$ modes included in the simulation. (The object of this thesis is to let the data determine the relative ratio $\Gamma_{\mathrm{WA}}/\Gamma_{btou}$, rather than imposing it externally.)

Put simply, the InclGen model of weak annihilation randomly generates the kinematics of a soft, light hadronic system according to a simple internal recipe, and then determines the W and lepton-neutrino kinematics that result under the assumptions that the weak coupling is of the $V - A$ form and that the hadronic system has spin $s = 0$. The hadronic four-vector is hadronized as usual into a multi-particle system using the same non-resonant decay handler already part of the InclGen package. Note that the algorithm does not depend on the flavor of the parent B: charged and neutral B decays are handled identically.

Simulation coverage by dataset is more uniform than for the standard $b \to u\,\ell\,\nu$ samples, as shown in Fig 5.4. The only dataset omitted is `4NH`.

5.3 Neutrino Reconstruction

The characteristic signature of a weak semileptonic decay is a charged lepton and a neutrino. Leptons are typically easy to detect and identify, but neutrinos present a much more difficult experimental challenge. As neutral weakly interacting particles, they have an extremely small cross section for interacting with matter. Consequently, for a typical collider detector such as CLEO weighing in at only a few kilotons, there is essentially zero probability that a neutrino will leave any indication of its passage. Thus, for instance, early studies of semileptonic $b \to u$ decays dropped the neutrino detection issue entirely and focused simply on the identification of high-momentum leptons. Without the confirming presence of a neutrino, however, it is harder to eliminate leptons from continuum processes, and

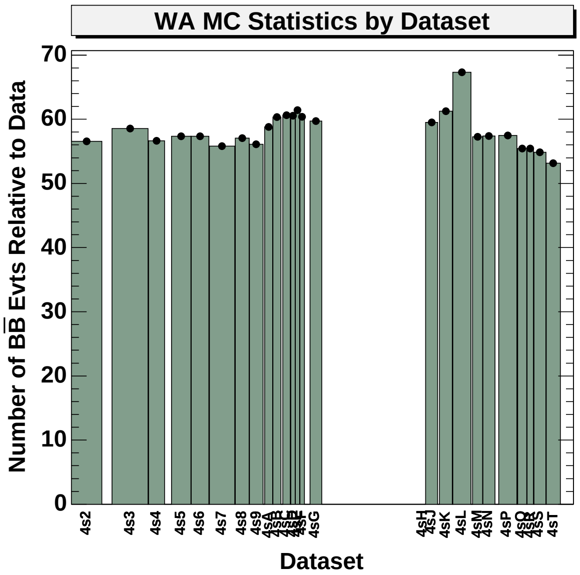


Figure 5.4: Ratio of WA Monte Carlo statistics to number of $B\overline{B}$'s in data, broken down by dataset. For display purposes, the WA sample is normalized assuming a branching fraction that is 3% of the $b \to u\,\ell\,\nu$ rate, where again we take $\mathcal{B}(b \to u\,\ell\,\nu) = 1.5\ \times 10^{-3}$. Note the only gap in the simulation coverage is in the `4sH` dataset, for which the corresponding amount of data is quite small, less than 0.4% of the total number of $B\overline{B}$ events.

without even basic knowledge of the neutrino energy and direction, any program for exclusive reconstruction of the B decay is impossible.

Unable to detect neutrinos directly, CLEO has developed a technique known as "neutrino reconstruction" that can deduce the presence and kinematics of a neutrino *indirectly*. The basic premise of the technique is simple energy conservation, but it also relies critically on the excellent hermeticity and resolution of the detector. Although not new with this work, the technology of neutrino reconstruction is the central feature of the analysis, so we devote some space here to its description.

5.3.1 How it Works

Reconstructing a neutrino is a straightforward consequence of applying momentum-energy conservation to the e^+e^- collision. The initial state is known very precisely, since the electron and positron beams collide under carefully controlled conditions in the center of the detector. At any point during the course of the interaction, a sum over all particles present, neutral and charged, detected and undetected, must give the same total momentum-energy four-vector as computed for the initial state. This requirement is basic four-momentum conservation. In particular, the constraint can be applied to the system after the decay of the initial B meson pair into the final assortment of light, long-lived particles that travel through the detector. Algebraically, the requirement appears as

$$\begin{aligned} p_{\text{initial}} = p_{\text{CM}} &= p_{\text{final}} \\ p_{\text{CM}} &= \sum_i p_i \\ &= \left(\sum_{\text{detected}} p_i\right) + \left(\sum_{\text{undetected}} p_i\right), \end{aligned}$$

where we've broken the sum over final-state particles into two pieces, separating the particles that were detected in some way from those that weren't. The total four-momentum of the undetected particles is then given formally by

$$p_{\text{undetected}} = p_{\text{CM}} - \sum_{\text{detected}} p_i \tag{5.1}$$

We now make the first of two crucial assumptions: we assume an exact 1:1 correspondence between the tracks and showers reconstructed in the event and the real particles actually present. In the same vein, we further assume that the reconstructed objects accurately represent the real particles' energy and momentum. This assumption rests on the approximate hermeticity of the detector and its high efficiency and excellent resolution: the claim is that all particles that are produced are subsequently detected and measured properly. Labeling the undetected four-momentum rather naturally as what's "missing" in the final state, we have then

$$p_{\text{miss}} = p_{\text{CM}} - \sum_{\text{obsvd}} p_i, \tag{5.2}$$

where the sum is now over *reconstructed* quantities, the only things really accessible to experiment.

Our second assumption is better characterized as a specialization: we consider the case where there there is only a *single* undetected particle, such as a neutrino. In this circumstance, the four-momentum of this single particle is given by the total missing four-momentum computed in Eqn 5.2, so

$$p_\nu = p_{\text{miss}} = p_{\text{CM}} - \sum_{\text{obsvd}} p_i. \tag{5.3}$$

To evaluate the right-hand side of Eqn 5.2, the momentum and energy of every track and shower included in the sum must be known. However, strings of hits in the drift chambers do not have an intrinsic energy, nor do clusters of energy hits in the crystals have an intrinsic three-momentum. In each case, additional detector information or other assumptions must be employed to "complete" the particle four-vector. For tracks, a mass assignment is sufficient to determine the energy

E from the momentum according to the usual relation $E = \sqrt{m^2 + p^2}$. Charged particle identification is thus another important element of the analysis, and it has to be done for every track in the event that will be included in the observed momentum sum. For showers, the situation is slightly different. All showers not matched to tracks are assumed to be from photons that originated in the primary decay at the center of the detector. Constructing the photon three-momentum amounts to little more than appropriately scaling the shower's position vector up to the shower energy.

As outlined above, the technique of neutrino reconstruction appears as a trivial algebraic exercise, but it disguises a rather complex experimental task. In a few words, the challenge is to guarantee that the two assumptions applied in the preceding derivation are valid: that the *only* undetected particle is a single neutrino, and that every other particle in the event is accurately reconstructed and properly identified. This demand is met in part with the use of a high-performance, well-calibrated detector, but it is also mitigated by the application of various consistency checks, such as confirming that the separately determined missing momentum and missing energy, when combined, are consistent with the mass of a hypothetical undetected neutrino. We elaborate on these and related issues in the following sections.

5.3.2 Limitations

Errors in the reconstructed momentum of the neutrino are often described as a "smearing" from its true value, and can generally be traced to detector or reconstruction effects (which can be mitigated to some extent), and physics effects (about which very little can be done). Ultimately, neutrino resolution is not limited by finite detector resolution but by the failure of either of the two fundamental assumptions on which neutrino reconstruction rests.

A key requirement in neutrino reconstruction is that there be one and only one reconstructed object per real particle in the event. For instance, if two distinct tracks are reconstructed for the same particle, the obvious double-counting will bias the inferred kinematics of the neutrino. This concern is addressed with a two-pronged approach. For each event, the track and shower lists are examined in an attempt to identify spurious or otherwise redundant entries that should be excluded from the observed four-momentum sum. Secondly, other event characteristics are checked for tell-tale signs of reconstruction mistakes that may not otherwise be apparent. For instance, the net charge of an event, computed from the charge of all reconstructed tracks, should be zero, matching the initial e^+e^- state. However, tracking mistakes arising from pattern recognition failures, inefficiencies, or material interactions can all lead to the creation of spurious tracks or the loss of real ones. Any event with a non-zero net charge clearly suffers from some sort of tracking error, and rather than trying to salvage the problem event, it is best to simply discard it entirely.

The presence of additional undetected or poorly detected particles is the dominant source of error in neutrino reconstruction, and the source least amenable to improvement. Briefly, particles can be missed due to any of several reasons:

- Acceptance limitations. Although the active coverage of the detector extends over 95% of 4π, it is not perfectly hermetic. For instance, soft charged particles may not pass through enough tracking layers to be confidently observed.
- Interactions of particles in detector material. Hard scatters, inelastic hadronic interactions, or other processes can cause particles to effectively "disappear" without leaving a detector signature complete enough to reconstruct what happened.
- Detector inefficiencies. Despite best efforts in detector element design and built-in redundancies, the detector doesn't respond 100% reproducibly or reliably to the passage of a particle. Occasionally, a detector sensor is simply not activated.[10]
- Tracking and calorimetry errors. Inefficiencies at the reconstruction or software level, arising from algorithmic limitations or mistakes, can prevent raw track and shower information from being recognized as evidence of a real particle.
- Characteristically small, unpredictable, or entirely absent detector signatures. Neutral hadrons such as neutrons and K_L mesons easily penetrate the crystals but do not regularly produce showers; when they do deposit energy, the showers are incompletely contained and unlikely to record the full energy of the incident hadron. Additional neutrinos beyond the one from the signal semileptonic decay will also leave no signal in the detector.

The last item (undetect*able* particles) is the largest contributor to neutrino smearing, both in frequency and in impact on neutrino resolution. The total error on the neutrino momentum has been estimated to be ~ 1.1 GeV per event on average [74], of which almost 0.8 GeV is due to the presence of undetectable neutrals, chiefly additional neutrinos and K_L's. A more detailed numerical breakdown of neutrino resolution will be presented in Sec 5.3.8.

Although there is obviously no method to directly observe the undetected particles in an event, various consistency checks in the analysis help screen out events

[10]Inefficiencies like these will contribute more to resolution than to actual loss of particles, simply because the probability for enough simultaneous sensor failures to completely miss a particle is vanishingly small. Mismeasurement due to calibration error or hardware failure is very real, however.

that contain unobserved neutrals.[11] For instance, if a single neutrino is the only missing particle, the invariant mass of the missing momentum vector should be consistent with the (vanishing) neutrino mass. Secondly, neutrinos are invariably produced in the weak semileptonic decay of quarks, so the presence of additional leptons in an event can signal the presence of additional neutrinos.[12] To take advantage of this correlation, leptons are identified across a wide momentum range, and events in which more than one lepton is found are subsequently rejected from the analysis. Past analyses and this one have also attempted to positively identify calorimeter showers caused by K_L's to trigger a more sophisticated event veto, but these efforts failed to bear fruit.

In signal $b \to u\,\ell\,\nu$ events, the largest source of undetected particles (particularly neutrals) is the decay of the other B, suggesting that another approach for making inroads is through an analysis of the other B in the event. At the data-laden B factories, neutrino reconstruction analyses indeed benefit from the ability to fully reconstruct the other B in an all-hadronic mode in which no particles are missing. The remaining tracks and showers in the event are then treated as a candidate semileptonic decay, with much smaller backgrounds. Unfortunately, the CLEO data set is not large enough to support neutrino reconstruction in conjunction with a second-B reconstruction attempt; the combined efficiency is simply too small to yield a statistically interesting number of candidate events. The new CLEO-c era, however, promises to usher in a new era of precision neutrino reconstruction in the D system, both with and without the reconstruction of the other D decay in the growing sample of $D\overline{D}$ events.[13]

(While full B reconstruction at CLEO is not statistically feasible, investigations have been made into reconstructing D's from the decay of the B on the other side of the event, as well as counting the number of reconstructed strange mesons. A reconstructed D eliminates one possible source of neutral K_L mesons, and the number of observed strange mesons is correlated to some extent with the presence

[11] As noted earlier, a net charge cut helps to reduce the frequency of undetected charged particles.

[12] Note that leptons can also arrive from other sources, such as decays like $J/\psi \to \ell^+\ell^-$ or photon conversions $\gamma \to e^+e^-$ in material. Nonetheless, the $\ell \Rightarrow \nu$ correlation is strong enough to make it a powerful veto of extra neutrinos.

