Improving IceCube's Sensitivity to Astrophysical Neutrino Sources

A dissertation
by

Alexander Andreas Harnisch
born in
Castrop-Rauxel, Germany

submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Physics
with Dual Major in
Computational Mathematics, Science, and Engineering

Michigan State University
2026

This dissertation follows the Chicago Manual of Style—intentionally.

The latest self-published version is freely available at
lightning-tracks.com/thesis

Revision 10922cc, built 2026-06-25

This work is dedicated to my nephew

Noah Romeo Harnisch

∗ March 11, 2026

† March 11, 2026

Abstract

This dissertation presents the most sensitive data sample assembled to date for astrophysical neutrino point-source searches, together with its first application. High-energy neutrinos point straight back to the astrophysical objects that accelerate cosmic rays, making them unique probes of the most energetic sources in the universe. Detecting them takes an enormous instrument: the cubic-kilometer IceCube Neutrino Observatory at the South Pole. The sample is a new selection of track-like events in IceCube, built with modern machine-learning and statistical methods and designed for reuse beyond this first application. It improves detection power over previous samples by up to a factor of four in the southern sky and by up to 30 percent in the northern sky, improvements that come largely from lowering the energy threshold to roughly 100 GeV. This application comprises an all-sky scan and a search over a catalog of candidate sources. The results remain sealed until IceCube collaboration approval, expected on July 2, 2026.

Acknowledgements

There are no individual achievements. I owe this one to many people.

Foremost among them is my advisor, Nathan Whitehorn. He supported me throughout, never more so than in navigating the complexities of a large scientific collaboration. When I got lost in mathematical or technical details, he reminded me of the physics behind them, drawing on a perspective I lacked.

I am deeply indebted to Mirco Hünnefeld. If it were not for him I might not have pursued a PhD in this field at all. I tried to follow in his footsteps; he taught me the basics of IceCube when I was new to all of it. He created the DNNC sample, from which I drew a great deal of inspiration.

Though he is at a different institution and owed me nothing, Michael Larson taught me more than almost anyone about how things are done in IceCube, patiently fielding an endless stream of questions.

I thank Thorsten Glüsenkamp for the angular reconstruction that the event selection presented here relies on, and for the discussions we had.

Several others deserve particular thanks. Chiara Bellenghi gave me extensive, genuinely helpful input through many productive discussions. I thank Ludwig Neste for our work together releasing the data sample and for our many exchanges. Christopher Weaver has a depth of software and IceCube experience I leaned on more than once. Alina Kochocki shared the struggle of graduate school with me. Bennett Brinson and Christopher Wiebusch reviewed my analysis, and I am grateful for their careful reading and the outside perspective they brought. I also thank the IceCube Collaboration as a whole, and the many people I cannot name individually, for the discussions and exchange that shaped this work.

I thank the rest of my dissertation committee for their time and service: Tyler Cocker, Claudio Kopper, Elizabeth Munch, and Remco Zegers. I am especially grateful to Claudio Kopper, who guided me through the initial years of my doctoral studies.

I am grateful to MSU's Institute for Cyber-Enabled Research (ICER), and to the staff who worked through every one of my support tickets; without these computing resources, this work would not have been possible.

I owe much to my friends, who stand by me even with an ocean between us.

To my parents and my grandparents, for their *unconditional* support: no matter how badly I stumbled, I could always just come back home.

And to my brother and his wife—the strongest people I know.

My final and deepest thanks go to Alejandra, for carrying me through the most demanding times of my life thus far.

Contents

List of Figures

List of Tables

List of Abbreviations

AdamW	Adam optimizer with decoupled weight decay
AGN	active galactic nucleus
AoS	array of structs
a.s.	almost surely
AUROC	area under the receiver operating characteristic curve
BDT	boosted decision tree
BFGS	Broyden–Fletcher–Goldfarb–Shanno
BG	background
BLL	BL Lacertae-type blazar
CC	charged-current interaction
CDF	cumulative distribution function
CI	confidence interval
CL	confidence level
CNN	convolutional neural network
CORSIKA	COsmic Ray SImulations for KAscade
CPU	central processing unit
CRNN	convolutional recurrent neural network
CUDA	Compute Unified Device Architecture
DAQ	data acquisition
DMA	direct memory access
DNN	deep neural network
DNNC	DNN Cascades event selection
DOM	digital optical module
DP	discovery potential
ELU	exponential linear unit activation function
ESTES	Enhanced Starting Track Event Selection
EVT	extreme value theory
FC	Feldman–Cousins
FDR	false discovery rate
FFT	fast Fourier transform
FP16	16-bit floating-point arithmetic
FPR	false positive rate
FSRQ	flat-spectrum radio quasar
FWER	family-wise error rate
GEV	generalized extreme value distribution

GeV	giga-electron-volt
GPU	graphics processing unit
HDF5	Hierarchical Data Format
HEALPix	Hierarchical Equal Area isoLatitude Pixelation
HEP	high-energy physics
HPCC	High Performance Computing Center
i.i.d.	independent and identically distributed
IC86	IceCube 86-string complete configuration
KDE	kernel density estimate
LAT	Large Area Telescope
LCSC	Lightning CNN Signal Classifier
LED	light-emitting diode
LF2I	likelihood-free frequentist inference
LOO	leave-one-out
LR	likelihood ratio
LRT	likelihood-ratio test
LT	Lightning Tracks
MC	Monte Carlo
ML	machine learning
MLE	maximum likelihood estimate
MLP	multilayer perceptron
MPEFit	multi-photoelectron track fit
MSU	Michigan State University
MuEX	muon energy extraction estimator
NC	neutral-current interaction
NT	Northern Tracks event selection
NuGen	Neutrino Generator simulation
PCA	principal component analysis
PDF	probability density function
PE	photoelectron
PeV	peta-electron-volt
PIT	probability integral transform
PMT	photomultiplier tube
POI	parameter of interest
POSIX	Portable Operating System Interface
PSF	point-spread function
PST	Point-Source Tracks event selection
PWN	pulsar wind nebula
RA	right ascension
RBF	radial basis function
RDG	radio galaxy
ReLU	rectified linear unit activation function
RMS	root mean square
RNG	random number generator
RNN	recurrent neural network

ROC	receiver operating characteristic curve
SBG	starburst galaxy
SF	survival function
SIBYLL	cosmic-ray hadronic interaction model
SIMD	single instruction, multiple data
SLT	Starting Lightning Tracks
SLURM	Simple Linux Utility for Resource Management
SoA	struct of arrays
SPICE	South Pole ICE optical-property model
STV	starting-track veto
TCP	Transmission Control Protocol
TeV	tera-electron-volt
TLT	Throughgoing Lightning Tracks
TNF	Transformer Normalizing Flows
TPR	true positive rate
TS	test statistic
UAT	universal approximation theorem
vMF	von Mises–Fisher distribution

List of Symbols

This list collects the recurring symbols used throughout the dissertation. Each is also defined where it is first introduced.

Detector and neutrino interactions

ν_e, ν_μ, ν_τ The three neutrino flavors, distinguished by the charged lepton produced in a charged-current interaction.

θ_c Cherenkov angle, the fixed angle at which a charged particle faster than the local phase velocity of light emits Cherenkov radiation (about 41° in ice).

Celestial coordinates and per-event observables

α, δ Right ascension and declination, the celestial coordinates of an event's reconstructed direction and of a tested source location. (In the hypothesis-testing chapters α also denotes the significance level; the two senses never appear together.)

$\Delta\psi$ Angular separation on the sphere between an event direction and a tested source location.

$\hat{\sigma}_i$ Per-event estimate of the angular error, used as the scale of the point-spread function for event i.

$\hat{E}_i$ Per-event reconstructed neutrino energy.

Source flux, effective area, and acceptance

γ Spectral index of the assumed source spectrum, a single unbroken power law $\mathrm{d}N/\mathrm{d}E \propto E^{-\gamma}$; fitted jointly with the signal count.

$\Phi(E)$, Φ_0 Differential neutrino flux as a function of energy, and the normalization of a power-law flux model.

A_{eff} Effective area, the equivalent cross-sectional area of an idealized detector that would observe the same event rate from a given flux.

$A(\gamma, \delta)$ Acceptance, the expected number of signal events from a source of spectral index γ at declination δ, integrating the effective area over the assumed spectrum and the livetime.

τ Detector livetime, the live observation time that scales the acceptance integral $A(\gamma, \delta)$.

Hypothesis testing and significance

H_0, H_1 Null and alternative hypotheses. Here H_0 is background-only (no source at the tested position) and H_1 adds a source.

T **(TS)** Test statistic, twice the log-likelihood ratio between the best-fit signal hypothesis and the background-only hypothesis; often abbreviated TS in prose.

Λ, $\Lambda_{\max}$ Likelihood ratio, and the likelihood ratio maximized over the alternative's parameter space.

p p-value, the null-hypothesis probability of a test statistic at least as extreme as the one observed.

$n\sigma$ A p-value expressed as the number of standard deviations of a one-sided standard normal fluctuation.

α Significance level (size) of a test, the probability of rejecting H_0 when it is true. (Distinct from the right ascension α above.)

q Degrees of freedom, the number of parameters constrained by H_0; under Wilks' theorem the test statistic approaches a χ^2_q distribution.

χ^2_q Chi-squared distribution with q degrees of freedom, and its critical values for confidence regions.

p_{pre}, p_{post} Pre-trial p-value at a single tested position, and the post-trial p-value after correcting for the search over many positions.

Likelihood, parameters, and densities

θ, Θ An unknown parameter indexing a parametric family of densities, and its parameter space.

$p(x; \theta)$ **vs.** $f(x \mid y)$ The semicolon indicates that θ indexes a parametric family of densities and is an unknown constant, not a random variable; the vertical bar is reserved for a genuine conditional density between random variables.

$\mathcal{L}(\theta)$ Likelihood function, the density of the observed data viewed as a function of the parameter; for independent events it factorizes into a product.

n_s, $\hat{n}_s$, $\hat{\gamma}$ Number of signal events (constrained to $n_s \geq 0$), and the maximum-likelihood estimates of the signal count and spectral index.

$\mathcal{S}$, $\mathcal{B}$, $\mathcal{D}$ Signal density, background density, and the data-derived estimate of the background used in its place.

$\mathcal{S}_{\text{space}}$, $\mathcal{S}_{\text{energy}}$ Spatial (point-spread) and energy components of the signal density.

$\mathcal{B}_{\text{space}}$, $\mathcal{D}_{\text{energy}}$ Declination-dependent background spatial density, and the data-derived background energy density.

Angular error and the point-spread function

f_{vMF}, κ Von Mises–Fisher point-spread model on the sphere, with concentration parameter κ setting how tightly it clusters around the source direction.

f_{Rayleigh}, σ Rayleigh point-spread model, the small-angle limit of the von Mises–Fisher form, with scale parameter σ.

Feldman–Cousins parameter estimation

$\hat{p}(x;\theta)$ Empirical sampling density of the estimator, built from pseudo-experiments and used in the calibrated construction.

$\tilde{\theta}$ Calibrated (bias-corrected) estimate, the parameter value maximizing the empirical sampling density evaluated at the observation; the empirical analog of the maximum-likelihood estimate.

$\tilde{\theta}_{\text{med}}$ Marginal-median estimate, a baseline alternative that is not adopted: the truth value at which the marginal medians of $\hat{\theta}$ over per-truth pseudo-experiments match the observed estimate component by component. Distinct from the adopted calibrated MLE $\tilde{\theta}$.

$R(x;\theta)$ Empirical likelihood-ratio rank used to order outcomes in the construction.

$A(\theta)$, $C(x_{\text{obs}})$ Acceptance region in data space for parameter θ, and the confidence region, the set of θ whose acceptance regions contain the observation. (Distinct from the acceptance $A(\gamma, \delta)$ above.)

1 Introduction

The high-energy sky has been observed for over a century through cosmic rays: charged nuclei arriving at Earth with energies extending far beyond anything achievable in accelerators. Where they are accelerated remains an open question at the highest energies, and it is a question cosmic rays themselves are poorly suited to answer: being charged, they are deflected by galactic and intergalactic magnetic fields, and their arrival directions no longer point back to their sources. Photons travel straight but have the complementary weakness: at the relevant energies the universe is not transparent to them, and gamma rays can also be produced by purely leptonic processes, so even a gamma-ray source is not by itself evidence of hadronic acceleration.[1]

Neutrinos evade both problems. They are neutral, so they point back to where they were made; they interact only weakly, so they escape dense source environments and cross cosmological distances unabsorbed; and they are produced, at high energies, essentially only in hadronic interactions (when accelerated cosmic rays collide with matter or radiation near their source). A detected flux of high-energy astrophysical neutrinos from an object is therefore direct evidence that the object accelerates hadrons.[2] The same weak interaction that makes the messenger clean makes it scarce: detecting astrophysical neutrinos requires instrumenting enormous volumes, which is the reason IceCube exists at its scale (Chapter 2). The production mechanism is the same everywhere: accelerated protons strike surrounding gas or radiation and make pions, whose charged members decay through muons into neutrinos. The identical chain runs when cosmic rays strike the Earth's atmosphere, making the *atmospheric* neutrinos and muons that dominate the background to every search in this work (Section 2.2); near an astrophysical accelerator the same process makes the signal (Section 13.1). When neutrinos interact, they produce two primary topological light patterns: elongated *tracks*, left by the muons from charged-current interactions of muon neutrinos, and roughly spherical *cascades*, left by the particle showers of the other interaction channels. By measuring this light, detectors like IceCube can *reconstruct* the neutrino's properties.

IceCube established that an astrophysical high-energy neutrino flux exists,[3] and subsequent measurements have steadily improved its characterization.[4] That measurement establishes existence: somewhere, objects accelerate hadrons to the required energies. It cannot say which ones, because the diffuse flux is summed over the whole sky and carries no information about the individual accelerators behind it. Resolving it into individual point sources identifies those accelerators: a localized excess from a fixed direction pins a specific object as a cosmic-ray source

[1] Halzen and Hooper 2002, "High-energy neutrino astronomy: the cosmic ray connection".

[2] Halzen and Hooper 2002.

[3] IceCube Collaboration 2013b, "Evidence for High-Energy Extraterrestrial Neutrinos at the IceCube Detector", IceCube Collaboration 2014, "Observation of High-Energy Astrophysical Neutrinos in Three Years of IceCube Data".

[4] IceCube Collaboration 2022b, "Improved Characterization of the Astrophysical Muon-Neutrino Flux with 9.5 Years of IceCube Data".

and ties it to everything else known about that object across the electromagnetic spectrum.

Identifying the individual sources of that flux is the program this dissertation contributes to: *neutrino astronomy* in the literal sense, doing astronomy with neutrino directions, embedded in the broader *multi-messenger* effort that combines photons, neutrinos, cosmic rays, and gravitational waves into one picture of the high-energy universe.

The flux is there. The question is what makes it. IceCube's scientific output extends well beyond that question, into atmospheric neutrino physics, oscillation measurements, and cosmic-ray studies. But neutrino astronomy is the experiment's primary objective—and three results frame where it stands.

The first compelling association of a high-energy neutrino with an individual object came in 2017: a realtime alert triggered by a high-energy track event led to the observation of the flaring blazar TXS 0506+056 in spatial and temporal coincidence, a multi-messenger campaign across the electromagnetic spectrum.[5]

The strongest steady individual-source result to date is an excess of 4.2σ significance at the active galaxy NGC 1068 in a time-integrated search for point sources in the northern sky.[6] This was a surprising finding: the measured photon emission of NGC 1068 falls far short of what is needed to account for the observed neutrino excess, even when its gamma-ray output is taken to be entirely hadronic. The neutrino flux exceeds the potential TeV gamma-ray flux by at least an order of magnitude,[7] consistent with neutrino production in a hidden, gamma-opaque region such as the corona of an active galactic nucleus (AGN).[8] The source belongs to a population of Seyfert galaxies, and this class is now regarded as among the most promising candidate classes of astrophysical neutrino emitters. A more recent analysis found evidence for neutrino emission from a sub-population of these sources: a binomial test over a Seyfert sample selected in X-rays, excluding NGC 1068 itself, yields a collective excess of 11 of 47 sources at a global significance of 3.3σ.[9] That analysis was limited to the northern sky. The strongest candidates—the brightest X-ray Seyferts, those most detectable under the corona scenario—lie predominantly in the southern sky, beyond the reach of the northern-sky analysis, which is exactly where the event selection developed in this dissertation delivers its largest improvement in sensitivity (Section 9.11). The same analysis also finds the NGC 1068 single-source significance slightly lower than before, at 4.0σ,[10] the decrease attributed to a shift toward lower neutrino energies where the atmospheric background is more prominent. NGC 1068 sits essentially on the celestial equator, just south of it, inside the northern-sky analysis range and in IceCube's most sensitive declination band, and it recurs throughout this dissertation as the benchmark source for parameter estimation (Chapter 12).

Beyond individual sources, the most recent defining result is the observation of neutrino emission from the Galactic plane at 4.5σ, obtained with a cascade sample selected and reconstructed with deep-learning methods.[11] This result carries a methodological message that the next section makes explicit. It also sets a concrete challenge that is one of the motivations for this work: resolving the morphology of the Galactic emission will take all the sensitivity the detector's data can sup-

[5] IceCube Collaboration 2017c, "The IceCube Realtime Alert System", IceCube Collaboration et al. 2018, "Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A".

[6] IceCube Collaboration 2022a, "Evidence for neutrino emission from the nearby active galaxy NGC 1068".

[7] IceCube Collaboration 2022a.

[8] Murase, Kimura, and Mészáros 2020a, "Hidden Cores of Active Galactic Nuclei as the Origin of Medium-Energy Neutrinos: Critical Tests with the MeV Gamma-Ray Connection".

[9] IceCube Collaboration 2026a, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube", Sec. 4.

[10] IceCube Collaboration 2026a, Sec. 4.

[11] IceCube Collaboration 2023, "Observation of high-energy neutrinos from the Galactic plane", Abstract.

port, including in the southern sky, where track selections have historically been weakest (Section 1.2). The southern Seyfert population described above is a second motivation of the same kind, and for the same reason.

1.1 Methodological advances and discovery

There is a pattern in the milestones above worth making explicit, because this dissertation is built on it. The physics questions of neutrino astronomy have been essentially stable for decades: does an astrophysical flux exist, which objects produce it, and what does that say about the origin of cosmic rays. What changed between no detection and 4.5σ Galactic-plane emission was not the questions but the instruments and methods used to answer them. The Galactic-plane observation was made possible by a deep-learning event selection and reconstruction that enlarged the usable cascade sample and sharpened its directional information[12].[13] The previous generation of cascade selections had reached only about 2σ on the same emission,[14] against the deep-learning sample's 4.5σ. The NGC 1068 evidence came from an analysis whose defining feature was methodological: improved data processing and analysis methods applied to the northern track sample.[15] The TXS 0506+056 association became possible when the realtime alert stream went live, an infrastructure that simply did not exist before.[16]

This dissertation takes that pattern as its operating principle: with the detector fixed and the questions fixed, sensitivity is advanced by technical work, on event selection and on the statistical machinery that turns events into measurements. The chapters that follow are deliberately organized around those *technical* contributions. The physics background is kept correspondingly brief: it is not where this work's contribution lies.

1.2 The Lightning Tracks event selection

The obstacle every IceCube event selection confronts is the atmospheric muon background: the detector records about 10^8 muons for every observed astrophysical neutrino.[17] IceCube's time-integrated search for point sources has historically rested on two kinds of track samples. *Northern Tracks* (NT) selections use the Earth and ice as a muon shield: restricted to declinations where events arrive through the planet, atmospheric muons are suppressed by construction and the background reduces to atmospheric neutrinos.[18] The NT sample in current use is the event selection that produced the NGC 1068 result.[19] The cost is coverage: the southern sky, and with it the inner Galactic plane and the densest part of the Galactic source population, is excluded.

Starting-track selections, such as the *Enhanced Starting Track Event Selection* (ESTES),[20] recover southern-sky coverage by demanding a contained neutrino interaction vertex, which vetoes entering atmospheric muons. ESTES was developed for a diffuse-flux measurement, and its central objective is extremely aggressive muon

[12] IceCube Collaboration 2023.

[13] IceCube Collaboration 2021a, "A Convolutional Neural Network based Cascade Reconstruction for the IceCube Neutrino Observatory", Hünnefeld 2023, "Observation of high-energy neutrinos from the Milky Way".

[14] Hünnefeld 2023, Table 13.1.

[15] IceCube Collaboration 2022a.

[16] IceCube Collaboration 2017c, IceCube Collaboration et al. 2018.

[17] IceCube Collaboration 2023, Introduction.

[18] IceCube Collaboration 2016, "Observation and Characterization of a Cosmic Muon Neutrino Flux from the Northern Hemisphere Using Six Years of IceCube Data", IceCube Collaboration 2017a, "All-sky Search for Time-integrated Neutrino Emission from Astrophysical Sources with 7 yr of IceCube Data".

[19] IceCube Collaboration 2022a.

[20] IceCube Collaboration 2024a, "Characterization of the Astrophysical Diffuse Neutrino Flux using Starting Track Events in IceCube".

rejection: the selection cuts hard and early, beginning with charge requirements that cost signal acceptance at low and medium energies.

The other track sample with full-sky coverage is the one known within the IceCube collaboration as *Point Source Tracks* (PST).[21] It shares its base processing with Northern Tracks but diverges heavily in the final event selection, and it has fallen behind NT in methodological refinement. Full-sky coverage means facing the southern muon background directly—and the one handle PST applies against it is energy. Energy spectra in this field are expected to be approximately power laws, $dN/dE \propto E^{-\gamma}$, because shock acceleration generically produces them.[22] The *spectral index* γ sets how steeply the flux falls with energy: a spectrum with large γ is called *soft* (it runs out of high-energy events quickly); one with small γ, *hard*. The atmospheric fluxes are much softer than the expected astrophysical flux, so in the southern sky PST imposes very strict high-energy cuts, which remove the background preferentially.[23] Those cuts do not select through-going events exclusively: high-energy starting tracks pass them as well. But PST does not treat the two topologies separately, and the starting events are not able to contribute meaningfully, owing to their significantly lower counts (Section 5.3).

No track sample there combined the statistics of an inclusive through-going selection with point-source sensitivity good enough to be competitive, and no sample combined through-going and starting tracks in a single selection at all. The distinction between rejection and sensitivity matters here: ESTES rejects the muon background extremely well, yet its point-source sensitivity does not follow (Section 9.11 quantifies this). Background rejection is not sensitivity, and purity (the fraction of signal events in a sample) alone is not a reliable proxy for it. That distinction is one of the core design principles of this work (Section 5.1). The horizontal and northern region is a different matter: nothing there is competitive with through-going tracks. The territory ceded to cascades is the genuinely southern sky. For extended emission such as the Galactic plane, cascade samples are expected to have substantially better sensitivity than track samples in the southern sky despite an angular resolution worse by an order of magnitude, because their lower background, better energy resolution, and lower energy threshold compensate for it.[24]

DNN Cascades (DNNC) made that advantage concrete. It improved substantially on the previous generation of IceCube cascade selections, in the northern and southern sky alike: relative to the precursor, the DNN Cascade sample improved the point-source sensitivity by up to a factor of four, and the Galactic-plane analyses by a comparable factor[25] (on the predecessor sample, the deep-learning reconstruction by itself improved point-source sensitivity at $\gamma = 2$ by about a factor of two[26]). In the northern sky, a better cascade sample changes little for point-source searches, because the through-going track samples dominate the sensitivity there. In the southern sky, cascades gave the best point-source sensitivity available before this work (the sample comparison in Section 9.11 quantifies this), so the same improvement was global for point-source searches: no track sample could compete. On the reading of Section 1.1, that is what delivered the Galactic-plane observation, since most of the plane's emission lies in the southern sky.[27]

[21] IceCube Collaboration 2020, "Time-integrated Neutrino Source Searches with 10 years of IceCube Data".

[22] Drury 1983, "An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasmas".

[23] IceCube Collaboration 2020.

[24] IceCube Collaboration 2023, Introduction.

[25] Hünnefeld 2023, Sec. 12.1, p. 143 (point sources); Fig. 12.1, p. 142 (Galactic plane).

[26] Hünnefeld 2023, Ch. 13, p. 150.

[27] IceCube Collaboration 2023.

Sensitivity is not the whole story for point sources, though. Even with the best point-source sensitivity in the south, cascade samples run into *source confusion*: tested source positions within the width of the point-spread function are strongly correlated, so sources closer together than the angular error cannot be separated. The source candidates close to the galactic center region, exactly where southern-sky sensitivity matters most, are dense enough for this to matter. The track samples available there—ESTES and PST—have much better angular resolution, so source confusion is much less of an issue for them. However, after the strict muon rejection cuts each applies, there is not enough sensitivity left to detect the sources that could be confused.

Why cascades win in the south deserves spelling out, because it also says how to beat them. A neutrino interaction produces a track or a cascade depending on the interaction type and the neutrino flavor (Section 2.1 explains the topologies and what creates them). What gives cascades their edge is not resolution and not statistics: it is purity, and purity sits naturally on their side. The channel counting already favors them, since the atmospheric background is track-dominated while an astrophysical $1:1:1$ flux is cascade-leaning (Section 2.2 makes this quantitative). On top of that, cascades are starting events by nature, whereas the dominant background, atmospheric muons, is through-going by nature, because a muon born in the atmosphere must enter the detector from outside, so the starting topology suppresses that background by construction, before any cut. Geometry splits the sky between them: a cascade spans only a few meters and must be contained, whereas a through-going muon is collected from kilometers away, so through-going tracks dominate the north, and only in the south, where the downgoing muon background closes that channel off, do cascades hold the edge. Starting tracks share the topological advantage in slightly weakened form: a track that appears to start inside the detector is more likely to be a disguised background event than a cascade is, because an entering muon that produces no detectable light in the outer detector layers takes on the appearance of a starting track (Section 2.1 returns to this). Where starting tracks make up the deficit is angular resolution: they point far better than any cascade, and angular resolution is the dominant lever on point-source sensitivity. So although purity favors cascades, we expect starting tracks to be competitive with them in the southern sky, at least in its more horizontal part, away from the straight-downgoing region (the southern sky is *up* from IceCube's perspective) where the muon background becomes overwhelming. As Section 9.11 shows, they are. This argument leans on properties of the detector, the backgrounds, and the neutrino interactions that the reader is not assumed to know yet. Part I develops all of them properly.

That reasoning motivated building a new track selection from the ground up, with the Seyfert-galaxy and Galactic-plane morphology follow-ups, which need every bit of usable data, as concrete targets. The result is *Lightning Tracks* (LT), the event selection at the center of this dissertation.

Its design rests on four principles, stated here once and executed chapter by chapter:

1. **Filtering taken to the extreme:** The filter stage is entirely topological and runs in a single step at the lowest accessible data level: no reconstructions, no multi-level cascade of computationally cheap cuts. Signal cut early can never be recovered later, so nothing is cut that does not have to be (Chapter 3).
2. **Purity is not the objective:** Past a crossover point, looser cuts give better sensitivity: in background-dominated regimes a harder cut cannot overcome the square-root background scaling, and cutting into irreducible background only removes signal. The selection is deliberately inclusive (Chapter 5).
3. **Empirical sensitivity optimization:** Final thresholds are set by directly computing the point-source sensitivity (declination by declination, Chapter 5, Section 3.6).
4. **One pipeline, two samples:** Processing is unified, but the selection resolves into disjoint starting (SLT) and through-going (TLT) Lightning Tracks components with per-topology likelihood ingredients, so each event class is modeled by probability density functions (PDFs) that match its physics (Chapter 4, Chapter 9).

Lightning Tracks inherits from DNNC the approach of selecting events with deep learning,[28] and that debt is acknowledged here and throughout. It is, however, much more than the same approach reapplied to tracks: the filtering philosophy, the sensitivity-driven cut functions, the calibration of the angular-error estimates, and the statistical machinery of Part II are this work's own, and where Lightning Tracks deliberately departs from its predecessors, the departures are documented in the prior-art discussions of the respective chapters. The selection was also built to outlast this dissertation. It provides the best point-source sensitivity of any neutrino sample to date, especially in the south (Section 9.11), and is reproducible from version-controlled code end to end, scalable to all years of detector data on CPU resources, and usable by analyses beyond the ones presented here.

[28] IceCube Collaboration 2021a, Hünnefeld 2023.

The dual-degree context of this work shows in the same places. The computational contributions, the filtering at scale (Chapter 3), the sensitivity optimization machinery (Chapter 5), the all-sky scan infrastructure (Chapter 10), and the Feldman-Cousins construction (Chapter 12), are computational-science work in their own right, and they are presented at a level of technical detail chosen accordingly.

1.3 How this dissertation is organized

The body of this dissertation is organized in three Parts.

Part I, Redefining Event Selection in IceCube (Chapter 2 through Chapter 6), builds the event sample, and it is the heart of this work: physics domain knowledge, of particle interactions and detection principles, combined with modern machine-learning methods for large-scale data processing under physically and statistically

motivated selection criteria. The Part title is a deliberate nod to the part titles of Huennefeld's thesis, among them *Redefining Event Reconstruction in IceCube*.[29] After the detector background (Chapter 2), Part I follows the data through the Lightning Tracks pipeline: the topological filtering stage (Chapter 3), the reconstructions, quality cuts, and final classifiers that define the sample (Chapter 4), the sensitivity optimization that sets the final thresholds (Chapter 5), and the data/Monte Carlo (MC) validation and systematic uncertainties of the finished sample (Chapter 6).

[29] Hünnefeld 2023, p. 3.

Part II, The Point-Source Analysis Framework (Chapter 7 through Chapter 12), is the statistics- and mathematics-heavy side of the dissertation, the side most closely aligned with the computational-mathematics half of the dual degree this work serves. It builds the machinery that turns the sample into measurements: the statistical foundations (Chapter 7), the calibration of the per-event angular uncertainties (Chapter 8), the point-source likelihood and the sample's resulting performance (Chapter 9), the all-sky search methodology (Chapter 10), the source catalog testing methodology (Chapter 11), and the Feldman-Cousins parameter estimation adopted for this analysis (Chapter 12).

Part III, Finding Astrophysical Neutrino Sources (Chapter 13 onward), carries most of the physics output, the neutrino astronomy itself: the all-sky and catalog searches (Chapter 13), the projected results (Chapter 14), and the conclusions and outlook (Chapter 15).

The weighting across these Parts is deliberate and follows the argument of Section 1.1: the physics background takes a back seat because it is not an original contribution and the physics questions are unchanged; the technical chapters carry the depth because that is where the contribution lies.

Part I

Redefining Event Selection in IceCube

2 The IceCube Detector and Neutrino Event Topologies

This chapter assembles the background the rest of this dissertation builds on: how neutrinos interact and the event topologies that follow from those interactions (Section 2.1), the detector that records them (Section 2.3), the natural ice that makes up most of it (Section 2.4), and the path from raw detector output to the dataset this work starts from (Section 2.5). The treatment is deliberately brief, with pointers to the primary literature throughout. It also assumes some familiarity with the Standard Model of particle physics, for which Griffiths[30] is a suitable introduction. Readers familiar with IceCube can skip ahead to Chapter 3, where the original work begins.

[30] Griffiths 2008, *Introduction to Elementary Particles*.

2.1 Neutrinos and how they interact

Neutrinos interact only through the weak interaction. At the energies relevant for this work (above roughly 100 GeV), a neutrino that encounters the detector interacts with a nucleon through deep inelastic scattering in one of two channels: in a *charged-current* (CC) interaction, mediated by a W boson, the neutrino converts into the charged lepton of its flavor and the struck nucleon fragments into a hadronic shower; in a *neutral-current* (NC) interaction, mediated by a Z boson, the neutrino survives, carrying away part of the energy invisibly, and only the hadronic shower is observable.[31]

The charged secondaries are visible through *Cherenkov radiation*: a charged particle moving through ice faster than the local phase velocity of light emits a coherent cone of optical photons at a fixed angle to its direction of travel (about 41° in ice).[32] IceCube detects these photons. Everything the detector knows about an event—its direction, energy, and identity—must be inferred from the spatial and temporal pattern of Cherenkov light recorded across the instrumented volume.

The event topologies that IceCube event selections are built around follow directly from the interaction channel and the neutrino flavor. A ν_μ CC interaction produces a muon, and a muon at these energies is highly penetrating, losing energy through three mechanisms as it propagates. Below roughly a TeV the dominant process is ionization, a quasi-continuous transfer of energy to atomic electrons described by the Bethe equation.[33] Above that, radiative losses take over, beginning with bremsstrahlung, the emission of photons as the muon is deflected in the field of

[31] Gandhi et al. 1998, "Neutrino interactions at ultrahigh energies".

[32] Frank and Tamm 1991, "Coherent Visible Radiation of Fast Electrons Passing Through Matter", p. 110, Eq. (1).

[33] Particle Data Group 2024, "Review of Particle Physics", Sec. 34, the Bethe equation.

a nucleus.[34] The remaining radiative channels—electron-positron pair production and photonuclear interactions—are stochastic, producing occasional large, localized energy deposits and coming to dominate the loss at the highest energies.[35] A muon can therefore travel kilometers through the ice. The result is a *track*: an extended, linear light pattern whose long lever arm makes the direction well measurable. Most other channels produce *cascades*: ν_e CC interactions, NC interactions of all flavors, and the majority of ν_τ CC interactions deposit their energy in electromagnetic or hadronic showers that extend only a few meters. The shower itself is nearly point-like on the scale of the detector, but the Cherenkov light it radiates spreads into a quasi-spherical light ball reaching much further, and it is this light ball that the detector records. The exception among ν_τ CC events is the roughly 17% of taus that decay to a muon,[36] which produce a track like a ν_μ CC interaction. Cascades calorimetrically measure energy well but resolve direction far more coarsely than tracks. At energies high enough for the tau decay length to span a detectable distance, a ν_τ CC interaction can instead produce a *double bang*: one cascade at the interaction vertex and a second one where the tau decays.[37] Tau-specific signatures and their energy dependence are discussed where they matter for the selection (Section 3.6).

Figure 2.1 shows these event topologies as abstract light patterns, independent of the detector instrumentation, with the relative arrival time of the Cherenkov light color-coded.

[34] Koehne et al. 2013, "PROPOSAL: A tool for propagation of charged leptons".

[35] Koehne et al. 2013.

[36] Particle Data Group 2024, Lepton Summary Table, tau branching fractions.

[37] Learned and Pakvasa 1995, "Detecting ν_τ oscillations at PeV energies".

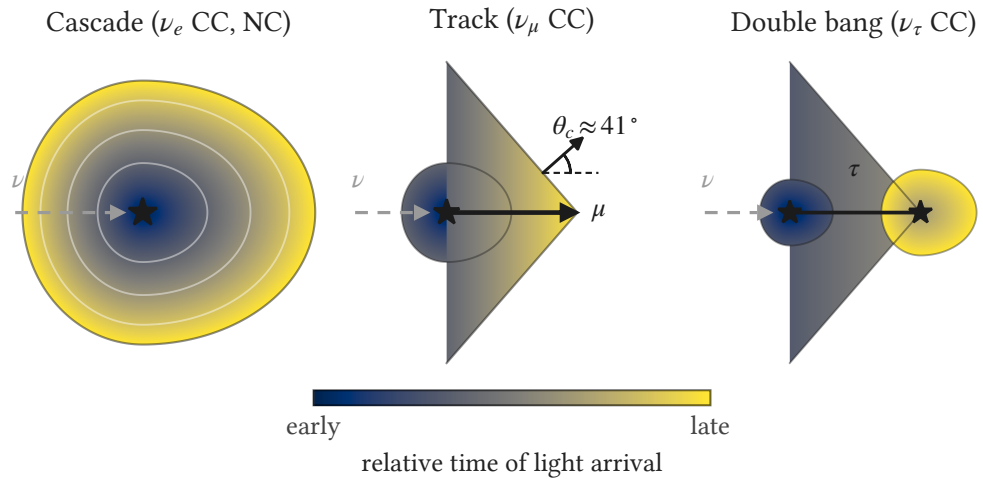

Figure 2.1: The neutrino event topologies drawn as abstract light patterns independent of the detector instrumentation. Color encodes the relative time at which Cherenkov light reaches each region, from early (dark) to late (bright). Left to right: a cascade from a ν_e charged-current or neutral-current interaction (a nearly point-like shower whose Cherenkov light expands outward as a quasi-spherical ball, slightly elongated forward by the boost of the shower); a ν_μ charged-current track, shown here as a starting event with a hadronic cascade at the vertex followed by an outgoing muon and the Cherenkov cone it sweeps at the Cherenkov angle $\theta_c \approx 41°$ in ice; and a ν_τ charged-current double bang, a first cascade at the interaction vertex, the tau track, and a second cascade where the tau decays.

For tracks, the geometry of the interaction point relative to the instrumented volume produces a further classification that the event selection in this dissertation is organized around. A *starting track* has its interaction vertex inside the detector: the neutrino interacted within the instrumented volume, and the outgoing muon exits it. A *through-going track* comes from a neutrino (or atmospheric muon) interaction outside the detector: the muon enters, crosses, and leaves. A *stopping track* enters from outside and ends inside, the muon decaying before it reaches the far boundary. A fully contained track both starts and ends inside. At the energies where full containment occurs, such a track can be hard to distinguish from a cascade. Figure 2.2 illustrates all four geometries on the actual detector volume. A further subclass is sometimes defined on top of these: *skimming tracks*, where the light-emitting particle passes outside the outermost optical modules of the detector rather than between the instrumentation (not shown in the figure).

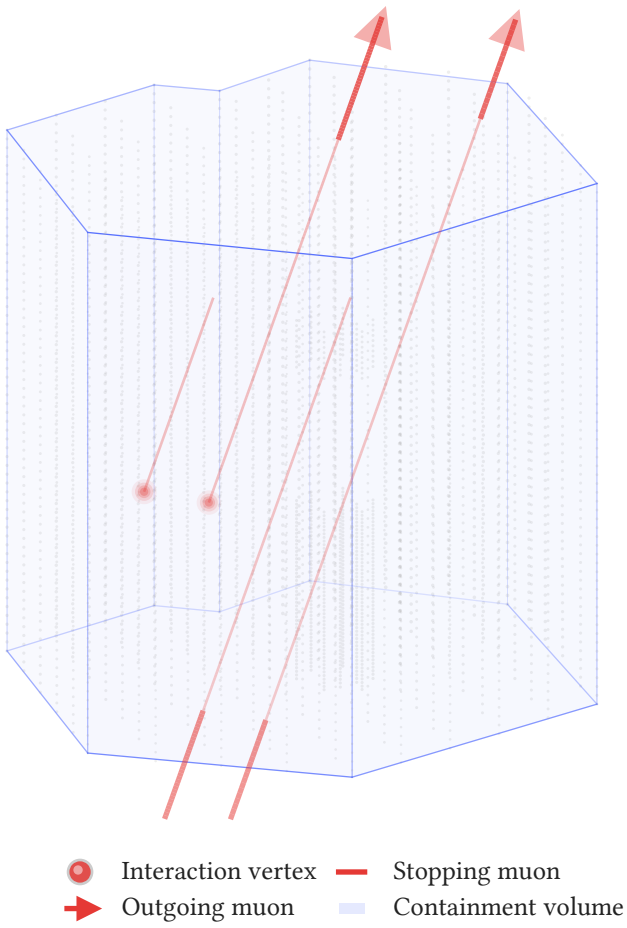

Figure 2.2: The four track geometries illustrated on the detector volume. For visual clarity, the wireframe surface is a simplified convex hull based on the corner DOM positions. DOM positions are overlaid as points. Left to right: a fully contained track (vertex inside, muon decaying before reaching the boundary), a starting track (vertex inside, muon exiting), a stopping track (muon entering, decaying inside), and a through-going track (entering and exiting). Interaction vertices are marked with spheres. Arrowheads indicate exiting tracks, while tracks without arrowheads terminate at the muon decay point. The same volume defines the containment criterion for the starting-track truth label in Chapter 3.

The distinction between starting and through-going tracks matters because it carries background information: an atmospheric muon can only enter from outside, so a track that visibly starts inside the detector cannot be a single atmospheric muon, while a through-going track is indistinguishable, event by event, from one. However, just because a track *looks* like it started inside the detector does not mean it actually did. The primary background for starting-track event selections consists of single atmospheric muons that produce no detectable light in the outermost layers of the detector (often called the *veto region* where explicitly defined), giving the event the topological appearance of a starting track despite its through-going

physical nature.

2.2 Atmospheric backgrounds

Most of what IceCube records originates in Earth's atmosphere rather than in the distant universe. Primary cosmic rays, protons and heavier nuclei, strike the upper atmosphere continuously and initiate hadronic *air showers*: cascades of secondary particles in which charged and neutral pions dominate, with a smaller admixture of kaons.[38] The charged mesons face a competition between decaying and reinteracting, and its outcome depends on energy. A higher-energy meson is the more likely to interact before it decays, so the fluxes of its decay products fall one power of energy more steeply than the primary cosmic-ray spectrum.[39]

The decays $\pi^{\pm} \to \mu^{\pm}\nu_{\mu}$ and the analogous kaon channels produce muons and muon neutrinos. The muons are the more conspicuous product. *Atmospheric muons* energetic enough to reach the deep ice pass through the detector in enormous numbers, outnumbering observed astrophysical neutrinos by a factor of order 10^8.[40] The same dominance is visible in the filter-level rates of Figure 3.13, where the experimental data rate is essentially the muon rate. Atmospheric muons are the dominant background to every neutrino event selection in this dissertation. Two features make them manageable. First, they arrive only from above: a muon cannot cross the Earth without ranging out, so atmospheric muons reach the detector as down-going events and leave the up-going sky free of them. And second, they are born outside the detector, entering as through-going tracks, so the containment and veto arguments of Section 2.1 apply against them.

The same decay chains produce neutrinos, a background of a different character. The direct meson decays $\pi^{+} \to \mu^{+}\nu_{\mu}$ (and the analogous kaon channels) yield muon neutrinos, and the muons in turn decay, $\mu^{+} \to e^{+}\nu_{e}\bar{\nu}_{\mu}$.[41] Unlike the muons, these *atmospheric neutrinos* arrive from every direction, since a neutrino crosses the Earth unattenuated at these energies, and they are indistinguishable event by event from astrophysical neutrinos of the same energy and direction. They are the *irreducible* background—no veto or containment cut removes them, and only their steeper energy spectrum and overall rate separate them statistically from a diffuse astrophysical signal. For the point sources this work targets, there is an additional and far more powerful handle: the local spatial clustering of events from a source stands out above the isotropic atmospheric background, even when the source's energy spectrum is indistinguishable from that of the background (Chapter 9).

The flavor content of the flux follows from this chain. The direct meson decay contributes a muon neutrino, and the muon decay contributes an electron neutrino and a second muon neutrino, so at low energy the flux runs about two parts ν_{μ} to one part ν_{e}.[42] The ratio grows with energy, because a higher-energy muon tends to reach the ground and stop before decaying, removing its contribution to the ν_{e} flux. Tau neutrinos are essentially absent: making them requires the decay of charmed hadrons, which are rare in air showers. This is the *conventional* atmospheric flux, from π and K decay. A harder *prompt* component from charm decay[43] is expected

[38] Gaisser, Stanev, and Tilav 2013, "Cosmic ray energy spectrum from measurements of air showers".

[39] Gaisser 2012, "Spectrum of cosmic-ray nucleons, kaon production, and the atmospheric muon charge ratio".

[40] IceCube Collaboration 2023, "Observation of high-energy neutrinos from the Galactic plane", Introduction.

[41] Barr et al. 2006, "Uncertainties in Atmospheric Neutrino Fluxes".

[42] Barr et al. 2006, Fig. 7.

[43] Enberg, Reno, and Sarcevic 2008, "Prompt neutrino fluxes from atmospheric charm".

to emerge at the highest energies but is subdominant and not relevant here. It is the target of dedicated searches in IceCube and has not yet been observed,[44] a difficult measurement because the astrophysical diffuse flux is itself the dominant background at the energies where the prompt component would appear.

[44] IceCube Collaboration 2022b, "Improved Characterization of the Astrophysical Muon-Neutrino Flux with 9.5 Years of IceCube Data", Sec. 4.1.

The flavor composition just given is the accounting the starting-track argument of Section 1.2 draws on (its expectation that starting tracks should reach point-source sensitivity comparable to cascades). The astrophysical flux this work targets is produced with a similarly muon-dominated flavor mix, but oscillation over cosmological baselines averages it close to an equal $1:1:1$ ratio of the three flavors at Earth.[45] Of a sample of events that start in the detector, the ν_μ charged-current interactions make tracks, while the ν_e and ν_τ charged-current interactions and the neutral-current interactions of all flavors make cascades.

[45] Learned and Pakvasa 1995, Sec. "Sensitivity to Neutrino Oscillations".

Comparing the two topologies as rates requires the *effective area*: the energy- and direction-dependent equivalent collecting area that converts an incident neutrino flux into a detected event rate, treated in full in Section 9.6. A contained-vertex selection requires every event, in every channel, to interact inside the fiducial volume, so the same volume collects them all and ν_μ gains no effective-area advantage from producing a track. The balance between cascades and tracks is therefore set by the fluxes and the interaction cross sections alone, and reduces to a count of channels.

Carrying out that count, the neutral-current cross section is about a third of the charged-current one,[46] so for a flavor ratio $f_e : f_\mu : f_\tau$ at Earth the ratio of cascades to tracks is $C/T = [\, f_e + f_\tau + r\,(f_e + f_\mu + f_\tau)\,]/f_\mu$, with $r = \sigma_{\mathrm{NC}}/\sigma_{\mathrm{CC}} \approx 1/3$. The ν_τ charged-current channel is counted here entirely as cascade. As noted in Section 2.1, it can itself yield a track, through the roughly 17% of taus that decay to a muon, but that track component is energy-dependent and turns on only at high energy, while the flux, a falling power law, is dominated by low energies, so we neglect it. For an equal $1:1:1$ flux the ratio is $C/T = 2 + 3r \approx 2 + 1 = 3$: an astrophysical flux produces about three cascades for every starting track.

[46] Gandhi et al. 1998, Tables I and II.