[13] Some of the gain in the analysis of semileptonic $D \to X\,\ell\,\nu$ decays in $D\overline{D}$ events is due to the fact that a sizable chunk of the D hadronic branching fraction can be reconstructed exclusively, in contrast to the situation for B's. Further gains are made because of the lower average particle multiplicity, which simplifies pattern recognition and reduces the possibilities of various inefficiencies associated with chance overlaps. In tagged events, where one D is fully reconstructed, the missing momentum resolution on the other side is fine enough to actually resolve the K_L contribution to the missing mass peak in non-leptonic events. This resolving power is unprecedented in traditional B neutrino analyses, where the error on the missing mass swamps any discernible features in the reconstructed mass spectrum.

of unobserved ones, *e.g.* K_L. Neither of these avenues proved particularly effective for this analysis, and so are not applied here.)

Particle identification mistakes and less than perfect energy and momentum resolution also contribute to finite neutrino resolution, but these effects are much smaller than the sources of error just discussed: tracking and calorimetry resolutions contribute 10–20 MeV to the error on the neutrino momentum, and PID mistakes perhaps another ~ 75 MeV. The efficacy of track-shower matching algorithms is also a factor (see Sec 5.3.6), but is again sub-leading in comparison to the problem of undetected particles.

Neutrino reconstruction in $B\overline{B}$ events is easy to do, but not easy to get right. As more event selection cuts are piled on to suppress mis-reconstructed or otherwise contaminated events, the efficiency drops. Almost inevitably, errors in the reconstructed neutrino momentum cause $b \to c$ events to smear into the $b \to u\,\ell\,\nu$ signal region defined for any variable based on the neutrino momentum. The final statistical power of the analysis rests on the balance between the efficiency necessary required to retain a significant number of signal events, and the minimum resolution required to separate the signal from the $b \to c\,\ell\,\nu$ background.

5.3.3 Building a Neutrino

The algebra of neutrino reconstruction can be applied to any event ($B\overline{B}$, continuum, or otherwise) in which there is good reason to believe there is a single undetected particle. To extract the best estimate of the missing particle's kinematics, we choose the tracks and showers that enter into the observed momentum sum carefully, selecting those that are most likely to correspond to the true particles generated in the primary decay. In addition, when practical, we take advantage of the additional constraints offered by secondary vertices to incrementally improve the resolution on the missing particle.

The observed four-momentum in an event is computed from a sum over the detector signatures of all charged and neutral particles in the event. Charged particles will leave a track in the drift chamber and some sort of shower in the calorimeter; a long-lived neutral particle will leave only an unmatched shower in the calorimeter.

Shorter-lived neutral particles with intermediate lifetimes may travel some distance from the primary event vertex before decaying, resulting in a secondary vertex displaced from the primary. Momentum-energy conservation implies that the sum of the outgoing four-momentum of the daughters matches the incoming four-momentum of the parent, so in principle, nothing is lost by ignoring the existence of the secondary vertex in favor of its daughters. However, treating the daughter tracks as prompt (originating from the primary vertex) rather than displaced can result in an error in their estimated momentum, through a mechanism described in more detail in Sec 5.3.4 below. Furthermore, the addition of a vertex constraint to the set of daughter tracks can significantly improve the resolution on their combined momentum. These two effects combine to make it worthwhile to consider the identification and use of secondary vertices in neutrino reconstruction.

By far the largest source of such secondary decays are K_S's, as shown in Table 5.2; these decays are therefore included as a separate contribution to the observed momentum sum.

The schematic expression below describes the final computation of the "observed" four-momentum as the sum of contributions from three categories: charged particles (tracks), prompt neutrals (showers), and longer-lived neutrals (K_S's):

$$p_{\mathrm{obsvd}} = \sum_{\text{good trks}} p_i + \sum_{\text{good shwrs}} p_i + \sum_{\text{good } K_S} p_i. \tag{5.4}$$

We emphasize again that it is crucial to the basic premise of neutrino reconstruction that the three sums be exclusive of each other, to avoid bias from double-counting.

Table 5.2: Multiplicity of secondary vertices in $B\overline{B}$ events. Standard sources of secondary vertices include $K_S \to \pi\pi$ and $\Lambda \to p\pi$ decays and photon conversions $\gamma \to e^+e^-$. The table shows the average number of candidate vertices (or "vees") identified in each category per (lepton-containing) Monte Carlo $B\overline{B}$ event. The vees are reconstructed with extremely loose cuts during `pass2` reconstruction. The last column is the most relevant, showing the mean number of identified vees that are actually "correct" based on Monte Carlo truth information. Properly chosen selection criteria can eliminate most of the spurious vees and retain most of the true ones. The bottom line of the table demonstrates the dramatic improvement possible for K_S candidates when the "NuRecon" criteria, described below in Sec 5.3.4, are applied. Most of the K_S's that are lost with the application of the these tighter criteria are in fact close to the origin, where the bias incurred by ignoring them is negligible.

Vee	$\langle N_{\mathrm{found}} \rangle_{\mathrm{evt}}$	$\langle N_{\mathrm{correct}} \rangle_{\mathrm{evt}}$
$K_S \to \pi\pi$	0.354	0.229
$\Lambda \to p\pi$	1.732	0.124
$\gamma \to ee$	0.051	0.019
NuRecon K_S	0.182	0.171

To compute the neutrino four-momentum, we subtract the observed momentum from the center-of-mass momentum p_{CM}. Accounting for a small crossing angle $\alpha \sim 2$ mrad at the IP, the center-of-mass conditions are given by:

$$p_{\mathrm{CM}} = p_{e^+e^-} = (\,-2E_{\mathrm{bm}} \sin\alpha,\ 0,\ 0;\ \ 2\,E_{\mathrm{bm}}\,)\,, \tag{5.5}$$

where $E_{\rm bm}$ is the (identical) energy of each accelerator beam.

For completeness, an error matrix for the observed four-momentum is accumulated during the summation from the error matrices of the constituent tracks, showers, and K_S's. The neutrino is assigned this same error matrix. Although there are other much larger sources of error in the neutrino momentum estimate, this matrix records the contribution of finite tracking and calorimetry resolution to the total neutrino error. The 4×4 matrix V is roughly diagonal, with $\sqrt{\mathrm{Tr}\, V} \approx 10$–20 MeV. It is not used in the present analysis.

Although not employed in later kinematic calculations, the energy component of the missing four-momentum vector is used as part of a check that the missing particle is consistent with a neutrino. By simultaneously measuring both the missing energy and momentum, we can compute an invariant mass for the missing particle, known as the "missing mass" for the event:

$$MM^2 = E_{\rm miss}^2 - p_{\rm miss}^2. \tag{5.6}$$

In the ideal case where there are no other sources of error, MM^2 is directly sensitive to the invariant mass of the system of missing particles; for an event where the missing particles consist of a single undetected neutrino, it should vanish, since $m_\nu = 0$.

Rather than a simple cut on MM^2, we apply a slightly more sophisticated cut on the so-called "v-ratio"

$$v_{\rm miss} \equiv \frac{MM^2}{2E_{\rm miss}}. \tag{5.7}$$

The relevance of this quantity is apparent if we consider the effect of an error Δp_ν on the calculation of the missing mass:[14]

$$\begin{aligned} MM^2 &= (p_\nu + \Delta p_\nu)^2 && (5.8)\\ &= 2E_\nu \left(\Delta E_\nu - |\Delta \vec{p}_\nu| \cos\theta_{\nu,\Delta\nu}\right) + \Delta M_\nu^2. && (5.9) \end{aligned}$$

If we assume that $\Delta M_\nu^2 \equiv \Delta p_\nu \cdot \Delta p_\nu$ is of second order, we find that

$$MM^2 \simeq 2E_\nu \left(\Delta E_\nu - |\Delta \vec{p}_\nu| \cos\theta_{\nu,\Delta\nu}\right), \tag{5.10}$$

that is, that the missing mass (squared) for the event scales with the neutrino energy. A straight cut on MM^2 thus has the potential to sculpt the neutrino

[14]Applying a first-order error analysis to Eqn 5.6 leads directly to the result $\sigma_{MM^2} \approx 2E_{\rm miss}\,\sigma_{E_{\rm miss}} + 2p_{\rm miss}\,\sigma_{p_{\rm miss}}$. The first term dominates the error (*i.e.* the departure of MM^2 from zero), since the resolution on the missing energy is 3–4 times worse than on the magnitude of the missing momentum. This reveals immediately that a cut on $v_{\rm miss}$ is therefore a cut on constant $\sigma_{E_{\rm miss}}$. The derivation that we follow in the main text, however, allows for an interesting observation about the impact of a mistake in the neutrino *direction* on the utility of the $v_{\rm miss}$ cut.

energy distribution unfairly, eliminating events with energetic neutrinos even when the error is small. A better measure of the error on the neutrino momentum is given by $MM^2/2E_\nu$: requiring that this ratio be small is roughly equivalent to limiting the error on neutrino reconstruction. Since the actual neutrino energy E_ν is unknown, we substitute the missing energy again to arrive at the v-ratio introduced above:

$$v_{\rm miss} = \frac{MM^2}{2E_{\rm miss}} \sim \Delta E_\nu - |\Delta\vec{p}_\nu| \cos\theta_{\nu,\Delta\nu}. \tag{5.11}$$

Another feature made clear in this derivation is that if $\Delta E_\nu \approx |\Delta\vec{p}_\nu|$, the missing mass can still be small if the error term $\Delta\vec{p}_\nu$ is in the same direction as the true neutrino. For instance, if the error term arises from the presence of a second undetected particle such as a K_L, no cut on MM^2 or any similar quantity will be able to eliminate the event in the case that the momentum of the K_L is along the neutrino flight direction.

The power of this cut is apparent in Fig 5.5, which shows distributions of $E_{\rm miss}$ versus MM^2 for $B\overline{B}$ Monte Carlo events containing leptons. In this plane, a cut on the v-ratio appears as a diagonal line, while a straight cut on MM^2 would be a vertical line. Events with extra neutrinos or K_L's generally fall outside a cut on $|v_{\rm miss}|$, while those with only a single undetected neutrino fall inside the cut boundary, as shown.

After event selection requirements have been satisfied for an event, the momentum four-vector of the neutrino is constrained to be massless by setting the energy component equal to the magnitude of the missing three-momentum:

$$p_\nu = (\, \vec{p}_{\rm miss}\, ; \quad |\vec{p}_{\rm miss}|\,)\,. \tag{5.12}$$

The resolution on the neutrino momentum is considerably better than on the neutrino energy for two main reasons: the vector nature of the three-momentum allows for partial cancelation of reconstruction mistakes, and the missing momentum measurement does not rely on particle identification. (The difference in energy and momentum resolution can be seen in Table 5.5, presented later in this chapter.) Thus it is the above expression for p_ν that is used, for instance, in the later calculation of kinematic quantities such as q^2.

We have detailed here the steps followed to build the neutrino estimator for an event, given a list of reconstructed objects "approved" for use in neutrino reconstruction. Subsequent sections will describe the preparation of these basic ingredients; discussion of event selection criteria is deferred until Sec 5.6.

5.3.4 K_S Selection

The K_S^0 lifetime is $c\tau = 2.68$ cm, and under the conditions at CLEO, these neutral particles can travel a measurable distance from the primary event vertex

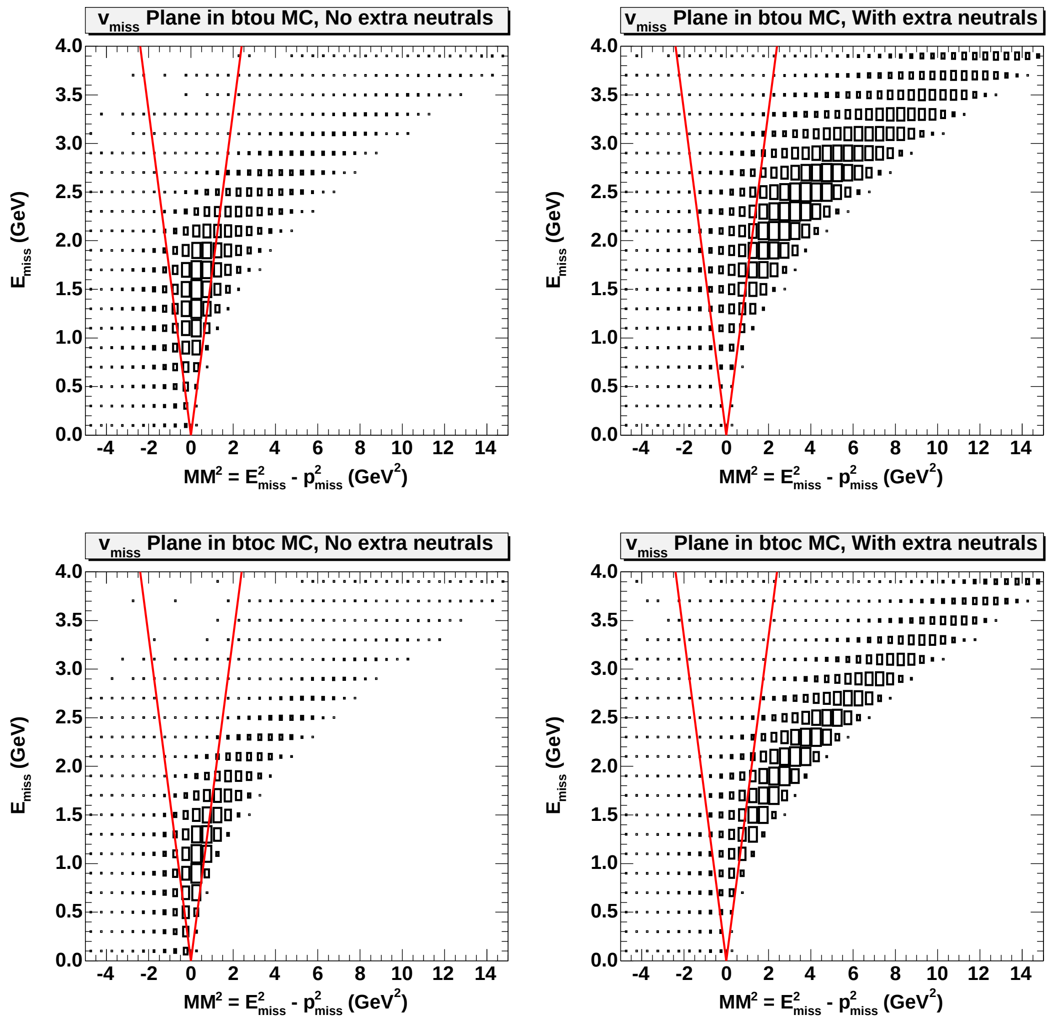


Figure 5.5: Utility of the v-ratio in eliminating events with extra missing particles, illustrated with signal $b \to u\,\ell\,\nu$ (top row) and background $b \to c\,\ell\,\nu$ (bottom) Monte Carlo. The E_{miss} *vs.* MM^2 plane is shown on the left for events with no additional ν's, K_L's, or neutrons; on the right is the distribution for events with missing particles, which generally results in a value for MM^2 that is inconsistent with zero. Note that the "error," as measured by the missing mass squared, scales with (reconstructed) neutrino energy, so a vertical cut $|MM^2| < M_0^2$ is too restrictive. A representative v-ratio cut of 0.30 is shown; events with missing particles largely fall outside these boundaries. (No other event cuts are applied in preparing this plot.)

before decaying into a pair of oppositely-charged pions. The decay point is called a "displaced" or "detached" vertex, since it represents the intersection point of two tracks (the $\pi^+\pi^-$ daughters) connected to the primary vertex only by the invisible flight path of the K_S. Treating the daughter tracks of a significantly displaced vertex as ordinary prompt tracks from the origin can potentially degrade the resolution on the reconstructed neutrino.[15] Fig 5.6 attempts to illustrate the mechanism by which this degradation occurs. Typically, the momentum of a track is calculated from its track parameters evaluated at the point of closest approach to the primary event vertex (usually taken as the beamspot, the run-averaged location of the event vertex). But for a track from a displaced vertex x_0, the proper point to evaluate its momentum is at the vertex, so that the sum $\sum_i p_i(x_0)$ over the vertex daughters properly represents the momentum carried by the parent particle. Even in the limit of no stochastic energy loss, an evaluation of the momentum elsewhere on the track helix, *e.g.* at the point of closest approach to the origin, will incur some error in the daughter's flight direction, simply due to the curvature of the track. This error only compounds in the presence of more daughters; the cartoon in Fig 5.6 shows that the errors can combine to distort the parent momentum estimated in this way from its true value. That is, the K_S momentum may be biased if it is evaluated simply as the sum $p_{K_S} = p_{\pi^+} + p_{\pi^-}$, with track momenta evaluated at the usual reference point, the primary event vertex, instead of at the secondary decay vertex. The magnitude of the bias clearly depends on the displacement of the K_S vertex from the primary vertex and the curvature of the daughter tracks, *i.e.* their transverse momentum.

The average K_S multiplicity in $B\overline{B}$ events is about 0.7 [1], making K_S's the largest source of significantly displaced vertices in the data sample. The majority of K_S's do not travel far before decaying, and so do not significantly impact neutrino reconstruction efforts. The small fraction that do have a significant (transverse) flight distance are relatively easily identified, and their daughter tracks are re-fit under the $K_S^0 \to \pi^+\pi^-$ assumption. The estimate of the true K_S momentum that results is used in the neutrino reconstruction sum in place of the daughter track contribution. To effect this substitution properly and to avoid double-counting, the tracks identified as K_S daughters are removed from the list of tracks employed in the similar track momentum sum.

High-quality K_S vertices for use in neutrino reconstruction are selected by applying several criteria to the list of standard K_S candidates identified in the `pass2` processing stage. The candidates in this default list consist of paired tracks of opposite charge that have been vertex-constrained to come from a common

[15]The quantitative extent of the degradation in resolution has not been measured. However, the effect is easy to reduce by identifying displaced vertices and handling them properly, as described in the main text. In addition, the constrained fits used to identify these vertices actually improve the resolution on the daughter track momenta, and therefore can also improve neutrino resolution.

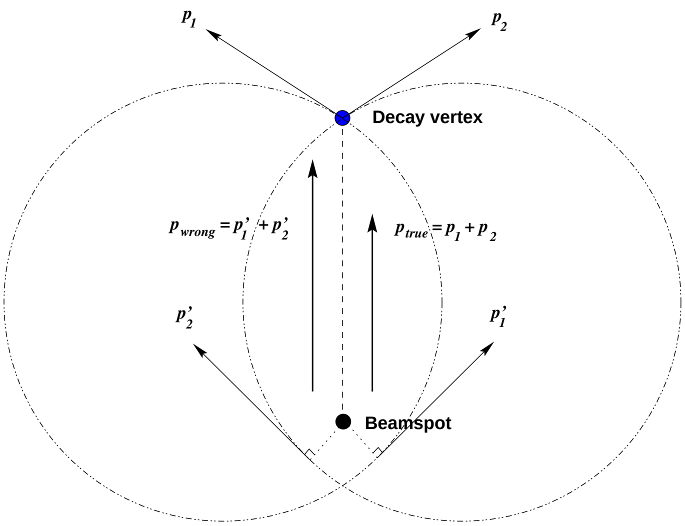


Figure 5.6: Potential error introduced by a displaced K_S decay vertex, when the daughter track momenta are evaluated at their point of closest approach to the origin rather than at the actual decay vertex. A K_S travels upwards in the figure from the beamspot, and subsequently decays into two charged tracks that start to curve along circular paths in the magnetic field immersing the tracking volume. A proper estimate of the parent K_S momentum is obtained by adding the daughter momenta $\vec{p}_1$ and $\vec{p}_2$ evaluated at the decay vertex to obtain $\vec{p}_{\text{true}}$. If instead the displaced vertex is ignored and the daughter momenta are evaluated at the respective points of closest approach between the helical projections and the beamspot, leading to momentum vectors $\vec{p}_1{}'$ and $\vec{p}_2{}'$, the sum $\vec{p}_1{}' + \vec{p}_2{}'$ will mismeasure the true K_S momentum, $\vec{p}_{\text{true}}$. (The vector sums have been carried out to scale in the figure, and the vectors $\vec{p}_{\text{true}}$ and $\vec{p}_{\text{wrong}}$ can be compared directly.)

point, and have passed a nominal cut indicating the quality of the constrained fit. Each candidate accepted from this list must:[16]

- Have a vertex-constrained invariant mass within 6.6 MeV of the nominal K_S mass. This cut is determined by the core K_S mass resolution measured in data [103].
- Have a flight significance $> 7.5\sigma$. Flight significance assesses the statistical viability of a true separation between the measured three-dimensional displaced vertex and the beamspot (from where the K_S is expected to have originated), given the finite resolution on each.

[16]Note that the daughters of the K_S are not required to be approved by Trkman (see next section).

- Satisfy $0 < \chi^2_{\text{bmspot}} < 15$, where χ^2_{bmspot} results from a fit that constrains the displaced vertex to actually come from the beamspot.
- Satisfy $\chi^2_{\text{m, bmspot}} > 0$, where $\chi^2_{\text{m, bmspot}}$ results from a simultaneous fit constraining the displaced vertex to come from the beamspot and the mass of the vertex to be consistent with the nominal K_S mass. The results of this final fit are taken as the best measurement of the K_S three-momentum.

These requirements were selected using an optimization procedure to maximize the ratio $S/\sqrt{S+B}$ for a K_S sample selected from data, developed as part of a related CLEO analysis using neutrino reconstruction to analyze the exclusive $b \to u$ channels $B \to \pi/\rho\,\ell\,\nu$ [79]. Because the cuts are made on quantities that scale out detector resolution, they are equally applicable across different datasets.

Fig 5.7 shows distributions of the K_S candidate mass, computed as the invariant mass $M_{\pi^+\pi^-}$ of the charged pion combination. On the left is the peak for all candidates in a Monte Carlo data sample, with the true K_S highlighted in blue. In data, the signal to background ratio is measured to be only $S/B \sim 1.53$, when the loose criteria applied during event reconstruction are used. After the application of the tuned selection cuts, the background level drops significantly, and the S/B improves to 6.58;[17] the improvement in a sample of CLEO III data is shown in the right-hand plot. The efficiency of these cuts for selecting all K_S's is estimated in Monte Carlo to be about 68%, with an estimated purity of 92%. Again, for neutrino reconstruction purposes, it is only the significantly displaced K_S's that it is important to identify.

5.3.5 Track Selection

The tracks identified in an event are expected to correspond in a 1:1 and onto fashion to the charged particles produced in the primary decay. Some charged particles are inevitably lost due to simple geometric acceptance (the particles don't pass through the active region of the tracking chambers), low transverse momentum (they curl around in the 1.5 T magnetic field on such a tight helical path that they don't even reach the innermost tracking chamber), or pattern recognition failures (the series of hits isn't recognized as resulting from the passage of a real particle, perhaps because it ends early in some material interaction). Tracking failures such as these are difficult or impossible to remedy, although some global event cuts can be used to reject events where these failures are more likely. Note that mistakes of this kind—where particles are lost—will tend to increase the inferred momentum

[17]The S/B figures quoted here are simply rough estimates computed by fitting the mass peak from `4sM` data with a Gaussian and a flat background. The signal is measured as the area under the Gaussian within 3 MeV of peak.

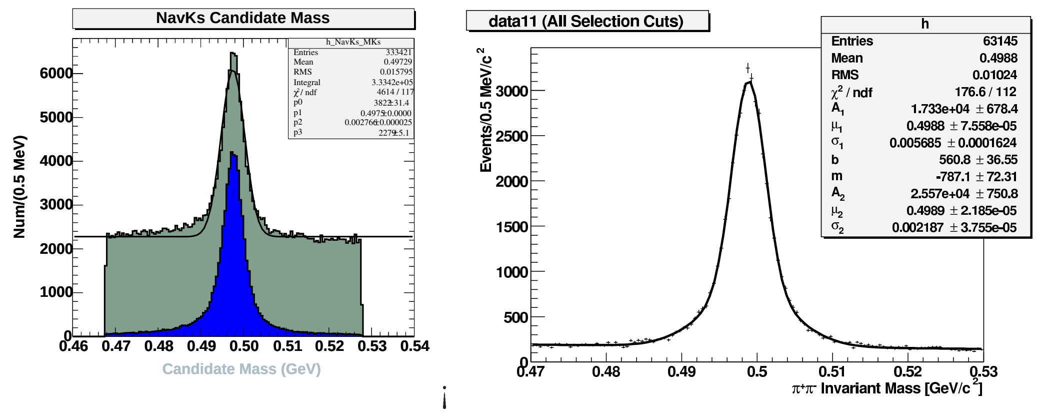


Figure 5.7: Invariant mass $M_{\pi^+\pi^-}$ of candidate $K_S^0 \to \pi^+\pi^-$ decays. Left: distribution for the loose K_S candidates identified during event reconstruction, from a $B\overline{B}$ Monte Carlo sample. The "correct" K_S candidates (as gauged from truth-matching information) are highlighted in blue. Right: the K_S mass peak in a sample of CLEO III data after all selection cuts except for the mass cut are applied [103]. The fit on the left employs a Gaussian with a flat background; on the right, a double Gaussian and a flat background.

of the neutrino, since the momentum of the lost particles is simply absorbed into the missing momentum four-vector.