The atmospheric background runs the other way. Its neutrino component is dominated by muon neutrinos, so charged-current muon neutrino interactions make it about as track-rich as cascade-rich at low energy and increasingly track-rich toward TeV energies. Layered on top is the atmospheric muon background, entirely tracks, which dominates the rate. The signal is therefore cascade-leaning while the background it must be separated from is track-dominated, so cascades begin with a structural advantage in purity against the atmosphere. How large that advantage is across the sky is set by the magnitude of the muon background and its steep dependence on arrival direction, a question taken up with the selection itself (Section 1.2) and quantified later (Section 9.11).

Through-going tracks earn their place by exactly the effective-area advantage a contained selection forgoes. Beyond the superior angular resolution that a long track affords, a muon radiates Cherenkov light over kilometers of path, so a ν_μ that interacts well outside the instrumented volume—even several kilometers away—is still recorded once its muon ranges into the detector. The distance a muon travels grows with its energy, so the through-going effective area rises far more steeply with

energy than that of any contained topology. Wherever enough ice or rock shields the detector from the atmospheric muon background, at the horizon and across the northern sky, which IceCube views through the Earth, this makes through-going tracks unbeatable. We return to this energy scaling and its consequences in Part II: the spectral dependence of the angular-error calibration (Section 8.3), the right-censoring of muon energy and the resulting recovery bias (Section 9.10), and the full sensitivity curves (Section 9.11). The same reach is a liability in the south, where it applies just as well to the atmospheric muons ranging into the detector from above. This is exactly what gives contained events their edge there, since a contained vertex rejects those entering muons.

One feature of the atmospheric neutrino flux makes part of it reducible after all. Because atmospheric neutrinos are produced in the same air showers as atmospheric muons, a down-going neutrino often arrives accompanied by muons from its parent shower, increasingly so at higher energies. A selection that vetoes on light near the detector edge then rejects the neutrino together with its accompanying muons, whether those arrive singly or as a bundle—the atmospheric *self-veto*, which suppresses the down-going atmospheric neutrino background that would otherwise mimic a contained signal.[47] At the energies relevant for this work, however, the self-veto is too weak to benefit a point-source analysis, as we will see in Chapter 6.

[47] Argüelles et al. 2018, "Unified atmospheric neutrino passing fractions for large-scale neutrino telescopes".

2.3 The IceCube array

IceCube instruments roughly a cubic kilometer of glacial ice at the geographic South Pole with 5,160 *digital optical modules* (DOMs): glass pressure spheres, each housing a downward-facing 10-inch photomultiplier tube together with its digitizing electronics.[48] The DOMs are deployed on 86 vertical cables (*strings*) lowered into drill holes and frozen in place, 60 DOMs per string instrumenting depths between about 1,450 and 2,450 m. Of the 86 strings, 78 form the main array, spaced roughly 125 m apart on an approximately hexagonal surface grid (Figure 2.3) with 17 m vertical DOM spacing. The remaining 8 are the more densely instrumented *DeepCore* infill strings in the deep clear ice at the bottom center, lowering the energy threshold there.[49] The detector is also being extended. The *IceCube Upgrade*, a denser low-energy infill currently being deployed in the central region, adds new strings instrumented with next-generation optical modules (multi-PMT mDOMs and D-Eggs).[50] It postdates the data used in this dissertation and contributes nothing to the analysis presented here.

[48] IceCube Collaboration 2017b, "The IceCube Neutrino Observatory: Instrumentation and Online Systems", Sec. 1.1.

[49] IceCube Collaboration 2017b, Sec. 1.1.2.

[50] IceCube Collaboration 2026c, "Physics potential of the IceCube Upgrade for atmospheric neutrino oscillations", Sec. II.

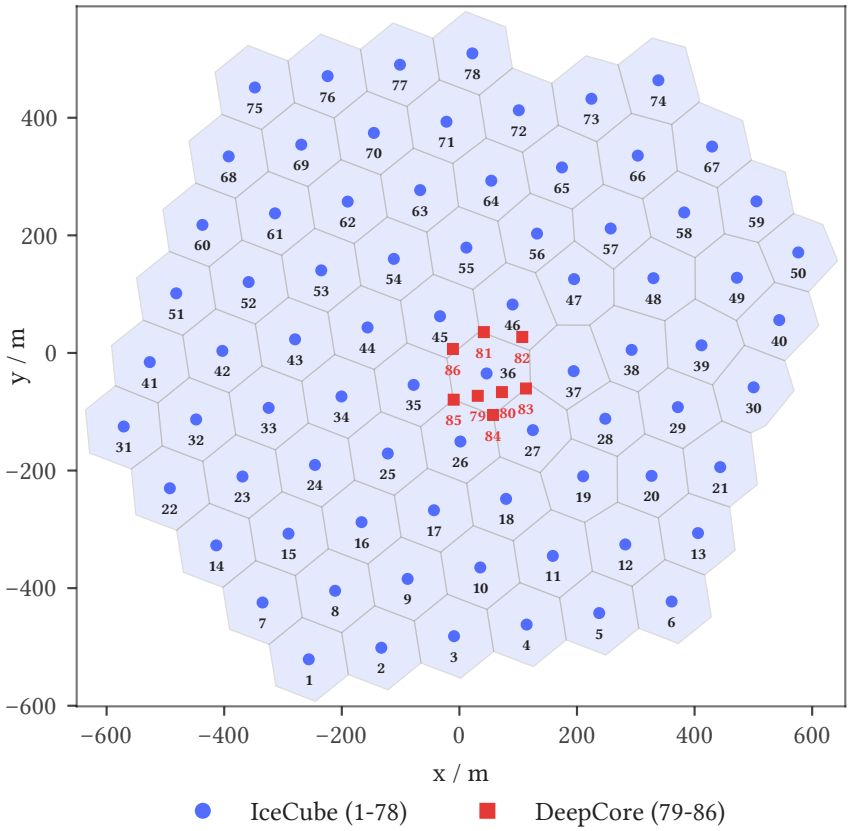

Figure 2.3: Top view of the IceCube string positions in the (x, y) plane. Strings 1–78 form the main array, whose Voronoi cells illustrate the approximately hexagonal tessellation of the surface grid. Strings 79–86 are the more densely spaced DeepCore infill near the array center. The strings of the IceCube Upgrade are deliberately not shown: the Upgrade is not part of any data used in this dissertation.

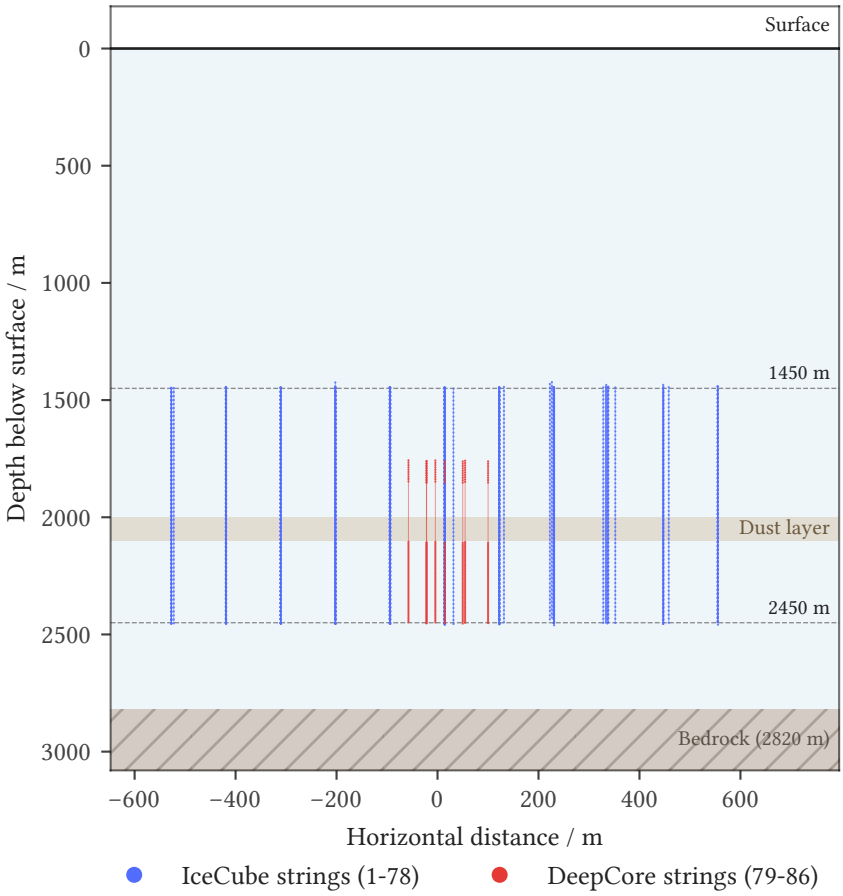

Figure 2.4: Side view of the IceCube array, generated from the detector geometry: the instrumented depth range, the dust layer (Section 2.4) with the uninstrumented gap it leaves in the DeepCore strings, and the bedrock below the array. DeepCore's denser vertical DOM spacing concentrates below the dust layer, with a smaller group of DOMs above it. The horizontal axis is a projection rotated in azimuth so the hexagonal-grid strings line up into columns. The surface layout is shown in Figure 2.3. Module coordinates are taken from the calibrated detector geometry (the IceCube-internal calibration data). In equatorial coordinates, up in this figure is south: up-going events enter from below, through the Earth.

Each DOM independently detects single photons, timestamps them to nanosecond precision, digitizes the waveform, and reports the result. The detector is in essence 5,160 independent photon counters whose combined output is a list of pulses, each with a position, a time, and a charge.[51] The electronics and readout chain are documented in the instrumentation paper and play no further role for us. What matters downstream is only the form of the data: per-DOM charge and timing patterns, from which all higher-level quantities are reconstructed.

[51] IceCube Collaboration 2017b.

The event topologies of Section 2.1 have characteristic detector-level appearances in these patterns. A track crosses the array as a moving light cone, since the muon travels faster than the phase velocity of light in ice: a line of DOMs

lights up in sequence, with a time gradient along the direction of travel and the charge concentrated in a cylinder around the trajectory. A cascade illuminates the array from a single region: a roughly spherical pattern, brightest at the vertex, with the timing expanding outward. At distance scales far below the detector spacing, though, the shower retains a strongly forward, almost track-like direction, and that forwardness imprints an asymmetry on the detected light pattern. The asymmetry still carries directional information, just much less of it than a track's long lever arm: cascade angular resolution is about an order of magnitude worse than for tracks at comparable energies. Bundles of coincident atmospheric muons from the same air shower, the dominant background at trigger level, also cross the array as a front entering from above, but with a wider light-emitting region than a single track: roughly a cylinder of larger radius rather than a thin line.[52] The machine-learning filters of Chapter 3 operate directly on per-DOM summaries of exactly these patterns, which is why a topological classification is possible without any reconstruction.

[52] IceCube Collaboration 2013a, "Cosmic Ray Composition and Energy Spectrum from 1–30 PeV Using the 40-String Configuration of IceTop and IceCube", Sec. 4.2.

2.4 The glacial ice as detection medium

The detector medium is not manufactured: it is natural glacial ice, accumulated over roughly a hundred thousand years, and its optical properties vary with depth. Scattering lengths change by up to a factor of seven and absorption lengths by about a factor of three across the instrumented depth range, following the dust content of the ice layers deposited in different climatic epochs.[53] The most prominent feature is the dust layer at roughly 2,000–2,100 m depth, introduced where it first matters for this work, in the discussion of position-dependent classification (Section 3.2). Within that band the layer resolves into a double structure.[54] The ice below it is the clearest in the detector, which is why the main part of DeepCore sits there, with a smaller group of its DOMs above the layer (Figure 2.4).[55]

For analysis purposes the ice enters through the *ice model*: a depth-dependent (and, in current generations, anisotropic) parameterization of scattering and absorption used by the photon propagation in simulation and by likelihood-based reconstructions. The models are calibrated in situ, primarily with the LED flashers carried by every DOM, and have been successively refined in the SPICE model family.[56] Their current state and remaining uncertainties are reviewed by Rongen and Chirkin.[57]

The simulation does not reduce the ice to an effective light-yield parameterization: IceCube simulates light propagation at the level of individual photons, each propagated through the modeled ice with GPU kernels in the collaboration's internal photon-propagation software.

The author of this dissertation contributed to a later port of the propagation kernel from OpenCL to CUDA, a joint project with NVIDIA.[58] In addition to the depth-dependent scattering and absorption and their azimuthal anisotropy, the current model generation also describes the birefringence (BFR) of the polycrystalline ice, which diffuses photon directions (most strongly for light traveling along the

[53] AMANDA Collaboration 2006, "Optical properties of deep glacial ice at the South Pole", Abstract.

[54] Rongen 2019, "Calibration of the IceCube Neutrino Observatory".

[55] IceCube Collaboration 2017b, Sec. 1.1.2.

[56] IceCube Collaboration 2013c, "Measurement of South Pole ice transparency with the IceCube LED calibration system".

[57] Rongen and Chirkin 2021, "Advances in IceCube ice modelling and what to expect from the Upgrade".

[58] Schwanekamp et al. 2022, "Accelerating IceCube's Photon Propagation Code with CUDA".

ice flow axis) and on average deflects them toward the flow axis,[59] and the tilt of the ice layers: the depth at which a given layer is encountered varies across the array footprint, by as much as ∼60 m over the kilometer scale of the detector.[60] The most complete single treatment of the calibration work behind these models is Rongen's PhD thesis.[61]

Two consequences of the ice matter for us throughout. First, simulation fidelity: every MC event is propagated through an ice model, so ice-model uncertainty is a detector systematic, and the SnowStorm ensemble[62] used for systematics (Chapter 6, Section 12.6) perturbs exactly the bulk-ice scattering and absorption scalings among its parameters. Second, position dependence: the same physical event produces systematically different light yields depending on the ice it traverses, which any position-aware method (the CNN filters of Section 3.4, the reconstructions of Chapter 8) must absorb.

[59] IceCube Collaboration 2024b, "In situ estimation of ice crystal properties at the South Pole using LED calibration data from the IceCube Neutrino Observatory", Sec. 4.2.

[60] IceCube Collaboration 2024b, Sec. 3.4.1.

[61] Rongen 2019.

[62] IceCube Collaboration 2019, "Efficient propagation of systematic uncertainties from calibration to analysis with the SnowStorm method in IceCube".

2.5 From the South Pole to a dataset

Triggered events leave the detector through a tiered data path. The data acquisition system merges DOM hits into events at a rate of 2.5–2.9 kHz, almost entirely atmospheric muons. The online processing and filtering system at the Pole then selects the subset passing at least one physics filter for daily satellite transmission north, while the full data stream is archived locally and shipped on physical media once per year.[63] In IceCube's conventional leveled data processing scheme, the satellite-transmitted stream is *Level 1*. *Level 2* adds offline reconstructions and bookkeeping in the North without removing events. The quantitative version of this path, with the rates and the archive details, opens the filtering chapter (Section 3.1), because it defines the boundary condition of this work: Level 2 is the lowest data level readily accessible on disk, and we start our event selection there, replacing everything above it. Since the selection relies on none of the Level 2 offline filters, this is in effect a start from Level 1.

[63] IceCube Collaboration 2017b, Sec. 6.4.

The simulated counterpart of this dataset is produced by a chain of dedicated generators. Neutrinos of all flavors are generated with NuGen, based on the ANIS event generator,[64] which injects neutrinos with configurable spectra, forces interactions in and around the detector, and records weights that allow reweighting to any assumed flux. Atmospheric muons are simulated with CORSIKA air-shower simulation,[65] weighted to cosmic-ray flux models such as the Gaisser parameterizations[66] with the SIBYLL family of hadronic interaction models.[67] Both are propagated through the detector simulation, including an ice model (Section 2.4), and processed through the same filtering and reconstruction chain as data. The current neutrino production uses the SnowStorm scheme, which continuously varies the dominant detector systematics (ice scattering and absorption, DOM efficiency, hole-ice angular acceptance) across the simulation set rather than producing discrete systematic variants.[68] This is the basis of the systematics treatment in Section 6.4 and Section 12.6.

[64] Gazizov and Kowalski 2005, "ANIS: High energy neutrino generator for neutrino telescopes".

[65] Heck et al. 1998, *CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers*.

[66] Gaisser 2012.

[67] Ahn et al. 2009, "Cosmic ray interaction event generator Sibyll 2.1".

[68] IceCube Collaboration 2019.

3 Filtering

The previous chapter ended at the boundary condition of this work: a 2.5–2.9 kHz trigger stream, almost entirely atmospheric muons, of which the offline analysis receives the Level 2 dataset. This chapter builds the first stage of Lightning Tracks on top of it: the filter that reduces the Level 2 event rate by 99.98% while keeping as much neutrino signal as possible. We start with the filtering problem itself and why conventional approaches give up too much signal (Section 3.1), review the background of convolutional neural networks (CNNs) that the filter models rest on (Section 3.2), and introduce machine-learning filtering on DOM-level features (Section 3.3). We then build the three filter models (Section 3.4, Section 3.5), combine them and measure what the combined filter keeps and rejects (Section 3.6), produce the filtered sample at scale (Section 3.7), and close by placing the approach against its prior art (Section 3.8).

3.1 The filtering problem

The filter stage reduces the input data rate, 481 Hz at Level 2, to a level where computationally expensive final-level reconstructions can be run on every surviving event. For Lightning Tracks, the input is the entire Level 2 dataset, the lowest data level that was readily accessible on disk. Level 2 contains the same events as Level 1 (the online-filtered data transmitted from the South Pole via satellite).[69] The Level 1 to Level 2 transition adds offline reconstructions and more computationally expensive filtering algorithms but does not remove any events. The only data reduction between the trigger level and Level 2 occurs at the online filtering stage (Level 0 to Level 1), where the online processing and filtering system reduces the trigger-level event stream (a merged event rate of 2.5–2.9 kHz) to a volume that fits the satellite bandwidth allocation.[70] As of 2016, approximately 15% of all triggered events are selected by one or more filters. Triggered events below the online filter threshold are not transmitted but are archived locally—formerly on tape (the tape system was retired in 2015) and since then on disks—with two copies of each archival stream transported north each year. The archive is retained indefinitely but not regularly reprocessed. Within that boundary, the Lightning Tracks filter completely replaces the standard IceCube filtering chain from Level 2 upward: filtering and final selection are both event selection at different scope, and the filter's job is to cheaply discard the clearly identifiable background so the expensive reconstructions can run directly on everything that remains.

[69] IceCube Collaboration 2017b, "The IceCube Neutrino Observatory: Instrumentation and Online Systems", Sec. 6.5.

[70] IceCube Collaboration 2017b, Sec. 6.4.

Starting from this Level 2 input, the Lightning Tracks filter must reduce the rate by several orders of magnitude while preserving as much neutrino signal as possible, since every event lost at the filter stage is lost permanently and can never be recovered by later, more sophisticated methods. The reduction is forced by the analysis itself: the point-source likelihood evaluation is computationally bounded, and the full Level 2 stream is far too large to carry into it. Conventional approaches to this problem rely on simple, computationally inexpensive methods: likelihood fits with few iterations, explicit charge thresholds (e.g., the $Q_{\text{total}} > 200$ pre-cut used by the previous starting track event selection ESTES), veto-based containment cuts, or combinations of early low-level filters. These are fast but lack discriminating power—and they significantly reduce effective area at low to medium energies before the actual selection begins. A sequence of one-dimensional threshold cuts, each placed on a single reconstructed variable, partitions the observable space into axis-aligned boxes and cannot follow the way signal and background separate along combinations of variables. A classifier that instead acts on the joint structure of many observables at once can place a decision boundary that follows those joint dependencies and separates the two classes far more cleanly. The fundamental tension is that the conventional methods efficient enough to run at full input rates lack the sophistication to distinguish signal from background reliably, while the methods that could make accurate classification decisions are too computationally expensive to run at those rates. One way to overcome these limitations is to use computationally cheap yet powerful machine-learning algorithms such as convolutional neural networks (CNNs).

3.2 Convolutional neural networks

CNNs are a class of neural network designed for data with spatial structure, most commonly images, but applicable to any data arranged on a grid. The following summarizes the mathematical foundations; readers familiar with CNNs may skip to the architecture.

The defining operation is the discrete convolution: a small learnable filter (or kernel) of fixed spatial extent is applied at every position on the input, computing a weighted sum of the local neighborhood. For a 3D input volume $X \in \mathbb{R}^{C_{\text{in}} \times D_1 \times D_2 \times D_3}$ (in our case, $C_{\text{in}} = 4$ per-DOM features on a $10 \times 10 \times 60$ grid), a convolutional layer with C_{out} filters, each with kernel $W_m \in \mathbb{R}^{C_{\text{in}} \times K_1 \times K_2 \times K_3}$ and bias b_m, produces the output volume $Y \in \mathbb{R}^{C_{\text{out}} \times D_1 \times D_2 \times D_3}$ by

$$Y_m = W_m * X + b_m, \qquad m = 1, \dots, C_{\text{out}}, \tag{3.1}$$

where $*$ denotes the 3D cross-correlation: at each spatial position r, the kernel is applied as an inner product $(W_m * X)(r) = \langle W_m, X_r \rangle$ over the local patch X_r centered at r. Each filter produces one output feature map, and the C_{out} maps are stacked along the channel axis. The same filter weights are reused (shared) across all spatial positions, which has two consequences. First, the number of learnable parameters per filter is $C_{\text{in}} \cdot K_1 \cdot K_2 \cdot K_3 + 1$, independent of the input volume size.

Second, the operation is translation equivariant: the same local pattern produces the same filter response regardless of where it appears in the input. Without padding, the output volume shrinks by $K_i - 1$ along each spatial axis, since the kernel cannot be centered on boundary positions. *Same* padding avoids this by zero-padding the input with $\lfloor K_i/2 \rfloor$ elements on each side, so that the convolution output retains the same spatial dimensions as the input. With same padding, the spatial dimensions change only through pooling (see below), giving explicit control over where in the network the resolution is reduced.

After each convolution, a nonlinear activation function σ is applied elementwise. Without nonlinearity, any sequence of convolutions would collapse to a single linear operation regardless of depth. Often this is misquoted as merely requiring a *nonlinear* activation, but nonlinearity is not strong enough: the activation must be non-polynomial. A polynomial activation, even a nonlinear one such as $\sigma(x) = x^2$, produces outputs that are themselves polynomials of bounded degree—at most d for a single hidden layer, d^L for L layers—which cannot approximate arbitrary continuous functions such as $|x|$ or $\sin(x)$, no matter how many neurons are added. All commonly used activations (ELU, ReLU, sigmoid, tanh) are non-polynomial. The filter models use the Exponential Linear Unit (ELU),[71]

$$\mathrm{ELU}(x) = \begin{cases} x & \text{if } x > 0, \\ \alpha(e^x - 1) & \text{if } x \leq 0, \end{cases} \tag{3.2}$$

with $\alpha = 1$ (Figure 3.1). ELU behaves like the identity for positive inputs and saturates smoothly to $-\alpha$ for large negative inputs—producing a nonzero mean activation that can accelerate convergence compared to the simpler ReLU ($\sigma(z) = \max(0, z)$), which clamps all negative values to zero.

[71] Clevert, Unterthiner, and Hochreiter 2016, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)".

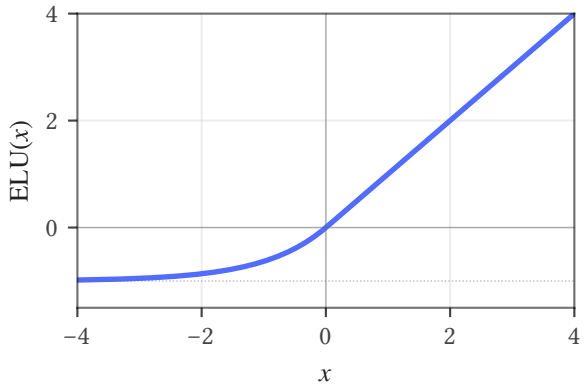

Figure 3.1: The Exponential Linear Unit (ELU) activation function with $\alpha = 1$. For positive inputs, ELU is the identity. For negative inputs, it saturates smoothly to $-\alpha$.

The theoretical foundation for the representational capacity of these networks is the universal approximation theorem (UAT).[72] The UAT states that a feedforward network with a single hidden layer of finite width and a non-polynomial activation

[72] Leshno et al. 1993, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function", Thm. 1, p. 863.

function can approximate any continuous function $f : \mathbb{R}^n \to \mathbb{R}$ on a compact subset $K \subset \mathbb{R}^n$ to arbitrary precision: for any $\epsilon > 0$, there exists a network g with N hidden neurons such that $\sup_{x \in K} |f(x) - g(x)| < \epsilon$. The theorem is existential: it guarantees the existence of such a network but says nothing about N, which may be impractically large for a single hidden layer. Deeper networks can represent the same functions with exponentially fewer parameters by composing simpler transformations hierarchically,[73] which is the practical motivation for using multiple layers despite the single-layer guarantee. For the convolutional layers specifically, the UAT applies to the fully connected classification head that operates on the learned feature representation; the convolutional layers themselves are a constrained (weight-sharing) feature extractor whose output is then classified by a universal approximator.

[73] Telgarsky 2016, "Benefits of depth in neural networks", Eldan and Shamir 2016, "The Power of Depth for Feedforward Neural Networks".

To stabilize the distribution of activations during training, batch normalization[74] is applied before each activation, standardizing the pre-activation values across the mini-batch to zero mean and unit variance:

[74] Ioffe and Szegedy 2015, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift".

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \qquad y = \gamma \hat{x} + \beta, \tag{3.3}$$

where μ_B and σ_B^2 are the per-channel mean and variance, ϵ is a small constant for numerical stability, and γ and β are learnable per-channel scale and shift parameters that allow the network to recover the optimal activation distribution if it is not standard normal. The statistics can be computed from the current mini-batch, from running averages accumulated over training, or from a combination of both, depending on the implementation and whether the model is in training or inference mode. At inference time, fixed statistics are used so that the output for a given event is deterministic and independent of what other events are in the same batch. Without the learnable parameters γ and β (the affine transformation), batch normalization reduces to a pure normalization layer, a property we exploit for input normalization (see the architecture below). Batch normalization reduces sensitivity to weight initialization and learning rate, acts as a mild regularizer, and generally allows faster convergence. Intuitively, by centering the pre-activation distribution around zero, it ensures that activations on average fall in the nonlinear regime of the activation function rather than in the saturated tails, reducing the incidence of neurons with near-zero gradients that effectively stop learning.

Between convolutional blocks, pooling layers reduce the spatial dimensions of the feature maps, decreasing computational cost and increasing the effective receptive field of subsequent convolutions. A pooling layer partitions each feature map into non-overlapping blocks of size $P_1 \times P_2 \times P_3$ and replaces each block with a single value, reducing the spatial dimensions by a factor of P_i along each axis while preserving the number of channels. The two most common variants are average pooling, which replaces each block with its arithmetic mean,

$$\mathrm{AvgPool}(X)_c(r) = \frac{1}{|\mathcal{B}_r|} \sum_{r' \in \mathcal{B}_r} X_c(r'), \tag{3.4}$$

and max pooling, which selects the maximum value within each block,

$$\mathrm{MaxPool}(X)_c(r) = \max_{r' \in \mathcal{B}_r} X_c(r'), \tag{3.5}$$

where $\mathcal{B}_r$ is the pooling block at position r and $|\mathcal{B}_r| = P_1 P_2 P_3$. Average pooling preserves the overall activation magnitude, while max pooling retains only the strongest response within each block, which can be more effective at propagating sharp features but discards information about the spatial distribution of activations. The filter models use average pooling exclusively.

The result of stacking multiple convolutional blocks with interleaved activations and pooling is a hierarchy of increasingly abstract feature representations: early layers respond to simple local patterns (edges, gradients), while deeper layers combine these into higher-order structures. After the final convolutional block, the 3D feature maps are flattened into a 1D vector and passed through fully connected (dense) layers. A fully connected layer computes $z = Wx + b$, where every input element is connected to every output neuron. Unlike convolutions, there is no weight sharing or spatial locality. The final layer produces $K = 2$ output logits (one per class), which are converted to a probability distribution by the softmax function:

$$p_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, \qquad k = 1, \ldots, K. \tag{3.6}$$

The output p_1 (class 1 probability) is the signalness score used for filtering.

The network is trained by minimizing the cross-entropy loss between the predicted class probabilities and the true labels. Cross-entropy originates from information theory, where it quantifies the expected number of bits needed to encode samples from the true distribution q using an encoding optimized for a predicted distribution p.[75] When $p = q$, the cross-entropy reduces to the entropy of the true distribution, the minimum possible encoding cost. Any deviation of the prediction from the truth increases the cross-entropy, so minimizing it forces the predicted distribution toward the true one. For a single event with true class $y \in \{0, 1\}$ and predicted probabilities (p_0, p_1), the loss reduces to

[75] Kullback and Leibler 1951, "On Information and Sufficiency".

$$\mathcal{L} = -\sum_{k=0}^{1} \mathbb{1}[y = k] \log p_k = -\log p_y, \tag{3.7}$$

where $\mathbb{1}[\cdot]$ is the indicator function. The loss equals zero when the predicted probability for the true class is 1 and diverges logarithmically as the prediction approaches 0, providing an increasingly strong gradient signal for confidently wrong predictions. The total loss over a mini-batch of N events is the mean $\frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_n$.

The learnable parameters $\theta = \{W^{(\ell)}, b^{(\ell)}\}_\ell$ across all layers are updated iteratively to minimize this loss via gradient-based optimization. The gradient $\nabla_\theta \mathcal{L}$ is computed by backpropagation,[76] a direct application of the chain rule through the computational graph. For a network with L layers, denoting the pre-activation at

[76] Rumelhart, Hinton, and Williams 1986, "Learning representations by back-propagating errors".

layer ℓ as $z^{(\ell)}$ and the post-activation as $a^{(\ell)} = \sigma(z^{(\ell)})$, the error signal propagates backward from the output:

$$\delta^{(L)} = \frac{\partial \mathcal{L}}{\partial z^{(L)}} = p - y, \qquad \delta^{(\ell)} = \left(\frac{\partial z^{(\ell+1)}}{\partial a^{(\ell)}} \right)^{\mathsf{T}} \delta^{(\ell+1)} \odot \sigma'(z^{(\ell)}), \tag{3.8}$$

where $p - y$, the difference between the predicted probabilities and the one-hot target, is the exact gradient of the cross-entropy loss composed with the softmax at the output layer. This remarkably simple form arises from the cancellation of the softmax Jacobian with the cross-entropy derivative, which is why ML frameworks evaluate the combined gradient directly rather than composing the individual gradients of softmax and cross-entropy separately (the latter would also be numerically less stable). $\odot$ denotes element-wise multiplication, and σ' is the derivative of the activation function. The parameter gradients follow as

$$\frac{\partial \mathcal{L}}{\partial W^{(\ell)}} = \delta^{(\ell)} \cdot \left(a^{(\ell-1)} \right)^{\mathsf{T}}, \qquad \frac{\partial \mathcal{L}}{\partial b^{(\ell)}} = \delta^{(\ell)}. \tag{3.9}$$

For convolutional layers, the weight gradient is computed as a cross-correlation between the incoming activations and the backpropagated error signal, summed over all spatial positions, a consequence of weight sharing. The filter models use AdamW,[77] an adaptive gradient method that maintains per-parameter running estimates of the first and second moments of the gradient and includes decoupled weight decay for regularization (see the training discussion below).

[77] Loshchilov and Hutter 2019, "Decoupled Weight Decay Regularization".

The physical motivation for using a CNN for this task is that the underlying physics is approximately translation invariant: a muon track passing through one region of the detector produces qualitatively the same light pattern as the same track passing through another region. Huennefeld emphasizes this use of symmetry to incorporate physics knowledge into the network design.[78] Weight sharing encodes this assumption directly. Early convolutional layers learn local visual patterns (lines, bright clusters, time gradients) that are useful everywhere in the detector, without needing to relearn them at each position. This is not strictly true—the DOMs are not identical, the string layout is only approximately hexagonal (Figure 2.3), and the glacial ice is optically inhomogeneous. But these local differences are second-order effects for a coarse topological classification task, and they can be captured by deeper layers, which combine the position-independent local patterns from early layers with position-dependent context from a larger receptive field, effectively allowing the network to weight the same local feature differently depending on where in the detector it appears. A concrete example is the dust layer: a region at approximately 2000–2100 m depth ($z \approx -150$ m to -50 m in detector coordinates) where glacial ice deposited during a cold period of the last glacial period roughly 65,000 years ago contains elevated dust concentrations, resulting in absorption and scattering coefficients several times higher than in the surrounding clean ice.[79] A muon track passing through the dust layer produces the same topological signature as one at a different depth, but the observed charge pattern is systematically dimmer. The early convolutional layers can still detect the

[78] Hünnefeld 2023, "Observation of high-energy neutrinos from the Milky Way", Sec. 2.1.1.

[79] AMANDA Collaboration 2006, "Optical properties of deep glacial ice at the South Pole", Abstract.

track pattern, while deeper layers can learn to account for the depth-dependent attenuation without requiring separate models for different detector regions.

For the filter models, the CNN operates in three spatial dimensions, two lateral dimensions corresponding to the hexagonal string layout and one depth dimension corresponding to the 60 DOMs along each string, making the convolution kernels 3D volumes rather than 2D patches.

3.3 Machine-learning filtering on DOM-level features

Throughout this chapter, *signal* and *background* refer to the machine-learning class labels defined by each model's training objective, not necessarily to astrophysical signal vs. atmospheric background in the physical sense, although the two may coincide for some models. For the starting track CNN, *signal* means events with starting track topology, which in nature are predominantly atmospheric neutrinos, not astrophysical in origin. The physical signal (astrophysical neutrinos) is a vanishingly small subset of the ML signal class. Similarly, *topology* refers to the observable spatio-temporal light pattern recorded by the detector, the spatial distribution and timing of charge deposits across the DOMs, not the true underlying particle interaction or light emission geometry, which may differ substantially from what is actually detected. The physical topology classes themselves are introduced in Section 2.1.

Lightning Tracks addresses the filtering tension with efficient ML models that operate directly on DOM-level pulse summary features, without requiring any reconstructions as input. For starting tracks, a single classification model handles the entire filtering decision, acting purely on event topology: whether an event exhibits a starting track signature. For upgoing tracks, the same approach applies: whether the event is consistent with an upgoing trajectory. Each CNN examines a single image-like representation of the event and produces a binary classification. There is no need for the filter to estimate an event's energy or reconstruct its direction beyond the upgoing versus downgoing distinction.

The design philosophy is analogous to a human physicist examining events in a visualization tool, tasked only with pre-selection: retaining anything that has a reasonable probability of being signal and rejecting only the clearly identifiable background. This is a comparatively straightforward visual classification task, which is why computationally inexpensive CNNs are effective. The objective at this stage is to maximize signal retention and address the residual background contamination at later stages, where more computationally expensive and more capable algorithms become available, and thus avoiding the aforementioned failure mode of overly aggressive early cuts based on computationally cheap reconstruction algorithms. Intuitively, we do not need to know the exact zenith angle[80] for a through-going track to classify it as upgoing, or the exact vertex of a starting track to classify it as starting. But attempting to reconstruct those quantities can more easily lead to misclassification when the estimates are unreliable.

[80] The zenith angle is the angle of an event's reconstructed arrival direction from the vertical. We define it relative to the particle's origin. So in IceCube's location and coordinate system a zenith angle of 0° describes a straight downgoing event from the southern sky.

Downgoing through-going tracks require a different approach. Single atmospheric muons and single neutrino-induced muons traversing the detector are topologically indistinguishable: a CNN operating on event topology alone cannot separate them. The downgoing through-going filter therefore uses existing L2 reconstructions (angular and energy estimates) and poses a distributional question: for an event at a given zenith angle, is it unusually bright compared to the expected atmospheric background? This amounts to a third step in a decision tree: the event does not exhibit a starting track signature, it is not upgoing, but at this angle and brightness, is it more consistent with a neutrino origin than an atmospheric muon? The MLP uses tabular features from existing L2 reconstructions, which the Lightning Tracks pipeline does not run but inherits. In that sense the downgoing through-going filter is more closely related to conventional filter algorithms, though it may find more efficient cuts through its physics-weighted training, where the explicit cuts of conventional methods are usually hand-picked.

The CNN filter models are referred to as LCSC filters (Lightning CNN Signal Classifier). The LCSC framework predates the full Lightning Tracks selection and can be seen as its origin—the development of these CNN-based filters was the starting point from which the rest of the selection was built.

Input features

Both CNN filters share the same input: 4 per-DOM summary features extracted by IceCube's ML Suite framework[81] from the pulse series. Before feature extraction, the `ChargeWeightedMedianTimePulseModifier` normalizes the event's time axis: it computes a single reference time—the charge-weighted median pulse time over all pulses in the event—and subtracts it from every pulse's time, placing all events on a common timing scale that is more consistent across events than the effectively arbitrary trigger time. Pulse charges and the per-DOM pulse series are unchanged. The per-DOM summarization into the four features happens afterward, in the feature extraction step. For each individual DOM, time windows flagged for unreliable data (e.g. PMT saturation) are excluded. The four features are the total charge deposited at the DOM, the time of its first pulse, and the charge-weighted mean and standard deviation of its pulse times (two values).

[81] ML Suite is part of IceCube's public icetray repository, with no accompanying publication. It extracts per-DOM summary features from pulse series for machine-learning applications.

With 4 features per DOM across 78 strings (DeepCore excluded) and 60 DOMs per string, the raw input consists of 18,720 values per event before grid transformation. This is a reduced feature set compared to earlier iterations that used 14–15 features per DOM. The reduction had no measurable impact on classification accuracy but substantially reduced computational cost (Section 3.7).

Input transformation

The raw ML Suite output has shape $(86, 60, 4)$ per event: 86 strings, 60 DOMs, and 4 features. Three transformations are applied before the data enters the CNN.

First, the 8 DeepCore strings are discarded, leaving the 78 main-array strings (see the prior-art discussion, Section 3.8, for the rationale). Second, a log transform

$\log(Q + 1)$ is applied to the total charge feature, compressing the dynamic range from several orders of magnitude into a scale more amenable to gradient-based optimization.

Third, the 78 string positions are mapped onto a 10×10 rectangular grid that preserves all n-th neighbor relations, effectively a coordinate transformation from physical positions into index space, so that convolution kernels of any size operate on physically neighboring strings. IceCube's main array is arranged in an approximately hexagonal pattern (Figure 2.3); the grid mapping exploits this regularity (Figure 3.2). Grid cells that do not correspond to a string position are zero-padded. The features are moved to axis 1 (PyTorch channel-first convention), yielding a final input shape of $(4, 10, 10, 60)$: four feature channels on a 10×10 lateral grid with 60 depth layers.

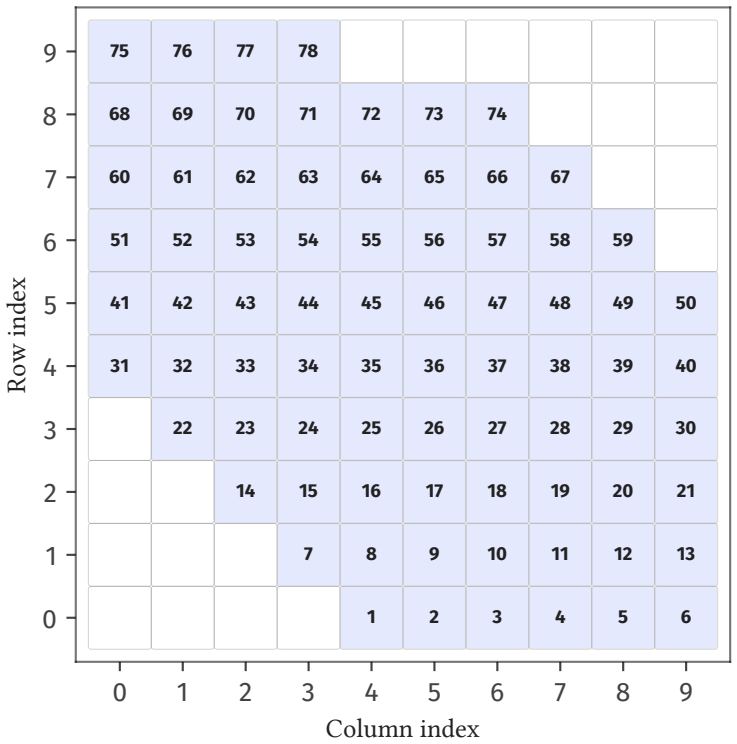

Figure 3.2: Mapping of the 78 string positions to the 10×10 rectangular grid. Filled cells correspond to instrumented string positions. Empty cells are zero-padded. The mapping preserves the spatial neighbor relations of the physical string layout so that convolution kernels of any size operate on physically neighboring strings.

Figure 3.3 demonstrates the neighbor preservation concretely: for any grid position and kernel size, the highlighted cells in the physical layout correspond to the kernel footprint in the grid representation. The convolution kernel always covers a spatially contiguous group of strings, even near the edges of the hexagonal array where zero-padded grid cells enter the receptive field.

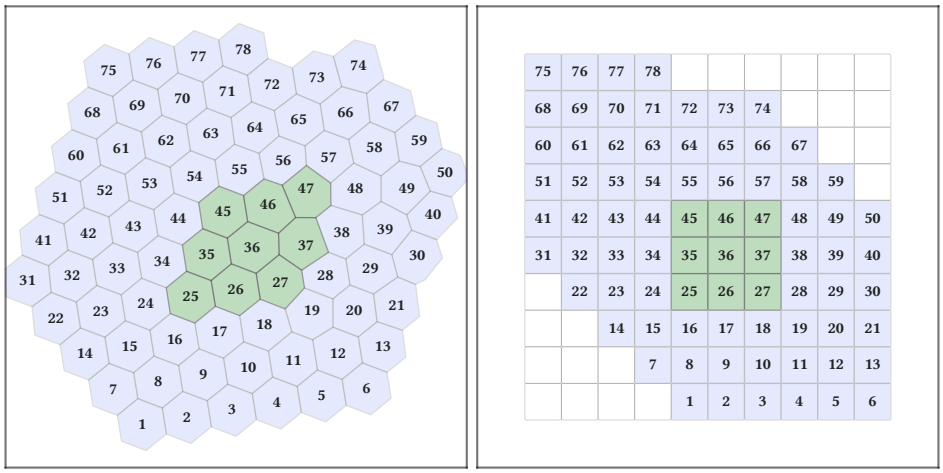

Figure 3.3: Neighbor preservation under the grid mapping for one example position: the highlighted cells in the physical layout (left) correspond to the footprint of a 3×3 convolutional kernel in the grid representation (right).

3.4 The starting- and upgoing-track CNNs

Classification objectives

Two separate CNN models are trained for two distinct binary classification tasks: one to identify starting tracks (contained interaction vertex) and one to identify upgoing tracks. They share the same architecture but differ in their training data and labels.

The choice of two binary classifiers over a single multi-class model is deliberate. Binary classification is a simpler task at every level: easier for the model to learn, easier to evaluate, and, critically, easier to prepare training data for, since the question of how to balance two classes is far more straightforward than balancing three or more. The two classification objectives are also not mutually exclusive: a starting track can simultaneously be an upgoing track, which makes a single multi-class model with disjoint categories a poor fit. Finally, since filter-stage performance is paramount (Section 3.7), two small, highly specialized models are more efficient than one larger model that attempts both tasks at once.

The starting track CNN classifies whether an event has a contained interaction vertex, a starting track topology (Section 2.1), or not. All training data comes from NuGen[82] simulation (~20.5 million events), split by MC truth into two classes.

The signal class consists of events with a contained interaction vertex: the neutrino interacts inside the detector volume and the outgoing muon exits it, producing a track that visibly starts within the instrumented region. The containment volume (Figure 2.2) is defined as the three-dimensional convex hull of all DOM positions, the minimal convex polyhedron enclosing every DOM center in the detector. In principle this volume is a simple hexagonal prism, but in practice the strings are

[82] As a reminder, NuGen is the part of IceCube's MC simulation chain that generates and propagates neutrinos (Section 2.5).

deployed into drill holes in the refrozen ice, and the drill depth varies across the array. Laterally, the DOMs are well aligned along each string (maximum deviation <0.5 m from the string axis), but the vertical extent differs substantially between strings: the topmost DOMs span a range of ∼33 m across the array ($\sigma \approx 5$ m), and the bottommost DOMs show a comparable spread. The true DOM positions are determined via in-situ calibration using the onboard LED flashers, which constrains them via trilateration from inter-DOM light travel times. The resulting convex hull is therefore not an ideal prism but has an uneven top and bottom surface that follows the drill depth at each string position.

A ν_μ CC event is labeled *starting* if the muon track originates inside this volume and crosses its boundary outward. Events where both the vertex and the muon endpoint are contained, fully contained tracks where the muon decays before reaching the detector boundary (Figure 2.2), are excluded from the signal class despite having a contained vertex. At the lowest energies, these events can be short enough to resemble cascade-like signatures, blurring the class boundary. Contained tracks with longer muon range do exist and are topologically distinct from cascades (the muon usually travels much further than an electron before losing all of its energy), but at the energies where full containment occurs the muon is faint enough that reliable separation from single atmospheric muons penetrating the detector becomes difficult.

Upgoing contained events, which are the more interesting subset since upgoing events are far less likely to be single atmospheric muons, are recovered by the upgoing track filter regardless—so excluding contained events from the starting track signal class does not result in a loss of coverage. More broadly, these are events that would almost certainly not survive the filter threshold in practice. Including them in the signal class would weaken the class boundary without contributing events that the filter could realistically retain. Excluding them produces a cleaner, more well-defined signal class, which translates directly into improved classifier performance. This is another instance of the general principle that training data curation has a far larger impact on filter quality than architectural choices (see the architecture discussion below).

Signal events are further restricted to those with at least 10 photoelectrons observed from the outgoing muon. The 10 PE threshold enforces consistency between the truth label and the observable detector response. An event can satisfy the geometric definition of a starting track (vertex inside the containment volume, muon exiting) while producing a detector signature that bears no resemblance to a starting track. At low energies or in geometrically unfavorable configurations, the outgoing muon may traverse only a short path through instrumented ice and deposit too little Cherenkov light to produce a recognizable track. Such events are indistinguishable from detector noise or faint cascades at the level of the per-DOM summary features available to the CNN. Labeling them as signal would degrade the classifier by associating the signal class with feature patterns that carry no topological information. The 10 PE requirement removes these ambiguous cases and restricts the signal class to events where the starting track topology is at least in principle identifiable from the observable charge and timing pattern. It

also excludes events where the neutrino interacts far from the detector and the outgoing muon never reaches the instrumented volume. In these cases there is no neutrino-induced light in the detector at all, and the event consists entirely of noise or injected coincident atmospheric muon background.