Another category of tracking mishaps occurs when *additional* tracks, beyond those corresponding to real particles, are reconstructed in an event. Although these additions happen in numerous ways, their existence reflects the general reconstruction philosophy that too many is better than too few: there is some hope of discarding extra tracks during the analysis stage, but it is difficult to recover particles for which tracks were never found. Neutrino reconstruction is particularly affected by such tracking mistakes because it relies on tracking performance across the entire event. To respond to these need for a general-purpose track filter, a CLEO software package called Trkman was developed in the mid-1990's, intended primarily for use in neutrino reconstruction. Its purpose is to sift through the default list of tracks found in an event and identify those that are redundant or otherwise spurious, and in the case that several tracks are determined to represent the same particle, it selects the track likely to represent the true particle's information most accurately.

The Trkman algorithms are described extensively elsewhere [93]; here we only review a few of the most common tracking pathologies and how they are addressed.

- **Curlers**
  Charged particles with small transverse momentum ($p_T \lesssim 225$ MeV) will complete a full circle in the 1 m radius of the tracking chambers without

reaching the stopping material of the calorimeter, and so can spiral many times under the influence of the detector's solenoidal magnetic field. This phenomenon is called a curler: the particle will loop back to the beam axis and out again several times until it stops due to energy loss or when it finally exits the chamber in the z direction. In any case, the long string of hits will appears as almost overlapping circles in the transverse plane, and the tracking algorithms will often fit each half of each circle as a separate track.[18] By comparing the z position, curvature, and azimuthal angle at the point of closest approach to the beam line for each semicircular track, it is possible to identify the tracks that belong to a common "curler group" and select the single outward-bound arc that best represents the original particle as it departed the origin for the first time.

On average, there is a curler in every other $B\overline{B}$ event [93], so even with the high curler identification efficiency attained with the use of Trkman, a considerable fraction of events still suffer from unidentified curlers. (Some of these events can be rejected by cuts on the net charge, but this costs further efficiency.) Even in events where curlers have been successfully identified and condensed into a single representative track, the large number of hits caused by even a single spiraling particle can obscure the hits from other tracks, leading to hit inefficiency and poor momentum measurements for the other tracks in the event.

- **Ghost Pairs**
  Ghost pair creation occurs when two tracks are reconstructed from a single set of hits. This "twinning" problem can occur when the initial estimate of the path of the particle as determined in the pattern recognition stage is rather poor, so that only some of the hits along the true flight path fall into the "road" within which hits are considered for addition to the track at fitting time.

  Ghost pairs can be identified by comparing tracks pairwise and examining their azimuthal separation and the pattern of hits on each. A typical ghost pair will have few layers in which both tracks have hits and many layers on which only one track has a hit, a natural consequence of dividing a single string of hits between two reconstructed tracks. (Note that hits are not shared between tracks in the CLEO tracking software.)

- **Scatters and Decays in Flight**
  Occasionally a particle will scatter by an appreciable angle off the gas molecules in the drift chambers or due to material interactions with the wires

[18] Recall that the basic track-finding assumption is that tracks originate from the IP at the center of the detector. One loop of a curler is essentially a circle tangent to the IP that is typically reconstructed as *two* tracks: a "front half" and "back half," one corresponding to the true outward-going and the other to the true inward-going trajectory of the particle.

or walls of the chambers. Alternatively, a charged pion or kaon may decay in flight to $\mu\,\nu_\mu$. In all of these cases, the initial trajectory of the particle is interrupted by a kink or vertex, and the subsequent particle(s) are not necessarily identical in nature or momentum to the incident one. These instances are generically called "scattered tracks," and are readily identified by the intersection of the "before" and "after" tracks, with the additional feature that only the "before" track will have hits inside the radius of the scatter, and only the "after" track will have hits outside it.

The inner track in these circumstances is the best available representation of the particle for the purposes of neutrino reconstruction.

Other tracking problems, such as "backsplash" tracks (when a particle interacts in the crystals and then it or some secondary re-enters the drift chamber volume), "low density" tracks (formed from the collection of random hits in a region where a curler has created an abundance of hits), and "tails" (created when a backsplash track curls close enough to the IP to be reconstructed, but as two separate tracks) are also identified and classified by the package. In addition, a separate list of tracks to use in identifying showers matched to charged particles is prepared. The entries in this list are selected with slightly different criteria, since the intention is to identify those tracks that represent the path of real particles at the calorimeter radius rather than at the IP.

Approval by the Trkman package is the most important requirement for tracks used in neutrino reconstruction, but in this analysis we apply a few other standard requirements as well. The final "good" track selection criteria are as follows:

1. Approval by the Trkman package.
2. Good z-fit information. The missing momentum can't be calculated reliably without a trustworthy measure of p_z.[19]
3. No part of any of the K_S decays earmarked for use in the observed momentum sum.

5.3.6 Shower Selection

Most particles, charged and neutral alike, deposit energy as they traverse the crystals of the calorimeter. Not all reconstructed showers, however, are relevant when reconstructing a neutrino. Showers caused by the interaction of charged particles must be removed from consideration, since the momentum of these particles

[19] For CLEO II and II.V data, this requirement is implemented by rejecting tracks with `kincd == -1, -3`.

has already been accounted for based on tracking information from the drift chambers. Neutral hadrons such as neutrons and K_L's produce showers via hadronic interactions in the crystals, but in these cases, the reconstructed shower energy is not a good indicator of the hadron's real energy; see Sec 4.3.2. The observed momentum sum in events with such particles is inevitably biased by the less-than-complete containment of the neutral hadronic showers, rendering neutrino reconstruction hopelessly ineffective; we have to rely on neutrino consistency checks to suppress such events.

The philosophy behind shower selection is to avoid double-counting of "charged" energy (crystal energy attributed to charged particles) as much as possible, rather than to exclusively identify photons with high efficiency. This goal is consistent with the broader neutrino reconstruction effort to include as much of the information reconstructed in an event as possible. Neutrinos are not reconstructed effectively if only the best-looking showers and tracks in an event are employed.

The basis for rejecting most charged particle showers is a set of track-matching algorithms developed over years of experience with the CLEO calorimeter. Although the process is quite complex, the general idea is to label a shower as "matched" to a track if the projection of the track onto the face of the crystals from the last layers of the drift chamber is "near" the shower in a three-dimensional sense. For electromagnetic showers, which have well-understood transverse and longitudinal profiles, a simple matching procedure is very efficient. The centers of hadronic showers, however, can be significantly offset from the projections of the tracks that cause them, and a looser type of matching must be employed to catch these instances. Studies of simulated data show that almost 90% of the energy deposited by charged particles is quite effectively eliminated this way, but about 3% of the energy from photons is also rejected due to chance overlaps between charged tracks and photon showers.

Charged hadrons can also produce more complicated artifacts in the calorimeter known as "splitoff" showers. These showers are created by secondary particles from the initial hadronic interaction, such as π^0's or nuclear spallation byproducts, that travel some distance laterally through neighboring crystals before depositing their energy, leading to the production of a secondary shower that may be separated or "split off" from the primary shower. The secondary shower may not be near the initial track projection but should properly be associated with the track and not counted as additional neutral energy in the calorimeter. Accurate neutrino reconstruction relies on correctly accounting for each particle's energy exactly once, and hence on identifying and then rejecting splitoff showers, but without mistaking real photons.

Identifying splitoffs is made challenging by the unpredictability of hadronic interactions,[20] but there is useful information in the separation between the parent

[20]If splitoff rejection sounds hard, try running the 15 miles across Mt. Pleasant on a

and daughter showers, their relative orientation, and features of the pattern of energy deposition in the secondary shower. A sophisticated shower evaluation package has been developed to employ this information for effective identification of splitoffs. The package, called *Splitoff*, is described in some detail in Appendix B, as well as Ref [74]. Fig 5.8 illustrates the basis for the splitoff assessment decision for a candidate splitoff shower that is found near a matched parent shower. Four scaled energy variables are computed for the candidate splitoff shower from the energy distribution in the 3×3 block of crystals centered on the most energetic crystal of the shower. The variables are constructed by summing the energies in the rows and columns of the block and normalizing by the total energy in the block. The sums are then classified by their orientation and distance relative to the hypothetical line of flight connecting the parent shower to the candidate shower. For splitoff showers, the energy deposition tends to be along the line of flight and closer to the parent shower, while for photons, the distribution should be more symmetric.

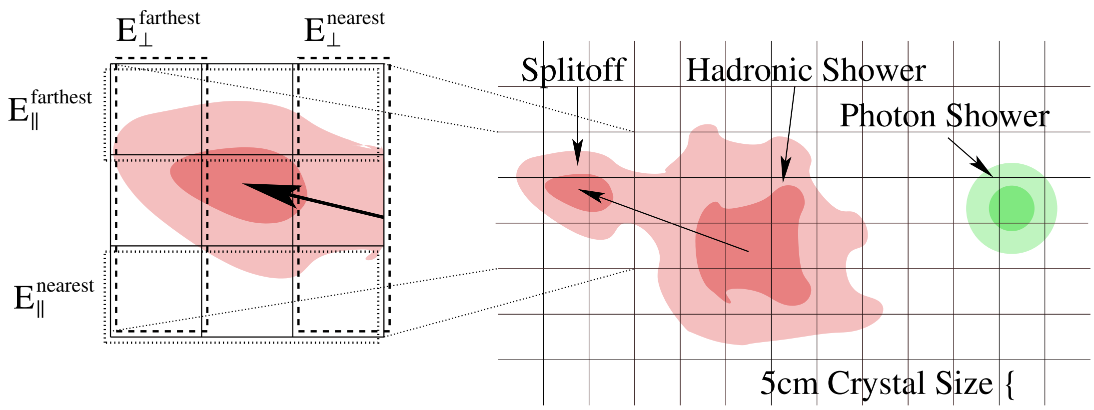


Figure 5.8: A cartoon illustration of a non-isolated splitoff shower, which has a possible track-matched parent hadronic shower within a cone of 25°. Scaled energy variables are constructed from the central 3×3 block of crystals in the splitoff shower as shown. The $\perp$ and $||$ notation refers to the orientation of the sum direction relative to the hypothetical line of flight from the parent shower. The two $\perp$ sums are expected to provide useful discrimination in splitoff determination. The figure also illustrates the differences in shower "shape" between hadronic and electromagnetic showers. Figure from Ref [78].

One of 26 different neural nets is then selected to combine these scaled energy sums with other information, including the separation angle between the two showers, the relative orientation of the splitoff candidate about the parent, and the ratio E_9/E_{25} of the energy in the central 3×3 block of crystals in the splitoff shower to the energy in the larger 5×5 block. The separate neural nets handle

wintry Sunday morning.

barrel versus endcap showers in different bins of splitoff shower energy.

For isolated showers that have no nearby track-matched parent shower, patterns of energy deposition can still suggest whether the initiating particle came from the origin or not. A similar procedure is used to calculate analogous inputs for another set of specially-trained neural nets.

The output of each two-layer net is subjected to a cut that varies with the energy of the splitoff candidate, its position, the quality of the best π^0 combination of which it is a part, and the species of the track that is matched to the parent shower.[21] For isolated showers, the cut is different but essentially just as sophisticated. An important feature of the design is that the neural nets themselves are independent of absolute shower energy, and any possible energy-dependent discrimination appears in the choice of cut placed on the net output.

Monte Carlo simulations of semileptonic B decays have shown that on average, almost 200 MeV of the energy that can be attributed to splitoffs is correctly rejected by the package, at a cost of less than 50 MeV of energy from true photons [78].

The *Splitoff* package is designed particularly for neutrino reconstruction efforts, and as such, the showers in the observed momentum sum are simply those approved by the package, with no additional requirements. Note that in addition to assessing the splitoff character of a shower with the use of neural nets, *Splitoff* also applies other internal cuts to discard showers less than 25 MeV[22] and those already matched to charged particles based on the standard track-shower matching algorithms.

A more detailed breakdown of the performance of shower selection criteria is shown in Table 5.3.

The shower selection criteria just described are designed to filter the list of reconstructed showers for those that most likely correspond to particles not observed in any other way in the event. We note here that even in the successful cases, when the selected showers do correspond in a 1:1 fashion with the neutral particles in the event, other factors still limit the accurate reconstruction of the total neutral energy. A reconstructed shower may not reflect the full energy of

[21] Hadrons are far more likely to produce splitoff showers than leptons. The splitoff decision incorporates this fact by applying a neural net cut that depends on the species of the track matched to the parent shower of a splitoff candidate. Internally, the relevant distinction is whether the track is an electron, muon, or hadron. The results of lepton identification performed elsewhere in the analysis (see Sec 5.4) are provided to *Splitoff* for use in choosing the proper cut; every track not positively identified as a lepton is considered a hadron. The splitoff decision is extremely insensitive to lepton identification simply because pions outnumber all other species by almost 10:1 in a typical event; independent tests have confirmed this result [105].

[22] Most of these low-energy showers are the result of noise, unidentified splitoffs, or other charged particles.

Table 5.3: Dissection of shower classification accuracy. Using a sample of 100 K generic $B\overline{B}$ Monte Carlo events, the classification of calorimeter energy and its true sources is cataloged. The figures show the mean energy per event that falls into one of three (exclusive) categories: matched to a track, rejected by *Splitoff*, and a third catch-all category. For each classification, the true source of the shower is shown, broken down into the cases where it is matched at truth-level to a charged or neutral particle. If matched to a charged particle, the shower is further classified as to whether a corresponding track was reconstructed. For showers caused by neutrals, the categorization is simply by the type of particle that initiated the shower. From Ref [105].

	$\langle E_{\mathrm{CC}}\rangle_{\mathrm{evt}}$ (GeV)			
	Tagged to Charged		**Tagged to Neutral**	
Shower Classification	Found	Not found	Photon	Other
Matched to track	2.333	0.061	0.135	0.020
Rej. by *Splitoff*	0.200	0.013	0.044	0.033
Other	0.268	0.109	2.117	0.209

the incident particle, or as remarked previously, a particle may not shower at all. Showers may be incompletely contained in the crystals, as already mentioned, particularly in the case of hadronic interactions, but crystal clustering algorithms can also miss some of the energy deposited around the periphery of a reconstructed shower. In addition, the detector geometry, while nearly projective in the barrel region, includes overlap regions at the barrel-endcap boundaries, where showers are often spread between two different segments of the detector and are both poorly modeled and poorly reconstructed. Energy resolution in the endcap is further degraded by a non-projective geometry and the presence of a significant amount of material in front of the crystals. In general, intervening material can initiate early showering that can dilute the amount of energy that reaches the crystals by either re-scattering it or spreading it over a larger area where individual crystal-by-crystal energy thresholds can begin to cut into the measurement of the total energy. Ultimately, we rely heavily on the accurate modeling of the detector material, clustering efficiency, and energy resolution in simulated data to account for these (and other) effects.

5.3.7 Particle Identification

To carry out neutrino reconstruction, particle identification (PID) assignments must be made for every acceptable track in the event; the assigned mass hypothesis is required to compute the associated particle's energy from the measured track momentum. In the CLEO II and II.V detectors, discrimination between the different possible hadron species (π, K, or p) is made using information from two independent sources: specific ionization (dE/dx) measurements made in the drift chambers, and time-of-flight (TOF) measurements made in the TOF scintillator system. These systems are described in more detail in Ch 4.

Standard event reconstruction routines compute a likelihood or Gaussian significance $Z(h)$ that a track is consistent with being of hadron species h. To make an overall best-guess PID assignment for a track, the probabilities from both detector systems are combined, and the result is weighted by relative particle abundances, using a technique described below.

The particle identification information available from each detector is combined using a χ^2 variable for each track,

$$\chi_i^2(h) = Z_{dE/dx,\,i}^2(h) + Z_{\mathrm{TOF},\,i}^2(h), \tag{5.13}$$

where $Z_i(h)$ are the species-dependent PID probabilities for track i to be of species $h = \pi, K$, or p, as measured with dE/dx or TOF information. This combination treats each detector measurement as an assessment of the Gaussian probability that the particle is consistent with being a hadron of species $h = \pi, K$, or p. (The detector PID quantities are constructed in such a way to make the Gaussian assumption a reasonable one.) If the information from one detector is not available due to acceptance or nominal quality cuts,[23] it is not included in the χ^2 sum.

A probability $P(h)$ that a given track is truly a hadron of species h is computed by translating the $\chi^2(h)$ variable into a Gaussian probability, which is then weighted by a particle abundance fraction w_h for species h, as shown below:

$$P(h) = w_h\, P_{\mathrm{gauss}}\left(\chi^2(h)\right) \qquad h = \pi, K, p. \tag{5.14}$$

The weighted probabilities for each hypothesis are compared and the largest determines the PID assignment for the track. If no usable detector information is available, or some part of the probability calculation fails (for instance, if the total weighted probability is vanishingly small), the PID assignment defaults to the pion hypothesis. In cases where the track has been identified as part of a curler group or ghost pair by the Trkman package (see Sec 5.3.5), a salvage attempt is made to use PID information from another track in the same group. In spite of these

[23] We use information from barrel and endcap TOF counters, and require that the raw pulseheight information pass basic consistency and quality checks. In CLEO jargon, this amounts to a requirement that `tfidql == 0`; see Ref [89] for more information. We only employ dE/dx information when more than ten hits are used in the calculation of the specific ionization. Note that dE/dx quantities are only available for tracks with good z information, which is required for computing the three-dimensional path length of a track across a cell, an important factor influencing total charge deposition.

Herculean rescue attempts, however, events with tracks identified in this fashion are effectively eliminated from the analysis by other selection cuts; see Sec 5.6.[24]

The production fractions w_h estimate the relative multiplicities expected in $B\overline{B}$ events for pions, kaons, and protons, and depend on track momentum p, as shown in Fig 5.9. They have been previously determined from measurements made in $B\overline{B}$ Monte Carlo [104], and are parameterized in the analysis as follows:

$$w_p(p) = 0.02 \tag{5.15}$$

$$w_K(p) = \begin{cases} 0.277p - 0.107p^2 & p < 2.0 \text{ GeV} \\ 0.06 & \text{else} \end{cases} \tag{5.16}$$

$$w_\pi(p) = 1 - w_p(p) - w_K(p), \tag{5.17}$$

The use of species abundances ensures that the PID decision algorithm respects known features of generic B decay, and also increase the robustness of the PID decision algorithm, raising the degree of confidence required before a particle is labeled as anything other than a pion.

The analysis machinery has been designed so that lepton identification naturally takes precedence over hadron PID by ensuring that no attempt is made to identify tracks that have already been identified as leptons; see Sec 5.4.

The performance of the identification algorithm just described is characterized by the plots in Fig 5.10 showing the efficiency and fake rate with which each of the three hadron species (π, K, p) are identified in Monte Carlo.

We trust that the Monte Carlo simulation of PID information for the purposes of hadron identification is accurate to first order. Any deviations from data are evaluated as part of a PID systematic described in Ch 7.

5.3.8 Neutrino Resolution

The final resolution on the neutrino kinematics that we achieve with the above machinery is shown in Fig 5.11 and tabulated in Table 5.5. For convenience, we consider the resolution on the magnitude of the neutrino momentum separately from that on its direction; for completeness we also show the resolution on the neutrino energy when using the missing energy in the event as the estimator.

The plots in Fig 5.11 show the resolution on these quantities as measured in a sample of $b \to u\,\ell\,\nu$ signal Monte Carlo, for all events, and then for the subset of events that pass event selection cuts specified later in Sec 5.6. To emphasize the

[24]The only tracks used in this analysis are those that pass the selection requirements imposed by the Trkman package and that have a good z-fit; except for a few rare pathologies, it will never be the case that the bad half of a good Trkman track will have PID information while the good half does not.

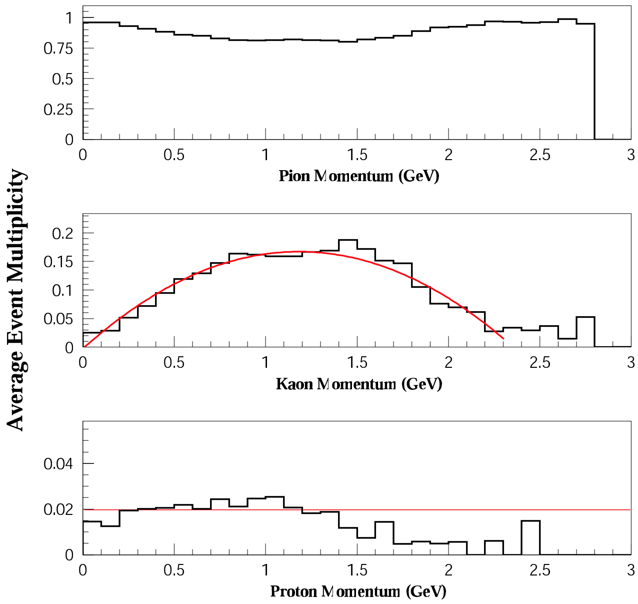


Figure 5.9: Relative particle abundances in $B\overline{B}$ events. The histograms shown are measured in a $(b \to u)$ Monte sample, and parameterizations previously obtained from generic $B\overline{B}$ Monte Carlo are overlaid in red. See the main text for the explicit formulation of these parameterizations. Pions clearly dominate the charged particle content of the event.

damage caused by the presence of undetected neutral particles such as K_L's, additional neutrinos, and neutrons, we also compare the resolution shape measured in events where these particles are absent at Monte Carlo truth level. It is immediately apparent that in these idealized events, the resolutions are significantly improved, and the shape reduces to a narrow and almost symmetric core.

Generally speaking, the resolution function has a roughly Gaussian shape with long non-Gaussian or "fat" tails. As discussed previously, many different event characteristics influence the ultimate neutrino resolution, from small and smoothly-varying effects such as finite detector resolution and particle identification mistakes, to more discontinuous problems such as missed particles or spurious tracks. All of these sources of error combine to determine the final shape of the resolution function. Separate studies of neutrino resolution indicate that the wide tails of the "all events" resolution distribution can be traced directly to those events that are contaminated by the presence of undesired neutrals. It turns out to be quite reasonable to model the neutrino resolution in a general event sample with a double Gaussian, where the mean and width of the outer Gaussian are expected to capture the bias and smearing effects of extra neutrinos and K_L's in the event, and the core Gaussian represents the contributions from the less dominant, smaller sources.