The background class includes all NuGen events from the same simulation that are not starting tracks, with no additional quality requirements: through-going tracks, stopping tracks, cascades, pure noise triggers, and topologically ambiguous events of any kind. The training signal is sampled approximately uniformly in deposited energy and zenith angle, so that the model focuses on topology rather than learning energy or angular distributions.

The upgoing track CNN classifies whether an event is upgoing or downgoing, using the same NuGen simulation split by the true neutrino direction. The signal class consists of true upgoing track events, defined by a true neutrino zenith angle $\geq 83°$, corresponding to declination $\geq -7°$, slightly below the *muon horizon*. We define the muon horizon as the declination at which the atmospheric neutrino and atmospheric muon rates cross, south of which the background is muon dominated, north neutrino dominated. It is a more natural hemisphere transition than the geographical horizon; its physical origin is taken up in Section 5.2. Signal events are additionally restricted to true tracks and at least 10 PE observed from the primary muon. The background class includes all true downgoing events (true neutrino zenith $< 83°$), with no additional quality requirements: all event topologies, not just tracks.

Both models are trained as standard binary classification with cross-entropy loss and output a softmax probability (signalness score) in $[0, 1]$.

Architecture

Both CNN filters use the same architecture with different trained weights. Table 3.1 summarizes the network, and Figure 3.4 shows its structure schematically.

Table 3.1: CNN architecture for both the starting track and upgoing track filter models.

Property	Value
Conv layers	8
Filters per layer	50, 50, 50, 50, 50, 50, 50, 10
Kernel size	$3 \times 3 \times 3$
Padding	Same
Pooling	Alternating: none, avg, none, avg, none, avg, none, avg
Pooling kernel	[1, 1, 2], [2, 2, 2], [2, 2, 2], [2, 2, 2]
Normalization	Batch normalization after each conv + pool + activation
Activation	ELU
FC layers	Flatten → 50 neurons (ELU) → 2 classes (ELU)
Output	Softmax probability (signalness = class 1 probability)

Input batch normalization[83] without learnable parameters (no affine transformation) is applied before the first convolutional layer, serving as an implicit input normalization step. Rather than computing global feature statistics in a separate preprocessing pass and maintaining a standalone normalization model, the network normalizes its own input using running batch statistics during training and frozen statistics at inference. This approach is slightly less efficient during the first few training steps, before the running statistics have converged, but this is irrelevant in practice given the training duration. The advantage is that all normalization is encapsulated in the exported model weights: no separate normalization function or statistics file needs to be maintained or shipped alongside the model, simplifying deployment. This differs from the prior work (Section 3.8), which uses a dedicated input normalization step with precomputed global statistics.

[83] Ioffe and Szegedy 2015.

The first pooling kernel is asymmetric, $[1, 1, 2]$, pooling only in the depth (z) dimension and preserving the lateral hex grid resolution in early layers. Subsequent pooling layers reduce all three spatial dimensions by a factor of 2.

The classification task is sufficiently straightforward that this minimal architecture is not a limiting factor. Substantially more complex designs, the same CNN with many more filters and layers, or deeper architectures such as ResNet,[84] yielded no measurable improvement in downstream performance when evaluated on data rates and effective areas rather than training metrics alone. The factor that had a much larger impact on the quality of the final selection was the curation of the training data: how to define truth labels that are consistent with what is actually observable in the detector, how to handle topological edge cases, and how to balance the training classes (see the classification objectives above). These choices propagate directly into the filter's signal retention and background rejection, whereas architectural complexity beyond a minimal threshold does not.

[84] He et al. 2016, "Deep Residual Learning for Image Recognition".

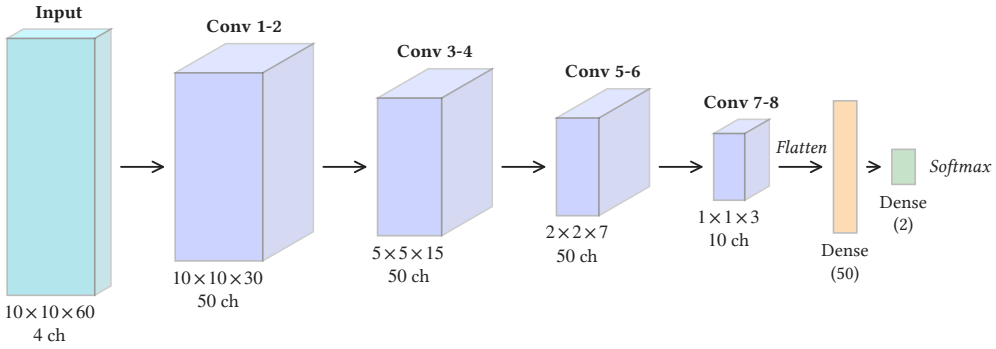

Figure 3.4: Schematic of the BasicNet 3D CNN architecture used for both the starting track and upgoing track filters. The input is a 4-channel $10 \times 10 \times 60$ hex grid representation of the event. Eight convolutional layers with alternating average pooling progressively reduce the spatial dimensions, followed by a fully connected classification head.

Evaluation metrics

Two standard metrics are used to evaluate classifier performance during training: accuracy and the area under the receiver operating characteristic curve (AUROC).

Accuracy is the fraction of correctly classified events at a given score threshold τ,

$$\text{Accuracy}(\tau) = \frac{N_{\text{correct}}(\tau)}{N_{\text{total}}}, \tag{3.10}$$

where an event is classified as signal if its score exceeds τ and as background otherwise. The default threshold is $\tau = 0.5$, but this is not meaningful for the filter models: the operating thresholds (0.99–0.999) are far above 0.5, and the relevant performance regime is deep in the tail of the score distribution. Accuracy at $\tau = 0.5$ is dominated by the bulk of easily classified events and provides no information about how well the model separates signal from background near the operating point.

The receiver operating characteristic (ROC) curve avoids this problem by characterizing classifier performance across all possible thresholds simultaneously. It plots the true positive rate (TPR, also called signal efficiency or recall) against the false positive rate (FPR) as the threshold varies from 0 to 1:

$$\text{TPR}(\tau) = \frac{N_{\text{signal}}(s \geq \tau)}{N_{\text{signal}}}, \qquad \text{FPR}(\tau) = \frac{N_{\text{background}}(s \geq \tau)}{N_{\text{background}}}. \tag{3.11}$$

A perfect classifier produces TPR = 1 at FPR = 0; a random classifier follows the diagonal TPR = FPR. The area under the ROC curve (AUROC) is the integral under this curve and ranges from 0.5 (random) to 1.0 (perfect). It is the probability that a randomly chosen signal event receives a higher score than a randomly chosen background event, independent of any particular threshold choice.

AUROC is the primary metric used for checkpoint selection during training. It is preferred over accuracy because the operating threshold is not known a priori during model development. For the filter models, it is determined afterward based on the achieved data rate and signal-to-noise ratio (Section 3.6). AUROC captures the model's discriminating power across the entire score range without requiring a threshold to be specified. That said, not all regions of the ROC curve are equally relevant for this application. Because the filter must reduce the data rate by several orders of magnitude, the operating point lies at very low FPR (high threshold), and what matters most is the behavior of the leading edge of the ROC curve: how steeply the TPR rises at small FPR. A model that achieves high TPR already at FPR $\ll 1$ retains most of the signal even at the aggressive thresholds required for rate reduction. A more targeted metric such as the TPR at a fixed low FPR would in principle be better suited to this regime, but in practice the two metrics are strongly correlated for well-performing classifiers, and AUROC has the advantage of being a standard, well-understood quantity. For all three filter models, both CNNs and the MLP, this leading edge is extremely steep, rising nearly vertically at very low FPR before gradually curving over. The operating thresholds are chosen to capture this near-vertical portion of the curve, just before the onset

of significant curvature, maximizing signal retention at the rate reduction required for downstream processing.

These metrics are used during training because they are computationally inexpensive and provide a useful signal for model selection and convergence monitoring. They do not, however, reflect the model's real-world performance on the actual filtering task, where the class balance is extreme (signal events are a vanishingly small fraction of the input data) and the evaluation must account for physics-weighted event rates. The real evaluation happens after training, by applying the trained classifiers to experimental data and weighted MC simulation and assessing the resulting data rates, effective areas, and ultimately point source sensitivity, as described in Section 3.6 and Chapter 5.

Training

Both CNN models are trained using PyTorch[85] with PyTorch Lightning[86] and Hydra[87] configuration management. All filter and final cut models in the Lightning Tracks pipeline are built on this stack, and, as an aside, the *Lightning* in Lightning Tracks is a nod to PyTorch Lightning rather than an acronym or a reference to the physics. The optimizer is AdamW with a learning rate of 10^{-3} and weight decay of 10^{-2}. Training uses FP16 mixed-precision arithmetic with a batch size of 512.

Events are drawn from an infinite iterator via rejection sampling with signal and background balanced at a 1:1 ratio, and each epoch consists of a fixed 100 batches (51,200 events) rather than a full pass through the data. The 1:1 class balance is the only sensible choice for binary classification training: any other ratio shifts the model's default prediction toward the majority class, reducing the information gained per training step. In the extreme case of using the true class balance at the filter input, roughly one atmospheric neutrino for every $\sim 10^4$ events entering the filters, the model would learn to predict the background class unconditionally, producing a useless classifier.

The choice of training dataset deserves comment. The NuGen dataset we use was generated specifically with high starting event statistics by forcing neutrino interactions inside and close to the detector volume and retroactively correcting the weights to account for the forced geometry. More recent NuGen productions do contain sufficient starting event statistics, but it has an additional practical advantage: it covers the full energy range (100 GeV–100 PeV) in a single continuous simulation with a hard spectral index ($\gamma = 1.5$), producing approximately uniform statistics across energies without elaborate reweighting. The standard modern productions are split into separate low-, mid-, and high-energy chunks that would each need to be loaded, weighted, and merged, substantially more preprocessing overhead for a training task that is purely topological and does not depend on spectral shape. The hard spectrum also strikes a reasonable balance for training: a perfectly uniform energy distribution would allocate equal capacity to low-energy events (which are both more common in nature and the harder classification cases) and to the high-energy tail (which is comparatively easy, since nearly all high-energy events pass regardless). A power law with $\gamma = 1.5$ naturally emphasizes the

[85] Paszke et al. 2019, "PyTorch: An Imperative Style, High-Performance Deep Learning Library".

[86] PyTorch Lightning is an open-source training framework for PyTorch, with no accompanying publication: `github.com/Lightning-AI/pytorch-lightning`.

[87] Hydra is an open-source framework for configuration management, with no accompanying publication: `github.com/facebookresearch/hydra`.

lower energies where the classifier needs to discriminate most carefully, without entirely neglecting the high-energy regime.

The dataset contains only ν_μ CC and NC interactions; the sibling ν_e and ν_τ datasets were not included, nor was any dedicated atmospheric muon simulation added to the background class. This is sufficient for both classification tasks. The core challenge for the starting track CNN is distinguishing true starting tracks from single through-going muons, the hardest instance of the classification problem. A through-going muon track in the detector is topologically identical regardless of whether it originates from a neutrino interaction or from an atmospheric muon, so ν_μ simulation alone fully captures this case without requiring dedicated atmospheric muon simulation. Easier variants of the same problem, separating starting tracks from cascades, or from incoming muon bundles, are solved implicitly by a classifier that handles the harder case, since cascades and bundles are more topologically distinct from starting tracks than single through-going muons are. Including these easier cases in the training data would not improve the classifier but would waste training capacity on instances the model already handles and slow convergence on the instances that actually matter. The NC interactions in the ν_μ dataset do produce hadronic cascades as part of the background class, but their inclusion is incidental rather than by design.

Although dataset 21217 uses an older ice model than the latest SnowStorm productions (used for all subsequent stages), we verified that this has no significant impact on the CNN classification task: the per-DOM summary features the CNNs operate on are robust to ice model differences at this level of topological classification. The use of an older dataset for the filter stage also avoids a more fundamental problem: any simulation used for training at one stage cannot be reused for training or validation at a later stage without risking overtraining bias. Reserving the modern SnowStorm NuGen for the final cut models (Chapter 4) and the analysis-level MC ensures clean statistical separation between stages.

Training was run for a fixed 24 hours, well beyond the point of convergence. The extended run was deliberate: it produces the full training curves shown in Figure 3.5, demonstrating that the models remain stable and never begin to overtrain. The exported model is the checkpoint with the highest validation AUROC observed during the full run.

The starting track CNN was trained on the aforementioned NuGen dataset split by MC truth containment; the best checkpoint was reached at epoch 3,931. The upgoing track CNN was trained on the same dataset split by true neutrino direction; the best checkpoint was reached at epoch 4,643. Both training curves are shown in Figure 3.5.

The training curves empirically confirm that overfitting is not a concern with the available data volume. For the upgoing track CNN and the downgoing MLP, the validation metrics track the training metrics almost exactly: the curves lie on top of each other, with the validation curves showing only higher variance from the smaller evaluation sample. This is in contrast to the typical behavior in data-limited ML applications, where a persistent gap between training and validation performance indicates that the model is memorizing training-specific patterns. In

fact, during the first few epochs the validation metrics slightly exceed the training metrics, an initially counterintuitive observation that is explained by the evaluation protocol: the training metrics are computed as a running average over each epoch while the weights are being updated, whereas the validation metrics are evaluated once at the end of the epoch using the final weights from that epoch. During early training, when the weights improve rapidly within each epoch, the end-of-epoch weights are substantially better than the epoch-averaged weights, producing a transient validation advantage. This effect is uncommon in conventional ML applications where per-epoch weight changes are small, but is expected here given the fast convergence rate. The starting track CNN does show a small persistent gap between training and validation metrics, presumably due to lower statistics in the high-energy regime where starting event rates are sparse, but exhibits no signs of overfitting: the training and validation curves remain parallel throughout convergence, with no point at which the validation metrics begin to diverge.

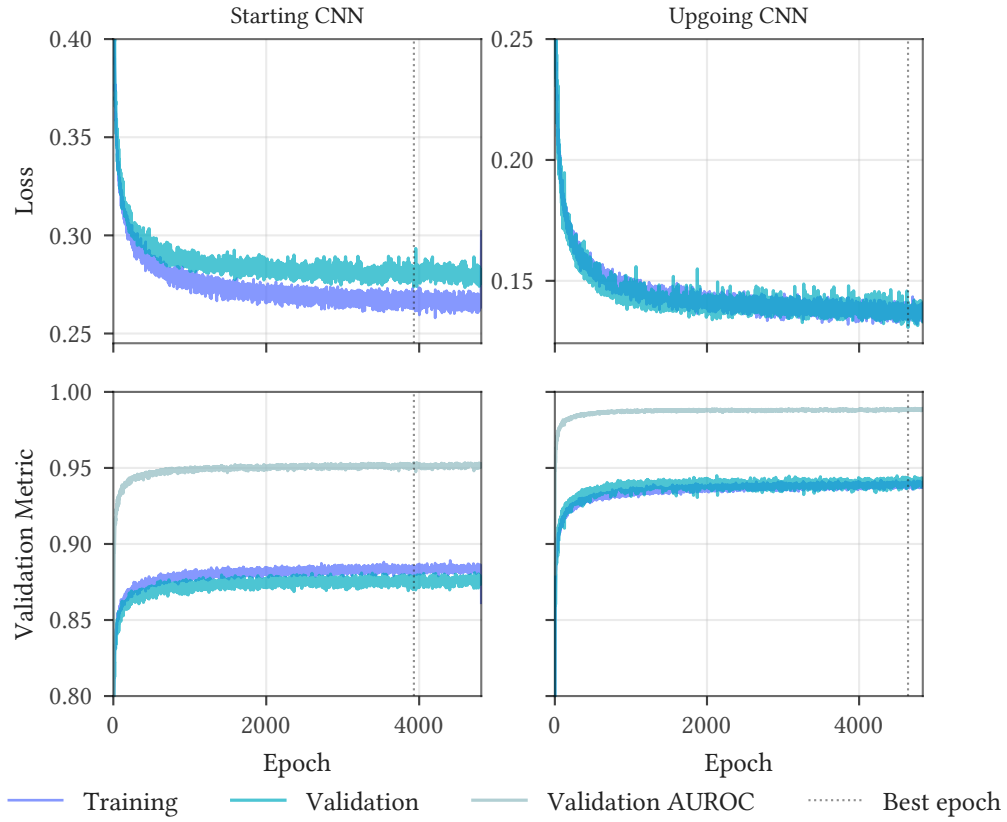

Figure 3.5: Training curves for the starting track (left column) and upgoing track (right column) CNN filters. Top row: training and validation loss. Bottom row: training accuracy and validation AUROC. Dashed vertical lines mark the best checkpoint selected by validation AUROC (epoch 3,931 for starting, epoch 4,643 for upgoing). Training was terminated by the SLURM wall time limit (23 h 45 min).

Performance and cut thresholds

The starting track CNN operates at a threshold of 0.999; the upgoing track CNN at 0.99. Figure 3.6 shows the score distributions for signal and background on held-out test data, and Figure 3.7 shows the corresponding ROC curves.

The thresholds were chosen to achieve the target data rates at which the reconstruction algorithms (Chapter 4) become computationally affordable, and no further than that. Although the filter models could be used to achieve an even lower data rate, the originally stated objective of the filtering applies: remove no more events than absolutely necessary.

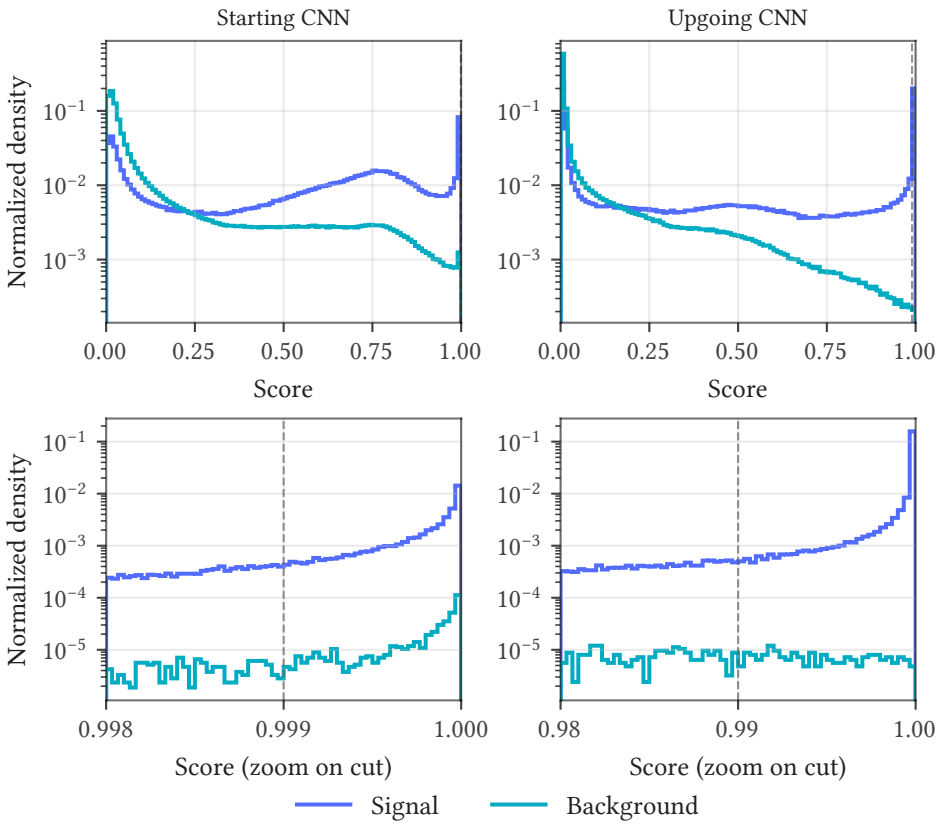

Figure 3.6: Score distributions for the starting track CNN (left) and upgoing track CNN (right) on unweighted NuGen test data. Signal and background classes are independently normalized. Vertical lines mark the operating thresholds (0.999 for starting, 0.99 for upgoing).

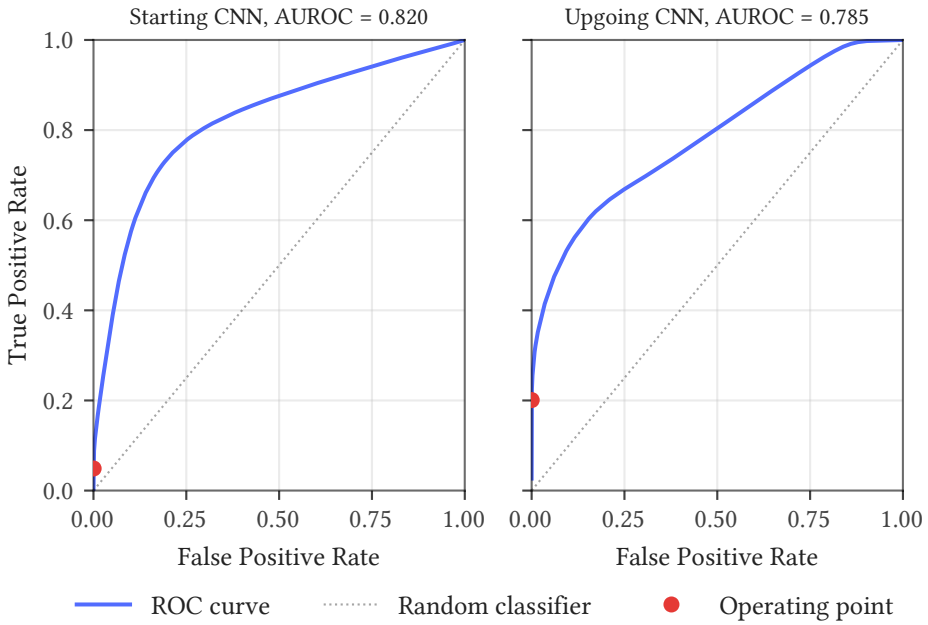

Figure 3.7: ROC curves for the starting track and upgoing track CNN filters, evaluated on held-out test data.

3.5 The downgoing through-going MLP

The downgoing through-going MLP is also a single binary classification task: for a downgoing through-going track, is this event more consistent with a neutrino origin or an atmospheric muon?

The signal class consists of NuGen simulation events with a true zenith angle $< 83°$ (a subset of the upgoing CNN's background class), restricted to downgoing tracks in truth that pass the same quality requirement (at least 10 PE observed from the primary muon). This ensures that the MLP is trained only on the event population it is designed to classify, without contamination from upgoing events that would already be captured by the upgoing CNN. Unlike the two CNN filters, which are trained to be spectrally independent and classify purely on topology, the downgoing MLP is explicitly trained to exploit the spectral difference between astrophysical neutrinos and atmospheric muons—the only discriminating lever available for downgoing through-going tracks, which are topologically identical. Signal events are weighted to a hard astrophysical power-law spectrum, roughly matching the diffuse astrophysical spectrum[88], while the background class consists of CORSIKA[89] atmospheric muon simulation. This teaches the model that a downgoing event must be increasingly bright (high charge or energy) to be considered signal-like, because the atmospheric muon rate rises steeply toward the downgoing direction while the astrophysical flux falls much less steeply; this spectral contrast is visualized in Figure 3.8. Weighting the signal class to a soft atmospheric neutrino spectrum was also tested but produced no useful classification, as expected:

[88] IceCube Collaboration 2022b, "Improved Characterization of the Astrophysical Muon-Neutrino Flux with 9.5 Years of IceCube Data".

[89] Heck et al. 1998, *CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers.*

atmospheric neutrinos and atmospheric muons have similar spectral shapes in the downgoing direction, leaving no spectral lever for the classifier to exploit.

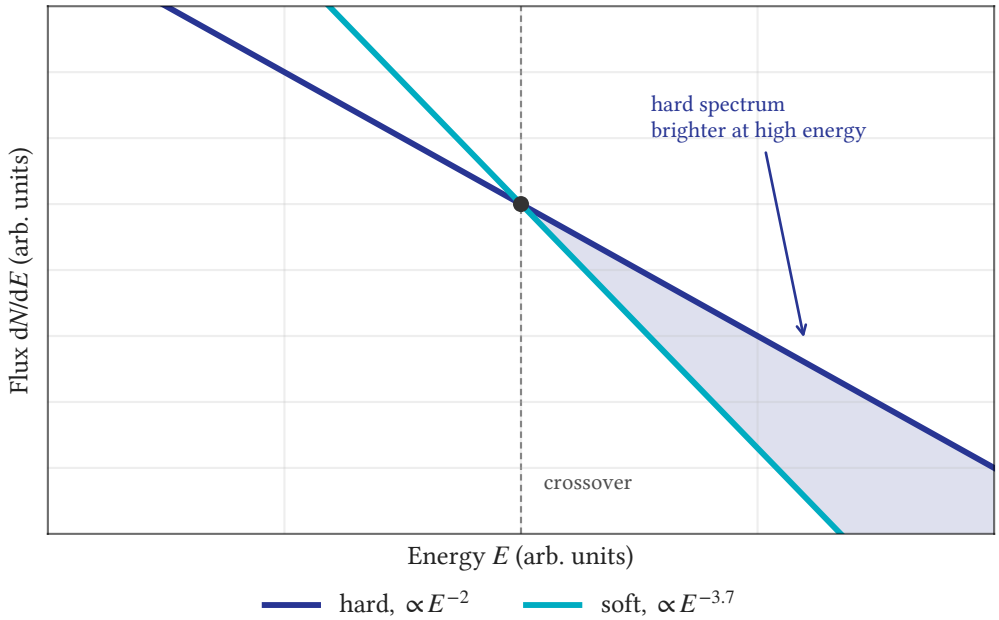

Figure 3.8: A hard ($\propto E^{-2}$) and a soft ($\propto E^{-3.7}$) power law on log-log axes, normalized to cross at a single energy (both axes in arbitrary units). Below the crossover the soft spectrum carries the larger flux; above it the hard spectrum does, so at high energy a hard source outshines a softer background (shaded region). This is a schematic illustration of the brightness discriminant only. It is not the actual atmospheric-muon flux, which is not a single power law.

Multilayer perceptrons

A multilayer perceptron (MLP) is a feedforward neural network that maps a fixed-length input feature vector to an output through a sequence of affine transformations interleaved with nonlinear activation functions. An MLP with L hidden layers computes

$$f(x) = \sigma_L \circ h_L \circ \sigma_{L-1} \circ h_{L-1} \circ \cdots \circ \sigma_1 \circ h_1(x), \tag{3.12}$$

where each layer $h_\ell(z) = W_\ell z + b_\ell$ is an affine transformation with learnable weight matrix W_ℓ and bias vector b_ℓ, and σ_ℓ is a nonlinear activation function (here, ReLU: $\sigma(z) = \max(0, z)$). Unlike CNNs, MLPs have no notion of spatial structure: every input feature is connected to every neuron in the first hidden layer, and the network treats the input as an unordered vector. This makes MLPs appropriate for tabular data (such as the L2 reconstruction features used by the downgoing through-going filter) but unsuitable for spatially structured data where local patterns matter. The MLP is the architecture directly addressed by the universal approximation theorem

(Section 3.2): a single hidden layer with a non-polynomial activation is sufficient in principle to approximate any continuous function on a compact domain.

The relationship between the two architectures is worth noting: the fully connected classification head of the CNN (Section 3.2) is itself an MLP that operates on the flattened output of the convolutional layers. In this view, the convolutional portion of the CNN serves as a learned feature extractor that transforms the raw spatial input into a compact, spatially invariant representation, which is then classified by the same type of fully connected network used here. The difference is that the CNN learns its features from spatial data, whereas the standalone MLP receives summary features from L2 reconstructions.

The final cut models (Section 4.3) use the same MLP framework with different input features and training objectives; the architectural and training details described here apply to both.

Table 3.2: MLP architecture for the downgoing through-going filter model.

Property	Value
Input dimension	30
Hidden layers	[128, 64]
Activation	ReLU
Normalization	Batch normalization per hidden layer
Regularization	Dropout (0.2)
Output	Sigmoid (single neuron)

Input features

The MLP uses 30 tabular event-level features from existing L2 reconstructions: the homogenized total charge (the total charge weighted by each DOM's light-detection efficiency); the reconstructed energy, direction, and fit-quality statistics of two simple likelihood-based track reconstructions; and the reconstructed direction from a line fit, a simple linear fit to the spatial and temporal pulse distribution. It operates entirely on these high-level event summaries, using no DOM-level or pulse-level information. Two features undergo logarithmic preprocessing before entering the network, $Q_{\text{total}} \to \log_{10}(1 + Q_{\text{total}})$ and $E_{\text{MPEFit}} \to \log_{10}(1 + E_{\text{MPEFit}})$. All features are then standardized via batch normalization. Note that the feature set includes the reconstructed zenith angles: as emphasized in Section 3.6, this by itself does not make the selection sensitivity-optimal. The sensitivity-driven, declination-dependent thresholding is applied on top of the score (Chapter 5).

This reliance on reconstructions seemingly goes against the stated design principle of avoiding them at the filter level. But for tracks that are both downgoing and through-going we simply have no other option, since they are topologically indistinguishable from the background (Section 3.3): a muon from a neutrino interaction and a muon produced directly in the atmosphere are physically the

same thing, a muon, and without the starting information there is nothing to distinguish them.

Training

[90] Kingma and Ba 2015, "Adam: A Method for Stochastic Optimization".

The MLP is trained using PyTorch Lightning. The optimizer is Adam[90] with a learning rate of 10^{-3}. The loss function is binary cross-entropy. For regularization we use only dropout (0.2 per hidden layer) here. No explicit L2 weight penalty is applied. The batch size is 2,048.

The epoch definition follows the same infinite-iterator approach as the CNNs: 100 batches per epoch with 1:1 balanced sampling between signal and background, drawing events with replacement proportional to their physics weights. A 90/10 train/validation split is used. The model was trained for 200 epochs and exported at the final epoch. Figure 3.9 shows the physics-weighted zenith distributions of the training data, and Figure 3.10 shows the training curves.

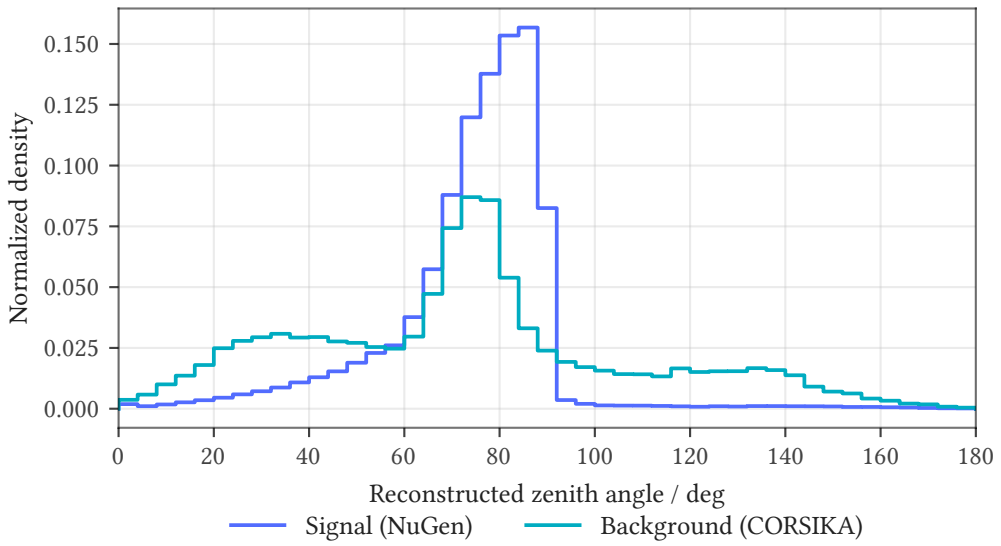

Figure 3.9: Physics-weighted MPEFit reconstructed zenith distributions of the MLP training data. NuGen signal and CORSIKA background are shown separately. The weighting teaches the model the prior signal-to-background ratio as a function of zenith.

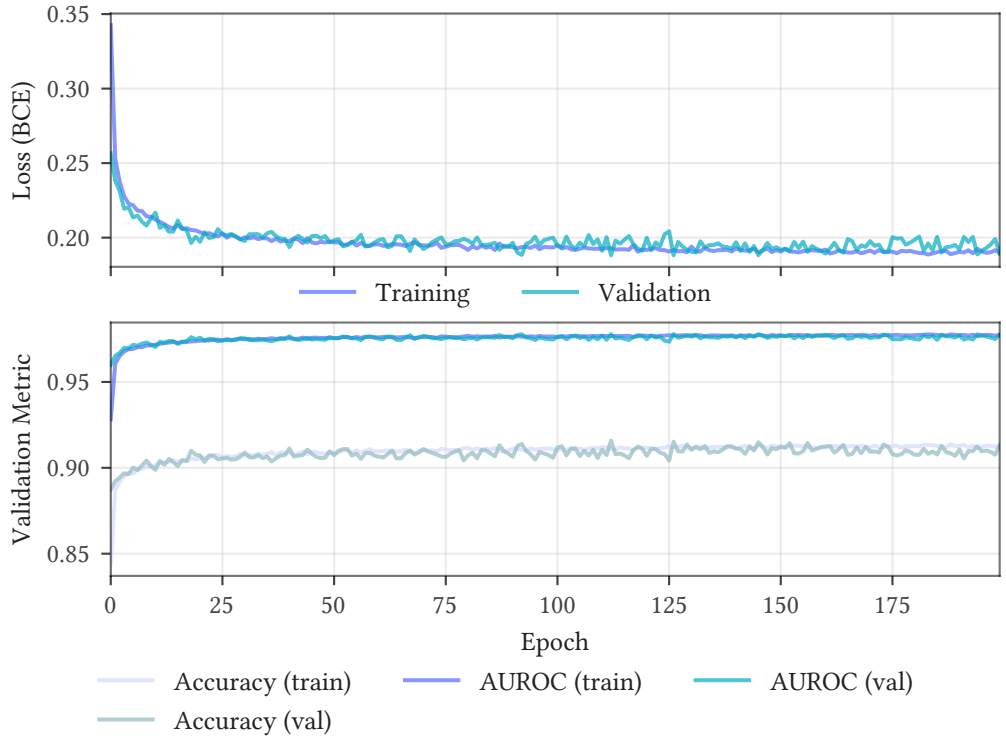

Figure 3.10: Training and validation loss curves for the downgoing through-going MLP filter. The model was trained for 200 epochs and exported at the final epoch.

Performance and cut threshold

The MLP operates at a cut threshold of 0.99. Figure 3.11 shows the score distribution and Figure 3.12 the ROC curve on held-out test data.

The ROC curve in Figure 3.12 shows two distinct bumps, an expected feature of binary classifiers operating on data with multiple distinct sub-populations. Both the signal and background classes contain populations with very different classification difficulty. On the background side, some CORSIKA events, such as those with misreconstructed upgoing zenith (compare the zenith distributions in Figure 3.9), are trivially distinguishable from signal and receive very low scores—the signal class explicitly excludes upgoing events (captured by the upgoing filter), and the NuGen events are on average much better reconstructed, producing fewer events misreconstructed as upgoing than CORSIKA. Genuine downgoing muons near the horizon with comparable brightness to neutrino events are much harder to separate. On the signal side, the MLP is trained with an astrophysical spectrum chosen to approximate the diffuse flux, so the high-energy tail where the hard astrophysical flux vastly exceeds the steeply falling atmospheric muon rate is trivially identifiable as signal, while lower-energy downgoing neutrinos are genuinely ambiguous.

The steep leading edge of the ROC curve (low FPR) reflects both of these easy regimes working together: trivially rejected background keeps FPR low, while

trivially identified high-energy signal drives TPR up. The bump around FPR $\approx$ 0.4 marks the transition into the hard regime, where the remaining signal and background overlap in the feature space that the MLP has access to. Above that point, the curve runs nearly parallel to the diagonal, not because the classifier is performing at random, but because the remaining populations are genuinely difficult to separate and lowering the threshold admits both at comparable rates.

For our purposes this is perfectly fine, since the operating threshold lies deep in the steep leading edge, well before the transition—and the primary objective of this filter is to capture the extremely high-energy downgoing through-going regime for sensitivity to hard source spectra.

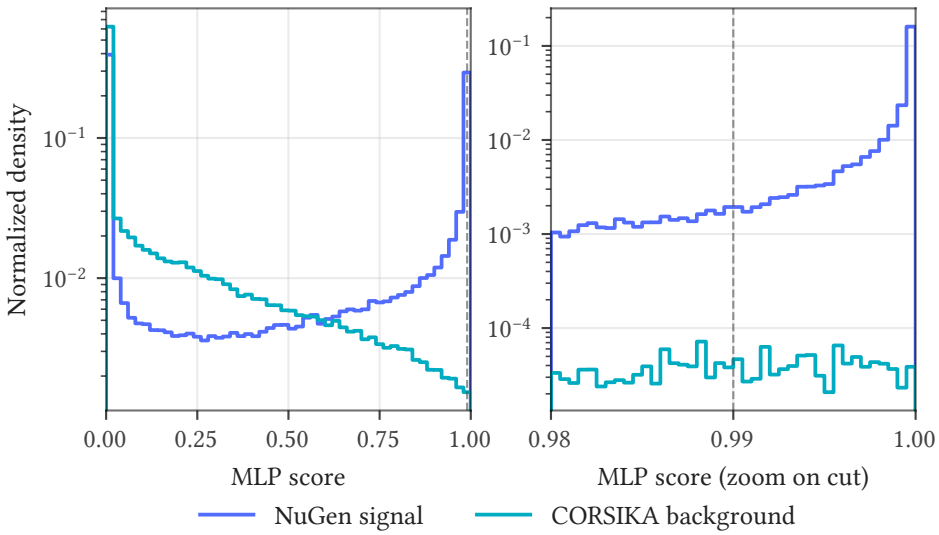

Figure 3.11: Score distribution for the downgoing through-going MLP filter. Solid histogram shows NuGen signal, dashed shows CORSIKA background.

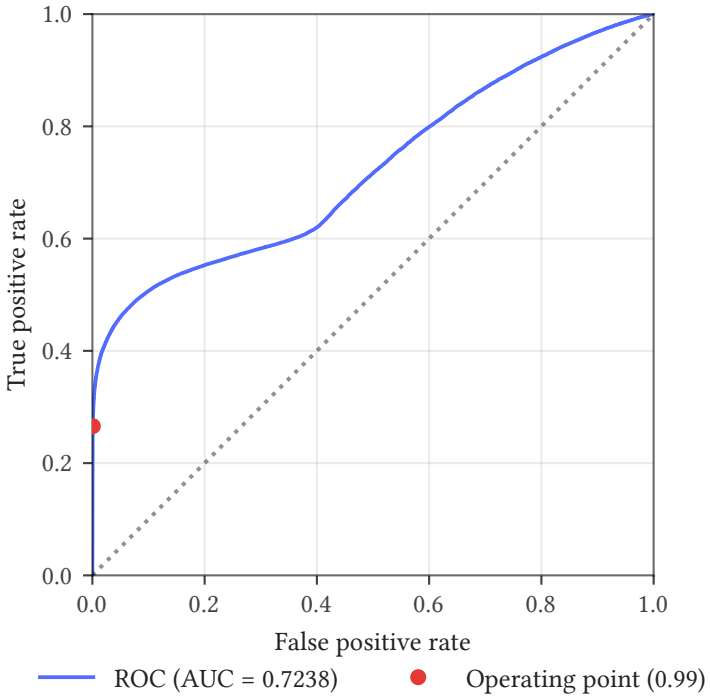

Figure 3.12: ROC curve for the downgoing through-going MLP filter, evaluated on held-out test data.

3.6 Filter combination and performance

The three filters are combined with an inclusive OR: an event enters the filtered sample if it passes any one of the three thresholds (Table 3.3). The resulting filter-level data rate is approximately 20 mHz for starting candidates and 70 mHz for through-going candidates.

Table 3.3: Filter model thresholds. An event enters the filtered sample if it passes any one threshold.

Filter	Threshold	Events targeted
LCSC starting track CNN	> 0.999	Starting track candidates (contained vertex)
LCSC upgoing track CNN	> 0.99	Upgoing through-going track candidates
LT downgoing through-going MLP	> 0.99	Downgoing through-going track candidates

The operating thresholds were chosen to reduce the data rate to $\mathcal{O}(100\,\mathrm{mHz})$, a level at which running the full reconstruction pipeline on every surviving event

becomes computationally feasible with the available resources. The thresholds were not optimized for maximum signal-to-noise: as visible in the $S/\sqrt{B}$ panel of Figure 3.13, the signal-to-noise ratio is still increasing at the operating threshold for all three filters, which is precisely the condition for cut power exceeding 2 (Section 5.1), meaning that harder cuts would further improve sensitivity and could in principle be applied. However, since the primary purpose of the filter is rate reduction rather than background rejection, there is no reason to cut harder than necessary. The more capable final cut models (Chapter 4) and sensitivity-optimized thresholds (Chapter 5) handle the remaining background rejection at the next stage, where the full reconstruction quality is available.

Classification loss is not sensitivity

A point worth emphasizing, because it is a recurring source of confusion: including zenith (or any other direction variable) among a classifier's input features is not sufficient for a sensitivity-optimal selection. The classifier is not, and cannot be, trained to optimize point-source sensitivity. It is a classifier: it minimizes its classification loss, usually a softmax cross-entropy. Whatever zenith dependence it learns is a function of the zenith-dependent event rates—nothing more. Optimal sensitivity is a different objective. As the cut optimization in Chapter 5 shows directly, different signal-to-background ratios require different cuts (the square-root background scaling and its consequences, Section 5.1), and full point-source sensitivity additionally depends on the local angular resolution and energy distributions, none of which enters a classification loss. The declination-dependent cut function on the score (Section 5.2) is therefore where the sensitivity optimization actually lives. The classifier only supplies the score. The training-weighting choices (Section 3.5, Chapter 4) matter by the same logic: classification loss, the counting-experiment signal-to-noise ratio, and full local sensitivity are three different objectives.

Filter performance

Figure 3.13 shows the experimental data rate and the expected atmospheric neutrino rate (NuGen, all three flavors, weighted with the Gaisser H3a[91] + SIBYLL 2.1[92] atmospheric model) as a function of score threshold for each filter model individually. The lower panel shows the signal-to-noise ratio $S/\sqrt{B}$, where S is the atmospheric neutrino rate and B is the data rate minus the expected neutrino rate (a residual dominated by atmospheric muons). This quantity tracks how effectively each filter enriches the neutrino component relative to the muon background as the threshold increases. The important difference from the score distribution shown for the downgoing filter (Figure 3.11) is that these performance plots apply the full physics weighting, including the extreme absolute normalization difference between the signal and background classes. Figure 3.14 provides a zoomed view of the same quantities around each filter's operating threshold.

[91] Gaisser 2012, "Spectrum of cosmic-ray nucleons, kaon production, and the atmospheric muon charge ratio".

[92] Ahn et al. 2009, "Cosmic ray interaction event generator Sibyll 2.1".

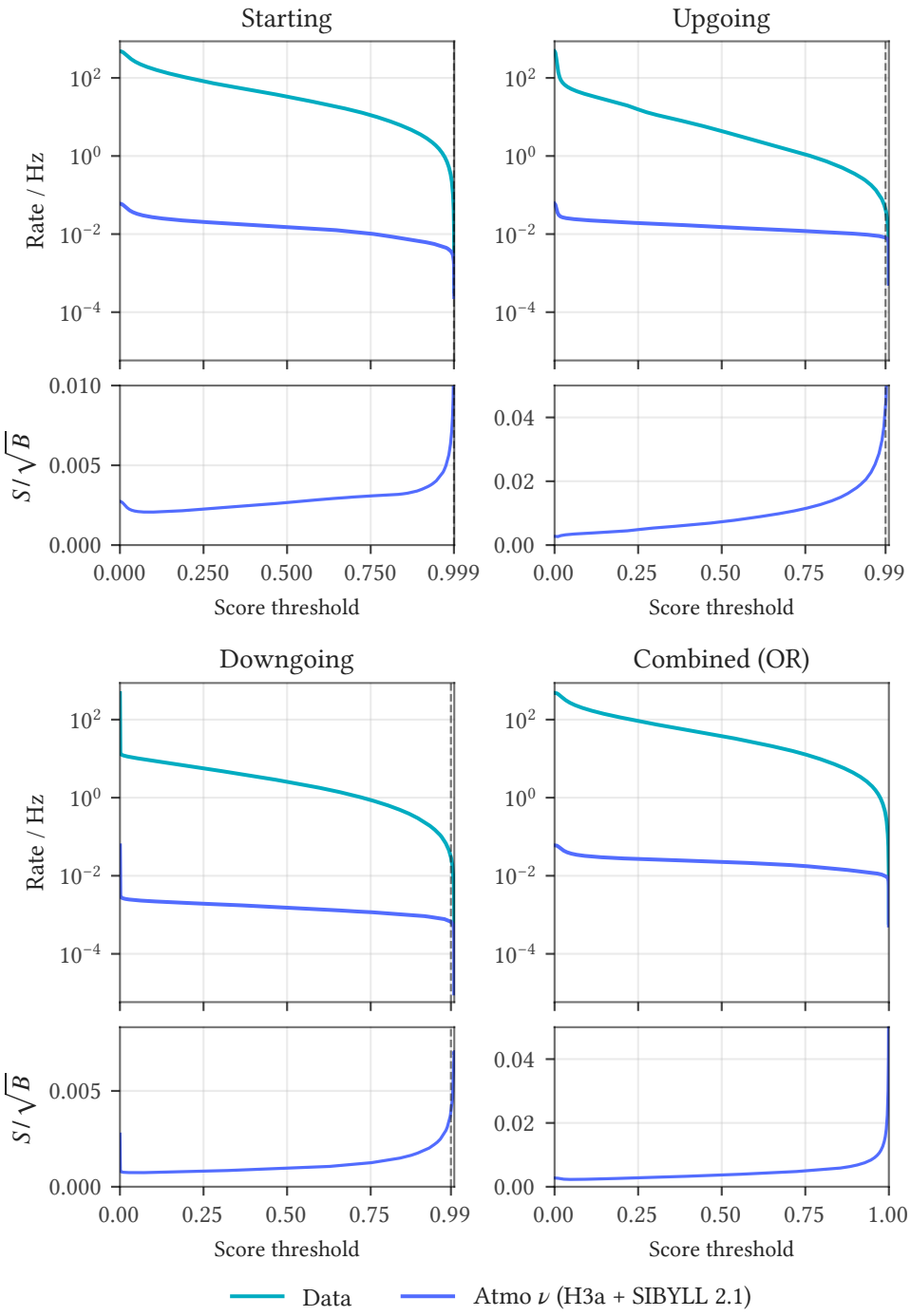

Figure 3.13: Experimental data rate and expected atmospheric neutrino rate as a function of score threshold, for each of the three filter models (starting-track CNN, upgoing-track CNN, and downgoing through-going MLP) and the combined OR selection, with the signal-to-noise ratio $S/\sqrt{B}$ in the lower row. The combined panel applies an inclusive OR: an event passes if any of the three filter scores exceeds the threshold on the x-axis. The dashed operating-threshold line is drawn only on the three individual panels: the final selection applies three separate per-filter thresholds rather than a single combined cut, so the combined panel carries no cut line.

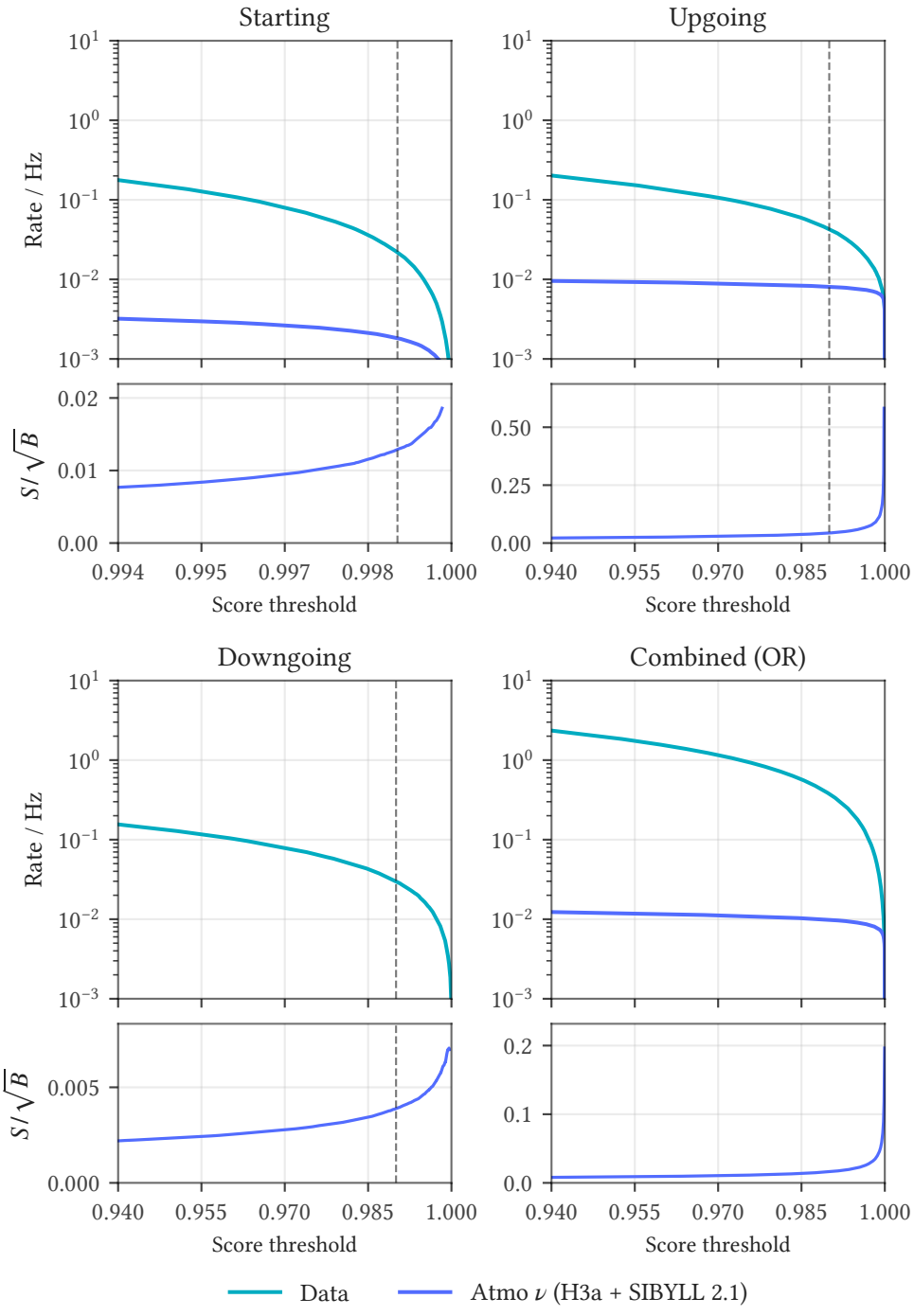

Figure 3.14: Zoomed view of Figure 3.13 around the operating thresholds, showing the data rate, expected atmospheric neutrino rate, and signal-to-noise ratio near each filter's cut. As in Figure 3.13, the combined panel applies an inclusive OR (an event passes if any of the three scores exceeds the threshold) and carries no cut line, since the final selection uses three separate per-filter thresholds.

The combined filter reduces the Level 2 data rate from 481 Hz to 92.1 mHz, a reduction of 99.98%. Despite this aggressive rate reduction, the filter retains approximately 50% of all true upgoing neutrino events and 20–25% of all true starting neutrino events at medium to high energies, as shown in the full-sky effective area retention curves (Figure 3.15; the per-filter breakdown into the starting and upgoing filters is shown in Figure 3.16). As noted in Section 3.3, *signal* here refers to all neutrino events, predominantly atmospheric neutrinos, not astrophysical.

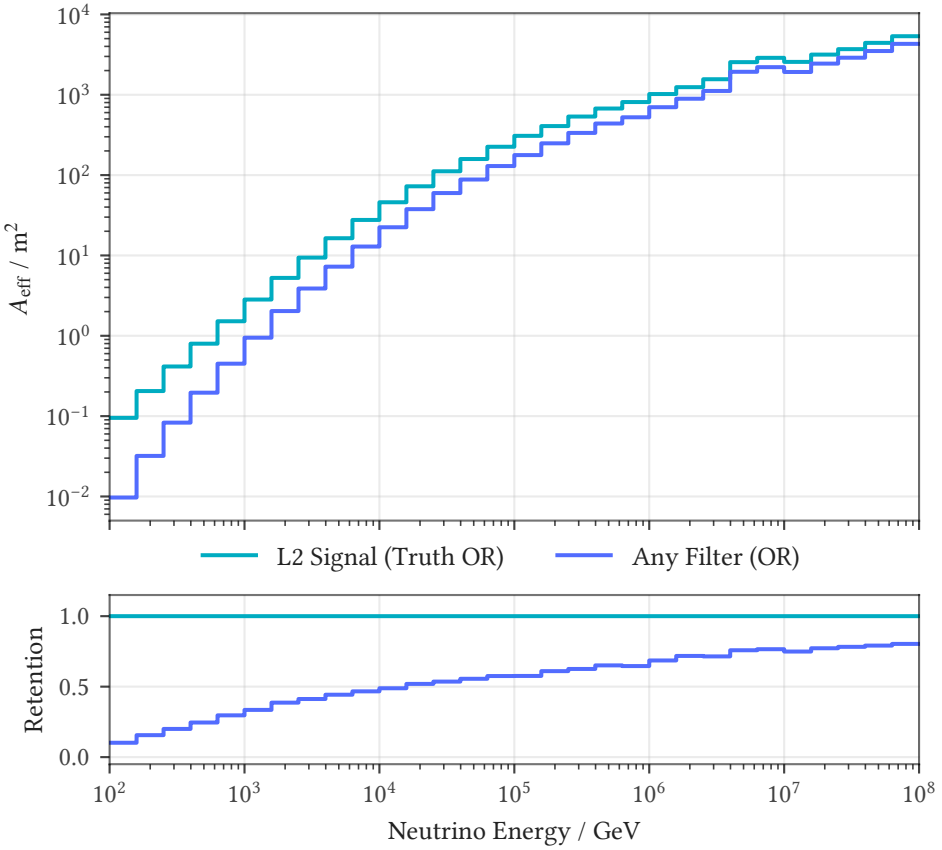

Figure 3.15: Full-sky effective area retention of the combined filter relative to Level 2, as a function of energy.

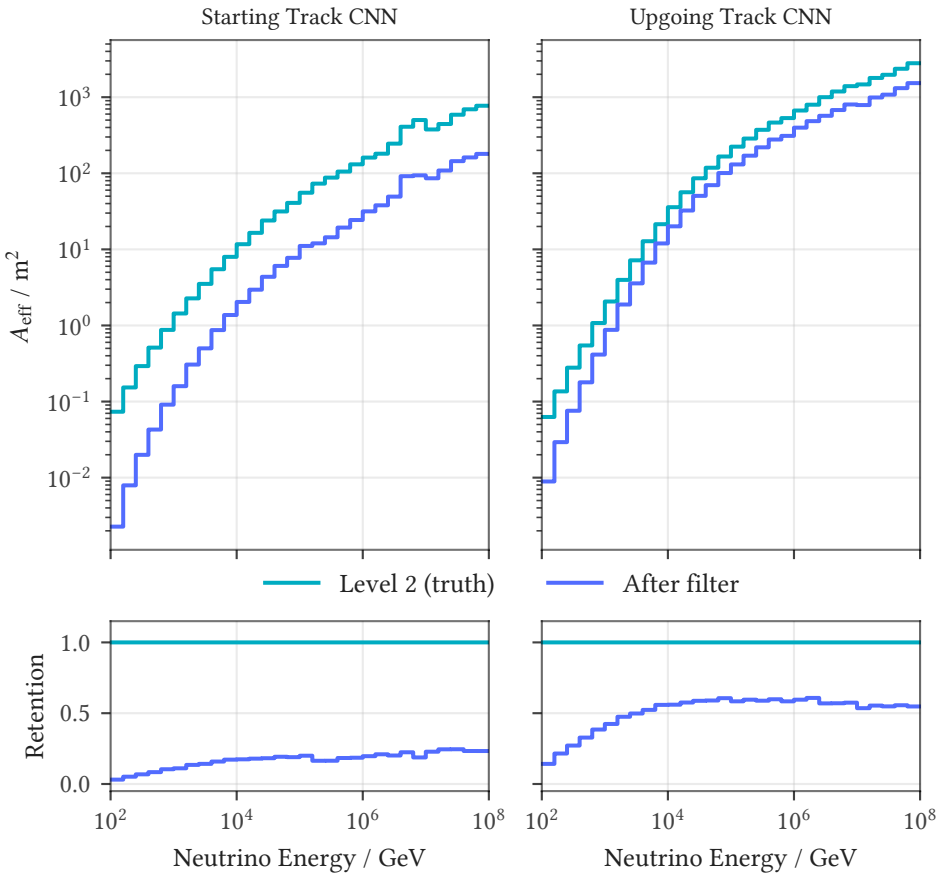

Figure 3.16: Effective area (top row) and retention relative to Level 2 (bottom row) as a function of energy, split by filter: the starting-track filter (left column) and the upgoing-track filter (right column). The downgoing through-going filter is excluded by design, as it operates on a distinct event population.

Flavor ratios

As discussed in the CNN training (Section 3.4), the CNN filters were trained exclusively on ν_μ CC and NC simulation, with no ν_e or ν_τ events in the training data. Figure 3.17 and Figure 3.18 validate this choice by showing the filter pass fraction as a function of energy and the integrated pass rates and flavor ratios for all three neutrino flavors.

As expected, ν_μ dominates the pass fraction across all filters and energies. At $\gamma = 2.5$, the ν_e pass fraction is ∼186× (starting CNN) and ∼35× (upgoing CNN) lower than the ν_μ pass fraction (Figure 3.18; each flavor's passing events are normalized to that flavor's own Level 2 total), confirming that cascade-like events are rejected even though cascades were never explicitly included as a background class during training. These factors are independent of the flavor ratio: they compare only what fraction of each flavor individually survives the filter. The downgoing through-going MLP, which operates on tabular reconstruction features rather than

topological information, achieves only ∼8× rejection—substantial ν_e contamination at the filter level, as expected for a classifier without access to spatial event topology. These remaining ν_e events are removed at the quality cut stage downstream (Chapter 4).

The ν_τ pass fraction is more nuanced and reflects the energy-dependent topology of tau decays (deferred to here from Section 2.1). The tau lepton has a mean decay length of $L \approx 49\,\mathrm{m} \times E_\tau/\mathrm{PeV}$,[93] so at the energies relevant for most of this selection ($E \lesssim 100\,\mathrm{TeV}$, $L \lesssim 5\,\mathrm{m}$), the tau decays almost immediately and the event topology is determined entirely by the decay channel. Approximately 83% of tau decays produce a second cascade (electronic or hadronic channels), making these events indistinguishable from ν_e CC or any-flavor NC cascades. The remaining ∼17% decay to a muon,[94] producing a track that is topologically identical to a ν_μ CC event. This is why the ν_τ pass fraction rises with energy for all three filters: at higher energies, the muonic decay channel produces increasingly energetic muons with full track signatures, and the ν_τ pass fraction approaches that of ν_μ. For the starting track CNN in particular, the two converge above $\sim 10^6$ GeV.

[93] Particle Data Group 2024, "Review of Particle Physics", Lepton Summary Table.

[94] Particle Data Group 2024, Lepton Summary Table.

At sufficiently high energies ($E_\nu \gtrsim 100\,\mathrm{TeV}$), the tau decay length becomes resolvable and ν_τ CC events can in principle exhibit their own distinctive topology, colloquially known as the *double bang* signature, consisting of two spatially separated cascades (one at the interaction vertex, one at the tau decay point) connected by a faint tau track. Identifying double bangs requires dedicated reconstruction algorithms operating on DOM-level waveform features and is the subject of dedicated tau appearance analyses. At the per-DOM summary feature level available to the CNN filters, double bangs are not distinguishable from elongated single cascades or starting tracks.

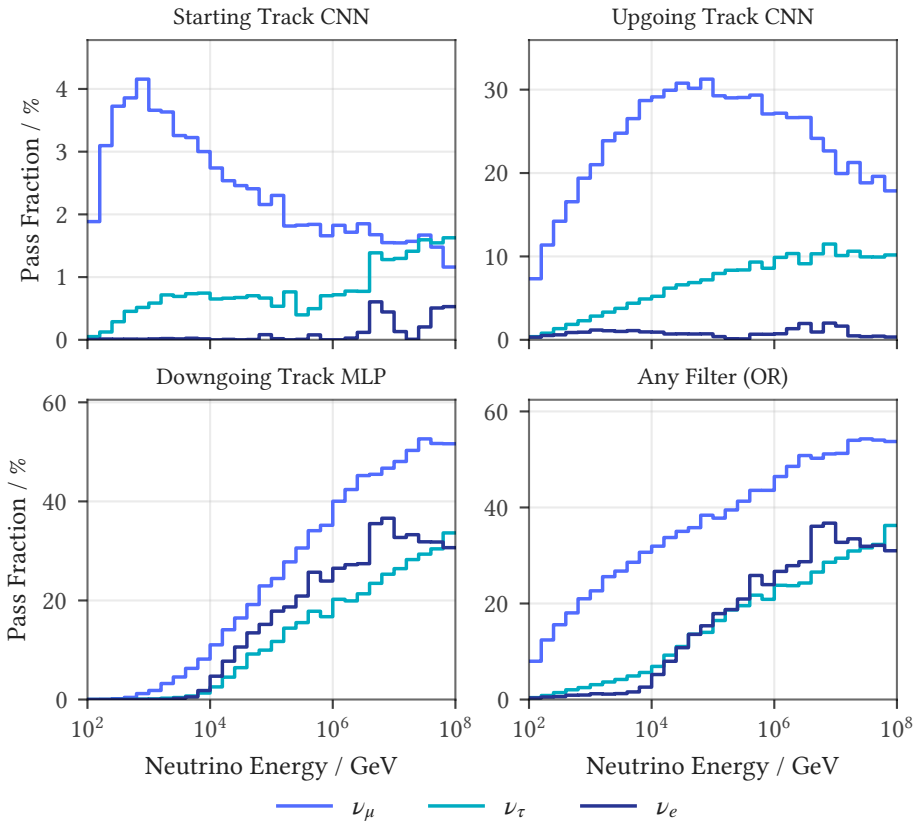

Figure 3.17: Filter pass fraction relative to Level 2 as a function of energy, per neutrino flavor, for each filter (starting, upgoing, downgoing) and their combination, each panel on its own vertical scale. Each bin shows the weighted fraction of events that pass, with events weighted to a uniform energy spectrum within each bin; reweighting within bins to power-law spectra with $\gamma = 2.0$–3.0 shifts the fractions by at most ∼1 percentage point, at the extreme-energy bins.

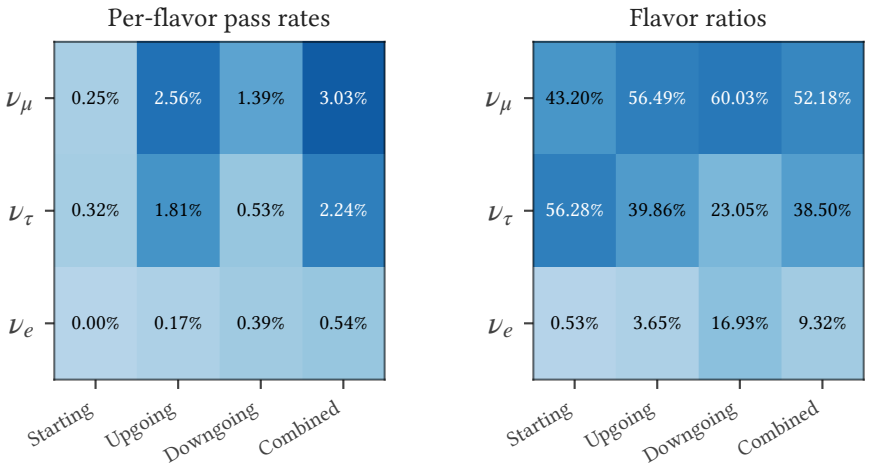

Figure 3.18: Per-flavor filter performance at the filter level, weighted to an $E^{-2.5}$ power-law flux ($\gamma = 2.5$), whole sky. Left: per-flavor pass rates relative to Level 2. Right: flavor ratios, with each column normalized to 100%.

A substantial fraction of the remaining signal loss is not recoverable in principle. Many of the neutrino events that fail the filter did not deposit a detectable signature in the first place: either the neutrino secondary produced too little Cherenkov light to rise above the detector noise floor, or the event triggered on noise hits (dark rate fluctuations) rather than on the neutrino-induced light.

Coincident events

Beyond pure detector noise, a significant source of unrecoverable signal is coincident events: atmospheric muons that share the same trigger readout window as a neutrino interaction, either temporally or spatially overlapping with it. IceCube simulates these coincidences explicitly. In many cases, the coincident atmospheric muon is substantially brighter than the neutrino event and dominates the recorded light pattern. This poses a fundamental problem for the reconstruction algorithms we use for Lightning Tracks (Chapter 4), which operate on the combined pulse pattern without any knowledge of which pulses belong to which particle: when the muon is much brighter, the reconstructions simply latch onto the brighter component and return the muon direction rather than the neutrino direction. When the two components have comparable brightness, the superposition of two unrelated light patterns produces a pulse pattern that is inconsistent with any single-track hypothesis, and the reconstructions produce unreliable results. In either case, the event would not survive the angular uncertainty quality cuts (Chapter 4).

Some previous selections (e.g., PST and NT) attempt to recover coincident events by employing a more sophisticated event splitter called HiveSplitter that separates temporally and spatially distinct sub-events within a single trigger window, rather than the simple trigger-time splitter used by Lightning Tracks. Lightning Tracks makes no such attempt, by design. If the neutrino and muon components are genuinely separable and the neutrino signature is the brighter of the two, the

reconstruction algorithms naturally latch onto the neutrino track without explicit splitting.

The only regime where explicit splitting could in principle recover additional signal is the case of a faint but spatially well-isolated neutrino event temporally coinciding with a much brighter muon elsewhere in the detector. HiveSplitter could separate these components, potentially allowing the neutrino event to be reconstructed independently. However, HiveSplitter is computationally prohibitive at this stage of the pipeline: its worst-case runtime reaches tens of seconds per event, making it infeasible to run at the full Level 2 input rate. Moreover, even if such events could be recovered, they would necessarily be faint (otherwise they would not require explicit splitting to be identified)—and faint events contribute minimally to point source sensitivity. This is a regime where marginal sensitivity gains may remain, but the computational cost and added pipeline complexity are not justified for the present analysis.

Application in IceTray

All three filter models are applied to every event in a single pass over the i3 files.[95] The two CNN scores and the MLP score are stored as frame objects on every event; no events are removed at this stage unless cuts are explicitly requested. When cuts are applied, any event passing at least one threshold is kept. Passing events are not separated into distinct selections at this point. They remain together in the same output i3 files, in order, with all three scores attached. The hierarchical separation into starting and through-going candidates happens later, when the final cut models are applied (Chapter 4).

This design keeps the filtered sample maximally flexible: different downstream analyses can apply different thresholds without reprocessing. Each subsequent processing stage (reconstructions, final cuts, export) also runs in a single pass over the i3 files, since all reconstructions are shared across both the starting and through-going channels.

[95] IceTray is IceCube's data-processing software. An open-source version is available at `github.com/icecube/icetray-public`, but it is only a subset of the full internal, closed-source version. Data are stored in i3 files: a series of *frames* of various types, of which the only ones relevant here correspond to temporal trigger windows of the detector. Each frame can hold an arbitrary number of objects, which store different types of data, such as the pulse series or reconstructed quantities.

3.7 Producing the sample at scale

The entire selection philosophy depends on the filter models being fast enough to evaluate on every input event at full data rates, using only CPU inference inside IceTray. All models are exported as TorchScript[96] and run single-threaded CPU inference with no GPU requirement.

Two design decisions were critical to achieving this. First, the CNN filters use a reduced 4-feature input (down from 14–15 per DOM in earlier iterations). Classification accuracy is indistinguishable between the reduced and full feature sets, but the model size and inference cost are dramatically lower. Each CNN has approximately 427,000 trainable parameters. Second, training in 16-bit mixed precision produces compact models that run efficiently on CPU. The MLP filter is far smaller still at approximately 13,000 parameters and uses only features already

[96] TorchScript is PyTorch's serialization and just-in-time compilation format for models, enabling Python-independent inference: `pytorch.org/docs`.

available at L2, adding negligible overhead to the per-event cost. The CNN models dominate the total inference time.

Applying all three filter models to i3 data processes approximately 16 events per second per core on an AMD EPYC 7763 (Zen3), stable across different IC86 seasons. This is the combined cost of all three models, not each individually, and the rate is hardware-dependent.

GPU inference was not used for production of the dataset but was explored using a shared-memory architecture developed for Lightning Tracks processing. A single GPU inference server runs alongside N CPU-based IceTray worker processes, each processing one input file sequentially. Each worker prepares batches and writes them to a pre-allocated, GPU-pinned block in shared memory, then signals the GPU server that a batch is ready. The server works through the batch queue by reading input data directly from shared memory, and the pinned allocation enables fast DMA transfers to GPU memory without network overhead. This design is fully compatible with IceTray's sequential processing: each worker maintains correct event ordering within its file.

On a single NVIDIA H200 GPU with 150 CPU workers for data loading, this architecture achieves a sustained throughput of approximately 26,000 events per second for a single CNN filter, an order of magnitude above the IceCube trigger rate. This measurement is CPU-limited (GPU utilization averages 86%); with faster data loading or pre-extracted features, the GPU throughput would be higher. For real-time processing, the data stream would need to be divided across multiple workers to buffer full GPU batches at sufficient rate, which is more involved but straightforward in principle.

Despite this throughput, the production processing was performed entirely on CPUs. The reason is resource availability: a single H200 GPU processes events roughly 800 times faster than a single CPU core—but GPU resources are scarce. As of June 2026, MSU's HPCC has 44 H200-class GPUs compared to 69,752 CPU cores. The aggregate CPU throughput at scale far exceeds what could be achieved with the available GPUs, even accounting for the per-device speedup, and GPU resources face substantially higher contention among users than CPU cores. For large-scale offline processing where wall time is the constraint, the CPUs are the more practical resource.

The three models together reduce the data rate by 99.98%, producing a filtered sample on which final-level reconstructions become computationally feasible.

The production campaign

The full production applies this pipeline to all 12 years of IC86 Level 2 data end to end. The processing is implemented as a Snakemake[97] workflow executing inside Apptainer containers, stepping each run through the filter models, the reconstructions and final cut models described in the next chapters, and the HDF5 export, with file-based rules managing the dependencies between steps and staging input data from IceCube's compute infrastructure located in Madison, Wisconsin

[97] Mölder et al. 2021, "Sustainable data analysis with Snakemake".

via the internet by each compute job. The large input files are then immediately deleted as soon as they are processed, to stay within local storage limits.

Scheduling this volume of work on a shared university cluster required dedicated infrastructure. MSU HPCC's scavenger partition offers large amounts of idle CPU capacity, but fragmented into slots too small for fixed large allocations. The production therefore ran on a purpose-built adaptive worker pool that decouples resource allocation from task execution: persistent SLURM workers of varying sizes hold allocations and pull tasks over TCP from a central server, so task dispatch is not throttled by SLURM scheduler overhead, preempted work is automatically re-queued, and worker sizes adapt to the fragmented capacity actually available.

In total, the production consumed at least 15 million CPU-hours on MSU's HPCC.

3.8 Prior art and departures

The ML-based filtering approach is inspired by DNN Cascades (DNNC), which demonstrated that ML classifiers on low-level detector data can achieve high signal retention at low computational cost. DNNC uses many small, computationally inexpensive models for filtering, each performing an explicit reconstruction or classification task (vertex and containment reconstruction, early directional and energy estimates, flavor classification, and so on), representing a middle ground between conventional and fully ML-based filtering.

A defining feature of the Lightning Tracks pipeline is that the very first reconstructions it runs are the final reconstructions used in the analysis (Chapter 4). There are no intermediate selection levels, no intermediate reconstructions, and no multi-stage pipeline between the filter and the final analysis-quality event characterization. This design avoids a failure mode common in conventional multi-stage selections (with the exception of the downgoing through-going component, which does rely on the L2 reconstructions, as discussed in Section 3.5), where events that would ultimately pass the final selection are removed at intermediate stages because the computationally inexpensive reconstructions available at those stages lack the sophistication to correctly identify them as signal.

The application of CNNs to IceCube DOM-level data was pioneered by Hünnefeld[98],[99] who developed a 3D CNN architecture for multi-label regression (direction and energy reconstruction). The LCSC filter models are heavily inspired by this work, in particular the grid transformation and the use of per-DOM summary features as input. However, the LCSC implementation differs in several respects, reflecting the much simpler requirements of a binary classification filter compared to a full reconstruction.

The prior work describes the grid transformation as "hexagonal convolution,"[100] but the term is slightly misleading. As described in the input transformation (Section 3.3), the hexagonal string layout is mapped to a rectangular grid via a coordinate transformation that preserves n-th neighbor adjacency, and standard rectangular convolution kernels are applied on the transformed grid. In the prior implementa-

[98] IceCube Collaboration 2021a, "A Convolutional Neural Network based Cascade Reconstruction for the IceCube Neutrino Observatory".

[99] Hünnefeld 2023.

[100] IceCube Collaboration 2021a, Sec. 4.1.

tion, the corner elements of each rectangular kernel are zeroed out to produce an effective hexagonal kernel shape. We omit even this masking: the full rectangular kernel is used. From the perspective of the network, the only thing that matters is preserving the neighbor relations in the grid mapping. Given that, the convolution itself is standard. The one scenario where true hexagonal convolution would matter is rotation equivariance: a hexagonal kernel can in principle be rotated by multiples of 60° within group convolutions to make the network equivariant under the sixfold symmetry of the hex grid. The prior work discusses this possibility and the code contains the infrastructure, but to our knowledge rotated kernels were never used in practice, and without kernel rotation, hexagonal convolution reduces to regular convolution on the transformed grid, which is what both implementations effectively do.

Several other architectural choices differ between the two implementations. The prior work introduces "unit variance maintaining layers"[101] that analytically rescale layer outputs by $1/\sqrt{n}$ (where n is the number of summed inputs) to stabilize activations, explicitly rejecting batch normalization on the grounds that its implicit regularization is too strong. We use standard batch normalization instead (Section 3.2), which achieves the same variance-stabilizing effect in practice while benefiting from highly optimized standard library implementations. We do not observe the excessive regularization reported for the regression task. The prior architecture also uses residual additions at every convolutional and most fully connected layers, with a trainable scale factor initialized near zero so that each layer begins as an approximate identity. We omit residual connections entirely, as we cannot find any measurable performance difference for our classification task—unsurprising given that the utility of residual connections scales with network depth, and the prior architecture has 20 convolutional layers for the main array alone compared to our 8.

[101] IceCube Collaboration 2021a, Sec. 4.4.

A more substantial difference concerns the treatment of DeepCore. The prior work processes the 8 DeepCore infill strings as two separate sub-arrays (upper and lower, split at the dust layer) with their own 1D convolutional stacks, which are concatenated with the main-array features before the fully connected head. This adds three independent convolutional pathways with different spatial dimensionalities. We exclude DeepCore entirely (Section 3.3), as we find that including it significantly complicates the architecture without measurable improvement for our classification tasks. The DeepCore strings are densely instrumented with 7 m vertical spacing (vs. 17 m for the main array) and irregular horizontal positions that do not fit the hexagonal grid. Accommodating them requires either a separate subnetwork or interpolation onto the main grid, neither of which proved worthwhile for a coarse topological filter.

On the training side, both approaches benefit from the same fundamental observation: the IceCube application is in the opposite regime from typical image recognition, where the available training data (hundreds of millions of simulated events) vastly exceeds the model capacity, so overfitting is not a practical concern. This holds even more strongly for our smaller models (∼427k parameters

vs. ∼6M[102] in the prior architecture). The training procedure that exploits this data-rich regime—an infinite-iterator epoch definition with rejection sampling, which removes any practical distinction between training and validation data—is described with the CNN training (Section 3.4).

Finally, the prior architecture performs full event reconstruction: multi-label regression (simultaneously predicting direction and energy) and, for each predicted quantity, an associated uncertainty. The latter motivates the dual sub-network design—one sub-network for the point estimates, a second for their uncertainties, with a gradient stop between them. Training is correspondingly involved: a multi-stage schedule that swaps loss functions (mean squared error early for robustness, a Gaussian likelihood later once the fit stabilizes) and adaptive per-label weighting to keep labels that are hard to predict from dominating the gradient. None of this is needed for binary classification with cross-entropy and a single softmax output. Similarly, the prior training procedure progressively adjusts learning rates, dropout schedules, and loss targets across 8 training steps. Our training uses a single fixed configuration throughout.

[102] IceCube Collaboration 2021a, App. A.

4 Final Event Selection

With the filtered sample in hand, this chapter defines the final Lightning Tracks event sample. We first introduce the reconstructions that the filter's rate reduction makes affordable (Section 4.1), then remove clearly misreconstructed events with quality cuts (Section 4.2), build the final classifiers (Section 4.3), and place the approach against prior practice (Section 4.4).

4.1 Event Reconstruction

For all filter-passing candidates, several dedicated reconstruction tools are run to obtain the observables used in the final selection. Gerrit Wrede developed[103] a convolutional-recurrent neural network (CRNN) reconstruction—a convolutional network for spatial feature extraction with a recurrent network on top for the temporal information—that provides a fast, full track reconstruction (direction, location, time, and uncertainty), used as one of the two main direction estimates for the quality cuts and as the track seed for the downstream tools. It is not used as the final directional reconstruction at analysis level. Its directional performance was independently verified using the moon shadow.[104] The Moon blocks the primary cosmic rays that would otherwise produce atmospheric muons, leaving a deficit of muons arriving from its direction. Because the Moon's position is known exactly, the location and angular width of that deficit give a direct, model-independent measurement of the detector's absolute pointing and angular resolution. That directness makes it the standard check for a directional reconstruction. Glüsenkamp's normalizing-flow reconstruction (TNF)[105] provides the directional fit and event-by-event angular uncertainty (the reconstruction's per-event estimate, denoted σ, of how far its fitted direction is likely to lie from the true one) used in the final cuts and as the final directional reconstruction for the source analyses. We revisit the reconstruction and its angular uncertainty in more detail in Chapter 8. The MuEX energy estimator, a conventional likelihood-based energy reconstruction included in IceTray, runs on the CRNN track and serves as the nominal reconstructed energy for cuts and analysis. During the development of Lightning Tracks, more sophisticated alternatives for energy reconstruction were considered but ultimately rejected, on the grounds of unnecessary complexity for very little to no gain, as we will see in Chapter 9. For starting-track candidates only, the ESTES Starting Track Veto (STV)[106] provides the miss probability p_{miss} (computed both with and without stochastic suppression), an input feature for the final starting-cut model. A high value indicates a likely entering muon that left no detectable light in the outer

[103] The recurrent (CRNN) track reconstruction is the subject of G. Wrede's forthcoming PhD dissertation (FAU / Erlangen Centre for Astroparticle Physics, expected 2026; ecap.nat.fau.de/index.php/research/publications/theses/).

[104] IceCube Collaboration 2021c, "Testing the Pointing of IceCube Using the Moon Shadow in Cosmic-Ray-Induced Muons".

[105] IceCube Collaboration 2026b, "Neural posterior estimation of the neutrino direction in IceCube using transformer-encoded normalizing flows on the sphere".

[106] IceCube Collaboration 2024a, "Characterization of the Astrophysical Diffuse Neutrino Flux using Starting Track Events in IceCube", Sec. V.B.

DOM layers of the detector and was therefore *misclassified* as a starting neutrino event. The STV performs similarly to alternative machine-learning models such as our starting-track filter CNN, but we include it as an *additional* input feature for the final classifier, on the idea that several algorithms sharing the same objective but built very differently can together outperform any one of them alone. This commonly falls under the umbrella of *ensemble methods* in machine learning, except that we extend it here to include conventional algorithms as well. We revisit this idea for the quality cuts in the next section.

Although TNF provides the final directional reconstruction and angular uncertainty used in the analysis, it produces a full azimuth–zenith posterior PDF but does not provide vertex position or timing information—it reconstructs the direction, not the full track. Both MuEX and STV instead depend on the complete 6D track hypothesis (3 vertex coordinates, 2 angles, 1 time) that only the CRNN provides. MuEX integrates the observed charge along the hypothesized track. STV, despite its name, requires no actual vertex reconstruction either: it defines its veto region directly from the track hypothesis, because what matters is where the track cuts through the outer layers of the detector—close to a string, or in a gap between DOMs. A muon that threads through such a gap can pass through the outer layers without triggering them and so escape the veto; we call these misclassified entering muons *sneaky muons*, since they sneaked through the outer DOMs undetected. The CRNN therefore serves as the track seed for the rest of the pipeline, in addition to providing an independent directional estimate for the quality cuts described below.

This combination ensures that each candidate has at least two independent direction estimates (CRNN and TNF), a simple yet reasonably robust energy estimate (MuEX), and, for starting tracks, a veto-based containment variable (STV's p_{miss}). In combination with the filtering model scores, those variables are used for the final sample cuts.

4.2 Quality cuts

Quality cuts act as a first pass to remove clearly poorly reconstructed events and improve data-MC agreement. An event is removed if either angular uncertainty estimate exceeds 5° ($\sigma_{\text{CRNN}} > 5°$ or $\sigma_{\text{TNF}} > 5°$), if the two directional reconstructions disagree by more than 5° ($\Delta\psi(\hat{n}_{\text{CRNN}}, \hat{n}_{\text{TNF}}) > 5°$), if the homogenized total charge[107] is below $Q_{\text{total}} < 20$, or if the reconstructed MuEX energy is below $E_{\text{MuEX}} < 10\,\text{GeV}$.

[107] As a reminder from the previous chapter, the homogenized total charge is the summed observed charge across all DOMs, weighted by each DOM's individual photon-detection efficiency.

The angular uncertainty thresholds remove events where at least one reconstruction reports low confidence in its own result. In principle, cutting on the angular uncertainty can only reduce point source sensitivity: if the uncertainty estimate is accurate, the likelihood correctly down-weights poorly resolved events, and removing them discards information. The problem is that the uncertainty estimates themselves become unreliable at large values. There is no second-order uncertainty available (no estimate of the error on the error), and for events with large σ, the true uncertainty on σ itself also grows, meaning the likelihood may assign these

events inappropriate spatial weight. Ideally, the cut would target this second-order uncertainty directly, but in its absence, cutting on σ serves as a practical proxy. An additional benefit is that large angular uncertainties are strongly correlated with non-track event topologies (cascades, coincident events, muon bundles), since the reconstruction algorithms are trained explicitly on single muon tracks and produce unreliable uncertainty estimates for event types outside their training domain.

As an additional proxy for this second-order reliability, we introduce the *reco separation cut*, which requires that the two independent directional reconstructions, TNF and the CRNN, agree to within 5° in their directional point estimates. The intuition is straightforward: at this point in the pipeline, both algorithms have already claimed angular uncertainties below 5° (events exceeding that threshold were removed by the preceding cut). If their point estimates nevertheless disagree by more than 5°, at least one of those claimed uncertainties must be substantially wrong—a direct indicator that the uncertainty estimates themselves are unreliable for that event. The exact relationship is probabilistic rather than binary, but the basic logic holds: large disagreement between two algorithms that both report high confidence is evidence of inaccurate error estimation. The approach is again conceptually related to ensemble methods in machine learning, but applied to quality assessment rather than prediction: instead of combining two regression outputs to improve accuracy, the relative agreement between two independent algorithms is used as a diagnostic of reconstruction reliability. When two fundamentally different algorithms, trained on different representations of the same data and with entirely different architectures, independently converge on the same direction, there is strong reason to trust the result. When they disagree substantially, at least one has produced an unreliable estimate, even if both individually report small angular uncertainties. This provides information that neither algorithm's uncertainty estimate alone can capture, precisely because the failure modes of the two algorithms are largely uncorrelated: they fail on different events and for different reasons. The cut is particularly effective at identifying cascade-like events and coincident events (Chapter 3) that are not well represented in the training data of either reconstruction and tend to produce inconsistent directional estimates. Including the reco separation cut substantially improved data-MC agreement. A specific population of corner-clipping CORSIKA[108] events at the bottom of the detector, misreconstructed as upgoing, showed poor data-MC agreement and was cleanly removed by this cut. Starting from filter-level events that pass both $\sigma < 5°$ thresholds, the reco separation cut removes 13.6% (SLT) and 7.4% (TLT) of remaining data events. The ν_e MC provides evidence that a large fraction of these are genuinely problematic: for ν_e weighted to an $E^{-2.5}$ spectrum, the cut removes 17.8% (SLT) and 63.5% (TLT) of remaining cascade-like events that the angular error thresholds alone miss. Within the TLT sample this removal is concentrated in the down-going component: the cut removes 73% of the down-going cascade-like events the angular error thresholds miss, against 19% of the up-going ones. The down-going filter operates on tabulated features and so lacks the topological awareness of the starting and upgoing filters, which is why its component is preferentially cleaned up here.

[108] As a reminder, CORSIKA is the atmospheric muon air-shower simulation software used here.

The charge and energy thresholds ($Q_{\text{total}} > 20$, $E_{\text{MuEX}} > 10\,\text{GeV}$) remove events

that are too dim for any reconstruction to produce a reliable result. These thresholds can be refined per analysis and are ultimately tuned to achieve the desired balance between efficiency, background rejection, and systematic robustness.

4.3 The final-cut MLPs

The final event selection for Lightning Tracks is defined by two multilayer perceptron (MLP) classifiers, one for Starting Lightning Tracks (SLT) and one for Through-going Lightning Tracks (TLT). Unlike the filter models, which identify events based purely on topological features, the final cut models incorporate prior knowledge about the expected signal and background distributions as a function of zenith angle and energy, just like the down-going filter model, but now at the final selection level where all topological information is already exploited and all that remains are the distributional differences. This physics-informed approach enables the models to learn the realistic signal-to-background ratio across the sky, rather than treating all directions equally.

The final cut models use the same MLP framework as the down-going–through-going filter (Section 3.5), reapplied at the final cut stage with more powerful input features derived from computationally expensive reconstructions (Table 4.2, Table 4.3) that are only available after the filter has reduced the data rate sufficiently for these reconstructions to be run on every surviving event.

They use 3–4 hidden layers with decreasing widths (e.g., $32 \to 16 \to 8 \to 4$), progressively compressing the representation into a single output value passed through a sigmoid function $\varsigma(z) = 1/(1 + e^{-z})$, yielding a probability in $[0, 1]$ that the event is signal.

Training philosophy

Physics-informed weighting

The central idea behind the final cut models is to train them on data that reflects prior knowledge of the signal and background distributions, informed by assumed flux models. Within each class, events are sampled according to their physics weights, so the model sees the physically motivated distribution shapes rather than the raw simulation statistics.

The signal class consists of NuGen neutrino simulation. For SLT, the weight combines both atmospheric and astrophysical neutrino fluxes, using the Gaisser H3a cosmic-ray primary spectrum[109] for the atmospheric component and a power law approximately matching the astrophysical diffuse spectrum for the astrophysical component. For TLT, only the astrophysical component is used: this teaches the model to favor upgoing events where astrophysical signal dominates, but also, in the southern sky where atmospheric muons are overwhelming, to retain only the high-energy tail of the distribution where the signal-to-background ratio is favorable for hard astrophysical spectra ($\gamma \approx 2$, well below the atmospheric spectral index of ~ 3.7[110]). The background class consists of CORSIKA atmospheric muon

[109] Gaisser 2012, "Spectrum of cosmic-ray nucleons, kaon production, and the atmospheric muon charge ratio".

[110] IceCube Collaboration 2017a, "All-sky Search for Time-integrated Neutrino Emission from Astrophysical Sources with 7 yr of IceCube Data".

simulation, weighted by the same cosmic-ray primary model mentioned above. This flux is strongly peaked toward the downgoing direction.

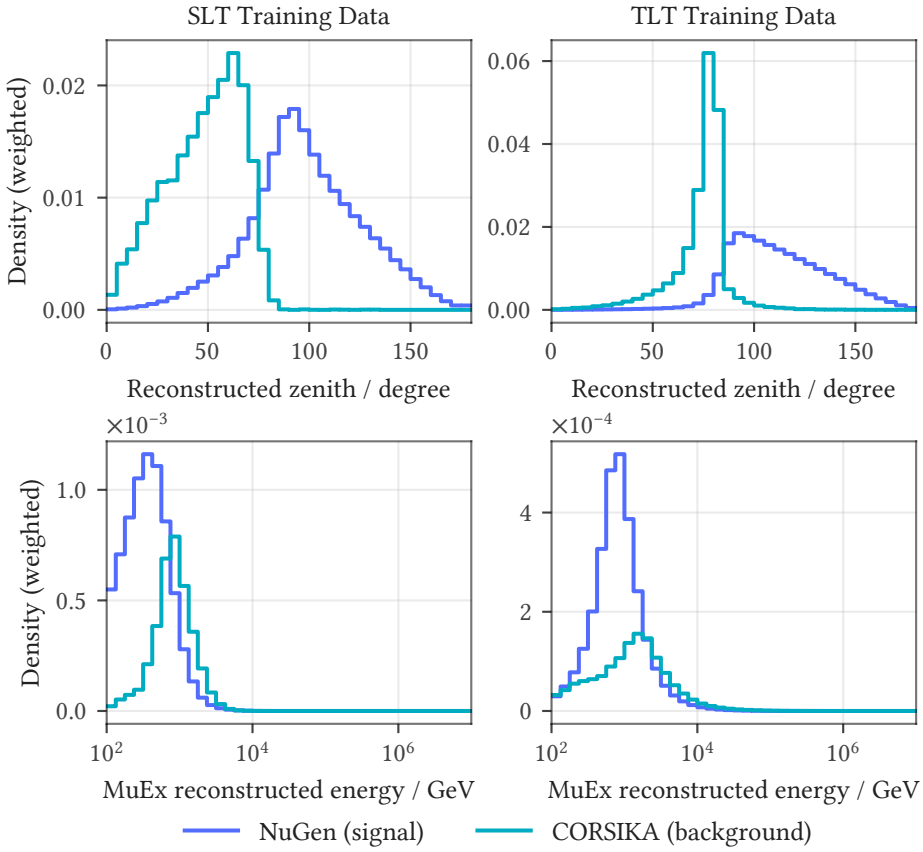

Figure 4.1: Physics-weighted reconstructed zenith and energy distributions (remaining features not shown) of the training data. Training on these distributions teaches the model the prior signal-to-background ratio as a function of zenith and energy, allowing it to calibrate its score based on where in the sky an event originates and how bright it is, not just its topological features.

Importantly, the two classes are balanced 1:1 during training (see the rejection sampling below): the correct absolute normalization between signal and background is not used, as the model would otherwise see almost exclusively CORSIKA events. Instead, the model learns the *relative* distribution shapes within each class: how signal and background are distributed differently across zenith and energy. This is sufficient for the model to learn where in the phase space signal is more prevalent relative to background, instead of just relying on the absolute normalization difference, and to calibrate its score accordingly.

Rejection sampling

In practice, the physics-based weighting is implemented through rejection sampling during training, following the same infinite-iterator approach used for the filter

models (Section 3.4). Each class maintains its own weighted sampler that draws events with probability proportional to their physics weights. Batches are assembled with a 1:1 signal-to-background ratio, and within each class events are drawn according to their physics weights so that the model sees a distribution reflecting the expected flux composition rather than the raw simulation statistics. This is necessary because the simulation samples have vastly different sizes (Table 4.1).

Table 4.1: Training sample sizes for the SLT and TLT final cut models.

Sample	SLT events	TLT events
NuGen (signal)	∼844,000	∼3,870,000
CORSIKA (background)	∼16,000	∼1,690,000

Without weighted sampling, the model would see NuGen events far more frequently than their true occurrence rate, leading to poor generalization.