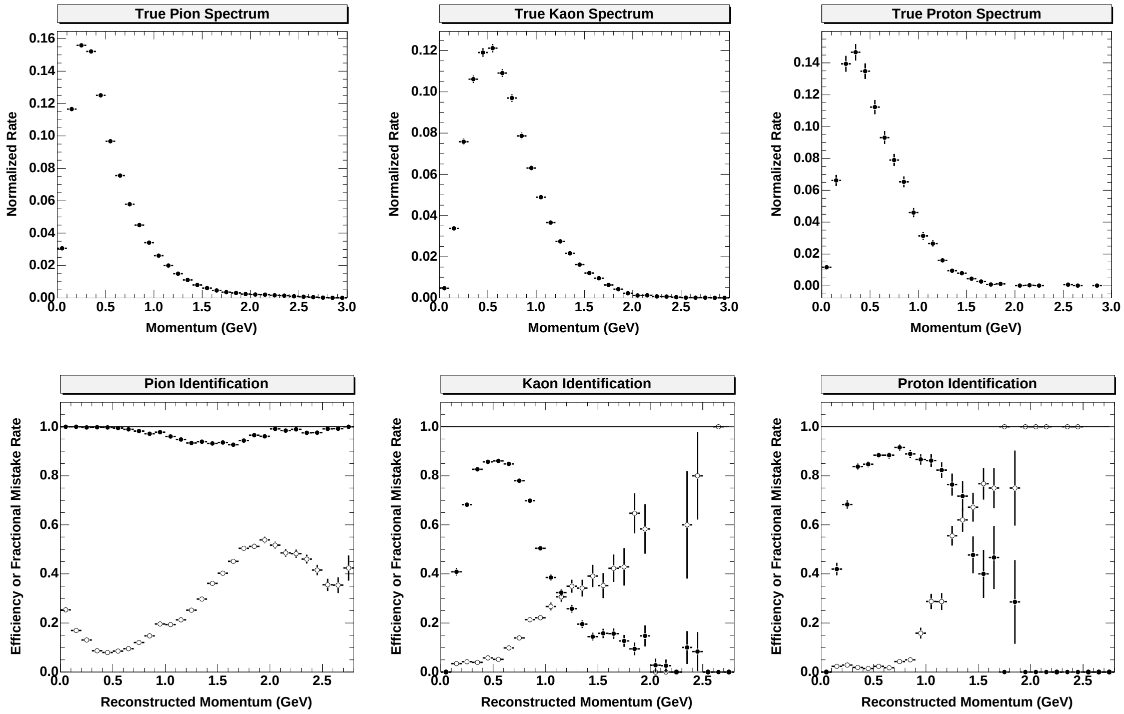


Figure 5.10: Hadron identification efficiency and fake rate for the algorithm described in the text, as measured in a sample of 50 K $b \to u$ Monte Carlo events. The top row of plots shows the (normalized) momentum spectrum for each species $h = \pi, K, p$. The second row shows efficiency in solid circles ($\bullet$) and mistake rate in open circles ($\circ$), as a function of momentum. Efficiency is defined to be the ratio of tracks correctly identified as species h that are matched to species h at generator-level. The fake fraction is taken as the impurity of the identified sample, *i.e.* the fraction of identified tracks that are not truly of species h; see Sec 5.4 for more discussion of these quantities.

One categorization of the contributions to the error Δp_ν on the magnitude of the neutrino momentum in events without additional neutrals is shown in Table 5.4. Here a cut on the v-ratio (defined in Eqn 5.7) has not been applied, allowing a more extensive probe of the sources of neutrino reconstruction errors.

Independent tests of the CLEO Monte Carlo have demonstrated that the simulation reflects the patterns in the data extremely well for a wide variety of distributions (*e.g.* v-ratio distributions), so we have confidence that the neutrino resolution as measured in Monte Carlo is a faithful representation of the actual resolution in data.

In Table 5.5, we compare numerical measures of the resolution under several different circumstances, and also include a comparison with a comparable sample of $b \to c\,\ell\,\nu$ Monte Carlo events. The mean and root-mean-square (RMS) are com-

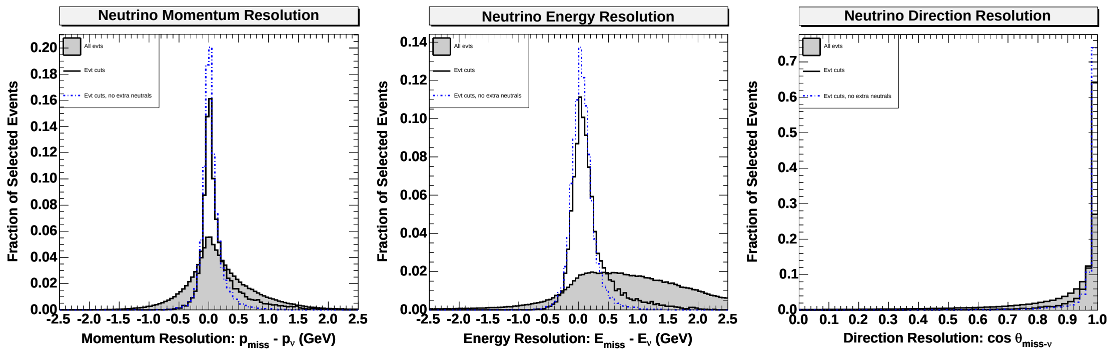


Figure 5.11: Neutrino resolution in a sample of 350 K $b \to u\,\ell\,\nu$ Monte Carlo events. The three plots show the resolution on the magnitude of the neutrino's momentum, its energy, and its direction. In the last case, the cosine of the angle between the missing momentum and the true neutrino direction is used to gauge the resolution. In each case, the shaded histogram shows the resolution distribution for all events in the sample; the open histogram in black outline shows the distribution for events that pass the nominal selection cuts described later (Sec 5.6), and the histogram in blue, dash-dot outline shows the resolution in those events that pass all event cuts and have no unwanted ν, K_L, or neutrons present at generator-level. The event cuts include a (reconstructed) lepton momentum cut of $|p_\ell| > 1.5$ GeV. For ease of comparing *shapes*, all histograms have been normalized to unit area.

Note the considerably better performance of the momentum estimator compared to the energy estimator. Both distributions demonstrate fat, non-Gaussian tails as well as an asymmetric excess on the high-side that can be traced to the fact that particles are *missed* more often than they are double-counted, leading to an over-estimate of the neutrino four-momentum. The non-Gaussian tails are well-represented by a second Gaussian resolution function taken from those events that contain extra undetected neutral particles.

Quantitative measures of the means and widths of these distributions, along with selection efficiencies for the cuts used here, are available in Table 5.5.

Table 5.4: A partial breakdown of the sources of error in neutrino reconstruction, based on a separate study of the missing momentum resolution in $b \to u\,\ell\,\nu$ events with no K_L or additional neutrinos, and all event cuts but a v-cut applied. From Ref [106].

Source	**RMS (GeV)**	**Core σ (GeV)**
Splitoff failures	0.213	0.223
Undetected photons	0.137	0.160
Undetected or multiply-counted charged particles	0.129	0.108
Finite resolution, PID, etc.	0.211	0.092
Quadratic sum	0.354	0.309

puted from all events in the sample that meet the specified cuts, and the quantity σ results from a simple Gaussian fit to the entire distribution, histogrammed as in Fig 5.11. The RMS resolution is sensitive to the extent of the long tails, while the resolution determined from the fit captures the width of the narrow core.

The resolution suffers in $b \to c\,\ell\,\nu$ events because the daughters of the semileptonic B decay in these events are charmed (and heavier) and are far more likely to yield additional neutrinos or K_L's than the corresponding signal-side daughters in $b \to u$ decays. (For instance, the X_c meson from $B \to X\,\ell\,\nu$ could decay semileptonically $X_c \to X_s\,\ell\,\nu$, yielding another neutrino; even if the decay isn't semileptonic, it is likely to include a $c \to s$ quark transition that could yield a K_L.) In particular, the fraction of events with extra neutrals present after the event cuts are applied is only 27% in the $b \to u\,\ell\,\nu$ sample, but much higher at 42% in the $b \to c\,\ell\,\nu$ sample. The consequence is that although $b \to c\,\ell\,\nu$ events survive the selection cuts with less efficiency than the signal, they are more likely still to be influenced by the presence of additional neutrinos and K_L's. Note further that these selection efficiencies do not reflect the actual relative abundance of these processes in the data, where $b \to c\,\ell\,\nu$ dwarfs $b \to u\,\ell\,\nu$ by almost two orders of magnitude before any cuts are applied.

We conclude our lengthy exposition of neutrino reconstruction with a few summary remarks about its efficacy and some foreshadowing of things to come. The bottom line is this: Neutrino reconstruction as a numerical exercise is almost trivial. Managing the inevitable physics, detector, and reconstruction challenges is where the real work lies, but these obstacles have been overcome with the the development of sophisticated reconstruction and filtering algorithms and an accurate, reliable detector simulation. These combined efforts make neutrino reconstruction a feasible avenue for exploring semileptonic B decays at CLEO [74–76, 78].

Table 5.5: Various measures of neutrino resolution, compared between Monte Carlo samples of $b \to u\,\ell\,\nu$ in (a) and $b \to c\,\ell\,\nu$ in (b). Δp_ν and ΔE_ν are measured in GeV. A 1.5 GeV lepton momentum cut is incorporated as part of the event cuts. Due to the one-sided shape of the $\cos\theta_{\Delta\nu}$ distribution, a Gaussian σ is not reported. The relative selection efficiencies for these cuts are recorded as well. The last portion of each table shows the resolution possible in the idealized case that no extra neutrals are present (*i.e.* $N_\nu = 1$, $N_{K_L} = 0$, $N_{n^0} = 0$). The $b \to u\,\ell\,\nu$ Monte Carlo sample has 350 K events, while the $b \to c\,\ell\,\nu$ sample has 482 K; both cover the CLEO II dataset in approximately the same proportions.

(a) $b \to u\,\ell\,\nu$ **Monte Carlo**			
Qty	**Mean**	σ	**RMS**
All evts ($\varepsilon_{\rm sel} = 100\%$)			
$p_{\rm miss} - p_\nu$	0.189	0.508	0.569
$E_{\rm miss} - E_\nu$	0.751	0.943	0.889
$\cos\theta_{{\rm miss}-\nu}$	0.794	–	0.334
Evt cuts ($\varepsilon_{\rm sel} = 3.5\%$)			
$p_{\rm miss} - p_\nu$	0.157	0.145	0.351
$E_{\rm miss} - E_\nu$	0.189	0.194	0.366
$\cos\theta_{{\rm miss}-\nu}$	0.949	–	0.141
Evt cuts, no extra neutrals ($\varepsilon_{\rm sel} = 2.5\%$)			
$p_{\rm miss} - p_\nu$	0.043	0.118	0.186
$E_{\rm miss} - E_\nu$	0.070	0.165	0.208
$\cos\theta_{{\rm miss}-\nu}$	0.968	–	0.102

(b) $b \to c\,\ell\,\nu$ **Monte Carlo**			
Qty	**Mean**	σ	**RMS**
All evts ($\varepsilon_{\rm sel} = 100\%$)			
$p_{\rm miss} - p_\nu$	0.282	0.510	0.584
$E_{\rm miss} - E_\nu$	0.874	1.029	0.921
$\cos\theta_{{\rm miss}-\nu}$	0.665	–	0.426
Evt cuts ($\varepsilon_{\rm sel} = 1.4\%$)			
$p_{\rm miss} - p_\nu$	0.252	0.211	0.415
$E_{\rm miss} - E_\nu$	0.294	0.240	0.427
$\cos\theta_{{\rm miss}-\nu}$	0.874	–	0.244
Evt cuts, no extra neutrals ($\varepsilon_{\rm sel} = 0.8\%$)			
$p_{\rm miss} - p_\nu$	0.075	0.138	0.215
$E_{\rm miss} - E_\nu$	0.108	0.180	0.233
$\cos\theta_{{\rm miss}-\nu}$	0.926	–	0.179

However, disentangling the minority $b \to u\,\ell\,\nu$ decays from the reigning background of $b \to c\,\ell\,\nu$ remains a challenge. Neutrino reconstruction as a technique does slightly favor events with $b \to u\,\ell\,\nu$ decays, simply because the signal B in these events is much less likely to produce undetectable particles that steer the missing momentum away from the kinematics of the true neutrino. But we've seen that the $b \to c\,\ell\,\nu$ events that do reconstruct successfully often still suffer from the troubles brought on by the presence of unobserved neutrals. The implication is that although a neutrino has been "identified" in these events, it is likely to be significantly biased (high) relative to the truth. In later sections, we shall see that the direction of this skewing acts to make $b \to c\,\ell\,\nu$ decays *further* resemble $b \to u\,\ell\,\nu$. While the discussion of data samples in Sec 5.2 made it clear that we have the statistics to easily afford the low efficiencies of neutrino reconstruction, it now becomes clear that to avoid being limited by systematic issues and too much reliance on the details of detector modeling, it will be essential to prepare an analysis that utilizes information directly from data as much as possible.

5.4 Lepton Identification

The presence of a high-momentum lepton[25] in a $B\overline{B}$ event is an almost ubiquitous signature of semileptonic B decay. As such, if neutrino reconstruction forms one pillar of this analysis, lepton identification surely forms the other. The combination of a well-reconstructed lepton and neutrino forms a powerful indicator of semileptonic decay, and offers enough kinematic information to allow for the calculation of kinematic quantities that can be used to further discriminate against $b \to c\,\ell\,\nu$.