Epoch definition

Unlike conventional machine learning where an "epoch" denotes one complete pass through the training data, the final cut models use a fixed-step epoch definition. Each epoch consists of a predetermined number of training batches (10 batches of 14,000 events, or 140,000 samples per epoch), regardless of the total dataset size.

This definition arises naturally from the rejection sampling approach: since events are drawn with replacement according to their weights, there is no concept of "seeing all the data once." Instead, training progress is measured by the number of gradient updates, and epochs serve as checkpointing intervals for monitoring convergence.

Model architectures

Both models share a common structure but differ in input features and network depth to accommodate the distinct physics of starting versus through-going tracks. The input features are derived from the LCSC filter scores, the CRNN and TNF reconstruction outputs, the reco separation, the MuEX energy and homogenized total charge, and, for starting tracks, the STV miss probabilities.

Table 4.2: Input features for the SLT final cut model.

Feature	Description
Q_{total}	Homogenized total charge
LCSC starting score	Starting track CNN filter score
LCSC upgoing score	Upgoing track CNN filter score
CRNN zenith	CRNN reconstructed zenith angle
E_{reco}	MuEX energy estimate
CRNN σ	CRNN angular uncertainty
TNF zenith	TNF reconstructed zenith angle
TNF σ	TNF angular uncertainty
Reco separation	Angular separation between CRNN and TNF
p_{miss}	STV miss probability
$p_{\text{miss}}^{\text{supp}}$	STV miss probability (stochastically suppressed)

The STV miss probabilities encode information about whether light was observed in the veto region, which helps identify entering muons masquerading as contained neutrino events.

Table 4.3: Input features for the TLT final cut model.

Feature	Description
Q_{total}	Homogenized total charge
LCSC upgoing score	Upgoing track CNN filter score
LT downgoing score	Down-going–through-going MLP filter score
CRNN zenith	CRNN reconstructed zenith angle
E_{reco}	MuEX energy estimate
CRNN σ	CRNN angular uncertainty
TNF zenith	TNF reconstructed zenith angle
TNF σ	TNF angular uncertainty
Reco separation	Angular separation between CRNN and TNF

Through-going tracks do not use containment-based features since they originate outside the detector by definition.

Table 4.4: MLP architecture comparison for the SLT and TLT final cut models.

Property	SLT	TLT
Input dimension	11	9
Hidden layers	[16, 8, 4]	[32, 16, 8, 4]
Activation	ReLU	ReLU
Regularization	Dropout (0.2), L2 ($\lambda = 0.02$)	Dropout (0.2), L2 ($\lambda = 0.01$)

The TLT model is deeper (4 hidden layers vs. 3) to capture the more complex decision boundary needed for through-going track classification, where the distinction between astrophysical and atmospheric backgrounds relies more heavily on zenith-dependent rate expectations.

Preprocessing

Two features undergo logarithmic transformation before entering the network:

$$Q_{\text{total}} \to \log_{10}(1 + Q_{\text{total}}), \qquad E_{\text{reco}} \to \log_{10}(1 + E_{\text{reco}}). \tag{4.1}$$

These transforms compress the dynamic range of charge and energy, which span several orders of magnitude, into a scale more amenable to gradient-based optimization. All features are then standardized using batch normalization, which learns the mean and variance from the training data and applies the same normalization at inference time.

Training and regularization

The models minimize binary cross-entropy with L2 regularization,

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right] + \lambda \sum_j w_j^2, \tag{4.2}$$

where $y_i \in \{0, 1\}$ is the true label, $\hat{y}_i$ is the predicted probability, and the L2 term penalizes large weights to prevent overfitting.

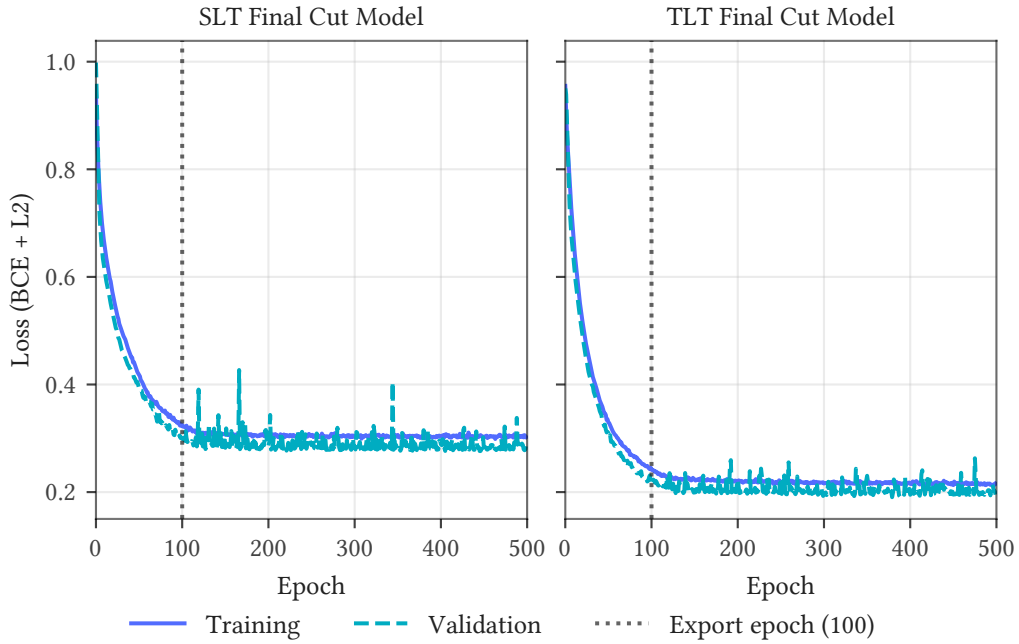

Figure 4.2: Training and validation loss curves for the SLT and TLT final cut models. The dotted vertical line marks epoch 100, where the final models were exported.

Both models were trained for 500 epochs, but the final exported models use the checkpoint from epoch 100 (Figure 4.2). The validation loss is computed on the same Monte Carlo as the training loss, so it would not necessarily reveal overtraining on simulation-specific features. Given the known data-MC mismatches (Chapter 6), a model fit harder to the specifics of the simulation may transfer worse to data, and exporting an early checkpoint is in that sense a robustness choice.

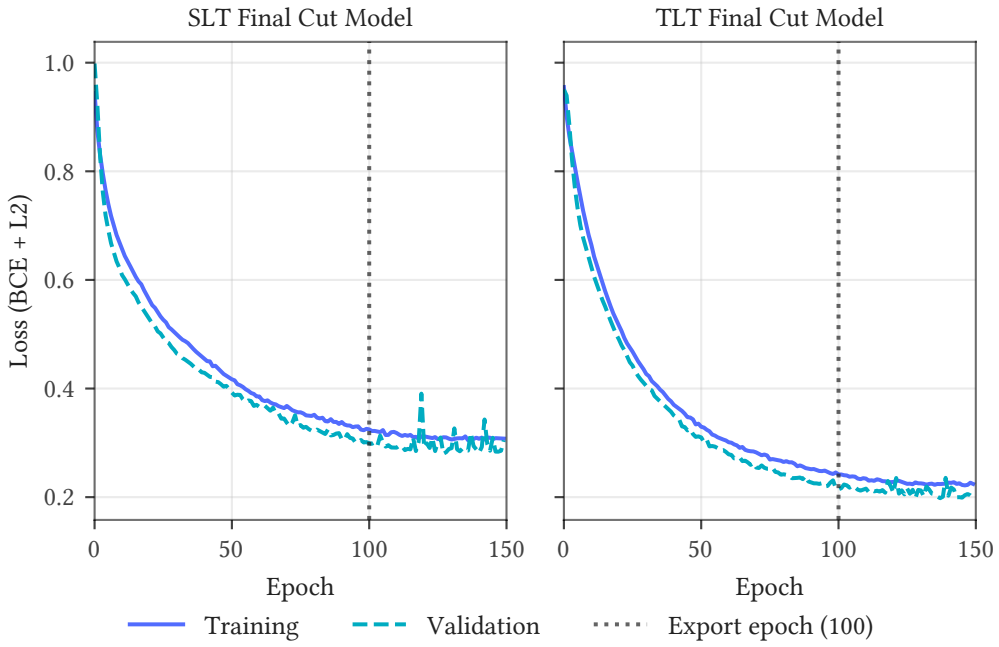

Figure 4.3: Zoomed view of the first 150 training epochs, showing the convergence behavior and the epoch selection point.

Figure 4.3 provides a closer look at the early training dynamics. Both models converge rapidly in the first 50 epochs, then continue to improve more gradually.

Table 4.5: Training and validation metrics at epoch 100 for both final cut models.

Model	Train loss	Val. loss	Train accuracy	Val. accuracy
SLT	0.325	0.299	92.2%	92.9%
TLT	0.242	0.214	94.8%	95.3%

The SLT training metrics at epoch 100 sit slightly below the corresponding validation metrics (higher loss, lower accuracy), consistent with the running-average-versus-end-of-epoch evaluation protocol described in Section 3.4.

Overtraining considerations

Comparing model output score distributions between training and test sets shows no serious signs of overtraining, except potentially for high-score CORSIKA events where statistics become sparse. The SLT training set contains only ∼8,000 CORSIKA events (half of the ∼16,000 total), meaning these events were resampled many times during training, which could lead to some memorization of individual high-weight events.

For NuGen, there is no meaningful difference between training and test: the model generalizes well to unseen neutrino events. Given the large NuGen sample

sizes, the models likely never saw most individual NuGen events more than once during training.

The first 50% of the CORSIKA sample (used for training) is effectively "burned" for further use in analyses that employ this selection, as the model has seen these events many times.

Score interpretation

The model outputs are *not* used as hard cuts during processing. Instead, the continuous scores are stored for each event, and final event selection is performed at the analysis stage by applying thresholds. This keeps the processing pipeline maximally flexible: different physics analyses can tune thresholds independently or work with continuous event weights.

For the default Lightning Tracks configuration, events with scores above a tuned threshold (determined by the sensitivity optimization, Section 5.2) are included in the final sample. The threshold can be adjusted to trade off between sample purity and effective area depending on the analysis requirements.

4.4 Prior art and departures

Prior to this work, boosted decision trees (BDTs) have been the predominant machine-learning tool for final-level event selection in IceCube point-source samples.[111] The physics-weighted MLP approach described in the previous section re-examines that default. The main reason to prefer an MLP here is regularization. A boosted decision tree builds its model from binary splits read directly off the training data, which makes it prone to overtraining. The available remedies—fitting an ensemble of trees to disjoint subsets of the data, for instance—are less straightforward than the standard regularization techniques for an MLP: dropout, an L_2 weight penalty, and early stopping. A second consideration is the nature of the input features. For the well-behaved, largely continuous detector features used here, multilayer perceptrons are competitive with gradient-boosted decision trees. The advantage of boosted trees concentrates on irregular, skewed, or high-cardinality categorical feature distributions, of which we have none.[112] There is therefore no clear upside to a BDT here, only potential downsides.

[111] IceCube Collaboration 2017a, IceCube Collaboration 2025, "All-sky Neutrino Point-source Search with IceCube Combined Track and Cascade Data", Hünnefeld 2023, "Observation of high-energy neutrinos from the Milky Way", Sclafani 2023, "Observation of Neutrinos from the Milky Way Galaxy".

[112] McElfresh et al. 2023, "When Do Neural Nets Outperform Boosted Trees on Tabular Data?"

5 Sensitivity Optimization

The final cut models of the previous chapter supply a score for every event; what they do not supply is where to cut it. This chapter sets the thresholds by optimizing point-source sensitivity directly. We build intuition with a counting picture that already shows when a cut helps (Section 5.1), then optimize the cuts against the full point-source sensitivity in practice (Section 5.2), discuss what the optimization buys and where it is limited (Section 5.3), and close by comparing the approach with prior cascade-selection work (Section 5.4).

5.1 The counting-experiment picture

Point source sensitivity in neutrino astronomy is fundamentally a signal-to-noise problem. The simplest useful model is a counting experiment: count every event that falls into a fixed search window and compare the count against the background expectation. Intuitively, someone first thinking about how to calculate the significance of such a counting experiment may naturally assume it is driven by the signal-to-background ratio or the absolute difference of signal and background counts. But thinking further they would quickly realize that for a fixed background one could simply subtract the background from the total count and would be left with only the signal at infinite significance, and for a truly constant background, this would trivially be correct. But any real such count is Poisson distributed, and for a Poisson distribution the variance equals the mean. With B expected background events in the window, the background-only count therefore fluctuates with standard deviation $\sqrt{B}$. A source adds S expected signal events on top. To be detected, that excess has to stand out against the background *fluctuations* that could fake it, so in the large-B regime, where the Poisson distribution is well approximated by a Gaussian, the statistical significance of the excess is the excess measured in units of the background fluctuation, and not the absolute background:

$$\text{Significance} = \frac{S}{\sqrt{B}}. \tag{5.1}$$

The background fluctuation in this context is also often called *noise*, and (5.1) is therefore the *signal-to-noise ratio*. Because the background enters under a square root, doubling the signal improves significance by a factor of 2, while halving the background improves it by only $\sqrt{2} \approx 1.4$.

A real point-source analysis is, of course, more complex: the unbinned likelihood, which we introduce in Section 9.1, incorporates energy and angular information, and sensitivity depends on the PSF width and spectral assumptions in addition to event counts. Nevertheless, at a fixed PSF width and fixed energy, the problem effectively reduces to a counting experiment.

We can now use this simplified picture to develop our understanding of when a cut that separates signal from background (such as a cut on our final classifier scores, Section 4.3) improves point-source sensitivity. First, we have to realize that any realistically available cut is imperfect: it reduces both signal and background. If a cut reduces signal by a factor f_S and background by a factor f_B,

$$\text{signal-to-noise ratio} = \frac{S/f_S}{\sqrt{B/f_B}} = \frac{S}{\sqrt{B}} \cdot \frac{\sqrt{f_B}}{f_S}. \tag{5.2}$$

It follows that the cut improves sensitivity only if

$$\sqrt{f_B} > f_S \quad \Rightarrow \quad f_B > f_S^2. \tag{5.3}$$

We define the *cut power* p such that $f_B = f_S^p$. Three regimes follow. When $p < 2$, the cut always hurts sensitivity (Figure 5.1, top panel): no matter how aggressively or conservatively it is applied, it removes proportionally too much signal relative to the background it rejects, the optimal cut strength is zero, and the cut should not be used at all. A classifier with $p < 2$ for a given signal hypothesis is simply not discriminating enough to be useful. When $p \approx 2$, the cut approximately preserves sensitivity but cannot substantially improve it (Figure 5.1, middle panel); it can still be useful for reducing the total event rate, and thus the computational cost, without degrading the analysis. When $p > 2$, the cut produces a clear sensitivity maximum at a nonzero cut strength (Figure 5.1, bottom panel), and the higher the cut power, the more aggressively the optimal cut removes background while retaining signal. The filter models of Chapter 3 operate in this regime: their signal-to-noise ratio is still increasing at the chosen operating thresholds, indicating $p > 2$, but the filter thresholds are set for rate reduction rather than sensitivity optimization, and the remaining discriminating power is captured by the final cut MLPs, which receive the filter scores as input features.

These cut-power curves are illustrative; they do not use simulation or data. They simply assume realistic component counts (atmospheric muons, atmospheric neutrinos, and astrophysical neutrinos) and apply a theoretical cut at the stated power, holding f_S, f_B, and therefore p constant as the cut strength varies. In a real selection p is not constant: it depends on how the topologies populate the reconstructed observables the cut acts on (Chapter 2, Chapter 3). We assume the cut can separate muons from neutrinos, but not atmospheric from astrophysical neutrinos.

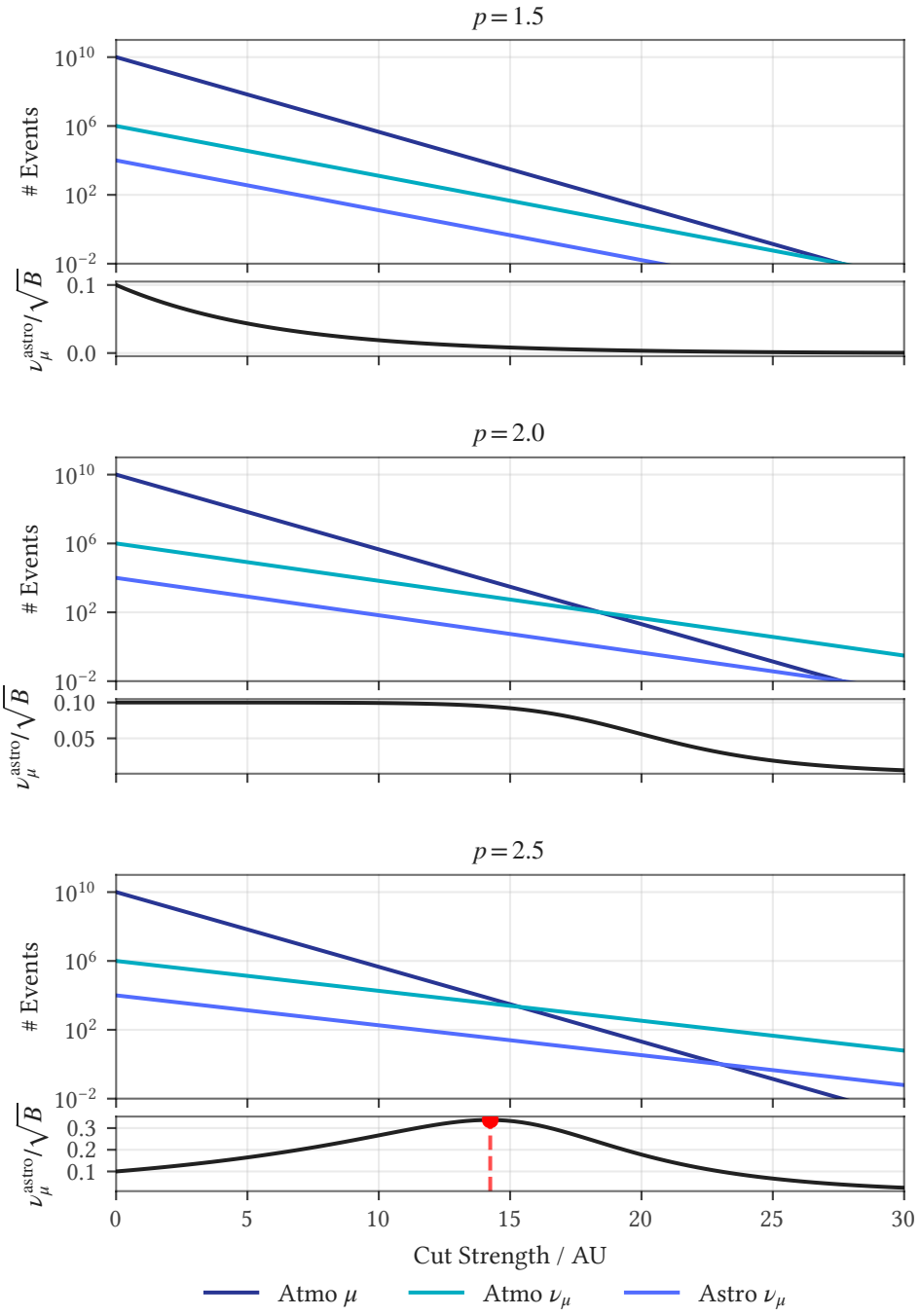

Figure 5.1: Signal-to-noise optimization of a classifier cut for three values of the cut power: $p = 1.5$ (top), $p = 2.0$ (middle), and $p = 2.5$ (bottom). Each block shows event counts and the signal-to-noise ratio as a function of cut strength, sharing a common cut-strength axis with a single legend below; the muon to atmospheric-ν ratio panel of the individual plots is omitted here, as it is flat across the cut. At $p < 2$ the signal-to-noise curve falls monotonically and the optimal cut is zero (no marker is drawn); at $p \approx 2$ it is approximately flat; at $p > 2$ a clear maximum appears at a nonzero cut strength, marked in red.

The cut power of a given classifier is not a fixed property: it depends on the background composition, which changes with declination. The same MLP score cut that achieves $p > 2$ near the horizon may have $p < 2$ in the deep southern sky, simply because the background it faces is different.

The critical factor is the relative rate of atmospheric muons and atmospheric neutrinos, exactly the separation the illustrative model assumes. Atmospheric neutrinos are genuine neutrino-induced muon tracks, indistinguishable from astrophysical signal by event morphology, so no morphology-based quality cut can separate them. Once the atmospheric muon rate has been reduced to match the atmospheric neutrino rate, further cutting removes signal and irreducible background equally, yielding no net sensitivity gain. This crossover point acts as a natural floor—cutting beyond it only reduces statistics. The baseline target of the Lightning Tracks sample design was therefore to reduce the atmospheric muon rate to approximately match the atmospheric neutrino rate before performing further sensitivity optimization.

The one mechanism that could in principle separate atmospheric from astrophysical neutrinos is the atmospheric self-veto (Section 2.2), and then only for down-going events, where an atmospheric neutrino can be tagged by the accompanying muons of its parent air shower. In Lightning Tracks this does not change the picture: the cut strength at which the self-veto becomes visible for down-going starting tracks already lies well past the global sensitivity maximum, where it produces a weak local maximum at most, so it does not move where the optimization places the cut.

The declination dependence then follows from how the ratio of muons to neutrinos varies across the sky. Near the horizon, the Earth's overburden attenuates the muon flux while remaining transparent to neutrinos; the ratio of muons to neutrinos is moderate, the topology-based classifier achieves high cut power ($p > 2$), and cutting harder yields substantial sensitivity gains up to the crossover point. In the deep southern sky, atmospheric muons overwhelm all other backgrounds: even for starting tracks, the enormous muon flux produces a significant population of single muons that penetrate several outer detector layers without depositing light, mimicking a starting event topology, so the classifier cannot reject muons efficiently enough relative to the signal it removes (the cut power falls below 2) and the optimal cut loosens rather than tightens. Higher purity could be achieved by cutting harder, but doing so hurts rather than helps sensitivity; this is visible in the data-MC composition for SLT (Chapter 6), where the atmospheric muon fraction increases steadily toward the south. In the northern sky, the Earth shields most atmospheric muons, the background is dominated by irreducible atmospheric neutrinos, and further cutting is counterproductive because it removes signal and background at the same rate.

The same principle applies to the observation time window: the intrinsic background rate determines where the optimal cut lies, and reducing the observation window reduces the effective background. For time-dependent searches (transient

follow-up, flares), the background scales with the observation window T,

$$\text{Sensitivity} \propto \frac{S}{\sqrt{B \cdot T}} = \frac{S}{\sqrt{B}} \cdot \frac{1}{\sqrt{T}}. \tag{5.4}$$

A shorter window lowers the effective background rate, shifting the optimal cut to looser values—exactly the same effect as moving to a declination with lower intrinsic background. For a 1000-second transient follow-up versus a steady-state search over 10 years, the effective background reduction is a factor of $\sqrt{10 \times 365.25 \times 24 \times 3600/1000} \approx 560$, which shifts the optimal cut substantially.

5.2 Cut optimization in practice

The quantity the optimization targets is the *sensitivity flux*: the minimum steady point-source flux that the analysis can distinguish from background at 90% confidence. Operationally, a source flux is injected at a fixed sky position and spectral index, and the sensitivity is the flux at which the analysis outcome exceeds the median background-only outcome in 90% of simulated experiments. Lower values are better.

For the purposes of this chapter, the machinery that produces this number is a black box: events pass the selection under a candidate cut, the full point-source analysis runs on the surviving sample, and a sensitivity flux comes out. Each candidate cut value therefore receives a figure of merit that folds in event counts, angular resolution, and energy information, rather than a counting-experiment proxy. How the analysis computes this number (the unbinned likelihood, the test statistic, and the injection procedure) is the subject of Chapter 9 (in particular the integrated sensitivity discussion, Section 9.11).

With this figure of merit defined, we drop the counting-experiment simplification and optimize the cuts against the full point-source sensitivity directly, which lets us check how well the counting picture holds up in practice.

To inform the cut function design, we perform a grid search over uniform cut values for the time-integrated search for point sources, using the full 12-year dataset. The grid spanned 19 cut values from 0.05 to 0.95 in steps of 0.05, 191 declination points uniform in $\sin\delta$ from $-95/96$ to $+95/96$ ($\approx\pm 0.990$), three spectral indices $\gamma \in \{2.0, 2.5, 3.0\}$, and 100,000 background trials per configuration. At each $(\sin\delta, \gamma)$ point, we compute the sensitivity for every cut value and identify the cut that gives the best (lowest) sensitivity. Figure 5.2 shows the resulting sensitivity curves for each uniform cut value, and Figure 5.3 shows the range between the best and worst cuts at each declination.

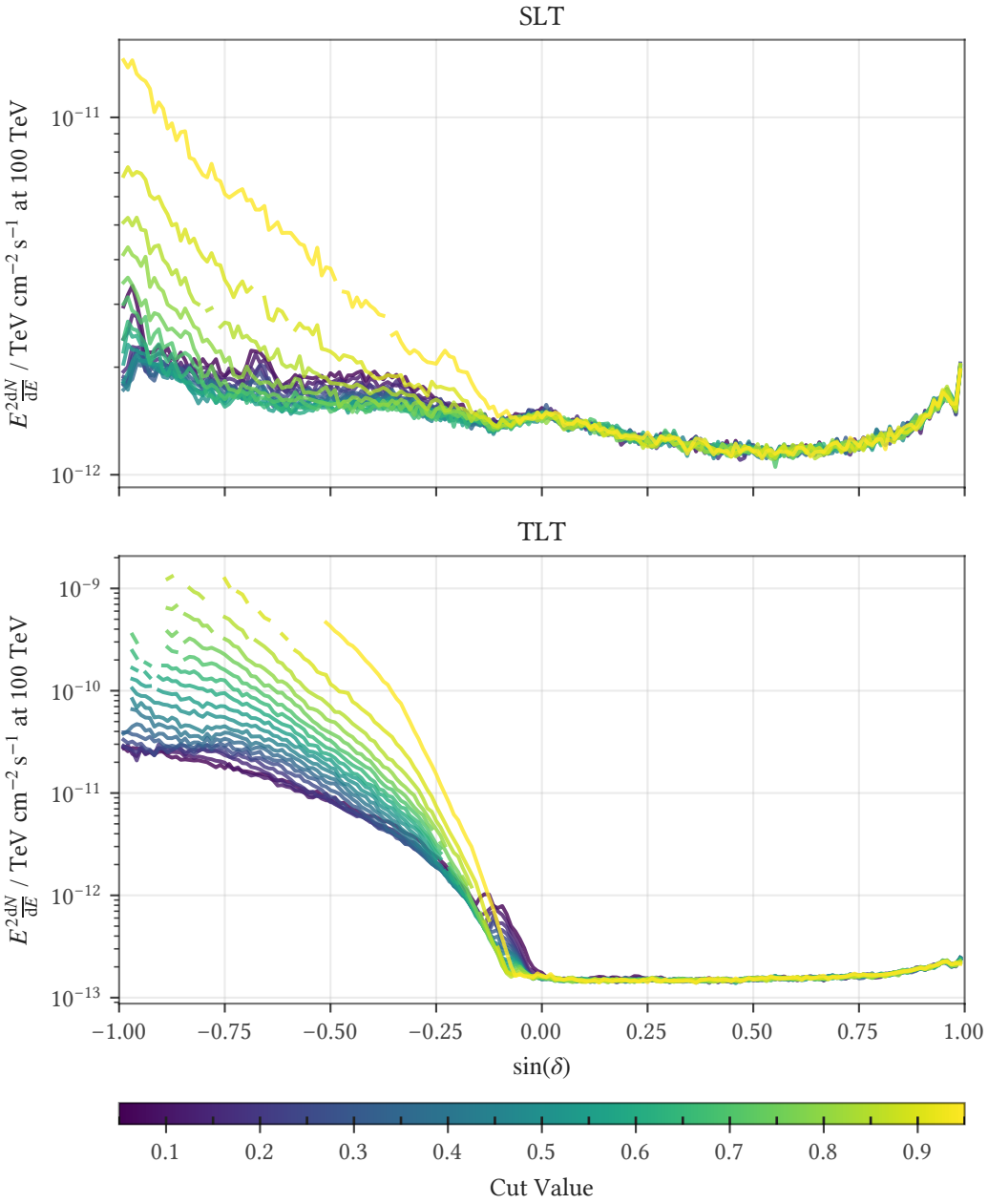

Figure 5.2: 90% sensitivity vs. $\sin\delta$ for a steady point source with an $E^{-\gamma}$ spectrum at $\gamma = 2.5$, for SLT (top) and TLT (bottom). Each line is a different uniform cut value (color scale from loose to tight), and the optimal cut varies with declination. In the TLT panel, at loose cut values the muon-horizon bump (Section 5.2) is visible as a local sensitivity degradation around $\sin\delta \approx -0.1$, its exact position shifting slightly with the cut value.

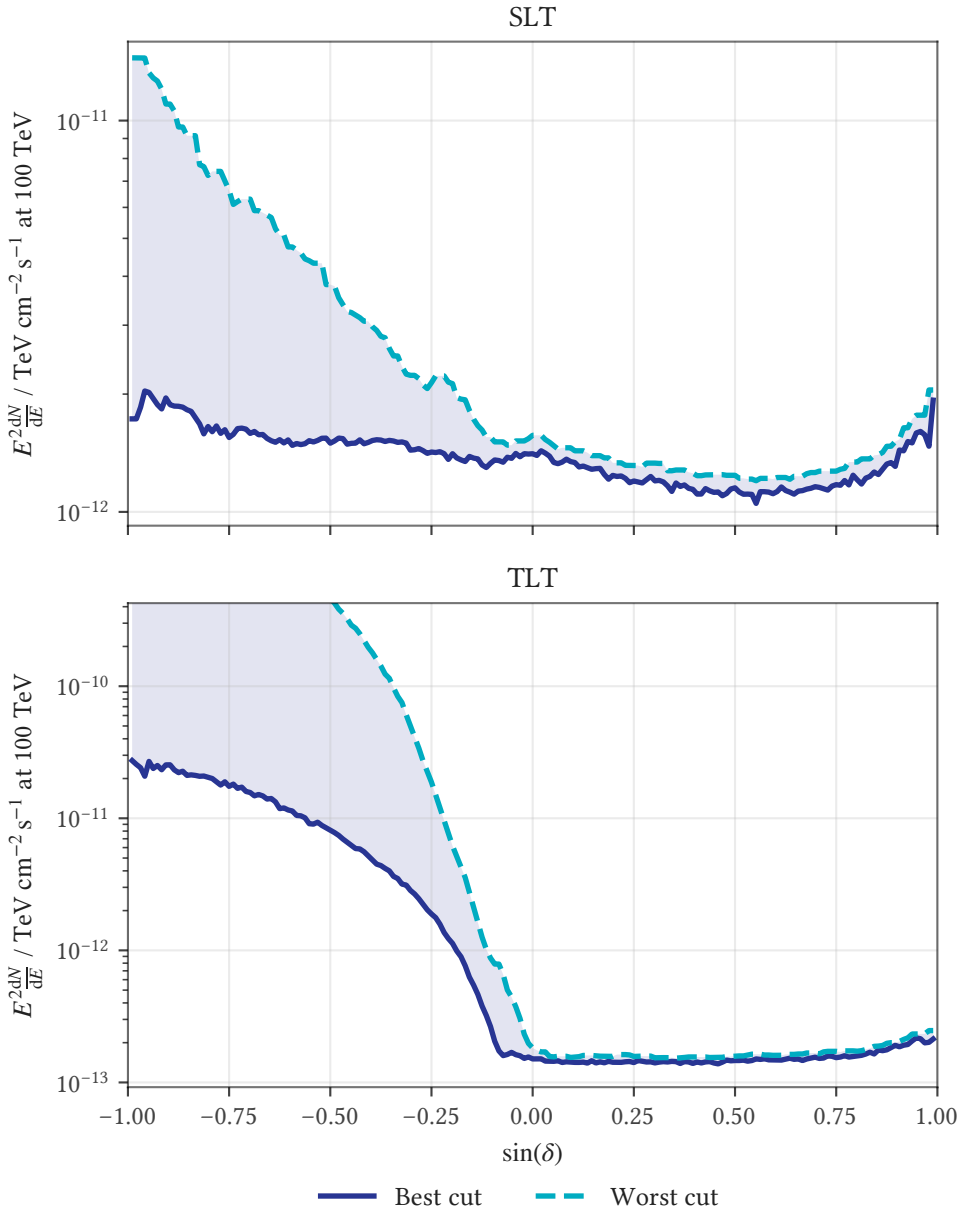

Figure 5.3: Sensitivity envelope at $\gamma = 2.5$ for SLT (top) and TLT (bottom). The shaded region shows the range between the best and worst cuts across the grid. The green line is the best achievable sensitivity and the red dashed line the worst. In the SLT panel, the muon-horizon bump (Section 5.2) is visible in the worst-cut curve but absent from the best-cut curve. The y range is deliberately chosen to cover the valid range: for $\sin\delta < -0.5$ the worst tested cuts have effectively no sensitivity (a diverging flux).

The optimal cut value at each declination depends on the assumed spectral index, and no single uniform cut is optimal everywhere. To strike a γ-agnostic compromise, we hand-fit simple analytic functions to the grid scan results, guided

by the optimal cut envelopes across all three spectral indices. For SLT,

$$\text{score} > 0.6 \cdot \sin(\theta_{\text{zenith}})^{1/3}, \tag{5.5}$$

and for TLT (the baseline cut),

$$\text{score} > \frac{0.3}{1 + e^{-6.5(\theta_{\text{zenith}} - 1.2)}} + 0.05. \tag{5.6}$$

Figure 5.4 compares these functions against the per-declination optimal cut values from the grid scan. For TLT, the plot also shows the muon horizon bump cut (red dashed line), which raises the threshold near $\sin\delta \approx -0.08$ to compensate for the switch from full to spatial-only signal subtraction.

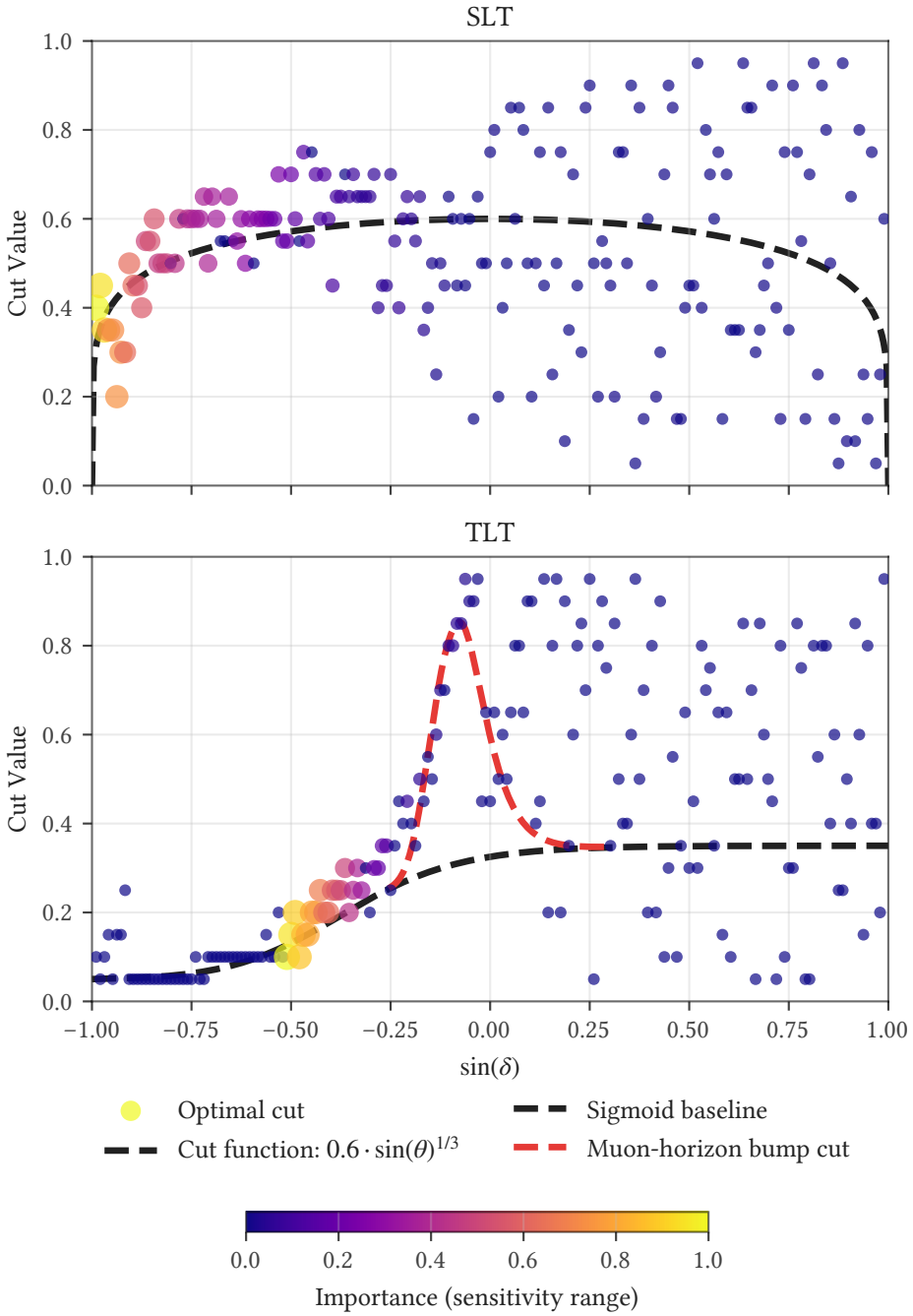

Figure 5.4: Optimal cut values against the deployed cut function at $\gamma = 2.5$, for SLT (top) and TLT (bottom). Scatter points show the optimal cut at each declination, with point size and color indicating the importance (the sensitivity range at that declination). In the TLT panel, the black dashed line is the sigmoid baseline and the red dashed line the bump cut added near the muon horizon.

The muon horizon bump cut for through-going tracks

For through-going tracks we need to give special treatment to a particularly challenging region, which we refer to as the *muon horizon* here. We define it as the region in reconstructed observable space where the atmospheric neutrino flux and atmospheric muon flux cross over. South of the horizon, atmospheric muons dominate the event rate. North of it, atmospheric neutrinos take over. The transition is steep and energy-dependent, coincides with extremely high event statistics, and is most pronounced in the 1–10 TeV energy range where point source sensitivity peaks. Any mismatch between the likelihood model and the true background composition in this region has an outsized effect on the analysis.

The baseline TLT cut function was optimized in conjunction with full signal subtraction (Section 9.7), which includes the energy term in the background correction. Full signal subtraction handles the muon horizon effectively because the energy-dependent weights suppress the muon-dominated regime relative to the signal hypothesis. However, full signal subtraction introduces problems at other declinations: high-energy events with poorly constrained energy PDF ratios produce extreme correction weights that distort the likelihood surface, resulting in TS distributions that are significantly narrower than those at neighboring declinations without such outliers (see the asymptotic expectation discussion, Section 9.8). Although statistically correct, this narrowing hurts sensitivity and manifests as conspicuous low-TS band artifacts in all-sky maps. For the all-sky scan (Chapter 10), where every declination is tested and such artifacts would be visible in the result, spatial-only signal subtraction is used instead. Dropping the energy term from the correction produces a likelihood that is well-behaved across the sky.

Spatial-only signal subtraction performs well everywhere except at the muon horizon, where the absence of the energy correction leaves the likelihood exposed to the sharp muon–neutrino transition. Without additional mitigation, sensitivity in this declination band degrades substantially. The bump cut addresses this by raising the MLP score threshold near the muon horizon, rejecting atmospheric muon contamination more aggressively. The bump is an asymmetric Gumbel function added to the baseline sigmoid:

$$\text{score} > \underbrace{\frac{0.3}{1 + e^{-6.5(\theta - 1.2)}} + 0.05}_{\text{sigmoid}} + \underbrace{1.471 \cdot e^{-(t + e^{-t})}}_{\text{bump}}, \tag{5.7}$$

where $t = (\theta - 1.4938)/w$ with asymmetric widths $w_L = 0.0801$ (south side) and $w_R = 0.0526$ (north side), and θ is the zenith angle in radians. The bump peaks at zenith $\approx$85.6° ($\sin\delta \approx -0.08$), exactly at the muon horizon, raising the total threshold to approximately 0.85. The SLT cut function is the same in both cases.

The bump cut restores nearly the same sensitivity as full signal subtraction at the muon horizon while allowing the rest of the sky to benefit from the more stable spatial-only likelihood. It also significantly improves data-MC agreement in the muon horizon region. Both cut functions are valid: the bump cut does not supersede the baseline. An analysis targeting sources specifically at muon horizon

declinations could achieve slightly better sensitivity using full signal subtraction with the baseline cut, but would need to verify that none of the tested declinations coincide with the energy PDF pathologies discussed in the asymptotic expectation section (Section 9.8).

5.3 Discussion

In addition to the threshold optimization, one further design choice is worth quantifying: Lightning Tracks treats its two topologies as separate likelihood components rather than merging them into a single sample (motivated in Section 1.2). Computing the point-source sensitivity both ways shows that keeping SLT and TLT separate gains roughly 25 to 45% in the southern sky and essentially nothing in the north (the sensitivity flux ratio between the merged and separated configurations is $\approx$1.25 to 1.45 in the south, near one in the north), measured on the full point-source sensitivity rather than point-spread alone.

With the optimization carried out, we can ask how well the counting-experiment picture of Section 5.1 held up in practice. For starting tracks (SLT), it holds closely: the MLP score primarily distinguishes starting event topology from penetrating muons, and the optimal cut follows the cut-power prediction at each declination, tightest at the horizon where the cut power is highest, loosest toward the south where it drops below 2 (Figure 5.2). These declination-dependent changes in the optimal cut produce the characteristic double-peak structure in SLT's background spatial distribution (see the background spatial PDF discussion, Section 9.4).

For through-going tracks (TLT), the situation is more complex. Through-going tracks lack topological information to distinguish single atmospheric muons from neutrino-induced muons (the two are indistinguishable by event morphology alone). The only available discriminant is the energy spectrum: astrophysical sources with harder spectra ($\gamma \lesssim 2.5$) produce events at higher energies than the softer atmospheric backgrounds. The MLP exploits this by effectively applying an energy-dependent cut in the southern sky, where topology provides no leverage. As a result, the optimal cut for TLT depends more strongly on the assumed spectral index than for SLT: softer spectra ($\gamma \geq 3$) favor looser cuts because the energy discriminant loses power, while harder spectra ($\gamma \leq 2.5$) favor tighter cuts that reject the soft atmospheric component more aggressively. This spectral dependence is clearly visible in Figure 5.5.

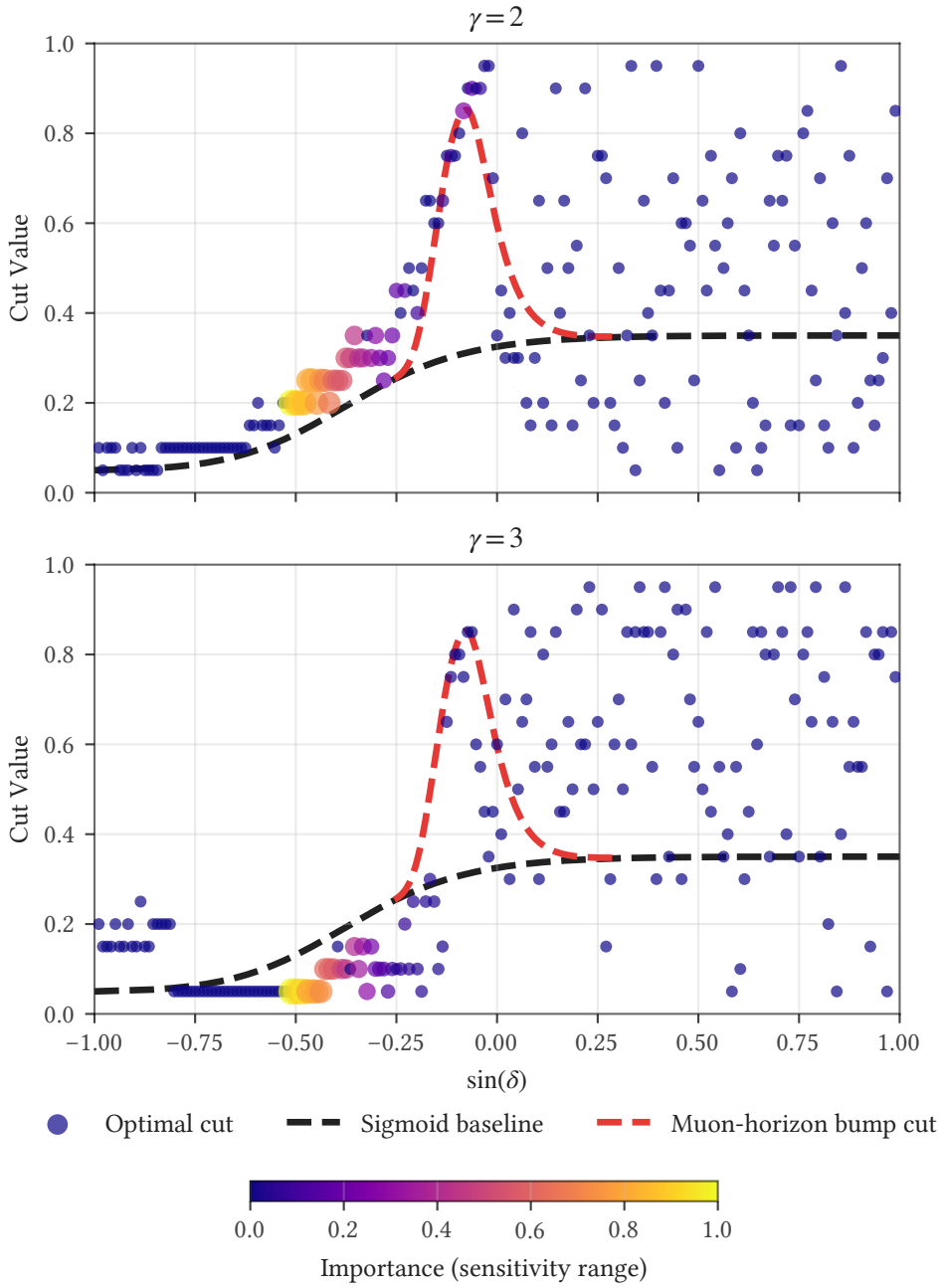

Figure 5.5: Optimal TLT cut values against the deployed cut function, for $\gamma = 2.0$ (top) and $\gamma = 3.0$ (bottom). The deployed cut (the sigmoid baseline plus the muon-horizon bump, shown dashed) is identical in both panels: it does not depend on the assumed spectrum. What the spectrum changes is the per-declination optimal cut. At $\gamma = 2.0$ the optimal cuts sit systematically tighter than the deployed function; at $\gamma = 3.0$ the high-importance points in the southern sky fall to looser values. Scatter point size and color indicate the importance, the sensitivity range at that declination.