Identified leptons play a second role in the analysis, however, beyond hinting at the presence of the signal semileptonic decay. Recall that neutrino reconstruction relies on the assumption that the only missing particle in the event is the signal neutrino, else the identification with the missing momentum fails completely. With few exceptions, neutrinos are always produced in conjunction with leptons, so the presence of additional leptons in an event can signal the presence of additional neutrinos.[26] In the analysis, then, we refer to two types of leptons:

[25] Unless otherwise explicitly stated, we use "lepton" to refer to the *charged* leptons e, μ, or τ in the Standard Model. When reasonably clear, we will occasionally also include neutrinos in the category of leptons, *e.g.* in a discussion of the "dilepton" pair in a semileptonic decay.

[26] At the $\Upsilon(4S)$ there are few particles that disappear entirely in a puff of neutrinos without leaving behind a signature lepton, but the reverse statement is far from true: leptons can result from annihilation decays such as $J/\psi \to \ell^+\ell^-$ and are produced copiously in material interactions (*e.g.* photon conversions $\gamma \to e^+e^-$), all without any associated neutrinos. Since the leptons are being used as a veto, however, this fact only results in a reduction in signal selection efficiency rather than an avenue by which fakes

- *Signal leptons* are high-momentum tracks expected to be from semileptonic B decay. Identifying a signal lepton is an indication that the event is signal and should be kept.

- *Veto* or *counting leptons* are leptons that are identified in addition to the signal lepton in an event, and serve as a basis for rejection of an event. To effectively eliminate other semileptonic decays, these lepton candidates are identified across as broad a momentum range as possible

In practice, tracks meeting the looser criteria of the second category are identified first. A set of tighter cuts are used to select signal leptons from this list. The event is rejected unless a signal lepton is identified; even if a signal lepton is found, the event is still rejected if any other leptons have been identified.

There are several ways to characterize the performance of any particle identification algorithm. To avoid later confusion, we spell out three quantities used in this work. *Efficiency* measures the success rate of an algorithm as the ratio of the number of tracks correctly identified as a particular species to the number that truly are of that species:

$$\epsilon \equiv \frac{N_{\rm correct}}{N_{\rm true}}. \tag{5.18}$$

Clearly we have $0 \leq \epsilon \leq 1$. Efficiencies can be measured readily in simulated data with the use of truth-matching information, which records the generator-level particles that contributed to each reconstructed track. In this case, the electron efficiency, say, is just the ratio of (the number of true electron tracks that are identified as electrons) to (the total number of true electron tracks), where the "truth" about a track is determined from Monte Carlo truth-matching or "tagging" information recorded during the creation of each simulated event. Efficiencies can also be directly measured in data by taking advantage of situations where other constraints can be used to determine the identity of the particle without applying the particle ID algorithm being tested; the success rate of the package on this sample clearly also measures the efficiency as defined above, since the denominator is determined independent of the PID package. For electrons, a popular track sample is that from radiative bhabha events $e^+e^- \to e^+e^-\gamma$, where one of the electrons loses some energy by radiating a photon. In fact, this environment provides a natural arena for measuring efficiency across a range of electron momenta.

Measures of the failure rate of a particular algorithm are more numerous. Here we define two: *(im)purity* and *fake rate*; the fundamental difference is merely the choice of denominator for each quantity. *Purity* $\wp$ measures the rate at which tracks are correctly identified, relative to the total number of identified tracks. It provides, in essence, the fractional "cleanliness" of the identified sample. The corresponding

---

can enter. The lepton–neutrino correlation is one of the few handles we have on extra neutrinos.

impurity ι records the analogous fractional contamination of a sample of positively identified tracks:

$$\iota \equiv 1 - \wp \equiv \frac{N_{\text{fake}}}{N_{\text{found}}}. \tag{5.19}$$

Again, it is obvious that $0 \leq \iota \leq 1$. The chief advantage of this quantity is that it immediately conveys the quality of any sample of identified tracks.

Fake rate, on the other hand, describes the probability that a track of a different species will be identified as the target species, *e.g.* the rate at which pions fake electrons:

$$f_e^{\pi} = \frac{N^e_{\text{true}\,\pi}}{N_{\text{true}\,\pi}}, \tag{5.20}$$

with $0 \leq f_e^{\pi} \leq 1$. In Sec 5.8, we will see that this latter quantity is useful when measuring the contribution to the analysis of events with mistakenly identified or "fake" leptons. The disadvantage, however, is that the natural relative abundances of the faking species and target species are not factored into this quantity. If the proton fake rate f_e^p is 50%, but electrons outnumber protons by 10:1, a sample of identified electrons will still be largely free of protons.[27]

Lepton candidates must pass some basic selection criteria to help ensure that they correspond uniquely to real particles and that they are well-measured. To this end, both electron and muon candidates must meet the following requirements:

- Approval by the Trkman track-filtering algorithm
- Have a good track-fit in the z dimension
- Pass a momentum cut
  - $|p_e| > 0.2$ GeV for electrons
  - $|p_\mu| > 1.0$ GeV for muons

The minimum momentum cut corresponds to the threshold at which useful efficiency for identifying the respective flavor turns on.

Electron candidates are also required to have a track hit fraction of at least 0.40. The hit fraction measures the fraction of layers in the tracking chambers traversed by the track that have hits reconstructed on the track. Past experience has shown that such a cut significantly reduces electron fake rates with little impact on efficiency [109].

Since electrons and muons leave quite different signatures in the detector, separate criteria are used to make the final lepton identification decision.

[27]This story rests on the assumption of reasonable electron efficiency. If true electrons are found rarely, chance-identified protons could be a significant fraction of the sample of found electrons.

5.4.1 Electron Identification

The most powerful variable for identification of electrons is the ratio of the energy deposition in the calorimeter to the momentum obtained from the drift chamber. Since electrons interact electromagnetically in the crystals, their showers are fully contained, and essentially all of the incident energy is transformed into scintillation light, making the approximation $E_{\mathrm{CC}} \approx E_e$ an extremely good one. And because at CLEO they are ultra-relativistic,[28] $E_e/p_e \approx 1$, implying that E/p as measured by the calorimeter and tracking chambers will indeed be about unity for these particles. In marked contrast, muon interactions in the crystals are dominated by much smaller ionization losses, and leave behind very little energy; so for muons, E/p is much less than one. Hadrons have much less predictable patterns of energy deposition but rarely give up all of their energy within the calorimeter. E/p for their showers is often less than unity, but can easily fluctuate high and even exceed one at CLEO.

Several other variables are available to combat the problems resulting from the E/p fluctuations of hadrons. These additional variables are combined with E/p into a likelihood ratio $\mathcal{L}$ as shown below.[29]

$$\mathcal{L} = \frac{\mathcal{P}(x|p,e)}{\mathcal{P}(x|p,h)} \tag{5.21}$$

Here $\mathcal{P}(x|p,H)$ describes the probability that a track of momentum p and species H leaves a detector response x. The detector signature vector x for a candidate electron track consists of the following information:

- E/p
- dE/dx information: the significance $\chi_e^{dE/dx}$, characterizing how consistent the track's ionization trail in the drift chamber is with the electron hypothesis
- w, a lateral shower shape variable; basically, the second moment of the shower energy about the shower center
- Δ_ϕ and Δ_θ, two angular track-matching variables that measure the quality of the match between the track projection and the shower center

[28]Recall typical track momenta in a $B\overline{B}$ event are $\mathcal{O}(250$ MeV$)$ or larger; in fact, we make no attempt to find electrons below 200 MeV. Any identified electron will thus have an energy E_e several orders of magnitude above its rest mass $m_e = 0.51$ MeV.

[29]It is well-known that a likelihood formulation provides the most effective identification possible for a selected set of variables. With this claim, we mean that in a world in which the distributions of the fundamental variables are known perfectly, a likelihood variable provides the highest efficiency for a given fake rate, compared to any other set of cuts or selection procedure involving the same variables.

In practice, the likelihood ratio is computed under the assumption that it can be factored into independent likelihoods $\ell_i(x_i|p,H)$ for each variable x_i. The probability distribution function (pdf) $\mathcal{P}_i^H(x_i)$ for each variable is obtained from shapes measured in real data. For electrons, the data sample consists of radiative bhabha events. The chief backgrounds are from pions and kaons, $H = \pi^{\pm}, K^{\pm}$. Pions and kaons are selected from a sample of D^*-tagged $D^0 \to K - \pi^+$ decays. Further fake rates studies also employ pions taken from clean $K_S^0 \to \pi^+\pi^-$ decays. To keep the problem manageable, the pdfs are determined in bins of electron momentum. In the bins where both pion and kaon fake rates are important, the pdfs are combined in a weighted fashion with coefficients tuned to offer the best discrimination in data.

A track is identified as an electron if it passes a cut on the value of the likelihood evaluated for the given track's detector signature x. The actual cut depends on both the track momentum and the location of the corresponding shower in the calorimeter. Because shower information is an important ingredient in the electron likelihood, the performance of the calorimeter in different angular regions (central barrel, good endcap, overlap region, etc.) cannot be neglected. The precise value of the cut for each bin is chosen to keep the pion and kaon fake rates acceptably small.

A likelihood-based approach to electron identification at CLEO was developed in the early 1990's, but only covered the range $p > 600$ MeV [108]. Since lepton identification across the full momentum spectrum is crucial to the multiple lepton veto, a new package known as `REId` was developed in 2002 to extend the coverage down to 200 MeV[107]. The new module re-implements the old algorithm at high momentum, but incorporates the use of another shower shape variable E_9/E_{25} and new pdfs for the new momentum region. The new likelihood cuts are chosen to limit the pion fake rate to 0.2% in the central or good barrel (GB) region ($|\cos\theta| < 0.71$), and 1.0% in the remainder of the detector (also known as the non-good barrel or NGB).

To reduce kaon backgrounds, a veto based on TOF information is used for tracks with momentum $0.4 < p < 1.0$ GeV. The requirement for electrons is $(\chi_e^{\rm TOF})^2 - (\chi_K^{\rm TOF})^2 \leq 10$. This additional veto is required because kaons in this momentum range look particularly like electrons. Around 900 MeV, they stop in the calorimeter material, leaving a shower with $E/p \sim 1$, and at the same end of this momentum interval, their specific ionization merges with that for electrons. Time-of-flight information, however, still offers good K/e discrimination; recall Fig 4.15.

The performance of the `REId` package is shown in Fig 5.12 which depicts the final efficiency as a function of (reconstructed) momentum, measured in data. Determining this efficiency accurately is critical since it must be incorporated into the Monte Carlo simulation, as described in Sec 5.4.4 below. Although the efficiency can be determined in an independent sample of radiative bhabhas with extremely high statistical precision, these numbers turn out not to represent the actual efficiency in $B\overline{B}$ events very well, due to significant differences in event environment. The input variables to the EID likelihood are—to differing extents—sensitive to the chance overlaps with other particles that become more likely in the crowded environment of a $B\overline{B}$ event. For example, overlaps can degrade dE/dx resolution when two tracks pass through the same cell, and they can distort shower shapes when adjacent showers are mistakenly merged. To correct for these effects, a specialized "embedding" technique is employed to measure the electron efficiency in $B\overline{B}$ events directly;[30] more specifically, it measures the impact event environment has on electron efficiency, and these corrections are then applied systematically to the radiative bhabha results. Put briefly, the embedding procedure lifts radiative bhabha tracks and showers from data events and inserts them into other hadronic events, also from real data. The efficiency can then be readily measured, since the "true" electron track is known.[31] These studies reveal a significant downward shift in electron efficiency of several percent, due solely to event environment effects, also indicated in Fig 5.12 by the shift from the radiative bhabha curve down to the hadronic data curve.

5.4.2 Muon Identification

The muon chambers (see Sec 4.3.2) are the obvious and sole source of muon identification information in the analysis. Recall that with this device, muons are identified by their chief feature: deep penetration into layers of dense iron absorber. The penetration depth is a function of muon momentum: the softest muons ($|p_\mu| < 1$ GeV) are blocked by the considerable material presented by the crystals, superconducting coil and cryostat, and the first layers of iron, while stiffer tracks will be able to reach the first layers of counters or even further.

In general, muons are identified by matching the hits recorded in the muon detector superlayers to the projections of drift chamber tracks. A muon hit (strip or wire) is matched to a track if the track projection is close to the location of the hit, given the resolution of the device ($\sigma \gtrsim \mathcal{O}(1$ cm$)$). A layer (strip and wire) is

[30] Conversion electrons arising from the process $\gamma \to ee$ are also a nice source of electrons in hadronic events, but identifying these decays cleanly can be difficult. Recently, it has been proposed to use photons produced in $\pi^0 \to \gamma\gamma$ decays, where one γ subsequently converts. This idea has been put to effective use in tuning a natural CLEO III extension of the `REId` package [107].

[31] There are additional complications to working with radiative bhabhas that this embedding process does not eliminate. In particular, if the radiated photon is close to the final electron direction, its shower can merge with the electron's and skew the E/p measurement of the combined shower. This particular problem can be mitigated with cuts on some other properties of the electron kinematics, but the moral remains: measuring something extremely accurately is extremely complicated.

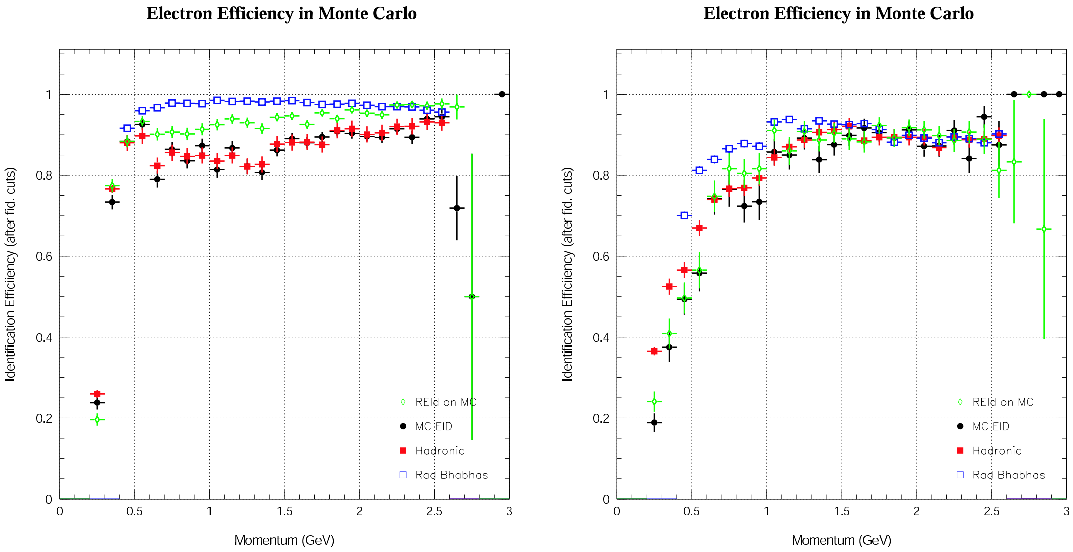


Figure 5.12: Electron identification efficiency of the `REId` package as a function of track momentum, for showers in the good barrel (GB) [left] and non-GB [right] regions of the calorimeter. Efficiency measured in radiative bhabha events is shown in open blue squares (□), and the true efficiency in hadronic events as determined by embedding studies in data is shown in solid red (■). The performance of the data-tuned package on $B\overline{B}$ Monte Carlo is shown in green (◊) as a check of the simulation of the various EID quantities. To eliminate the evident data-MC systematic, the Monte Carlo is forced to deliver the data-determined efficiency; the success of this implementation is shown by the agreement between the solid red squares (data) and the solid black circles (MC). The error bars indicate statistical errors in all cases.

matched based on a two-dimensional χ^2. Finally, a superlayer is designated as "hit" if two of its three layers have hits. For each track projected into the muon chambers, an expected penetration depth is computed based on the track momentum. Hits are then expected in all (active) counters located at depths smaller than and matching this estimate. For each track, the maximum penetration actually achieved is also computed. The basic definition of a good muon is that the track reached the depth expected for its momentum, and all inner superlayers have hits.

Taking into account the momentum-dependent penetration, we identify muons with the following criteria:

- $|\cos\theta| < 0.85$, respecting the geometric coverage of the detector
- For tracks with $|p| > 1.20$ GeV:
  - Penetration depth $d_\mu \geq 5$ interaction lengths
  - Hits are found everywhere along the track's projection where they are expected, *i.e.* in all superlayers of depth equal to and less than the expected depth.[32]
- For tracks with $1.0 \leq |p| < 1.75$ GeV:
  - Penetration depth $3 \leq d_\mu < 5$ interaction lengths
  - Superlayers at depths less than the calculated penetration depth are allowed to have up to two (of three) hits missing per superlayer.[33]

The second set of criteria effectively identifies softer muons for the purposes of the multiple lepton event veto.

Fig 5.13 displays the resulting muon identification efficiency using this algorithm. The efficiencies shown are measured in simulated events, but they compare reasonably well with independently-measured efficiencies from radiative "mu-pair" events, $e^+e^- \to \mu^+\mu^-\gamma$.

5.4.3 Signal Lepton Criteria

Signal leptons are selected from the lists of electrons and muons just described. Since these are to be the leptons produced in the $b \to u\,\ell\,\nu$ decay and play a critical role in the rejection of $b \to c\,\ell\,\nu$ background, the requirements are significantly more demanding. Briefly, the criteria are:

---

[32] CLEO-educated readers will realize that this cut is implemented by the oblique statement that `muqual == 0`. It is an unfortunate truth that even in the new software environment of object-oriented C++, the muon detector jargon of bygone eras yet lingers.

[33] This requirement is equivalent to the even more impenetrable statement that, for these tracks, we require `(muqual % 10000) == 0`.

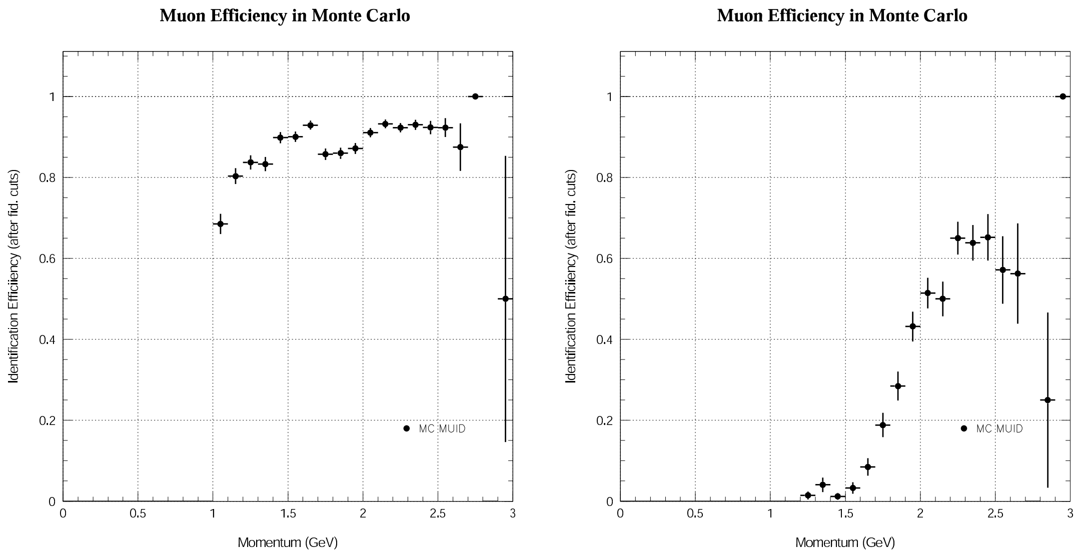


Figure 5.13: Muon identification efficiency as a function of momentum, measured in Monte Carlo. On the left, efficiency in the barrel; on the right, efficiency in the endcap region. The dip in the efficiency at 1.75 GeV marks the upper limit for the looser counting muon criteria. Error bars are based on the statistics of the measurement sample.

- The track must be already identified as a loose electron or muon.
- $|p_\ell| > 1.5$ GeV. This cut is set at 1.2 GeV during early skim stages, and is only raised to its final value of 1.5 GeV in the final analysis pass in preparation for fitting.
- Hit fraction $hf > 0.4$, ensuring that the drift chamber signature is based on a significant number of hits and undistorted.
- We compute the point of closest approach between the track and the beamspot and require the radial distance to satisfy $|d_{\rm bm}^{\rm PCA}| < 2$ mm, and the axial coordinate to agree as well: $z_{\rm bm}^{\rm PCA} < 5$ cm.[34] The radial impact parameter $d^{\rm PCA}$ is signed according to charge and curvature conventions; hence the $|\cdot|$.
- We require the RMS residual between the track fit and the reconstructed wire chamber hits to be less than 0.4 mm.[35] This quantity gives an overall

[34]The CLEO-indoctrinated reader will recognize these as cuts on the old roar quantity `DBCD` and a modified version of `ZOCD` that is relative to the beamspot rather than the origin. We re-compute the quantities at analysis time rather than resuscitating them from whatever CLEO II common block they're stored in.

[35]In the CLEO catechism, this is a cut on the CLEO II/II.V version of the quantity

assessment of the quality of the fit.

- We require that the z-momentum fit be of exceptional quality.[36]
- For CLEO II data, we require that the sum of hits in the PT and VD exceed four. This cut is inherited from previous inclusive neutrino reconstruction analyses and leads to another incremental improvement in track quality.

The intent is basically to keep only pristine tracks of the highest quality.[37]

Signal muons are subjected to slightly more conservative muon chamber criteria than those implicit in the muon criteria listed previously:

- Penetration depth $d \geq 5$ interaction lengths
- All hits expected in the muon chambers are found

Any slight reduction in efficiency resulting from the application of these additional cuts is assumed (and known) to be well-modeled in the Monte Carlo.

We scale up the signal lepton momentum in off-resonance events by the ratio of the on- and off-resonance beam energies. The lepton energy is then re-computed using the mass hypothesis assigned during lepton identification. For a more complete discussion of the purpose and application of these re-scalings, see the description of the continuum subtraction procedure in Sec 5.5.2.

5.4.4 Simulation of Lepton Efficiency

A significant portion of the analysis relies on the use of simulated data. However, relying on the Monte Carlo for accurate simulation of something as critical as lepton efficiencies and fake rates can cripple the analysis by making it sensitive to any systematic discrepancies between the efficiencies as modeled in the simulation and the true efficiencies in data. As we observed above in Fig 5.12 showing electron efficiencies in data and Monte Carlo, there is in fact a systematic spread between simulated and measured efficiency. Rather than trust the Monte Carlo on this point, we instead eliminate all fake leptons from the simulation and force the electron efficiency to match that observed in data.

`RESICD`.

[36]Only CLEO readers inculcated in the vast arcana of the roar era will recognize this as the requirement that `kincd` ≥ 0, ruling out the previously allowed value of `-2` which, in spite of a long history, actually has little real meaning. Nonetheless, we eliminate tracks that even look like they might smell funny.

[37]Yup, these tracks are the inscrutable, good-looking kind that you'd want your darling youngsters to hang out with.

A special implementation of lepton identification is used when analyzing Monte Carlo data. Muons are identified as those tracks that are matched to true muons and that pass all muon identification cuts. Thus, no fake muons are allowed in the simulation, but the efficiency of the muon identification criteria is assumed to be correct.

For electrons, efficiencies from data are enforced by using random numbers to accept/reject each track tagged to a real electron in the same proportion as the measured efficiency for that momentum interval. This method both guarantees the replication of the observed efficiency (to within statistical errors from the dice-throwing) and eliminates any possibility of fake electrons. The expected agreement appears in Fig 5.12 as the consistency between the electron efficiencies measured in hadronic data and the efficiencies from the specialized "MC EID" procedure just outlined.

The handling of fake leptons is significantly more complicated and will be described in a later section.

As a tentative consistency check, some of the radiative bhabha embedding measurements were repeated using the particular EID cuts applied in this analysis. The efficiency measurements reported for the long-established, high-momentum portion of the `REId` package were confirmed on an independent sample of hadronic events with radiative bhabhas embedded in them. Fig 5.14 compares the results of the new study to the standard results, with and without embedding corrections applied to the latter. The results are indeed consistent within statistical errors; there is no clear case for modifying the existing efficiency table.

We have now completed description of the core elements for an inclusive analysis of semileptonic B decay: neutrino reconstruction and lepton identification. Many pieces remain to be assembled, however, before a coherent understanding of the data can be attempted. We survey these other concerns in the remainder of this chapter.

5.5 Continuum Suppression

The source of the so-called "continuum" background was discussed previously: e^+e^- annihilation into $q\bar{q}$ final states with $q = u, d, s, c$ far above threshold. We also outlined a general technique for addressing this background in a general B physics analysis:

- Employ event selection cuts to reject as much of the continuum as possible.
- Use direct measurements of continuum processes collected below the $\Upsilon(4S)$ peak to quantitatively account for their contribution to data taken on the resonance. This process is known as "off-resonance subtraction."