The grid search results (Section 5.2) confirm this picture across all declinations

and spectral indices. It was also verified that the expected $1/\sqrt{T}$ scaling of sensitivity with observation time (Section 5.1) holds, but a time-dependent treatment is out of scope for the time-integrated search for point sources here, so we do not include it.

The hand-fitted functions approximately follow the optimal cut envelope across spectral indices, providing a reasonable compromise without being tuned to any single γ. Much of the scatter in the per-declination optimal cut values is driven by Monte Carlo statistical uncertainty in the sensitivity estimates rather than genuine structure, so the smooth analytic functions effectively average over this noise.

These cut functions are optimized for the time-integrated search for point sources, using the full 12-year dataset. As discussed in the observation window dependence (Section 5.1), shorter observation windows shift the optimal cut to looser values. The muon horizon bump cut is another example of the same limitation: the optimal cut depends on the signal subtraction configuration, and each configuration favors a different threshold near the horizon. More broadly, the current IceCube point source sample approval process calls for fixed final cuts, which forces compromises that no individual analysis would make on its own. In Figure 5.4, the TLT sigmoid is visibly tuned to $\gamma = 2.5$ as a general compromise, while the optimal cut values for $\gamma = 2$ scatter systematically above it and those for $\gamma = 3$ scatter below (Figure 5.5). Ideally, each analysis would determine and apply the final cuts that yield optimal sensitivity for its specific source hypotheses, rather than relying on a single compromise function (a future-work direction taken up in Chapter 15). The time-integrated case (the most common use case) is therefore the natural optimization target.

5.4 Prior art and departures

The closest prior selection to compare against is DNN Cascades (DNNC), the deep-learning cascade selection behind the 4.5σ Galactic-plane observation.[113] The DNNC final cuts were likewise optimized for sensitivity: the selection applies boosted decision trees (BDTs) whose cuts were optimized for its source hypotheses, candidate thresholds for the muon BDT were compared before the final value was chosen, and a looser cut on the cascade BDT was adopted because it was found to improve sensitivity.[114] How that optimization was evaluated is only partly documented, however. The cut-selection figures show counts of remaining events above 1 TeV as a function of the BDT threshold, and it is unclear which source spectrum and declinations the sensitivity evaluation assumed, or what figure of merit beyond these count curves informed the chosen values.

The training of the BDTs is described at the algorithmic level: events misclassified by earlier trees receive higher weight in later ones, the reweighting intrinsic to boosting.[115] Whether the training events additionally carried a physics weighting, such as a flux model or a spectral assumption, is not documented. By the argument of Section 3.6, the training weighting is part of the objective the classifier minimizes, so this choice shapes the score before any threshold is placed on it.

[113] IceCube Collaboration 2023, "Observation of high-energy neutrinos from the Galactic plane", Sclafani 2023, "Observation of Neutrinos from the Milky Way Galaxy".

[114] Sclafani 2023, Sec. 3.2.6, p. 48.

[115] Sclafani 2023, Sec. 3.2.6, p. 47.

The structural difference that is documented lies in the form of the final cut. DNNC places a single global threshold on each BDT score, 5×10^{-3} for the muon BDT and 0.1 for the cascade BDT.[116] Zenith is among the BDT input features, so the score itself carries directional information, but as argued in Section 3.6, a direction-aware score does not make the threshold sensitivity-aware: the classifier minimizes its classification loss, while the sensitivity optimum moves with declination, which the grid search of Section 5.2 measures directly. The selection built here makes that dependence explicit through the declination-dependent cut function on the MLP score (Section 5.2).

[116] Sclafani 2023, Sec. 3.2.6, p. 48, Figs. 3.6–3.7.

6 Data-MC Agreement and Systematic Uncertainties

The sample is now fully defined. Before Part II builds statistical machinery on top of it, this chapter validates it: we compare data against simulation across the observables the selection and the likelihood depend on (Section 6.1 through Section 6.2), and then turn to the detector systematics that the later analysis chapters must carry (Section 6.3, Section 6.4).

6.1 Weighting and flux assumptions

Lightning Tracks is designed for point-source analyses. In this regime, the background for the test statistic is obtained from data scrambling, so the validity of the analysis results is immune to systematic mismodeling of the atmospheric background (see Section 9.8 for the full discussion). Nevertheless, to validate the selection behaves well on real data we need to make assumptions about the atmospheric fluxes. The data-MC comparisons in this chapter use the *burn sample*: experimental data designated for analysis development under a much lower blindness standard, making up about 10% of the full dataset.

For the atmospheric neutrino flux we use the model of Schöneberg,[117] computed with a direct CORSIKA air-shower simulation, not the commonly used analytic solution of the cascade equations (as in MCEq[118]), using the Gaisser H3a cosmic-ray spectrum[119] and the SIBYLL 2.1 hadronic interaction model.[120] We do not model the self-veto effect. At tighter cut levels than those used for the point source final selection, this produces visible under-fluctuations of starting events in the southern sky at high energies (the MC overpredicts the neutrino rate because it does not account for accompanying muons vetoing the event). At the cut levels shown here (the point source final selection), the southern sky is strongly muon-dominated and the self-veto effect is not prominently visible.

The atmospheric muons were simulated with CORSIKA using the SIBYLL 2.1 hadronic interaction model[121] and we weight them with the Gaisser H3a primary cosmic-ray flux model.[122] In practice, the absolute normalization of this flux does not match the data perfectly. We perform a simple chi-square fit on the histogrammed burn data to find the best-fit normalization scaling factor of 0.48.

For the astrophysical neutrino flux we use the most recent IceCube northern-tracks diffuse measurement:[123]

[117] Schöneberg 2016, "The spectrum of atmospheric neutrinos above GeV energies", Sec. 2.6.

[118] Fedynitch et al. 2015, "Calculation of conventional and prompt lepton fluxes at very high energy".

[119] Gaisser 2012, "Spectrum of cosmic-ray nucleons, kaon production, and the atmospheric muon charge ratio".

[120] Ahn et al. 2009, "Cosmic ray interaction event generator Sibyll 2.1".

[121] Ahn et al. 2009.

[122] Gaisser 2012.

[123] IceCube Collaboration 2022b, "Improved Characterization of the Astrophysical Muon-Neutrino Flux with 9.5 Years of IceCube Data".

$$\Phi_\nu^{\mathrm{astro}} = 1.44 \times 10^{-18} \left(\frac{E}{100\,\mathrm{TeV}}\right)^{-2.37} \mathrm{GeV}^{-1}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mathrm{sr}^{-1}, \tag{6.1}$$

quoted per flavor ($\nu + \bar{\nu}$) at a 100 TeV pivot and applied identically to all three flavors in a 1 : 1 : 1 ratio.

6.2 The data/MC ratio

Each data-MC figure consists of two main panels. The top panel shows the burn-normalized rates of the individual MC components (atmospheric μ, atmospheric ν, astrophysical ν), their sum, and the burn data with statistical error bars. The lower panel shows the bin-wise data/MC ratio, the ratio of burn data to the total MC expectation, with uncertainties from standard error propagation.

Conceptually, the error treatment is simple: data are treated as Poisson counts, MC uncertainties are obtained from the sum of squared event weights in each bin, and both are propagated to the data/MC ratio using standard Gaussian error propagation. Burn data are histogrammed without weights, yielding counts N_j^{data} in bin j, with the statistical uncertainty treated as Poisson:

$$\delta N_j^{\mathrm{data}} = \sqrt{N_j^{\mathrm{data}}}. \tag{6.2}$$

When converting to rates using the burn livetime T_{burn},

$$R_j^{\mathrm{data}} = \frac{N_j^{\mathrm{data}}}{T_{\mathrm{burn}}}, \qquad \delta R_j^{\mathrm{data}} = \frac{\sqrt{N_j^{\mathrm{data}}}}{T_{\mathrm{burn}}}. \tag{6.3}$$

These rate errors are the vertical error bars on the data points in the top panel.

For each MC component and each bin, we sum the event weights w_i to obtain the expected contribution

$$H_j = \sum_{i \in \mathrm{bin}\ j} w_i, \tag{6.4}$$

and estimate the statistical variance as the sum of squared weights,

$$\left(\delta H_j\right)^2 = \sum_{i \in \mathrm{bin}\ j} w_i^2. \tag{6.5}$$

This is the standard treatment for weighted Monte Carlo, and it follows because the simulated sample is itself a random draw. Consider first a bin receiving n simulated events of identical weight w. The estimate is $H = nw$, and the only random quantity in it is n: rerunning the simulation would produce a different count, fluctuating as a Poisson variable with $\mathrm{Var}(n) = n$. The variance of the estimate is therefore $\mathrm{Var}(H) = w^2\,\mathrm{Var}(n) = nw^2$, which is exactly $\sum_i w_i^2$ over the n events in the bin. General weights follow by the same logic applied per event: each simulated event stands for an independent Poisson-distributed occurrence entering the sum scaled by its weight w_i, so it contributes a variance of w_i^2, and

the independent contributions add. Contributions from different components are added in bin j, and their uncertainties are combined in quadrature, so that the total MC expectation

$$H_j^{\mathrm{MC}} = H_j^{\mathrm{CORSIKA}} + H_j^{\mathrm{atmo}} + H_j^{\mathrm{astro}} \tag{6.6}$$

has an uncertainty

$$\delta H_j^{\mathrm{MC}} = \sqrt{\left(\delta H_j^{\mathrm{CORSIKA}}\right)^2 + \left(\delta H_j^{\mathrm{atmo}}\right)^2 + \left(\delta H_j^{\mathrm{astro}}\right)^2}. \tag{6.7}$$

In the top panel, the grey band around the total MC curve corresponds to $H_j^{\mathrm{MC}} \pm \delta H_j^{\mathrm{MC}}$ (normalized to the burn livetime).

For each bin in which both data and MC are defined and non-zero, we form the ratio

$$\mathcal{R}_j = \frac{R_j^{\mathrm{data}}}{R_j^{\mathrm{MC}}}, \tag{6.8}$$

where $R_j^{\mathrm{MC}} = H_j^{\mathrm{MC}}/T_{\mathrm{burn}}$. The relative uncertainty on $\mathcal{R}_j$ is obtained by standard error propagation assuming independent errors on data and MC:

$$\left(\frac{\delta \mathcal{R}_j}{\mathcal{R}_j}\right)^2 = \left(\frac{\delta R_j^{\mathrm{data}}}{R_j^{\mathrm{data}}}\right)^2 + \left(\frac{\delta R_j^{\mathrm{MC}}}{R_j^{\mathrm{MC}}}\right)^2. \tag{6.9}$$

The ratio panel shows $\mathcal{R}_j \pm \delta \mathcal{R}_j$ together with a reference line at data/MC = 1.

Finally, Table 6.1 and Table 6.2 integrate over all bins, summarizing the per-component event counts and rates for the two selections: the Monte Carlo expectations normalized to the burn livetime, the burn-sample data for comparison, and the flavor and topology breakdown.

Table 6.1: Event counts and rates for the starting-track (SLT) selection at the final cut, whole sky. The Monte Carlo expectations are normalized to the burn livetime (1.20 yr), so the Monte Carlo total and the burn-sample data row are directly comparable as counts. The rate and MC-fraction columns are livetime-independent.

Component	Events	Rate (mHz)	MC fraction (%)
Atmospheric ν_μ	44 889	1.1875	71.4
Atmospheric ν_e	1.0	< 0.0001	< 0.1
Atmospheric ν_τ	0	0	0
Astrophysical ν_μ	64.0	0.0017	0.1
Astrophysical ν_e	< 0.1	< 0.0001	< 0.1
Astrophysical ν_τ	7.3	0.0002	< 0.1
Atmospheric μ (CORSIKA)	17 887	0.4732	28.5
Total Monte Carlo	62 849	1.6626	100
Burn data	65 589	1.7351	—

Table 6.2: Event counts and rates for the through-going track (TLT) selection at the final cut, whole sky. Conventions as in Table 6.1.

Component	Events	Rate (mHz)	MC fraction (%)
Atmospheric ν_μ	171 488	4.5366	84.3
Atmospheric ν_e	3.5	0.0001	< 0.1
Atmospheric ν_τ	0	0	0
Astrophysical ν_μ	477.0	0.0126	0.2
Astrophysical ν_e	1.3	< 0.0001	< 0.1
Astrophysical ν_τ	45.3	0.0012	< 0.1
Atmospheric μ (CORSIKA)	31 382	0.8302	15.4
Total Monte Carlo	203 397	5.3807	100
Burn data	210 058	5.5569	—

For the full evaluation, Data-MC figures were produced for all observables entering the selection and the likelihood: reconstructed energy and direction, CNN and MLP filter scores, angular uncertainties, reconstruction agreement, containment probabilities, and the final cut model predictions themselves, across cut levels, sky regions, energy ranges, and alternative atmospheric flux assumptions, 55,080 in total, which is far too many to include in a dissertation. Instead, Figure 6.1 and Figure 6.2 show full-sky distributions of the most important reconstructed energy and angular observables.

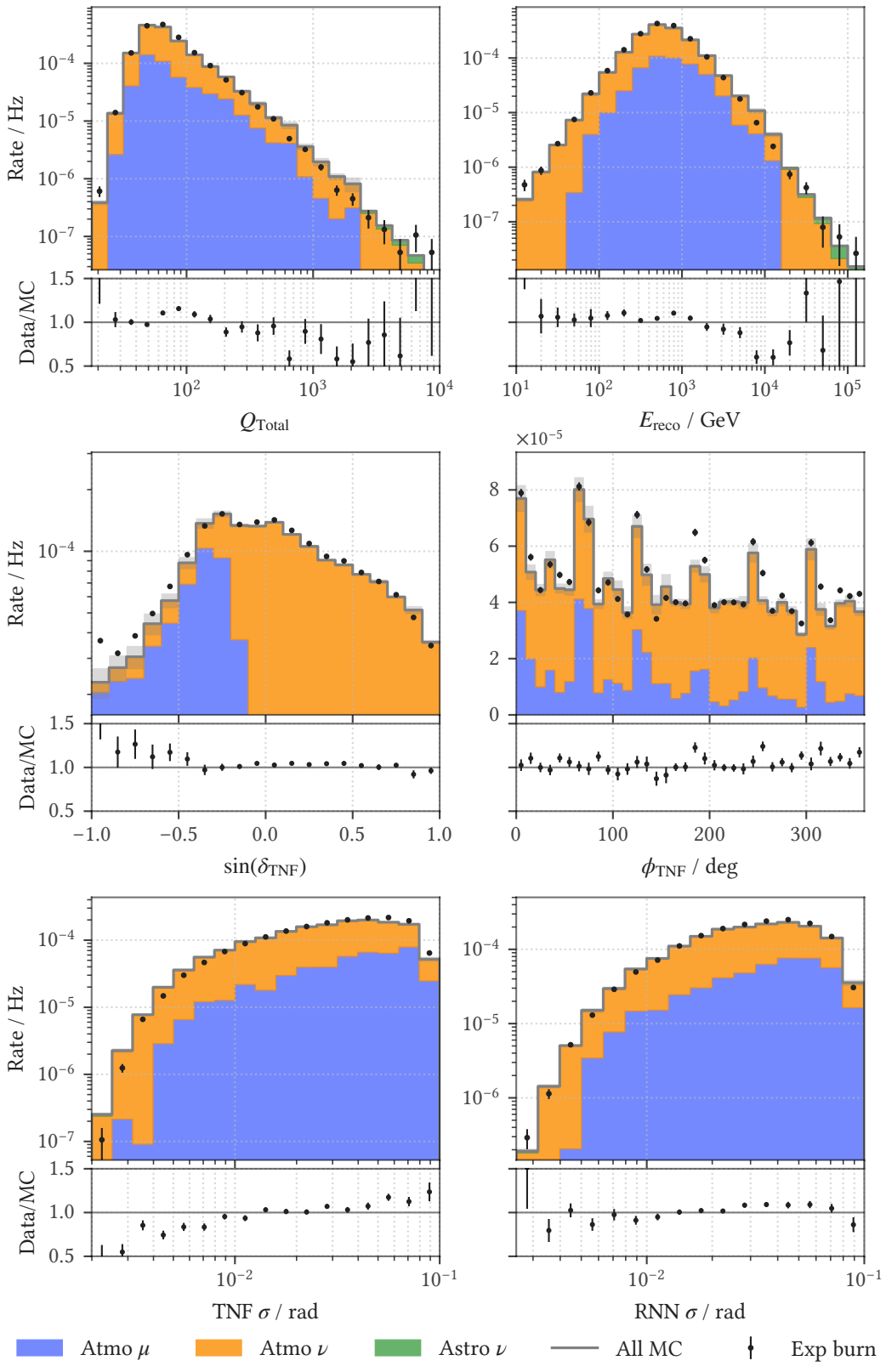

Figure 6.1: Data-MC comparison grid for the starting-track selection, full sky, at the final cut. Each of the six panels compares burn-normalized rates (top) against the total MC expectation, with the data/MC ratio below (Section 6.2).

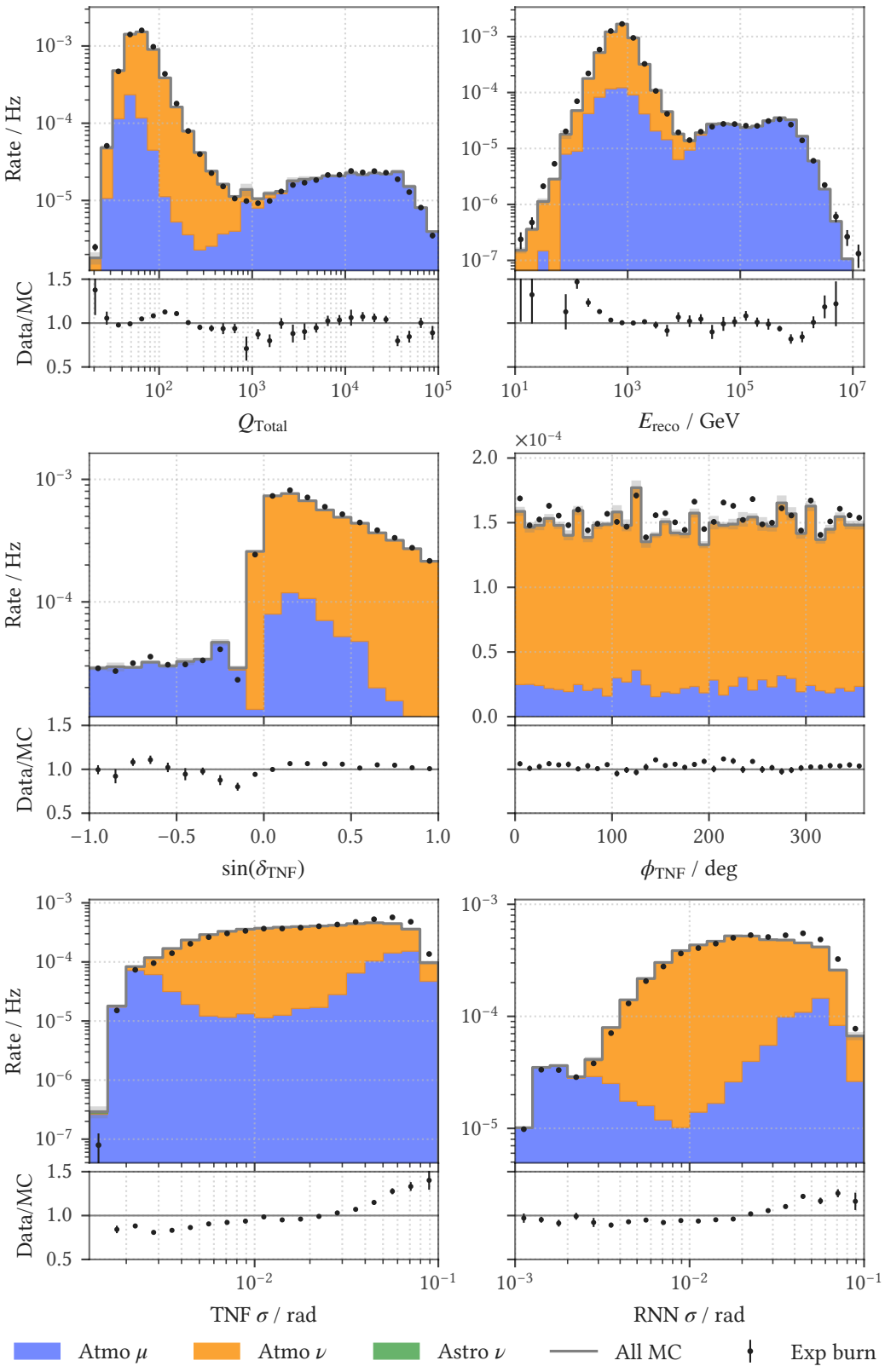

Figure 6.2: Data-MC comparison grid for the through-going track selection, full sky, at the final cut. Panels as in Figure 6.1.

Across all distributions, the dominant source of data-MC discrepancy is the modeling of the atmospheric background, for both muons and atmospheric neutrinos. This is a consequence of the selection itself: the Lightning Tracks cuts are looser than those of the dedicated diffuse analyses, so far more background survives them, many more muon bundles and many more low-energy events. The simulation reproduces this background less accurately, because it samples the deep tails of the atmospheric flux models, the regime of largest uncertainty. The limitations of the CORSIKA dataset are immediately visible: statistics are poor at low and high energies, an unavoidable constraint given the computational cost of simulating atmospheric muons at scale. Even for neutrinos, where statistics are not the limiting factor, the choice of atmospheric flux model has an enormous impact on the resulting distributions (not shown here). Treating the atmospheric parameters that drive this (the spectral-shape difference $\Delta\gamma$ and the Barr parameters for the kaon fractions[124]) as nuisance parameters would improve the agreement, but remains challenging for a selection with cuts as loose as ours. Nevertheless, the overall data-MC agreement is reasonable.

[124] Barr et al. 2006, "Uncertainties in Atmospheric Neutrino Fluxes".

Many features of these distributions could be discussed in detail. We limit our attention to two. First, the SLT azimuth distribution (Figure 6.1) shows a clear periodic structure that is absent from the TLT sample: six peaks, at azimuths near 10.5°, 68.6°, 128.7°, 189.6°, 248.2°, and 310.6°, aligned with the principal axes of the hexagonal string lattice (the directions connecting nearest-neighbor strings, near $9° + n \times 60°$). A starting event aligned with a string axis, in both direction and position, is most confidently identified as starting. When it is aligned, the instrumentation density is highest and the outermost DOMs veto best. When the track is offset in direction or position, lower-energy muons more readily pass through undetected, so the filter and final-cut models learned to cut harder in those cases. The second feature is the only apparently problematic mismatch: the TNF angular-error estimate shows a significant disagreement, with events of small estimated angular error less common in the data than the simulation predicts. It is the subject of the next section.

6.3 Angular-reconstruction systematics

We use simulation in three roles: to build the signal probability densities that enter the point-source likelihood, to calibrate the parameter estimation of Chapter 12, and to predict the analysis power, the sensitivity and discovery potential of Chapter 9. A data-MC mismatch enters these differently. In the first it can only cost analysis power, never the validity of a result: the significances are calibrated empirically, with no simulation entering the null distributions (Section 9.8). In the other two it biases the corresponding estimates. The one mismatch large enough to matter is in the reconstructed angular error: events with a small estimated error are less common in the data than the simulation predicts.

This can be traced to the algorithm that turns the recorded signals into photon counts. Light reaching a module is registered by its photomultiplier as a current

pulse, which the readout electronics record as a digitized waveform. Recovering the number and arrival times of the photons that produced it requires unfolding that waveform.[125] The simulation and the experimental data were processed with different settings for this unfolding. The exact difference does not matter here. What matters is its effect on the reconstructed angular error, shown in Figure 6.3.

[125] IceCube Collaboration 2017b, "The IceCube Neutrino Observatory: Instrumentation and Online Systems", Sec. 3.2, IceCube Collaboration 2010, "Calibration and Characterization of the IceCube Photomultiplier Tube".

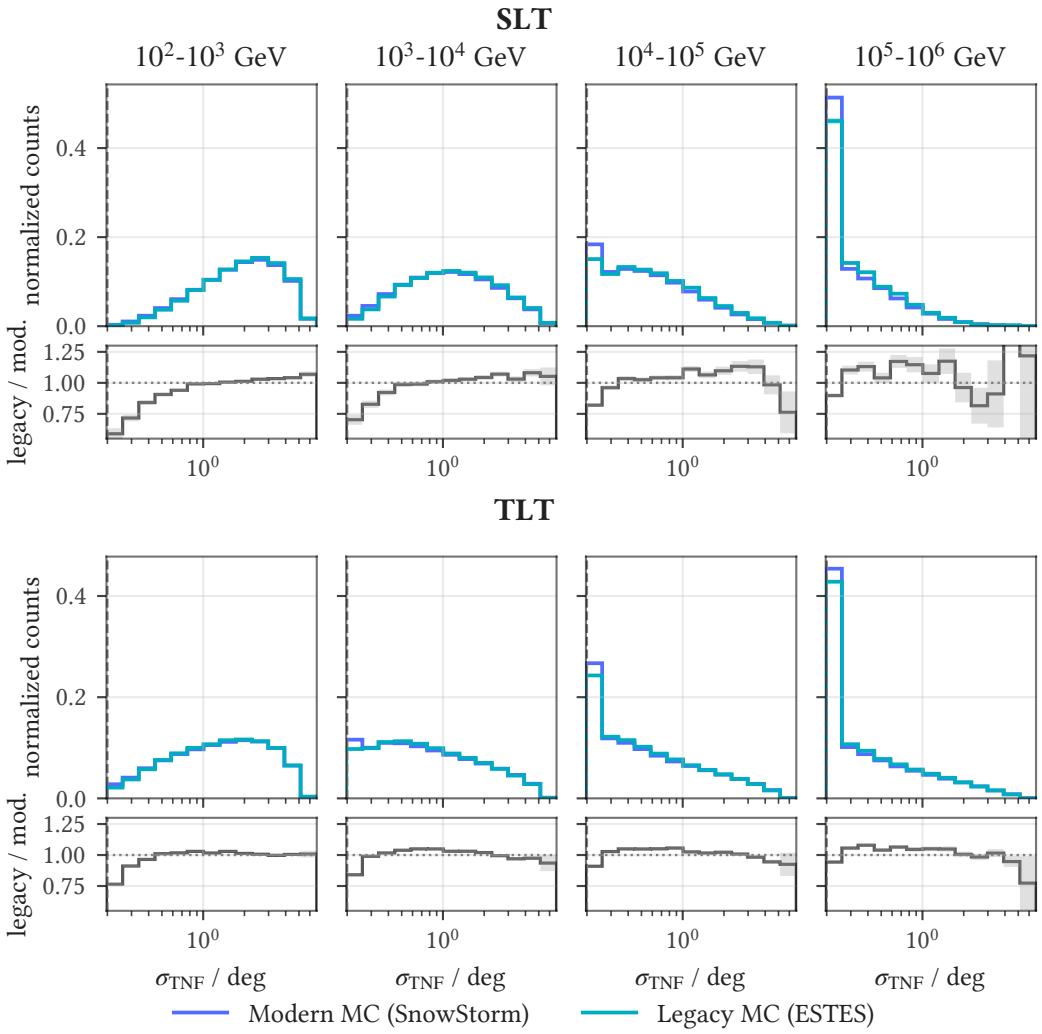

Figure 6.3: Effect of the modern processing on the pull-corrected angular error, in bands of true energy. The modern SnowStorm production is compared against an earlier production that matches the data, both carried through the identical Lightning Tracks chain to the final selection and weighted to the same power law. The modern production has more events at small reconstructed angular error. Angular error floor of 0.2° applied, causing visible pileup in the lowest angular-error bins at high energies.

To gauge how far this could shift the analysis, we repeat the relevant calculations with an earlier simulation, one processed with the same settings as the data. That

earlier production is not a clean control, since it also carries an older ice model and other issues since corrected, but comparing it against the modern simulation (the figure above) and re-deriving the sensitivity and discovery-potential fluxes from it (Section 9.11) bounds the size of the effect. We find a ∼5% upward change in the required fluxes and a ∼2% upward bias in signal flux recovery. This means that under the legacy signal model the analysis is estimated to be ∼5% less powerful than under the modern model with its mismatched unfolding settings.

The parameter-estimation bias is small, and it is partly treatable. The Feldman-Cousins construction folds detector systematics directly into its pseudo-experiments (Section 12.7), and the SnowStorm detector variations of Section 6.4 reach the same observables: the DOM efficiency and the ice absorption shift the reconstructed angular error in much the way the mismatch does, so marginalizing over them absorbs part of it.

A further mitigation is already built into the likelihood. We floor the reconstructed angular error at 0.2°, so that an event reconstructed with an implausibly small error cannot take on an extreme spatial weight in the point-source likelihood when that error is not to be trusted. The floor is not free: for the events whose small error is genuine, it inflates the angular uncertainty and degrades the coverage of the point-spread function, at some cost in power (Section 8.5). Here it is clearly the right choice, and it also absorbs much of the mismatch: we apply the floor identically to data and simulation, so the events that pile at it carry the same weight in both, and the difference among them is neutralized.

This dominant data-MC mismatch was avoidable. The only modern simulation available, built on the latest ice model, was produced with updated processing settings intended for a reprocessing of the experimental data as well. That reprocessing has been postponed for years and has still not been carried out, so the modern simulation and the data it is meant to model were prepared with different settings. The discrepancy follows from that delay, not from any deficiency of the reconstruction or the selection.

[126] IceCube Collaboration 2019, “Efficient propagation of systematic uncertainties from calibration to analysis with the SnowStorm method in IceCube”.

6.4 Detector and ice systematics

This section quantifies how the dominant detector and ice systematic uncertainties propagate to the two quantities the point-source analysis uses: the effective area and the pull-corrected angular error. For these events we use the SnowStorm MC ensemble,[126] in which each batch of approximately 2,000 simulated events is generated with an independent draw of detector systematic parameters from a multivariate prior. Conceptually, SnowStorm replaces the older practice of generating discrete *perturbed sets* at fixed offsets from nominal (followed by hyperplane interpolation across results) with a proper prior-marginalization scheme, so the ensemble marginalizes over the systematic uncertainty by construction. Each batch draws five parameters from independent uniform priors (Section 6.4).

tableSnowStorm-perturbed detector parameters and their per-batch uniform priors. The scattering and absorption parameters are global scalings of the bulk-ice scattering and absorption coefficients. Their ranges, with that of the DOM efficiency, are multiplicative factors applied to the nominal best-fit value of each parameter ($[0.9, 1.1] = \pm 10\%$ of nominal). The hole-ice (p_0, p_1) row gives absolute parameter values of the unified angular-acceptance parameterization.

Parameter	Prior	Range
Scattering	uniform	$[0.9, 1.1]$
Absorption	uniform	$[0.9, 1.1]$
DOM Efficiency	uniform	$[0.9, 1.1]$
Unified Hole Ice (p_0, p_1)	uniform	$[-0.1, 0.6] \times [-0.12, 0]$

The hole-ice DOM angular acceptance is the unified parameterization, and the central ice model is SPICE FTP-v3[127] with anisotropy axis, tilt, and depth-dependent dust profile fixed at their best-fit values. SnowStorm in this configuration does not marginalize over ice anisotropy, depth-dependent ice structure, or any other simulation-level uncertainty beyond the five parameters above. It also does not address the particle-physics models entering the neutrino and lepton propagation chain (neutrino-nucleon cross sections, bremsstrahlung, pair production, ionization losses, photonuclear interactions, among others).

[127] Rongen and Chirkin 2021, "Advances in IceCube ice modelling and what to expect from the Upgrade".

To isolate the effect of each parameter, we compare slices of the ensemble taken at the outermost edges of each prior range against the baseline simulation. An edge slice keeps only the events whose drawn value of that parameter falls in the extreme tail of its prior, so it is a small subsample—about 5% of the full-range ensemble, barely enough for stable statistics—but it maximizes the separation from baseline and makes each parameter's effect visible. Figure 6.4 shows the result for the effective area and Figure 6.5 for the pull-corrected angular error, each as the ratio of the edge slice to the baseline, with starting tracks (SLT) above and through-going tracks (TLT) below.

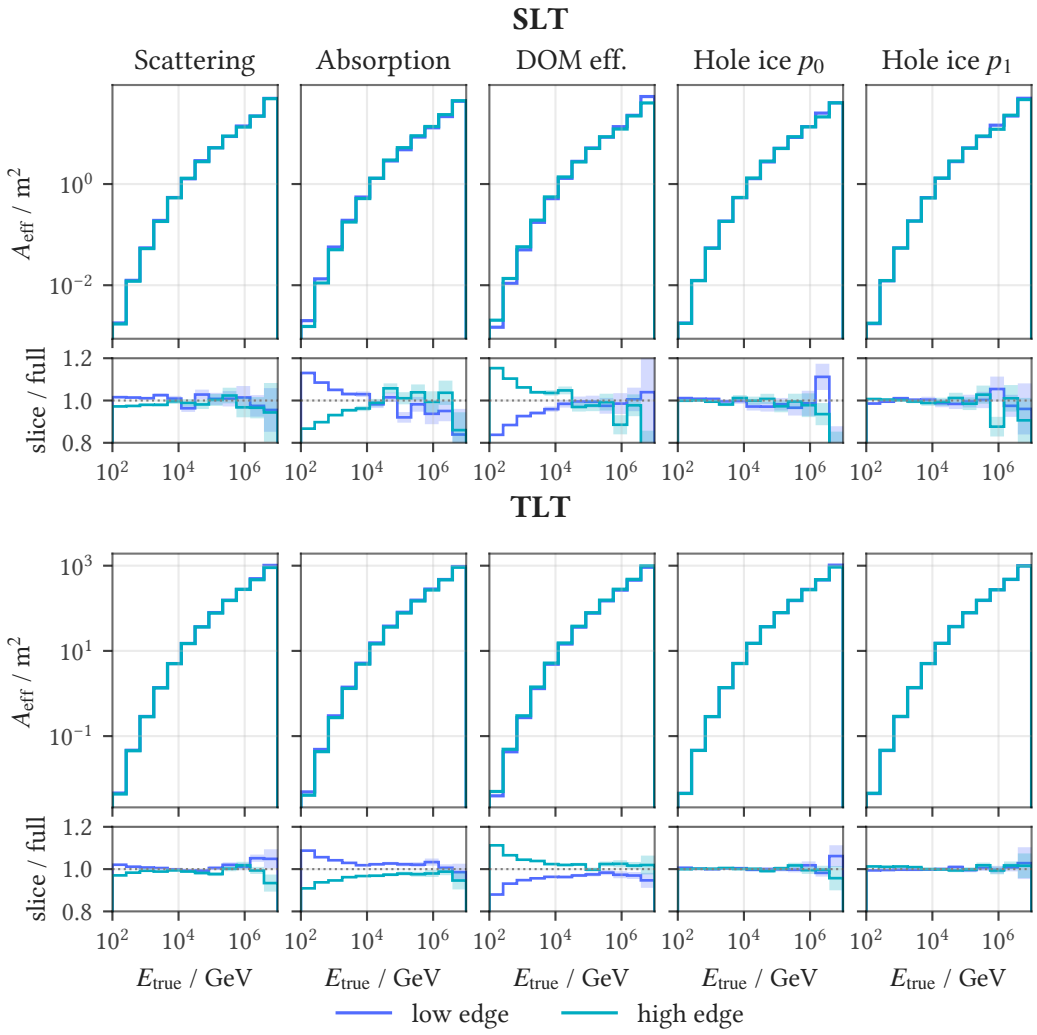

Figure 6.4: SnowStorm systematic slices for the effective area: the low and high edge slice of each of the five continuous parameters (scattering, absorption, DOM efficiency, hole-ice p_0 and p_1), shown as the ratio to the baseline simulation. Starting tracks (SLT) above, through-going tracks (TLT) below.

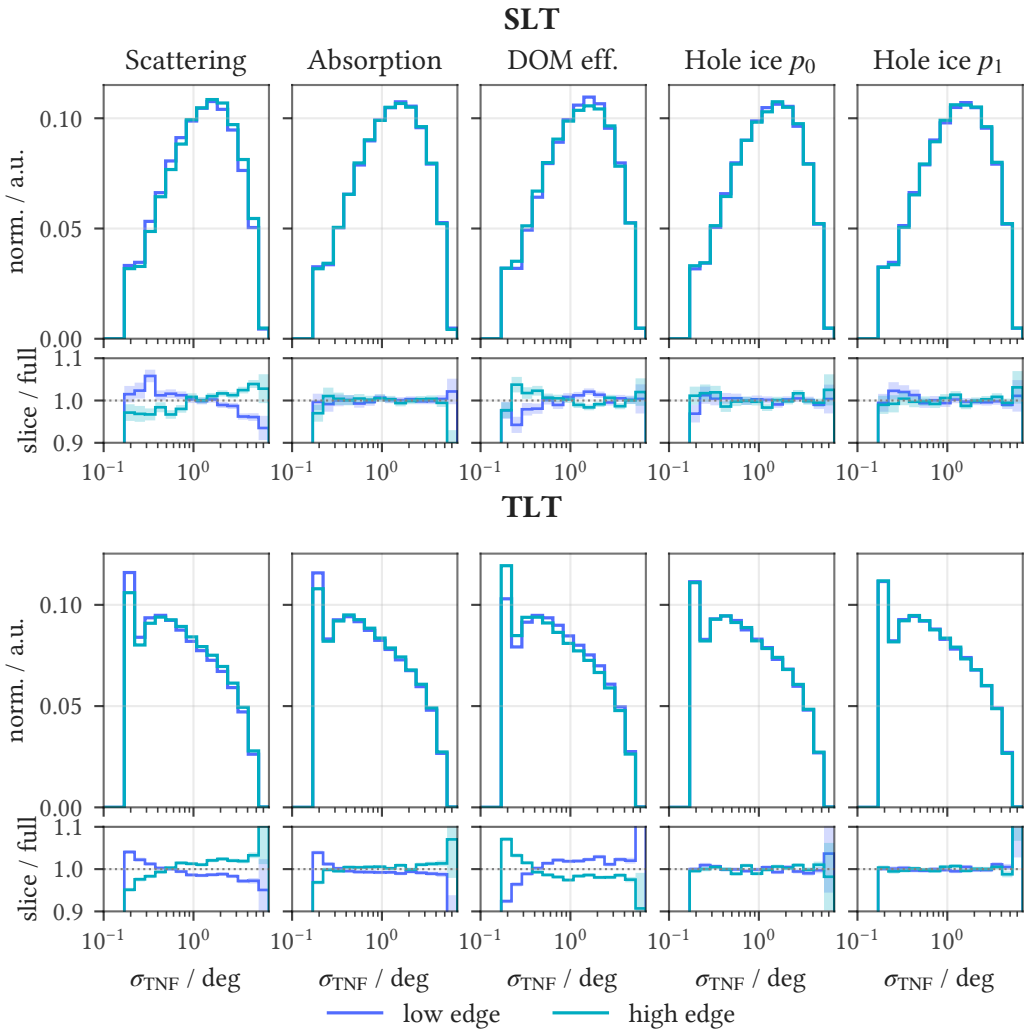

Figure 6.5: SnowStorm systematic slices for the pull-corrected angular error: the low and high edge slice of each of the five continuous parameters, shown as the ratio to the baseline. Panels as in Figure 6.4. Angular error floor of 0.2° applied, causing visible pileup in the lowest angular-error bins.

Across all five parameters the effects on both observables are modest. The two hole-ice parameters, p_0 and p_1, have no measurable effect. Hole ice is the column of refrozen, bubble-rich ice in each drill hole, which scatters light strongly and reshapes the angular acceptance of the DOM it surrounds;[128] over the prior range used here its variation does not measurably change either the effective area or the angular error for these track selections. The DOM efficiency has the largest effect, most pronounced at low energy: it directly scales the amount of detectable light, so it moves both the angular uncertainty (fewer photons give a poorer directional fit) and the effective area (events near threshold cross it or fall below it). Bulk-ice absorption acts the same way and for the same reason, since it too changes the total collected light and so shifts both observables. Bulk-ice scattering, by contrast,

[128] Rongen 2019, "Calibration of the IceCube Neutrino Observatory", Ch. 9.

affects mainly the angular error: it redistributes the arrival times and directions of the detected photons without changing the total light yield, so it degrades the directional reconstruction while leaving the effective area essentially unchanged.

6.5 Example events

We close Part I with a look at three real events from the final sample, one for each of the three track topologies: a starting track (Figure 6.6), an upgoing through-going track (Figure 6.7), and a downgoing through-going track (Figure 6.8).

The figures visualize the first two of the four filter model input features (Section 3.3): total charge (marker size, log-scaled) and first pulse arrival time (color). The remaining two features (charge-weighted mean and standard deviation of pulse times) are not shown. Each DOM is plotted at its physical (x, y, z) position in orthographic projection. To preserve the timing structure of the event, the colormap is anchored to a primary charge window: the shortest contiguous time interval containing 85% of the total deposited charge, found by a sliding window algorithm over the time-sorted hit sequence. DOMs with first pulse times inside this window are rendered with full colormap saturation. DOMs outside the window are progressively desaturated toward neutral gray, remaining visible but clearly distinguished from the primary event timing. This adaptation ensures that the colorbar reflects the physically relevant time scale of each event (compact cascades produce a tight window while extended tracks produce a wider one) without being stretched by late scattered photons in the Pandel tail. The only difference from the CNN's actual input is the coordinate system: the CNN sees the hex grid representation, not the physical geometry.

Since these are static 2D projections of a 3D volume without the ability to rotate the view interactively, the choice of viewing angle matters considerably for interpretability. The viewing azimuth is seeded perpendicular to the RNN-reconstructed track direction in the horizontal plane, choosing the side that places the event's charge-weighted center of gravity closer to the camera. The viewing elevation is set to half the track's vertical inclination, avoiding the extreme angles that a fully perpendicular view would produce for steep tracks. This seed is then adjusted by up to 30° using charge-weighted PCA on the brightest DOMs (top 85% of total charge, with squared charge weights to emphasize the highest-charge deposits): the first principal component identifies the direction of maximum projected spatial spread, and the viewing angle is shifted toward it within the allowed range. Camera roll is held fixed to avoid disorienting tilted views.

To convey depth in the orthographic projection, two additional cues are applied. First, DOM markers and the reconstructed-direction overlay (the track line) are z-ordered by distance to the camera, so that foreground elements occlude background elements. Second, the absolute opacity of each marker is scaled with distance to the camera, causing more distant DOMs to appear progressively fainter. Together, these provide a sense of spatial depth that partially compensates for the absence of interactive rotation.

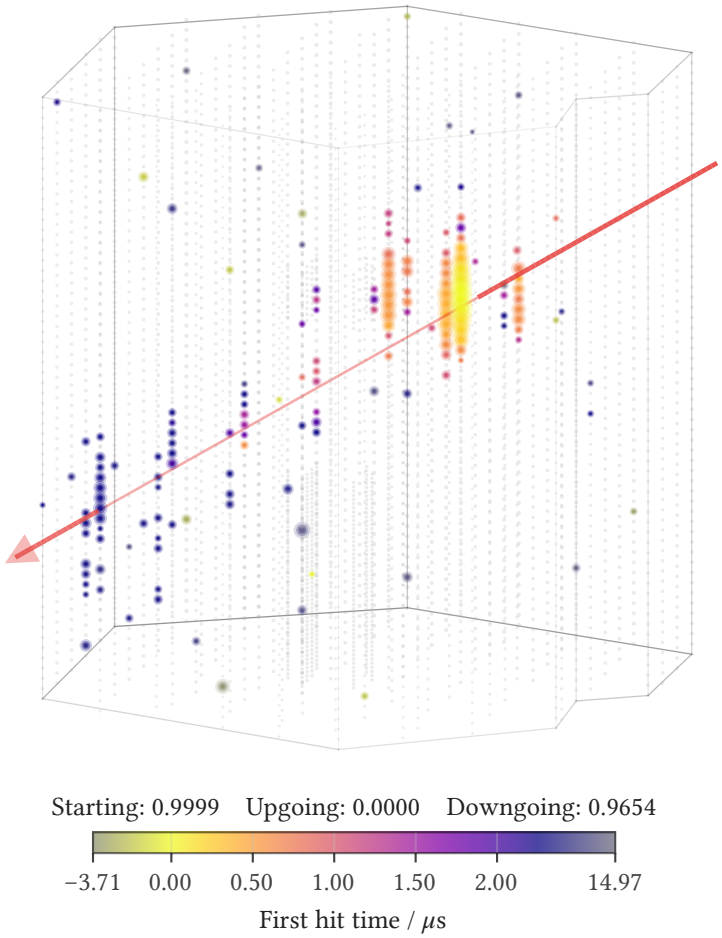

Figure 6.6: A starting-track event from the final sample. This is real experimental data from the burn sample at the final selection level. Each marker is a DOM that recorded light, sized by the logarithm of its collected charge and colored by its first-hit time (colorbar). The footer lists the three filter classifier scores. The overlaid line is the RNN-reconstructed track direction, and the event has a reconstructed energy of about 12 TeV and a reconstructed declination of about 0°, near the horizon. The RNN does not reconstruct the interaction vertex, so the overlay always shows a through-going track regardless of the true event topology.

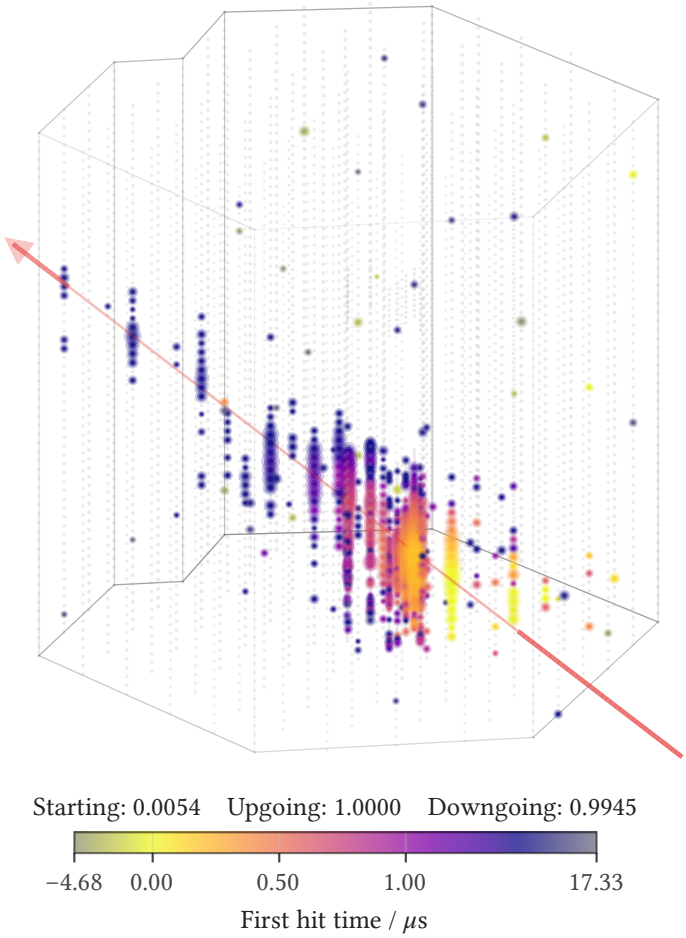

Figure 6.7: An upgoing through-going track from the final sample: real burn-sample data at the final selection level, with a reconstructed energy of about 10 TeV and a reconstructed declination of about +39°. Marker size and color and the footer scores follow Figure 6.6. The overlaid line is the RNN-reconstructed direction.

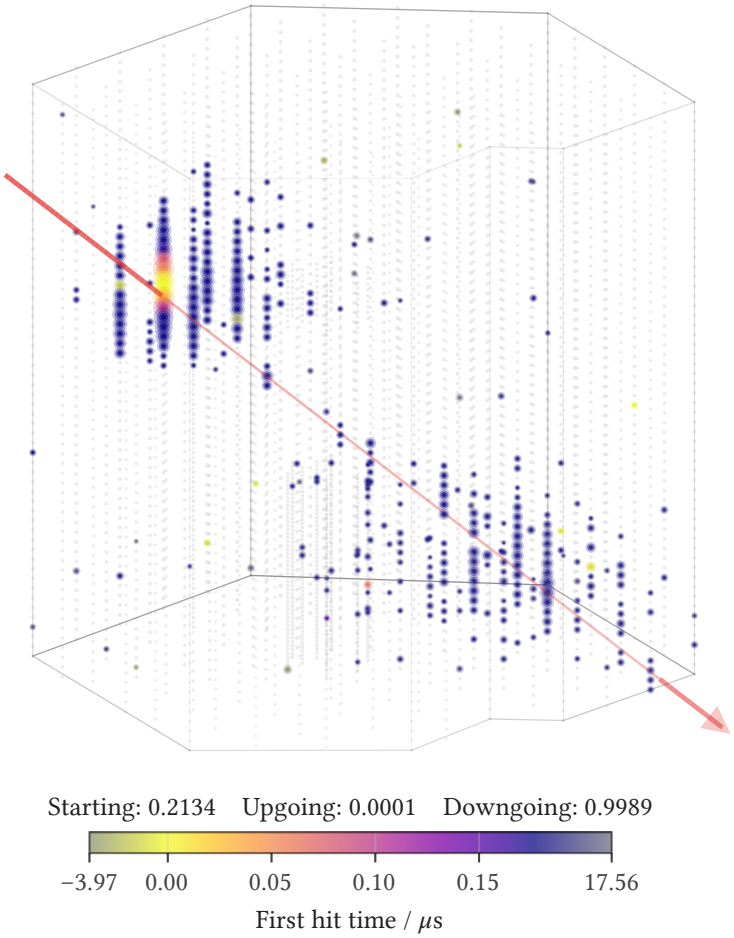

Figure 6.8: A downgoing through-going track from the final sample: real burn-sample data at the final selection level, with a reconstructed energy of about 15 TeV and a reconstructed declination of about −29°. Marker encoding and footer as in Figure 6.6. The overlaid line is the RNN-reconstructed direction.

Part II

The Point-Source Analysis Framework

7 Statistical Foundations

Part II builds the statistical machinery that turns the event sample of Part I into measurements. This chapter collects the general foundations: the frequentist testing framework (Section 7.2), the likelihood-ratio test and its optimality (Section 7.3), maximum-likelihood estimation (Section 7.4), Wilks' theorem and the regularity conditions it rests on (Section 7.5), and the Neyman construction of confidence sets (Section 7.6). None of this material is original, and readers fluent in it can skip ahead. It is collected here because the later chapters lean on it in a specific way: we use these classical results where their assumptions hold, measure where they do not, and replace them with empirical constructions exactly there.

7.1 Notation and parametric families

Throughout Part II, data are modeled as draws from a *parametric family* of probability densities: a collection $\{p(x;\theta) : \theta \in \Theta\}$ of densities over the observation space, indexed by a parameter θ from a parameter space Θ. The semicolon in $p(x;\theta)$ is deliberate. It marks θ as an index into a family of distributions, not as a random variable: in the frequentist framework used throughout this dissertation, parameters are unknown constants, and probability statements are made about the data, never about the parameters. A vertical bar, as in $f(x|y)$, is reserved for a genuine conditional density between random variables. Both notations appear in this dissertation because the analysis genuinely mixes the two situations: the directional reconstruction produces event-conditional posteriors (Section 8.1), while the densities entering the point-source likelihood and the Feldman-Cousins construction are parametric families indexed by signal parameters.

Two pieces of standard shorthand recur. The *likelihood function* is the density of the observed data viewed as a function of the parameter, $\mathcal{L}(\theta) = p(x_{\text{obs}};\theta)$, and for n independent observations it factorizes into a product over events, so its logarithm is a sum. A *statistic* is any function $T(X)$ of the data alone, computable without knowledge of θ.

7.2 Hypothesis testing

A *hypothesis test* asks whether observed data are compatible with a stated hypothesis about the data-generating process. The hypothesis under test is the *null hypothesis* H_0. The *alternative hypothesis* H_1 specifies what is tested against it.[129]

[129] Casella and Berger 2002, *Statistical Inference*, Sec. 8.1, Defs. 8.1.1–8.1.2.

In this dissertation, H_0 is background-only (no astrophysical point source at the tested position), and H_1 adds a source with some signal strength and spectrum.

A test is built from three ingredients. A *test statistic* $T(X)$ compresses the data into a single number, chosen so that larger values indicate stronger evidence against H_0. A *rejection region* declares H_0 rejected when T exceeds a threshold. The *size* (or significance level) α of the test is the probability of rejecting H_0 when it is true, computed under the distribution of T implied by H_0. The complementary performance measure is *power*: the probability of rejecting H_0 when the alternative is true. Fixing α and maximizing power is the design principle behind the optimality result of the next section.

Instead of reporting only the binary decision, an observed value T_{obs} is conventionally summarized by its *p-value*: the probability, under H_0, of obtaining a test statistic at least as extreme as the one observed,

$$p = P(T \geq T_{\text{obs}} \,;\, H_0) = 1 - F_T(T_{\text{obs}}; H_0), \tag{7.1}$$

where $F_T(\cdot\,; H_0)$ is the cumulative distribution function of T under the null hypothesis. A p-value is a statement about the data under H_0. It is not the probability that H_0 is true: in the frequentist framework, hypotheses are not random variables and carry no probabilities. All probability statements in this dissertation are of the form *in repeated experiments with this procedure, outcome A occurs with frequency f*.

One property of the p-value carries much of Part II on its back.

Claim 7.1. If T has a continuous distribution under H_0, then p is uniformly distributed on $[0, 1]$ under H_0, regardless of the choice of T.

Proof. This is the probability integral transform:[130] for a continuous random variable T with CDF F_T, the variable $F_T(T)$ is uniform on $[0, 1]$,[131] and therefore so is $p = 1 - F_T(T)$. □

[130] Rosenblatt 1952, "Remarks on a Multivariate Transformation".

[131] Casella and Berger 2002, Thm. 2.1.10.

Uniformity is what makes p-values from different tests comparable, and it is the property the empirical calibration of Section 10.2 is built to guarantee: there, the null distribution F_T is not assumed but estimated from background ensembles, precisely so that Equation (7.1) holds by construction.

Remark 7.1. Small p-values are conventionally translated into Gaussian $n\sigma$ significances: we quote a p-value as the number of standard deviations n for a one-sided fluctuation of a standard normal, so that 3σ corresponds to $p \approx 1.35 \times 10^{-3}$ and 5σ to $p \approx 2.87 \times 10^{-7}$. The fluctuation is one-sided because we look for an excess. The $n\sigma$ convention is convenient and very common in high-energy physics, and we adopt it because $n\sigma$ is more intuitive than a tiny negative power of ten; the p-values remain the fundamental quantities.

7.3 The likelihood-ratio test

For a given size α, tests differ only in their power, so the natural question is which test statistic extracts the most evidence from the data. For the simplest case the answer is complete. A hypothesis is *simple* if it fixes the data distribution entirely, leaving no free parameters. When both $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ are simple, the optimal statistic is the *likelihood ratio*

$$\Lambda(x) = \frac{p(x;\theta_1)}{p(x;\theta_0)}, \tag{7.2}$$

and the optimality statement is the Neyman–Pearson lemma.[132]

[132] Neyman and Pearson 1933, "On the Problem of the Most Efficient Tests of Statistical Hypotheses", Sec. III(a), p. 300, Casella and Berger 2002, Thm. 8.3.12.

Claim 7.2. Among all tests of H_0 against H_1 with size at most α, the test that rejects H_0 when $\Lambda(x) > k$, with k chosen so that the size equals α, has the largest power.

Proof. Write a test as a function $\phi(x) \in \{0, 1\}$ indicating rejection, and let ϕ be the likelihood-ratio test and ϕ' any competitor with size $\alpha' \le \alpha$. For every x, the integrand

$$[\phi(x) - \phi'(x)]\,[p(x;\theta_1) - k\,p(x;\theta_0)]$$

is non-negative: where $p(x;\theta_1) > k\,p(x;\theta_0)$ the first factor is $1 - \phi'(x) \ge 0$, and where $p(x;\theta_1) < k\,p(x;\theta_0)$ it is $-\phi'(x) \le 0$. Integrating over x gives $(\beta - \beta') - k(\alpha - \alpha') \ge 0$, where β and β' are the two powers. Since $\alpha' \le \alpha$ and $k > 0$, it follows that $\beta \ge \beta'$. □

The lemma needs two qualifications. For discrete data, achieving size exactly α can require randomizing on the boundary $\Lambda = k$. That technicality plays no role here. More importantly, any strictly monotone transformation of Λ defines the same rejection regions and is therefore equally optimal, which is why test statistics are conventionally quoted as $2\ln\Lambda$ without loss.

Real alternatives are rarely simple: the point-source hypothesis of Section 9.1 leaves the signal strength and spectral index free. The standard generalization replaces the fixed-alternative ratio with the *maximum likelihood ratio*

$$\Lambda_{\max}(x) = \frac{\sup_{\theta\in\Theta_1} p(x;\theta)}{p(x;\theta_0)}, \tag{7.3}$$

in which the alternative is represented by its best-fitting member.

Remark 7.2. Casella and Berger define the ratio inversely, as the supremum over the null divided by the supremum over the full parameter space;[133] in the nested point-null case the two conventions are monotone transformations of one another ($2\ln\Lambda_{\max} = -2\ln\lambda$) and define the same tests.

[133] Casella and Berger 2002, Def. 8.2.1.

The Neyman–Pearson optimality does not carry over exactly: against a composite alternative there is in general no uniformly most powerful test, and the

maximum likelihood ratio is the standard pragmatic choice rather than a guaranteed optimum. What does carry over is the qualitative lesson, and we lean on it repeatedly—power is determined by how faithfully the densities in the ratio describe the data. The point-source test statistic (Section 9.1) is exactly of the form Equation (7.3), and the per-event PDFs entering it are the reason the angular-error calibration of Chapter 8 feeds directly into analysis power (Section 8.1).

7.4 Maximum-likelihood estimation

Testing asks whether a hypothesis is compatible with the data. Estimation asks which member of the family describes it best. The *maximum-likelihood estimator* (MLE) is the parameter value that maximizes the likelihood of the observed data,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta), \tag{7.4}$$

equivalently the maximizer of $\ln \mathcal{L}$.[134]

[134] Casella and Berger 2002, Def. 7.2.4.

How good an estimator can be is governed by the *Fisher information*. The *score* is the gradient of the log-likelihood, $\partial_\theta \ln \mathcal{L}(\theta)$. Under regularity conditions that permit differentiating $\int p(x;\theta)\,\mathrm{d}x = 1$ under the integral sign, the score has zero mean at the true parameter, and the Fisher information is its variance,

$$I(\theta) = \mathrm{Var}_\theta\big[\partial_\theta \ln \mathcal{L}(\theta)\big] = -\,\mathrm{E}_\theta\big[\partial_\theta^2 \ln \mathcal{L}(\theta)\big], \tag{7.5}$$

where the second equality is the information equality, valid under one more order of the same differentiability.[135]

[135] Casella and Berger 2002, Lemma 7.3.11.

The information measures how sharply the likelihood distinguishes neighboring parameter values: a peaked likelihood carries much information, a flat one little. The Cramér–Rao bound makes this quantitative: any unbiased estimator $\tilde{\theta}$ satisfies

$$\mathrm{Var}_\theta(\tilde{\theta}) \geq I(\theta)^{-1}, \tag{7.6}$$

so the inverse information is the variance floor no unbiased estimation procedure can beat.[136]

[136] Casella and Berger 2002, Thm. 7.3.9.

The MLE attains this floor asymptotically. For n independent, identically distributed observations and under regularity conditions (an identifiable, correctly specified family, smooth in θ, with the true parameter in the interior of Θ),

$$\sqrt{n}\,(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\big(0,\, I_1(\theta)^{-1}\big), \tag{7.7}$$

where I_1 is the information per observation: the MLE is consistent, asymptotically normal, and asymptotically efficient,[137] with the rigorous version due to van der Vaart.[138] In this asymptotic regime the curvature of the observed log-likelihood at its maximum estimates the variance of $\hat{\theta}$, which is how fit uncertainties are conventionally read off a likelihood surface.

[137] Casella and Berger 2002, Sec. 10.1.2.

[138] Vaart 1998, *Asymptotic Statistics*, Theorem 5.39.

Two warnings attach to Equation (7.7), and both matter later. First, the guarantee is asymptotic: at finite sample size, and especially near the boundary of Θ, the

sampling distribution of the MLE can be far from Gaussian. The signal-strength estimate $\hat{n}_s$ of the point-source likelihood lives at exactly such a boundary ($n_s \geq 0$), and its empirical sampling distributions (Section 12.3) carry a discrete atom at zero. Second, under model misspecification the MLE converges not to the truth but to the pseudo-true parameter of the misspecified family—a point developed in the regularity discussion of Section 12.1. Both are reasons Part II ultimately replaces asymptotic arguments with empirical calibration.

7.5 Wilks' theorem

The maximum likelihood ratio of Equation (7.3) is only useful for inference if its distribution under H_0 is known. Computing that distribution exactly is rarely possible, and the classical resolution is an asymptotic result due to Wilks.[139]

[139] Wilks 1938, "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses", p. 62, Casella and Berger 2002, Thm. 10.3.3.

Claim 7.3. Let the null hypothesis H_0 restrict q of the parameters of a family satisfying the regularity conditions below, with the data consisting of n independent, identically distributed observations. Then, under H_0,

$$2 \ln \Lambda_{\max} \xrightarrow{d} \chi^2_q \qquad (n \to \infty), \tag{7.8}$$

the chi-squared distribution with q degrees of freedom.

The proof idea is a Taylor expansion of the log-likelihood around the MLE: asymptotic normality of the MLE (Equation (7.7)) turns the quadratic term into a sum of q squared standard normal variables[140] (the modern general version is given by van der Vaart[141]). The practical appeal is enormous—the null distribution of the test statistic becomes known, analytic, and independent of the model details, so p-values come from a chi-squared table rather than from simulation.

[140] Casella and Berger 2002, Sec. 10.3.1.

[141] Vaart 1998, Theorem 16.7.

The same asymptotic result inverts to give confidence regions. The limit (7.8) also holds when θ is taken as the truth, so the set of parameter values it does not reject at level α,

$$\{\theta : -2 \log[L(x;\theta)/L(x;\hat{\theta})] \leq \chi^2_q(\alpha)\}, \tag{7.9}$$

is a $1-\alpha$ confidence region centered on the MLE $\hat{\theta}$. This likelihood-ratio confidence region is the standard asymptotic alternative to the Feldman–Cousins construction for parameter estimation, and it inherits the same regularity conditions as the test above.

The conditions doing the work in the background are the same ones behind the MLE asymptotics, and each can fail in practice: the true parameter must be identifiable and the model correctly specified, the log-density must be smooth enough for the quadratic expansion, the true parameter must lie in the interior of the parameter space rather than on a boundary, and every parameter must remain meaningful under H_0 rather than becoming undefined when another parameter is switched off. The point-source likelihood violates several of these at once, most prominently the boundary condition (the background-only hypothesis sits at $n_s =$

0, the edge of the parameter space) and the identifiability condition (with no signal, the spectral index is undefined). The mapping of each condition onto the IceCube likelihood is laid out in Section 12.1, and the measured consequences appear in the background test-statistic distributions of Section 9.8, which deviate from Equation (7.8) in exactly the ways the violated conditions suggest. That failure, observed rather than assumed, is why the analysis calibrates its null distributions empirically instead of relying on Wilks' theorem, and why parameter estimation is built on the Feldman-Cousins construction (Chapter 12) rather than on asymptotic confidence regions.

7.6 The Neyman construction

A point estimate without an uncertainty is not a measurement. The frequentist expression of estimation uncertainty is the *confidence set*: a data-dependent region $C(X)$ of parameter space constructed so that

$$P(\theta \in C(X);\, \theta) \geq 1 - \alpha \qquad \text{for every } \theta \in \Theta. \tag{7.10}$$

This property is *coverage*, and its quantifier deserves emphasis: the probability statement holds for every possible true parameter value, and the randomness lives in $C(X)$, not in θ. In repeated experiments, the constructed region contains the truth in at least a fraction $1 - \alpha$ of them, whatever the truth is.

Neyman's construction[142] produces such sets by inverting a family of hypothesis tests. For every candidate parameter value θ_0, build an *acceptance region* $A(\theta_0)$ in data space containing at least probability $1 - \alpha$ under θ_0. After observing x, collect every parameter value whose acceptance region contains the observation:

$$C(x) = \{\theta_0 \in \Theta : x \in A(\theta_0)\}. \tag{7.11}$$

Coverage then holds by construction, with no asymptotics and no distributional assumptions beyond the ability to compute (or simulate) the data distribution at each θ_0: the truth θ is in $C(X)$ exactly when X falls in $A(\theta)$, which happens with probability at least $1 - \alpha$ by the way $A(\theta)$ was built. We state and prove this duality formally where it carries the analysis (Section 12.6).

The construction leaves one choice completely free: which probability-$(1 - \alpha)$ subset of data space to take as $A(\theta_0)$. That freedom does not affect coverage, but it determines the shape, the size, and the practical usefulness of the resulting confidence regions. Exploiting it well is a design problem, and choosing the acceptance regions by a likelihood-ratio ordering is the contribution of Feldman and Cousins,[143] whose construction, adapted to empirical sampling distributions, is the subject of Chapter 12.

[142] Neyman 1937, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability", Eq. 20.

[143] Feldman and Cousins 1998, "Unified approach to the classical statistical analysis of small signals".

8

Angular Error Calibration

Four reconstruction tools run on every filter-passing candidate. Their pipeline roles, the credits, and the seeding relationships are laid out with the final sample (Section 4.1). What this chapter needs is the directional estimate and its uncertainty, and those come from the *transformer normalizing flow* (TNF),[144] Glüsenkamp's reconstruction: a neural network that, conditioned on the observed pulses, models the full two-dimensional posterior over arrival directions with conditional normalizing flows (Section 8.1 develops this in detail). TNF supplies both the directional fit and the per-event estimate of the angular error that the rest of this chapter calibrates. The TNF paper is, at the time of writing, close to submission, so the citation is a forward reference.

The conventional alternative is SplineReco, known within IceCube as SplineMPE: a purely likelihood-based track reconstruction that maximizes the likelihood of the observed photon arrival times under a track hypothesis, with the arrival-time distributions obtained from high-dimensional splines fitted to simulated photon propagation.[145] Wrede's convolutional-recurrent neural network (CRNN) supplies the track seed and an independent direction estimate for cross-checks, MuEX the energy estimate, and the Starting Track Veto (STV) the containment variable for starting candidates. All of them are introduced, with credits and the empirical moon-shadow verification of the CRNN, in Section 4.1. None of these tools were modified in this work.

Everything the point-source likelihood of Section 9.1 extracts from an event hinges on that event's angular error estimate: the per-event σ that sets its spatial weight. This chapter builds and calibrates that estimate, starting from the directional posterior the reconstruction provides (Section 8.1) and ending with the coverage validation of the calibrated result.

[144] IceCube Collaboration, "Neural posterior estimation of the neutrino direction in IceCube using transformer-encoded normalizing flows on the sphere" (In preparation, 2026).

[145] IceCube Collaboration, "A muon-track reconstruction exploiting stochastic losses for large-scale Cherenkov detectors", *Journal of Instrumentation* 16, no. 08 (2021): P08034, AMANDA Collaboration, "Muon track reconstruction and data selection techniques in AMANDA", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 524, no. 1-3 (2004): 169–194.

8.1 The point-spread function

The signal spatial PDF $\mathcal{S}_{\text{space}}(x_i; x_s, \hat{\sigma}_i)$ is the point-spread function (PSF), which describes the probability density for observing an event at position x_i given a true source at x_s. Throughout this chapter (and only in this chapter) we denote it with the shorthand f. The fidelity of the PSF model directly impacts analysis power. By the Neyman–Pearson lemma (Section 7.3), the likelihood ratio constructed from the true signal and background distributions is the most powerful test statistic. Any deviation from the true PSF reduces statistical power. We start our discussion by returning to event reconstruction.

Remark 8.1. In this chapter the semicolon in $f(\Delta\psi; \sigma)$ marks parameters (a frequentist density family indexed by σ), and the vertical bar in $\tilde{f}_{\text{TNF}}(\varphi, \vartheta|\text{pulses})$ marks a genuine conditional density between random variables (a Bayesian posterior). Both notations appear because the angular-error problem genuinely mixes the two frameworks: the directional reconstruction (TNF) is Bayesian and produces a full event-conditional posterior, while the PSF actually used in the likelihood is a parametric family indexed by a scale parameter.

The full directional posterior

For any event, the most fundamental quantity encoding its directional information is a normalized probability density for the true arrival direction,

$$\hat{f}(\hat{n}_{\text{true}}|\text{event}), \tag{8.1}$$

where $\hat{n}_{\text{true}}$ is a unit vector on the sphere. This function is the *spatial probability density* for the event's true direction: it is the angular factor of the per-event signal PDF S, which factorizes into an angular component and an energy component. Ideally this full density would enter the signal term directly; in practice the likelihood uses the reduced one-dimensional parametric form derived below, so we keep $\hat{f}$, the ideal posterior, distinct from that reduced f throughout.

Since any such direction can be parameterized by two angles, this PDF is equivalently a fully two-dimensional function

$$\tilde{f}(\varphi, \vartheta|\text{event}), \tag{8.2}$$

defined over the sphere. In principle, the ideal point-source analysis would evaluate a model of this entire 2D directional posterior directly within the likelihood ratio.

Glüsenkamp's normalizing-flow reconstruction (TNF)[146] is a novel IceCube algorithm that explicitly approximates this full 2D surface,

$$\tilde{f}_{\text{TNF}}(\varphi, \vartheta|\text{pulses}), \tag{8.3}$$

using conditional normalizing flows. TNF does not merely provide a direction and uncertainty estimate: it produces a full, event-specific posterior over all possible arrival directions on the sphere.

[146] IceCube Collaboration 2026b.

Using $\tilde{f}_{\text{TNF}}$ directly in a point-source likelihood would be statistically ideal.

However, a typical point-source analysis requires millions of likelihood evaluations (e.g., scrambled background trials, iterative maximizations, grid scans, and stacking evaluations). Evaluating a high-dimensional normalizing-flow model for every event–source pair across millions of likelihood calls is computationally infeasible with current methods.

Reduction to a parametric PSF

To make point-source searches tractable, the full 2D posterior is replaced by a parametric PSF,

$$\hat{f}(\hat{n}_{\text{true}}|\text{event}) \longrightarrow f(\Delta\psi; \theta_{\text{event}}), \tag{8.4}$$

where $\Delta\psi = \arccos(\hat{n}_{\text{reco}} \cdot \hat{n}_{\text{true}})$ is the angular separation, and θ_{event} is a small set of event-level parameters.

This step imposes rotational symmetry about the reconstructed direction: having fixed $\hat{n}_{\text{reco}}$ and the event properties, the PSF is assumed to depend on the source location only through $\Delta\psi$. A genuinely two-dimensional directional density on the sphere is thus collapsed into a one-dimensional radial profile, and this reduction is only accurate when the underlying posterior is itself approximately radially symmetric.

However, the detector geometry directly produces cases where this assumption fails. A particularly salient example occurs for steeply up- or down-going tracks: because the detector's 17 m vertical DOM spacing (along strings) is ∼7× finer than its 125 m horizontal string spacing,[147] many such events illuminate only a narrow vertical band of DOMs. The zenith angle can then be tightly constrained while the azimuth remains weakly constrained, producing posteriors that are elongated in azimuth, in the most extreme cases closing into a full ring on the sphere, as the TNF paper demonstrates for an event passing upward close to a single string.[148] A related degeneracy appears for tracks that are close to horizontal: when most of the light is deposited in a single layer of DOMs, the lever arm in the x–y plane is large while the vertical sampling is limited, yielding (in our understanding, though we have not quantified this) zenith uncertainties that are broader than those in azimuth. In both regimes, the posteriors are intrinsically anisotropic and, in some cases, even multimodal. No radially symmetric function of $\Delta\psi$, regardless of how many parameters it carries, can fully reproduce such shapes. Even if one could choose a scale parameter that perfectly matches the radial width of the posterior (e.g. yielding an ideal Rayleigh pull distribution), a one-dimensional PSF $f(\Delta\psi; \theta_{\text{PSF}})$ would still fail to capture the true two-dimensional structure.

This is a fundamental and irreducible approximation introduced by the radial symmetry assumption itself. It is independent of any subsequent calibration: pull correction can improve the radial scaling of the PSF but cannot remove the loss of information caused by compressing an intrinsically 2D directional posterior into a 1D function.

[147] IceCube Collaboration 2017b, "The IceCube Neutrino Observatory: Instrumentation and Online Systems", Sec. 1.1.

[148] IceCube Collaboration 2026b, Fig. 14.

The von Mises–Fisher and Rayleigh PSFs

[149] Fisher 1953, "Dispersion on a Sphere".

The standard choice for parametric PSF modeling is the von Mises–Fisher (vMF) distribution.[149] In general, the vMF family describes probability densities on the surface of a unit $(p-1)$-sphere S^{p-1} embedded in $\mathbb{R}^p$, with the concentration parameter κ controlling how tightly the distribution clusters around a preferred direction. For the case relevant to directional reconstruction in IceCube (that is, directions on the two-dimensional sphere S^2 embedded in $\mathbb{R}^3$), the vMF density, expressed per unit solid angle, reduces to

$$f_{\mathrm{vMF}}(\Delta\psi;\kappa) = \frac{\kappa}{2\pi(e^{\kappa}-e^{-\kappa})}\exp\bigl(\kappa\cos\Delta\psi\bigr), \tag{8.5}$$

where $\Delta\psi$ is the angular separation between the reconstructed and hypothesized source directions.

The distribution of the angular separation $\Delta\psi$ itself follows from the spherical area element: integrating the per-solid-angle density over the azimuth about the reconstructed direction attaches the Jacobian factor $2\pi\sin\Delta\psi$, giving the radial density

$$g_{\mathrm{vMF}}(\Delta\psi;\kappa) = 2\pi\sin\Delta\psi\; f_{\mathrm{vMF}}(\Delta\psi;\kappa) = \frac{\kappa\sin\Delta\psi}{e^{\kappa}-e^{-\kappa}}\exp\bigl(\kappa\cos\Delta\psi\bigr). \tag{8.6}$$

In the small-angle limit ($\sin\Delta\psi\approx\Delta\psi$, $\cos\Delta\psi\approx 1-\frac{1}{2}\Delta\psi^2$, and $e^{\kappa}-e^{-\kappa}\approx e^{\kappa}$ for large κ), this radial density reduces to the Rayleigh form

$$f_{\mathrm{Rayleigh}}(\Delta\psi;\sigma) = \frac{\Delta\psi}{\sigma^2}\exp\left(-\frac{\Delta\psi^2}{2\sigma^2}\right), \tag{8.7}$$

with the scale parameter fixed by

$$\sigma \approx \frac{1}{\sqrt{\kappa}}. \tag{8.8}$$

Thus f_{vMF} is the density on the sphere, per unit solid angle, while f_{Rayleigh} is the corresponding radial density of the angular error in the small-angle regime. The PSF model transitions between the Rayleigh and vMF forms at a configurable angular error estimate threshold `kent_min` (default 7°). Our coverage studies (Section 8.5; Figure 8.1) show that for the track samples, which exhibit sufficiently small angular errors, the Rayleigh and vMF models indeed produce indistinguishable coverage. Only the DNN Cascade sample, with its substantially larger angular-error estimates, exhibits the expected slight deviation between the two descriptions.

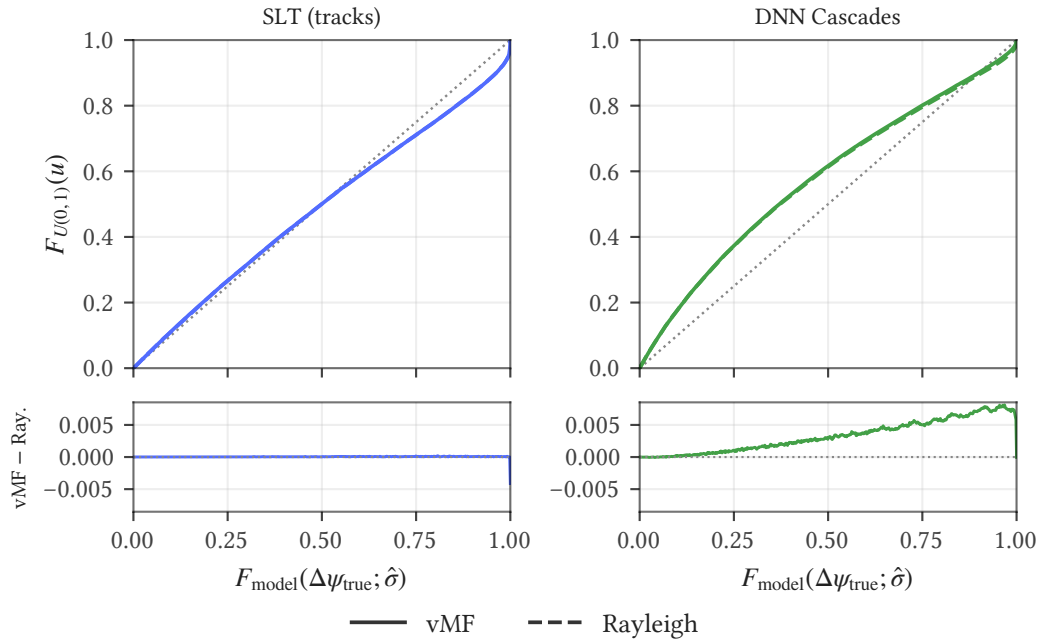

Figure 8.1: Coverage comparison between the Rayleigh and vMF PSF models for SLT and DNNC. The DNNC panel's dominant visual is the overall pre-floor off-diagonal bow, the cascade sample's σ miscalibration as a whole. The deviation between the vMF and Rayleigh models is the small effect on top, isolated in the bottom difference strip. The SLT strip is flat at zero—the two models are indistinguishable, while the DNNC difference grows to $\sim$+0.007, the expected slight deviation.

Intuitively, the vMF distribution is the spherical analogue of a 2D Gaussian: it defines the most symmetric, single-parameter density concentrated around a preferred direction, but crucially it does so on the curved surface of the sphere: it is the maximum-entropy density on the sphere for a fixed mean vector (equivalently, a fixed mean direction and resultant length),[150] formalizing the sense in which it is the most symmetric single-parameter choice. The normalization and shape of the distribution explicitly account for spherical geometry, which is why the vMF PDF remains accurate even when the support of the PSF spans several degrees.

[150] Mardia and Jupp 2000, *Directional Statistics*, Sec. 9.3, p. 172.

The Rayleigh form, by contrast, arises as the radial distribution of a 2D Gaussian on its tangent plane. It ignores curvature entirely, treating small angular displacements as Euclidean distances. This is exactly why it is an excellent approximation in the small-angle regime: for sufficiently small $\Delta\psi$, the sphere is locally flat, the difference between geodesic and Euclidean distances is negligible, and the vMF distribution reduces to a Gaussian whose radial part is Rayleigh. In this limit, the Rayleigh and vMF models become nearly indistinguishable.

Attempts to improve the 1D PSF: KDEs and King functions

The considerations above address the fundamental limitation of using a 1D, radially symmetric PSF: such a model can never reproduce genuinely anisotropic

or multimodal posteriors. However, even within the restricted class of 1D radial profiles, the standard vMF/Rayleigh approach is further constrained by using only a single scale parameter (σ or κ) to characterize the entire shape of the angular-error distribution. This can lead to imperfect coverage, especially at low energies where the true $\Delta\psi$-distribution develops broader tails.

Several analyses have therefore explored more flexible 1D-in-$\Delta\psi$ PSF models that retain radial symmetry but relax the single-parameter assumption. Two such approaches are the kernel density estimate (KDE)-based PSF and the King function.

The KDE-based PSF originates with the IceCube NGC 1068 analysis,[151] which replaced the analytic vMF/Rayleigh spatial term with a non-parametric 1D KDE built for an assumed spectral index γ. The Northern Tracks implementation[152] uses a KDE in the variables $(\Delta\psi, E_{\text{reco}}, \sigma)$. More on why the angular-error distribution must be built for the assumed spectral index γ to remain physically consistent is discussed later in Section 8.3.

[151] IceCube Collaboration 2022a, "Evidence for neutrino emission from the nearby active galaxy NGC 1068".

[152] IceCube Collaboration 2026a, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube", Appendix.

Mathematically, the KDE-based PSF is a kernel density estimate of the conditional distribution of the opening angle,

$$f_{\text{KDE}}(\Delta\psi|E_{\text{reco}}, \sigma; \gamma) = \sum_k w_k \, K_h\big(\Delta\psi - \Delta\psi_k\big), \tag{8.9}$$

with MC-derived weights w_k, bandwidth h, and a 1D kernel K_h (Gaussian in the NT implementation), evaluated in the radial variable $\Delta\psi$.

This KDE is not a 2D angular posterior. It still assumes a radially symmetric PSF and models only the distribution of $\Delta\psi$. The KDE approach improves the shape of the radial PSF (especially in the tails) but does not remove the radial-symmetry assumption.

The *King function* provides a more flexible two-parameter radial model of the form

$$f_{\text{King}}(\Delta\psi; \alpha, \beta) = N(\alpha, \beta) \left(1 + \frac{\Delta\psi^2}{2\beta\,\alpha^2}\right)^{-\beta}, \tag{8.10}$$

where α controls the core width, β governs the tail slope, and N is the normalization constant satisfying $\int f_{\text{King}}(\Delta\psi)\, d\Omega = 1$. This form allows a narrow core together with independently adjustable heavy tails and is used, for example, in Fermi-LAT PSF modeling.[153] King-function PSF support is under active, ongoing development within the collaboration.

[153] Fermi-LAT Collaboration 2012b, "The Fermi Large Area Telescope on Orbit: Event Classification, Instrument Response Functions, and Calibration", Sec. 6.1.1, Eq. (38).

The functional form is isomorphic to a Student-t distribution and is known in optical astronomy as the Moffat profile.[154] The name "King function" is historical, from King's empirical star-cluster surface-density law,[155] applied to PSF modeling by the X-ray community and adopted by Fermi-LAT.[156]

[154] Moffat 1969, "A Theoretical Investigation of Focal Stellar Images in the Photographic Emulsion and Application to Photographic Photometry".

[155] King 1962, "The Structure of Star Clusters. I. An Empirical Density Law".

[156] Fermi-LAT Collaboration 2012b, Sec. 6.1.1, Eq. (38).

Like the KDE approach, the King function remains a purely radial one-dimensional model in $\Delta\psi$. It can improve coverage by providing a more flexible radial profile than a single-parameter Rayleigh or vMF, while retaining the strong practical advantage that it is much cheaper to fit and evaluate than a full KDE. However, it also cannot resolve the fundamental mismatch that arises when the true directional posterior is anisotropic or otherwise not radially symmetric.

Calibration defaults

All plots in this chapter are generated after loading the data into the likelihood, so they reflect exactly what the likelihood sees during analysis. For the Lightning Tracks samples we calibrate the PSF to a 2.5 spectrum, while NT and PST calibrate for $\gamma = 2$. An angular error floor of 0.2° is applied to all samples. This is the default and is used throughout all sensitivity, discovery potential, and TS distribution results. For a more detailed discussion of the PSF calibration, see Section 8.3.

8.2 Pull calibration

There are two conceptually distinct reasons why a pull calibration is required, even when using the best available directional reconstructions. First, any angular-error estimator may itself be imperfect. Second, even an ideal estimator characterizes the observable final state rather than the neutrino, so the calibrated angular-error distribution depends on the assumed spectrum (Section 8.3).

Classical methods such as Paraboloid[157] (as used in PST) underestimate the angular error for bright events, and machine-learning-based regressors (as used in ESTES, DNNC, and NT when not using KDEs) tend to absorb biases from the distributional structure of the MC training sample (e.g. in visible energy, zenith, or containment). Even if they avoid event-level overtraining, these models can systematically misestimate the scale parameter (σ or κ) because the loss function is minimized over the ensemble of training events rather than the posterior of each individual event.

[157] Paraboloid estimates the angular error by fitting a paraboloid to the reconstruction likelihood surface near the best-fit direction and reading off its curvature.

In contrast, TNF derives κ from the event's own reconstructed posterior $\tilde{f}_{\mathrm{TNF}}(\varphi, \vartheta|\mathrm{pulses})$, not from a global regression task. This would ideally make the κ estimate spectrum-agnostic, but in practice it is not: the spectral dependence documented in Section 8.3, together with the TNF coverage studies discussed there, is the evidence. This residual spectrum dependence is part of why pull calibration remains necessary even for TNF. The proposed extension of the signal posterior to include reconstructed energy explicitly (Section 8.1) is expected to improve exactly this.

With one motivation already in hand, we first set out the general calibration procedure, then return to the second reason it is needed (Section 8.3). The goal of the calibration is to make the PSF used in the likelihood consistent with the empirical distribution of angular errors seen in Lightning Tracks, conditional on the reconstructed observables and assumed spectrum. To formalize this, we introduce explicit notation. For each event, let $\hat{n}_{\mathrm{reco}}$ be the reconstructed unit direction and $\hat{n}_{\mathrm{true}}$ the true unit direction (available only in MC). The true angular error is

$$\Delta\psi \equiv \arccos\left(\hat{n}_{\mathrm{reco}} \cdot \hat{n}_{\mathrm{true}}\right) , \tag{8.11}$$

and the raw TNF-derived angular error estimate (after $\kappa \to \sigma$ conversion) is

$$\sigma_{\mathrm{TNF}} . \tag{8.12}$$

For the likelihood, we assume a parametric PSF family, which can be written schematically as

$$\Delta\psi \sim f(\Delta\psi;\, \sigma)\,, \tag{8.13}$$

where f is Rayleigh-like with scale parameter σ for small angles, and for larger angles the same scale is mapped to a vMF concentration parameter $\kappa(\sigma)$ internally. For the purposes of this discussion we take this family to be the true angular-error distribution. In other words, all the angular-error information is summarized by a single scale parameter (Rayleigh σ or equivalently vMF κ), and the likelihood assumes that the distribution of $\Delta\psi$ is fully described by that one parameter.

Given an event with true angular error $\Delta\psi$ and estimated scale σ_{est}, we define the *pull*

$$r \equiv \frac{\Delta\psi}{\sigma_{\text{est}}}\,. \tag{8.14}$$

Under the idealized assumption that the PSF model is correct and σ_{est} equals the true scale parameter σ_{true} for all events in a given population, then in the small-angle Rayleigh approximation

$$\Delta\psi \sim f_{\text{Rayleigh}}(\Delta\psi;\, \sigma_{\text{true}})\,, \tag{8.15}$$

where

$$f_{\text{Rayleigh}}(x;\sigma) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad x \geq 0 \tag{8.16}$$

denotes the Rayleigh PDF, and therefore

$$r = \frac{\Delta\psi}{\sigma_{\text{true}}} \sim f_{\text{Rayleigh}}(r;\, 1)\,, \tag{8.17}$$

i.e. the pull follows a Rayleigh distribution with unit scale.

The Rayleigh($\sigma = 1$) distribution has analytic quantiles. In particular, the median is

$$r_{\text{med}}^{\text{Rayleigh}} = \sqrt{2\ln 2} \approx 1.1774\,. \tag{8.18}$$

Thus, in the ideal case where σ_{est} exactly matches the true PSF scale σ_{true}, the pull distribution at fixed reconstructed observables would satisfy

$$r|(E_{\text{reco}}, \sin\delta_{\text{reco}}) \sim f_{\text{Rayleigh}}(r;\, 1) \qquad \forall\,(E_{\text{reco}}, \sin\delta_{\text{reco}})\,, \tag{8.19}$$

and

$$\text{median}\,[r|(E_{\text{reco}}, \sin\delta_{\text{reco}})] = r_{\text{med}}^{\text{Rayleigh}} \qquad \forall\,(E_{\text{reco}}, \sin\delta_{\text{reco}})\,. \tag{8.20}$$

With the pull and its ideal distribution defined, we turn to measuring and correcting it in reconstructed space. For a given region of reconstructed space R (for example, a cell in $(E_{\text{reco}}, \sin\delta_{\text{reco}})$), we can look at an ensemble of MC events and empirically measure the pull distribution

$$r_R = \frac{\Delta\psi}{\sigma_{\text{TNF}}} \quad \text{for events in } R\,. \tag{8.21}$$

Let the empirical median pull in that region be

$$q_{50}(R) \equiv \text{median}\,[r_R]\ . \tag{8.22}$$

If $q_{50}(R) > r_{\text{med}}^{\text{Rayleigh}}$, then in that region the TNF errors are too small on average (overly optimistic); if $q_{50}(R) < r_{\text{med}}^{\text{Rayleigh}}$, they are too large (overly conservative).

Importantly, overestimation is not safer than underestimation: both represent a mismatch between the assumed PSF and the true angular-error distribution. Either direction of bias makes the likelihood suboptimal in the Neyman–Pearson sense (Section 7.3) and therefore reduces analysis power. The goal is not to err high or low, but to match the true scale.

We now introduce a *multiplicative correction factor* $\alpha(R)$ such that the corrected angular-error estimate

$$\sigma_{\text{corr}}(R) = \sigma_{\text{TNF}} \cdot \alpha(R) \tag{8.23}$$

produces a corrected pull

$$r_{\text{corr}}(R) = \frac{\Delta\psi}{\sigma_{\text{corr}}(R)} = \frac{\Delta\psi}{\sigma_{\text{TNF}}\,\alpha(R)} = \frac{r_R}{\alpha(R)}\,, \tag{8.24}$$

whose median matches the Rayleigh unit-scale median in that region:

$$\text{median}\,[r_{\text{corr}}(R)] = r_{\text{med}}^{\text{Rayleigh}}\,. \tag{8.25}$$

Since $\text{median}(r_R) = q_{50}(R)$, choosing

$$\alpha(R) = \frac{q_{50}(R)}{r_{\text{med}}^{\text{Rayleigh}}} \tag{8.26}$$

implies

$$\text{median}\,[r_{\text{corr}}(R)] = \frac{q_{50}(R)}{\alpha(R)} = \frac{q_{50}(R)}{q_{50}(R)/r_{\text{med}}^{\text{Rayleigh}}} = r_{\text{med}}^{\text{Rayleigh}}\,. \tag{8.27}$$

In words, we measure (on simulation) how far the median pull in each reconstructed region deviates from the Rayleigh unit-scale median, rescale σ_{TNF} by that ratio, and thereby enforce that the corrected pull has the correct median in that region.

In practice, we implement the regions R as discrete bins in the two-dimensional reconstructed-observable space of energy and declination, chosen large enough to provide sufficient MC statistics for a stable median estimate. We then construct the fitted correction field $\alpha(E_{\text{reco}}, \sin\delta_{\text{reco}})$ as a continuous function by interpolating between the bin-center values. This avoids discontinuities in the likelihood: for example, the artificial jump in PSF scale that would occur when probing a source location very close to a bin boundary, causing nearby events to be evaluated with abruptly different scales on either side of that boundary.

8.3 Spectral dependence

The second—and more fundamental—reason why pull calibration is required is that even an ideal angular-error estimator would only be ideal with respect to the observable final-state lepton or hadronic shower, not the underlying neutrino. Because the true neutrino energy, flavor, and interaction channel are intrinsically unobservable, the mapping from neutrino kinematics to the outgoing final state introduces an irreducible, energy- and flavor-dependent angular smearing. As a consequence, the conditional angular-error distribution depends on the assumed spectral index and flavor composition, even in the limit of perfect detector response and perfect reconstruction. The calibrated PSF is therefore unavoidably tied to the spectral assumptions used to construct it.

Any PSF is parameterized by observable quantities, like the reconstructed energy. Because the mapping from true neutrino energy and interaction channel to reconstructed observables is broad and non-bijective, a fixed region of reconstructed space corresponds to a spectrum-dependent mixture of true energies and interaction types. Changing the assumed spectral index therefore changes the underlying mixture of neutrino kinematics populating that region, which in turn changes the distribution of physics-driven scattering angles and inelasticities that set the irreducible neutrino-to-lepton (or hadronic-system) angular smearing. In this sense, a PSF calibration is strictly valid only for the spectral index used to construct it.

Consequently, the effective PSF used in point-source likelihood analyses is the convolution of

1. an irreducible *physics PSF*, describing the distribution of final-state directions relative to the true neutrino direction (set by weak-interaction kinematics and dependent on true energy, flavor, and interaction channel), and

2. a *detector PSF*, describing the distribution of reconstructed directions relative to the true final-state direction (set by detector geometry, light transport, and the reconstruction algorithm).

The detector PSF depends only on the final-state particle and could, in principle, be made spectrum-independent. The physics PSF, however, depends on unobservable neutrino-level quantities and thus on the assumed spectral model and flavor composition used to represent them. As a result, even a perfectly operating detector and a perfectly specified reconstruction cannot produce a spectrum- and flavor-universal angular-error model at the level required by the likelihood. Conditioning the PSF on the assumed spectrum and flavor mix is therefore necessary to enforce correct conditional coverage in reconstructed space.

Ideally, the point-source likelihood would therefore adopt a γ-dependent spatial term, using the same spectral index that appears in the signal hypothesis. If γ is treated as a free parameter, a self-consistent treatment would require either a family of pull-correction surfaces or a smooth interpolation of the correction factor as a function of γ.

As discussed earlier, the KDE-based PSF approach for Northern Tracks effectively takes this route: it constructs γ-conditioned KDEs for the angular-error distribution, thereby folding the neutrino-physics contribution into the spatial term in a way that is explicitly dependent on the assumed spectral index. This achieves spectral consistency at the cost of increased model and computational complexity.

The standard IceCube framework does not implement these more elaborate treatments for the default vMF PSF. Instead, a pragmatic compromise is adopted here: the pull correction is derived from MC events reweighted to an effective spectral index $\gamma = 2.5$, chosen as an intermediate between $\gamma = 2$ (high-energy samples) and $\gamma = 3$ (low-energy samples). This choice does not attempt to match every possible spectral index; it simply provides a stable and representative calibration point.

These considerations become more pronounced when the assumed signal spectrum differs significantly from a simple power law. Broken power laws, exponential cutoffs, and more structured spectra induce stronger variations in the true-energy mixture within each reconstructed region than modest changes in spectral index. In such cases, a pull correction tuned to a single effective spectrum may introduce more noticeable spectrum-induced mismodeling of the PSF and therefore a greater loss of analysis power. Quantifying these effects requires targeted injection studies tailored to the specific spectral hypotheses under consideration.

TNF is trained on events weighted to a flat distribution in the logarithm of deposited energy.[158] In the ideal limit TNF would provide a spectrally neutral estimate of the detector-only angular response. However, even a perfect detector-level posterior does *not* remove the need for pull calibration: the default point-source likelihood still uses a single PSF, and this PSF must implicitly fold in the additional, irreducible broadening from neutrino–lepton kinematics. That physics-induced component depends on the assumed spectral index and flavor composition and therefore lies outside what TNF can approximate from the detector response alone. As a result, pull correction remains necessary to obtain the correct conditional coverage for Lightning Tracks events, even when TNF provides a good estimate of the underlying event-level posterior. The TNF coverage studies show good agreement overall, better than the B-spline baseline, with slight undercoverage above a few hundred TeV.[159]

[158] IceCube Collaboration 2026b, p. 17.

[159] IceCube Collaboration 2026b, Fig. 9.

TNF reconstructs the entire 2D likelihood surface $\tilde{f}_{\mathrm{TNF}}(\varphi, \vartheta|\mathrm{pulses})$, but, at least for now, the point-source likelihood requires a 1D radially symmetric PSF. For that purpose TNF also provides an effective vMF concentration κ_{TNF}, obtained by sampling the full estimated posterior $\tilde{f}_{\mathrm{TNF}}(\varphi, \vartheta|\mathrm{pulses})$ and finding the best-fitting vMF for those samples on an event-by-event basis. The quantity

$$\sigma_{\mathrm{TNF}} = \frac{1}{\sqrt{\kappa_{\mathrm{TNF}}}} \tag{8.28}$$

becomes the input to the pull-correction procedure.

8.4 The two-dimensional pull correction

The correction is constructed in the space of reconstructed observables

$$(E_{\mathrm{reco}}, \sin\delta_{\mathrm{reco}}), \tag{8.29}$$

where E_{reco} is the MuEX energy proxy and δ_{reco} is the reconstructed declination. We work in $\sin\delta_{\mathrm{reco}}$ rather than δ_{reco} (or equivalently $\vartheta_{\mathrm{reco}}$) itself to obtain approximately uniform solid-angle (area) sampling in sky coordinates.

All calibration is performed on NuGen events reweighted to an effective $E^{-2.5}$ spectrum, as discussed in Section 8.3. In practice, this is implemented by accumulating event weights instead of counts during the pull quantile computation. The resulting median-pull sample points $q_{50}(R_{ij})$ are therefore strictly matched to this assumed spectrum.

Starting and through-going tracks are treated separately throughout the procedure, producing two independent calibration fields and two independent correction factors. By not forcing a single calibration surface to absorb both contained-vertex and entering-track regimes simultaneously, the pull model has far less variability to marginalize over (e.g., differences in neutrino interaction kinematics and energy resolution). The benefit is asymmetric: through-going tracks dominate the calibration sample (82% of events), so a merged calibration surface essentially describes the through-going population. Merging therefore does not hurt the through-going topology, but the starting tracks are drowned out: they cannot meaningfully contribute, and their calibration quality suffers. The same asymmetry recurs at the likelihood level among all samples of the analysis (see the samples discussion, Section 9.1).

A direct comparison (refitting the calibration with both topologies pooled and evaluating each topology's coverage under the merged surface) bears this out. Under the merged calibration, starting-track uncertainties come out roughly 10% too small across the northern sky, costing the starting topology about 4 percentage points of median coverage (the weighted fraction of events contained within the PSF-predicted median radius drops from 0.499 to 0.463), concentrated in the starting sample's statistical core, $\log_{10} E_{\mathrm{reco}} \approx 2.5$–$4.0$. Through-going coverage is equivalent at the sub-percent level everywhere it has events. South of $\sin\delta_{\mathrm{reco}} \approx -0.3$, where starting tracks are the only population, the merged and per-topology calibrations coincide, independently confirming that the northern degradation reflects the through-going-dominated pool. Above $\log_{10} E_{\mathrm{reco}} \approx 6.5$ the merged calibration is mildly better for starting tracks, plausibly because the abundant through-going statistics stabilize the surface in a region where starting events are scarce. The weight of this region in the starting sample is negligible.

Step 1: Robust adaptive binning in energy and declination

We first define a set of *declination bands* in $\sin\delta_{\mathrm{reco}}$ via fixed edges $\{\sin\delta_i\}_{i=0}^{N_{\mathrm{bands}}}$, and within each band construct an *adaptive binning in* $\log_{10} E_{\mathrm{reco}}$. For a given

declination band i (with $\sin\delta_i \le \sin\delta_{\text{reco}} < \sin\delta_{i+1}$), we introduce log-energy bin edges $\{\log_{10} E_{i,j}\}_{j=0}^{N_{\text{cells},i}}$, and define the corresponding cells (regions) $R_{i,j}$ as

$$R_{i,j} = \Big\{(\log_{10} E_{\text{reco}}, \sin\delta_{\text{reco}}) \,\Big|\, \begin{array}{l} \sin\delta_i \le \sin\delta_{\text{reco}} < \sin\delta_{i+1}, \\ \log_{10} E_{i,j} \le \log_{10} E_{\text{reco}} < \log_{10} E_{i,j+1} \end{array}\Big\}. \tag{8.30}$$

Let the number of events from the chosen calibration sample that fall inside $R_{i,j}$ be $N_{i,j}$. The adaptive-binning algorithm enforces two constraints for every cell $R_{i,j}$ inside declination band i. The first is a minimum number of events per cell,

$$N_{i,j} \;\ge\; N_{\min} \qquad\qquad \forall i, j\,, \tag{8.31}$$

and the second is a minimum bin width in $\log_{10} E$,

$$\Delta \log_{10} E_{i,j} \equiv \log_{10} E_{i,j+1} - \log_{10} E_{i,j} \;\ge\; \Delta^{\min}_{\log E} \qquad \forall\, i, j\,. \tag{8.32}$$

Within each declination band, the algorithm proceeds as follows. It selects all events whose reconstructed $\sin\delta_{\text{reco}}$ falls in the band, sorts their $\log_{10} E_{\text{reco}}$ values, and sweeps from low to high energy, growing each bin until both the minimum-count and minimum-width thresholds are satisfied. If the final high-energy bin fails the thresholds, it is iteratively merged backward with its neighbor until the constraints are met.

This produces non-uniform energy binning for each declination band, with fine granularity where statistics are abundant and coarse bins where statistics are sparse. Only raw event counts are used for binning; MC weights enter only at the later quantile stage.

Step 2: Weighted per-cell pull quantiles

For each cell $R_{i,j}$ we compute the *weighted median pull*. Let $I_{i,j}$ denote the set of event indices whose reconstructed observables fall inside $R_{i,j}$, and let $\{r_k\}_{k\in I_{i,j}}$ be the corresponding pull values with weights $\{w_k\}_{k\in I_{i,j}}$. The weighted median pull in cell (i, j), which we denote

$$q_{i,j} \;\equiv\; q_{50}(R_{i,j})\,, \tag{8.33}$$

is defined as the value r^* satisfying

$$\sum_{k\in I_{i,j}:\, r_k \le r^*} w_k \;\ge\; \frac{1}{2}\sum_{k\in I_{i,j}} w_k \qquad \text{and} \qquad \sum_{k\in I_{i,j}:\, r_k < r^*} w_k \;\le\; \frac{1}{2}\sum_{k\in I_{i,j}} w_k\,. \tag{8.34}$$

From all valid cells $R_{i,j}$ we obtain a corresponding set of scattered control points

$$\{\,(\log_{10} E^{\text{ctr}}_{i,j},\ \sin\delta^{\text{ctr}}_i,\ q_{i,j})\,\}, \tag{8.35}$$

where $(\log_{10} E^{\text{ctr}}_{i,j},\ \sin\delta^{\text{ctr}}_i)$ denotes the geometric center of region $R_{i,j}$. Each point represents the empirical weighted median pull in a well-populated region of reconstructed space.

Step 3: Smooth 2D surface fit via RBF thin-plate spline

To obtain a continuous median-pull field $\tilde{q}_{50}(E_{\text{reco}}, \sin\delta_{\text{reco}})$, we fit a smooth 2D surface to the scattered control points

$$\{(\log_{10} E^{\text{ctr}}_{i,j},\ \sin\delta^{\text{ctr}}_{i},\ q_{i,j})\}. \tag{8.36}$$

Let each control point be denoted $x_k = (x_{1,k}, x_{2,k}) = (\log_{10} E^{\text{ctr}}_{i,j},\ \sin\delta^{\text{ctr}}_{i})$, and let $q_k = q_{i,j}$ be the corresponding median pull, with $k = 1, \dots, M$.

The radial-basis-function (RBF) thin-plate spline[160] interpolant has the form

$$\tilde{q}_{50}(x) = \sum_{k=1}^{M} c_k\, \phi(\|x - x_k\|) + a_0 + a_1 x_1 + a_2 x_2, \tag{8.37}$$

[160] Duchon 1977, "Splines Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces".

where $\phi(r)$ is the *thin-plate spline kernel*

$$\phi(r) = r^2 \log r, \tag{8.38}$$

the c_k are the RBF coefficients, the triple (a_0, a_1, a_2) gives the polynomial tail required for well-posedness in 2D, and $x = (x_1, x_2)$ is any query point in reconstructed-observable space.

The coefficients are obtained by solving the block linear system

$$\begin{bmatrix} K + \lambda I & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} c \\ a \end{bmatrix} = \begin{bmatrix} q \\ 0 \end{bmatrix}, \tag{8.39}$$

where K is the $M \times M$ kernel matrix

$$K_{jk} = \phi(\|x_j - x_k\|), \tag{8.40}$$

and P is the $M \times 3$ polynomial matrix with entries

$$P_{k0} = 1, \qquad P_{k1} = x_{1,k}, \qquad P_{k2} = x_{2,k}, \qquad k = 1, \dots, M, \tag{8.41}$$

so that each row k is $(1, x_{1,k}, x_{2,k})$. The solution vectors are $c = (c_1, \dots, c_M)^T$ and $a = (a_0, a_1, a_2)^T$, the right-hand side is $q = (q_1, \dots, q_M)^T$, and λ is a non-zero smoothing parameter that regularizes the fit.

Intuitively, the thin-plate spline can be viewed as the smoothest possible surface that passes near the scattered control points while respecting a global notion of curvature. The RBF kernel $\phi(r) = r^2 \log r$ penalizes rapid changes in second derivatives, so the fitted surface bends only as much as needed to accommodate the data. Each control point contributes a radially symmetric "influence field" whose strength is determined by the solved coefficients c_k, and the polynomial tail accounts for the overall linear trend that cannot be represented by radial functions alone. The smoothing parameter λ controls the trade-off between fidelity to the control points and global smoothness: smaller values force the surface to interpolate the median pulls exactly, while larger values allow the fit to ignore local noise and produce a more regular, well-behaved correction field. In effect, the procedure

constructs a smoothly varying approximation of the per-cell median pulls that avoids discontinuities and captures only the large-scale structure supported by the statistics of the calibration sample.

This procedure could, in principle, be generalized to include γ as an additional variable. If the number of control points M is chosen wisely (that is, sufficiently small), the RBF evaluation might be efficient enough to be performed directly in the likelihood.

The fitted surfaces can be visualized either as cell plots in $(E_{\text{reco}}, \sin\delta_{\text{reco}})$, showing the observed cell-wise q_{50} values with overlaid spline contours, or as 1D slices at fixed declination bands, showing the RBF prediction as a function of energy with the underlying cell medians as points.

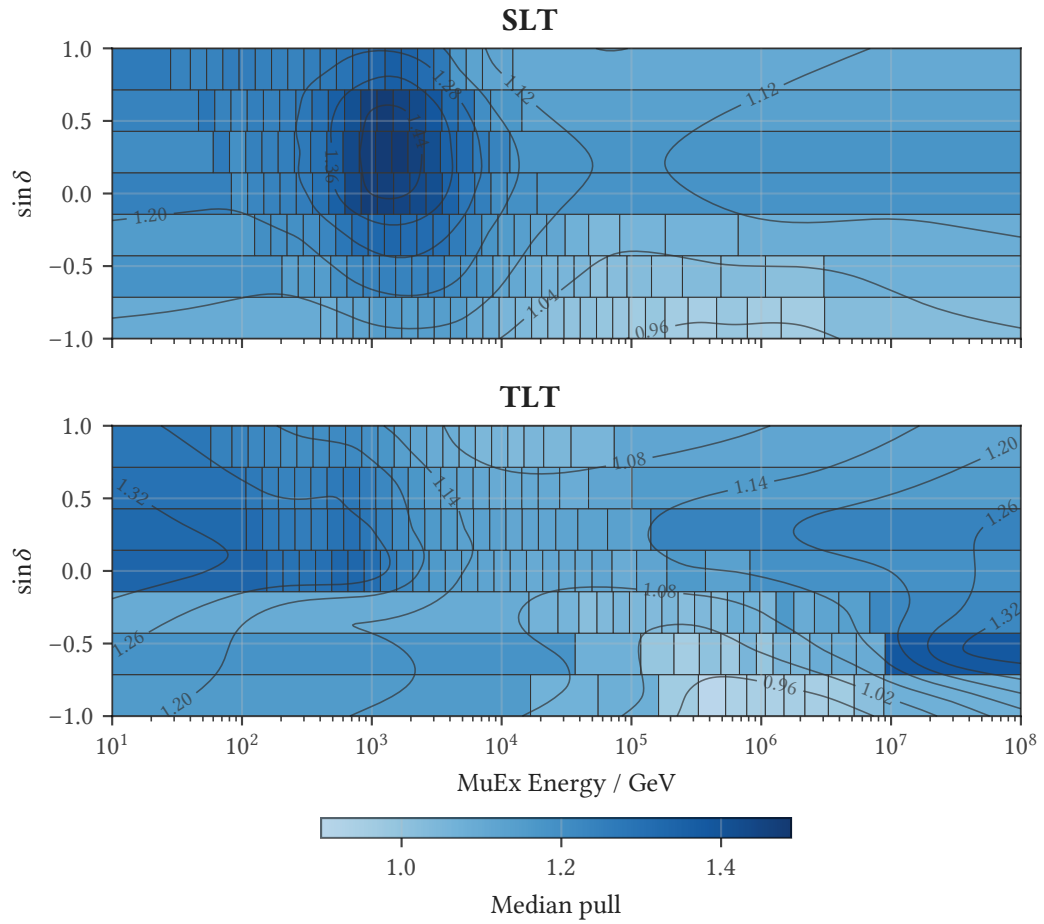

Figure 8.2: Median pull values in adaptive 2D bins, for starting tracks (SLT, top) and through-going tracks (TLT, bottom). Contours show the fitted RBF thin-plate spline surface.

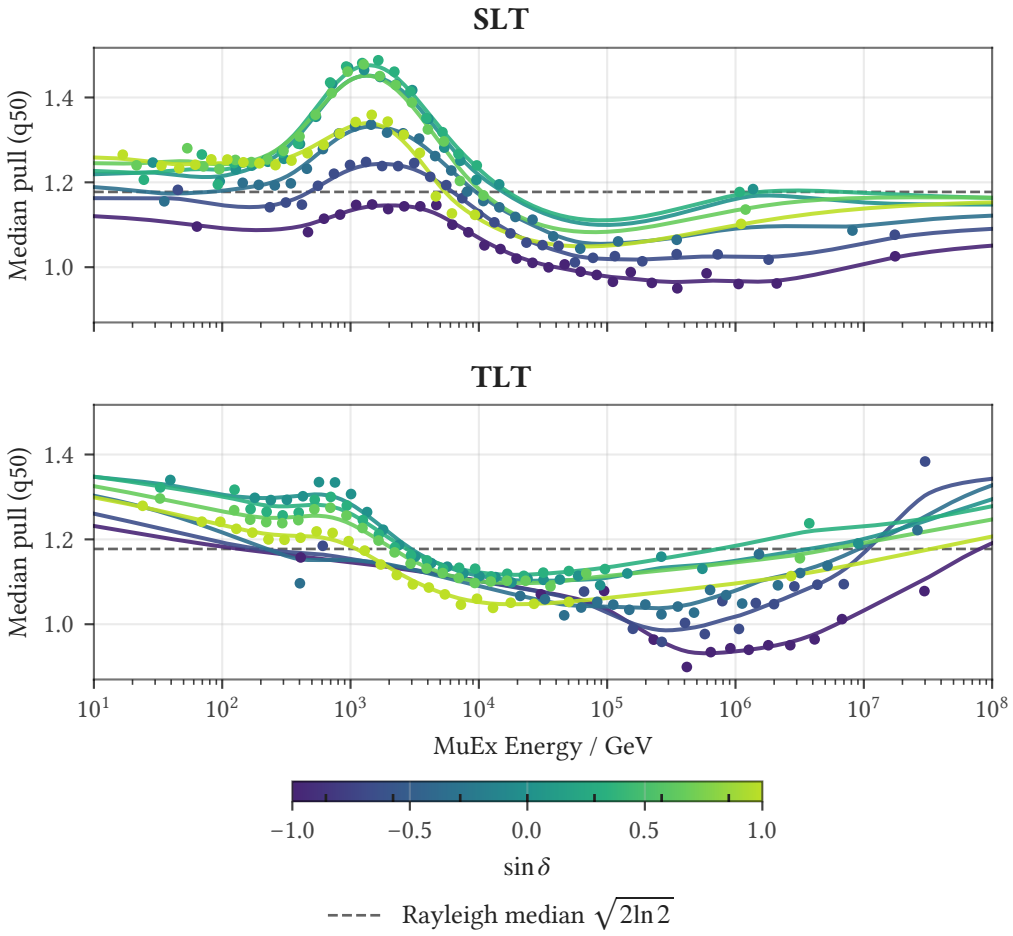

Figure 8.3: RBF surface slices at $\sin\delta$ band centers, for starting tracks (SLT, top) and through-going tracks (TLT, bottom). Tick marks on the colorbar indicate evaluation points. Scatter points show the underlying cell medians. The dashed horizontal line marks the Rayleigh median $\sqrt{2\ln 2}$.

These diagnostic plots provide a direct check that the fit is smooth and consistent with the underlying per-cell medians.

Step 4: Applying the correction

Once the median-pull surface has been constructed, the final correction factor is defined as

$$\alpha(E_{\text{reco}}, \sin\delta_{\text{reco}}) = \frac{\tilde{q}_{50}(E_{\text{reco}}, \sin\delta_{\text{reco}})}{r_{\text{med}}^{\text{Rayleigh}}}. \tag{8.42}$$

For each event, both in simulation and in data, we evaluate

$$\sigma_{\text{corr}} = \sigma_{\text{TNF}} \cdot \alpha(E_{\text{reco}}, \sin\delta_{\text{reco}}), \tag{8.43}$$

and finally use σ_{corr} as the PSF scale parameter in the point-source likelihood. Intuitively, regions where the uncorrected median pull is too large (overly optimistic

σ_{TNF}) receive a correction factor $\alpha > 1$, inflating the errors, while regions where the median pull is too small (overly conservative σ_{TNF}) receive $\alpha < 1$, shrinking the errors.

By construction, this drives the median of the corrected pull distribution towards the Rayleigh unit-scale value in each region of reconstructed phase space, while leaving the detailed shape of the pull distribution (for example tail behavior) as determined by TNF.

8.5 Coverage

A coverage plot provides a visual diagnostic of whether a probabilistic model (here, the PSF) correctly describes the distribution of the quantity it models. The construction proceeds as follows. For each MC event, we compute the CDF of the assumed PSF evaluated at the true angular error:

$$p_i = F_{\text{PSF}}(\Delta\psi_i; \hat{\sigma}_i), \tag{8.44}$$

where F_{PSF} is the cumulative distribution function of the PSF model (Rayleigh or von Mises–Fisher) with estimated scale parameter $\hat{\sigma}_i$, and $\Delta\psi_i$ is the true angular separation between the reconstructed and true directions. If the PSF model is correctly specified, that is, if $\Delta\psi_i$ is indeed drawn from the distribution $F_{\text{PSF}}(\,\cdot\,; \hat{\sigma}_i)$, then by the probability integral transform,[161] these PIT values should be uniformly distributed: $p_i \sim \text{Uniform}(0, 1)$.

[161] Rosenblatt 1952, "Remarks on a Multivariate Transformation", Casella and Berger 2002, *Statistical Inference*.

To visualize this, we bin events by some observable (e.g., reconstructed energy) and, within each bin, compute the empirical CDF of the p_i values. If the model is well-calibrated, the empirical CDF should follow the diagonal $F_{\text{empirical}}(p) = p$. Deviations from the diagonal indicate miscalibration: curves lying above the diagonal indicate overcoverage (the PSF is too wide; $\hat{\sigma}$ is overestimated), while curves below the diagonal indicate undercoverage (the PSF is too narrow; $\hat{\sigma}$ is underestimated). The magnitude of the vertical deviation at any point p gives the fraction of events for which the model assigns a CDF value less than p when the correct fraction should be exactly p.

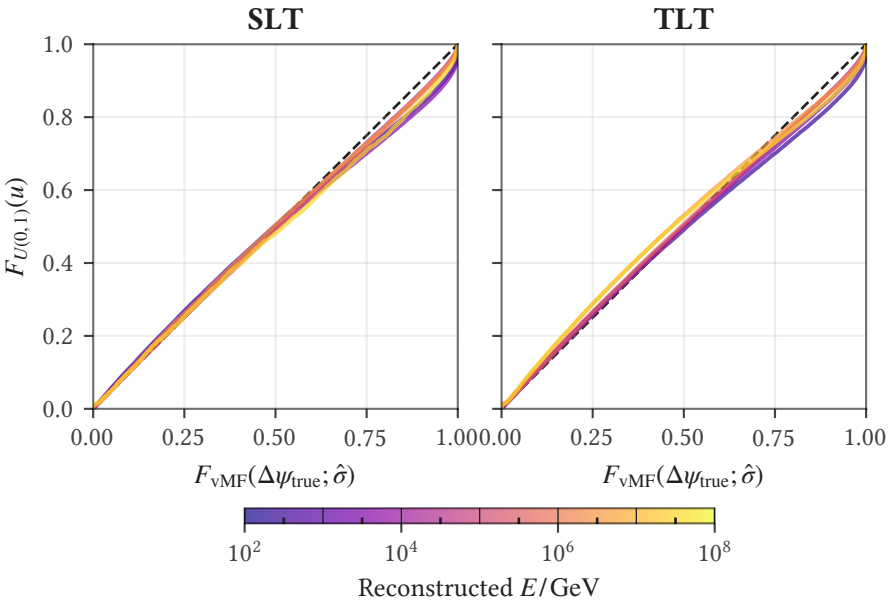

Figure 8.4: PSF coverage: the empirical CDF of the probability-integral-transform values p_i, binned in reconstructed energy, for starting tracks (SLT, left) and through-going tracks (TLT, right). A well-calibrated PSF follows the diagonal.

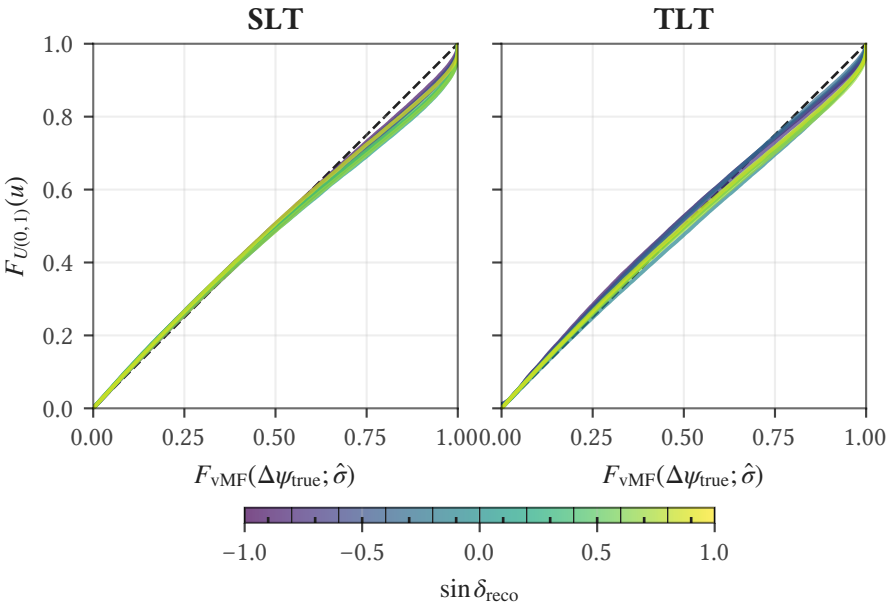

Figure 8.5: PSF coverage binned in reconstructed declination $\sin\delta_{\mathrm{reco}}$, for starting tracks (SLT, left) and through-going tracks (TLT, right). As with the energy binning, a well-calibrated PSF follows the diagonal.

The PSF scale estimator $\hat{\sigma} = \sigma_{\mathrm{TNF}}$ used in this selection is calibrated to achieve accurate coverage conditional on the reconstructed observables, namely the re-

constructed energy proxy and reconstructed zenith (transformed to $\sin\delta$). The nominal estimator is subjected to a two-dimensional pull correction in the space $(E_{\text{reco}}, \sin\delta_{\text{reco}})$, as described in Section 8.4. This correction enforces that, everywhere on this reconstructed-observable plane, the median pull approaches the desired target value. As a consequence, the point-spread function (PSF), modeled by a von Mises–Fisher distribution (or Rayleigh distribution in the small-angle regime), is explicitly calibrated in the space that the likelihood function evaluates, ensuring internal self-consistency.

The success of this procedure is illustrated in Figure 8.4, which shows that the PSF is well-calibrated across reconstructed energy. The same check binned in reconstructed declination is equally well-calibrated (Figure 8.5).

Figure 8.6 repeats the reconstructed-energy coverage of Figure 8.4 with the 0.2° angular-error floor applied. The floor inflates the highest-energy bins into visible overcoverage, because it raises the error estimates that the calibration would otherwise leave below 0.2°. This is a deliberate trade-off. The affected bins sit at high energy, where the angular errors are already small and the spatial weights already high, so widening them slightly changes little. In return, the floor protects against overconfidence: it keeps the likelihood from assigning an error so small, and a spatial weight so large, that a modest data-simulation discrepancy in the angular-error model (Chapter 6) would produce a strongly mis-weighted event. Guarding against such overconfident weights, particularly for lower-energy events, protects the downstream analysis estimates from a bias that no later step could recover.

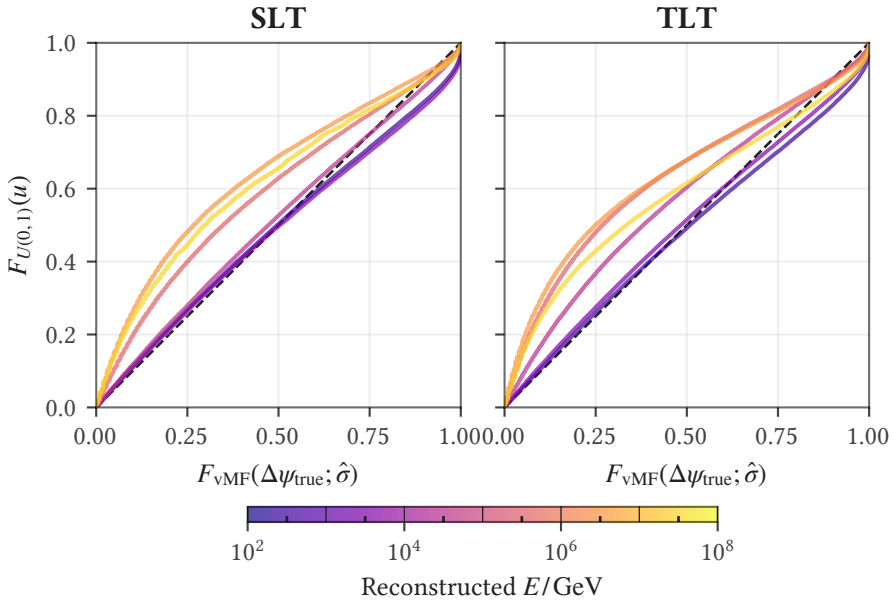

Figure 8.6: Reconstructed-energy PSF coverage with the 0.2° angular-error floor applied, for starting tracks (SLT, left) and through-going tracks (TLT, right). Compared with Figure 8.4, the floor inflates the highest-energy bins into overcoverage.

The post-floor coverage is also shown binned in reconstructed declination (Figure 8.7).

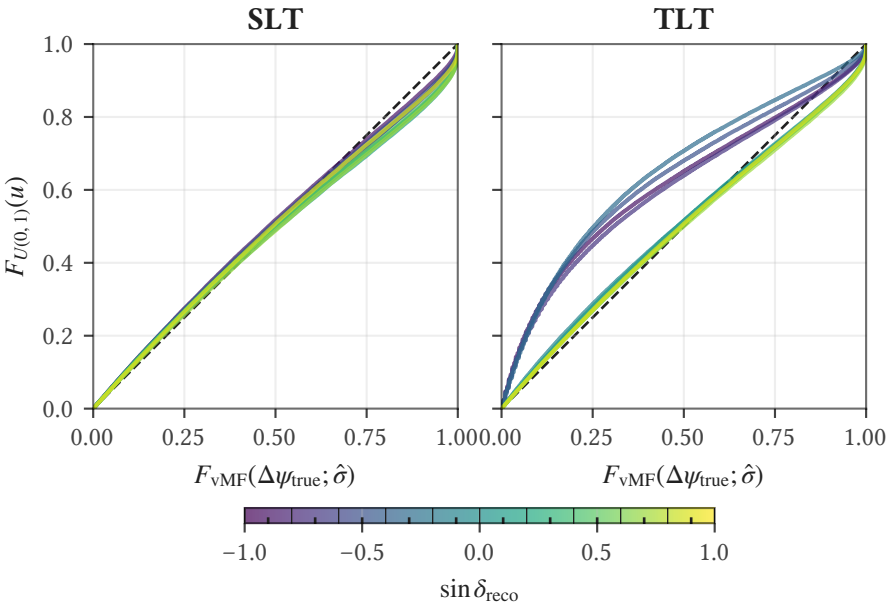

Figure 8.7: Reconstructed-declination PSF coverage with the 0.2° angular-error floor applied, for starting tracks (SLT, left) and through-going tracks (TLT, right). The floor's overcoverage is concentrated in the southern through-going bins.

Limits of reconstructed-space calibration

However, when the resulting angular-error model is evaluated as a function of true neutrino energy, the coverage degrades noticeably, as can be seen in Figure 8.8. This behavior is expected and does not indicate a flaw in the calibration procedure. The pull correction is performed solely with respect to reconstructed observables, since the true neutrino energy is not available for data and therefore cannot be used as a conditioning variable in the angular error model. In an idealized scenario where the mapping between true and reconstructed energy were one-to-one (i.e., bijective over the physically populated domain), conditioning on E_ν^{true} or on E_{reco} would be equivalent. In that limit, a PSF calibrated in reconstructed space would automatically exhibit identical coverage as a function of true energy, $f(\Delta\psi|E_\nu^{\mathrm{true}}) = f(\Delta\psi|E_{\mathrm{reco}})$.

In reality, the mapping between true and reconstructed energy is broad and non-bijective: fluctuations in inelasticity, hadronic light production, muon energy-loss stochasticity, and (for through-going tracks) the right-censoring of the pre-detector muon range, all contribute to a many-to-many relationship between E_ν^{true} and E_{reco}. Since the pull correction enforces agreement between the PSF model and the mixture-averaged distribution of angular errors present at each point in reconstructed space (for the utilized MC sample), it does not enforce that the angular error distributions corresponding to individual true-energy subpopulations

match the PSF separately. When the sample is regrouped in slices of true energy, the mixture of reconstructed energies is reweighted according to $p(E_{\mathrm{reco}}, \mathrm{latent}|E_\nu^{\mathrm{true}})$, which generally differs from the mixture implicit in the calibration sample. As a result, the pull distribution in true-energy slices no longer coincides with the calibrated PSF, and the coverage as a function of true energy degrades even though the model remains correctly calibrated in the reconstructed observables that the likelihood actually uses.

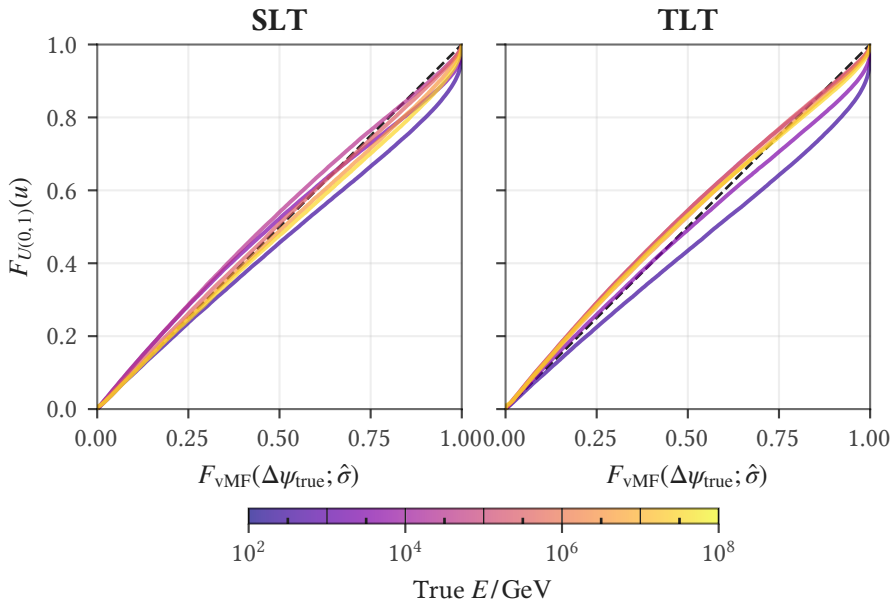

Figure 8.8: PSF coverage evaluated as a function of true neutrino energy, for starting tracks (SLT, left) and through-going tracks (TLT, right): the calibration, performed in reconstructed observables, degrades here because the true-to-reconstructed energy map is broad and non-bijective.

The same diagnostic against true declination is also shown (Figure 8.9).

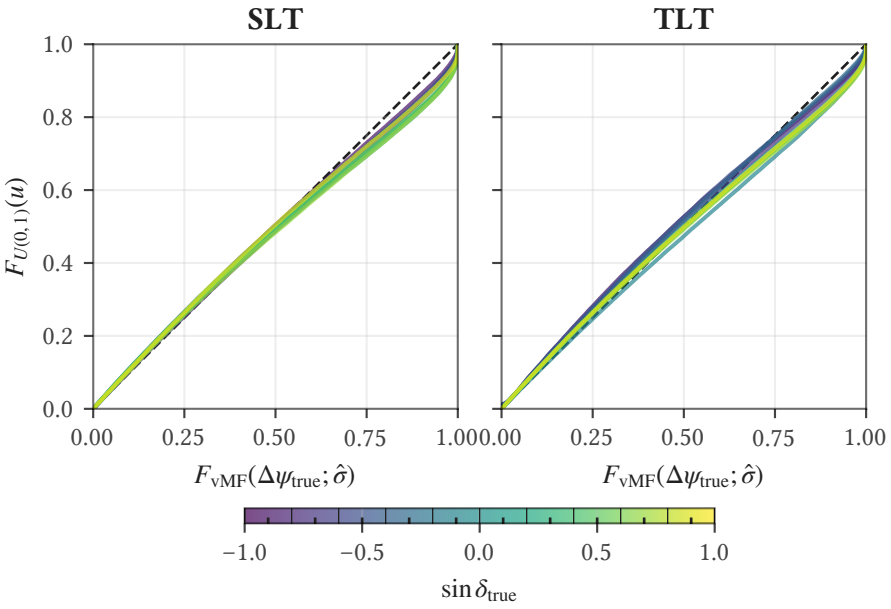

Figure 8.9: PSF coverage evaluated as a function of true declination, for starting tracks (SLT, left) and through-going tracks (TLT, right). In contrast to the true-energy case, the calibration holds across true declination because declination is well reconstructed.

Comparison to 1D corrections and systematics

For completeness, we also tested a simple 1D correction scheme as conventionally applied, in which the median pull was corrected as a function of energy only. While this removes the leading energy dependence, it leaves visible declination-dependent structure in the median pull maps and in the coverage plots (Figure 8.10), stronger in the south, where the energy distributions shift more, and growing toward the poles. The energy-only correction leaves a declination-dependent miscoverage that the full two-dimensional $(E_{\mathrm{reco}}, \sin\delta_{\mathrm{reco}})$ calibration removes, largest in the south and growing toward the poles. At $\sin\delta_{\mathrm{reco}} = -0.87$ (about 60° south) the energy-only correction leaves a peak coverage deviation of about +0.09 for starting tracks and about +0.045 for through-going tracks, which the 2D correction reduces to about +0.02 and +0.015 respectively. The two schemes nearly coincide at the horizon, both leaving an upper-tail deviation of roughly −0.06 to −0.07. At $\sin\delta_{\mathrm{reco}} = +0.87$ (about 60° north) a smaller residual remains, about +0.03 under the energy-only correction and essentially zero under the 2D correction. This asymmetry is the expected behavior rather than a uniform improvement: the mixture of true energies populating a fixed reconstructed-energy bin shifts most strongly with declination in the south (Section 8.3), which is where conditioning the pull correction on declination matters most, and the gain is correspondingly larger for starting tracks than for through-going tracks.

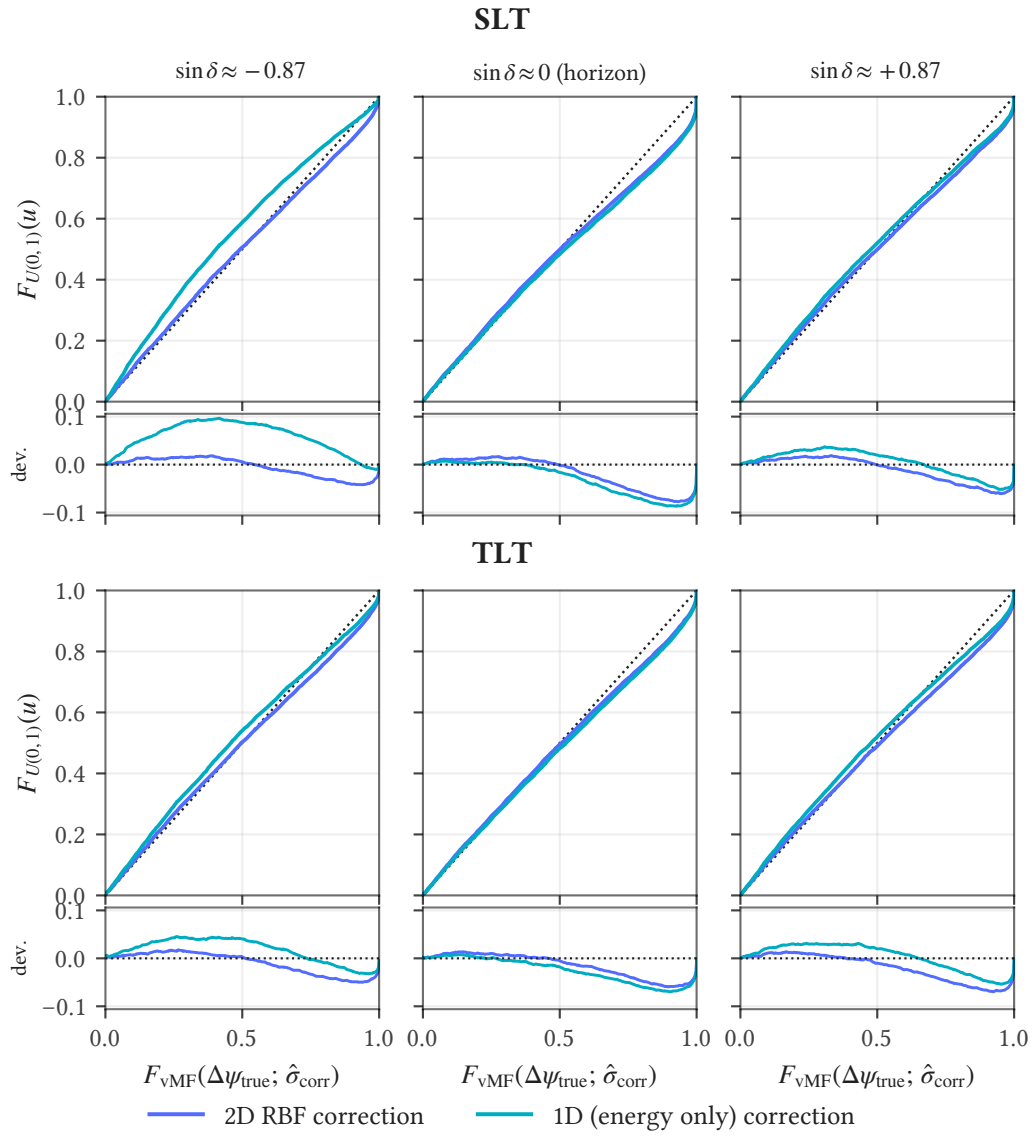

Figure 8.10: Reconstructed-energy PSF coverage under the energy-only (1D) and full $(E_{\mathrm{reco}}, \sin\delta_{\mathrm{reco}})$ (2D) pull corrections, for starting tracks (SLT, top) and through-going tracks (TLT, bottom). The 2D correction improves coverage in the south and is nearly indistinguishable from the energy-only correction at the horizon and in the north.

9 The Point-Source Likelihood and Sample Performance

We evaluate the performance of point-source event samples component by component, following the structure of the point-source likelihood. Each section addresses one component of the statistical model, and the diagnostics cover Lightning Tracks alongside earlier IceCube samples for comparison.

To carry out the computations in practice, we use csky, the most commonly used framework for unbinned point-source likelihood analyses in the IceCube collaboration. It is internal software with no citable published reference. The framework constructs the signal and background probability densities, evaluates the point-source likelihood, and performs the maximum-likelihood fits. It also generates the RA-randomized background trials, runs the all-sky and catalog scans, and converts test statistics to p-values. Essentially every numerical result in this part of the dissertation passes through it.

It was originally written by Mike Richman. Since he left the collaboration it has had no single maintainer, growing instead into a collective effort to which many contributors have added samples and analysis methods over the years. The high-performance reimplementation of its fitting core described in Section 9.13 is part of that lineage.

csky is not the only such framework. The principal alternative is SkyLLH, the framework behind the Northern Tracks point-source results, including the 2022 evidence for neutrino emission from NGC 1068[162] and its more recent extension, a search for neutrinos from X-ray active galactic nuclei.[163] SkyLLH is intrinsically tied to the Northern Tracks selection and provides full support for kernel density estimates of its probability densities. It shares much of its development history with csky, and the two differ mainly in emphasis: SkyLLH is built more around Monte Carlo background modeling and csky around data-driven randomization, though both support either approach.

[162] IceCube Collaboration 2022a, "Evidence for neutrino emission from the nearby active galaxy NGC 1068".

[163] IceCube Collaboration 2026a, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube".

9.1 The unbinned point-source likelihood

The unbinned point-source likelihood tests for an excess of events clustered around a fixed position on the sky, the *test location* $x_s = (\alpha_s, \delta_s)$. The data are the full set of observed events,

$$x = \{x_i\}_{i=1}^{N}, \qquad x_i = (\alpha_i, \delta_i, \hat{\sigma}_i, \hat{E}_i), \tag{9.1}$$

each event x_i carrying a reconstructed right ascension α_i, declination δ_i, angular-error estimate $\hat{\sigma}_i$, and reconstructed energy $\hat{E}_i$.

The signal hypothesis localizes events around the test location through the angular separation between each event and x_s,

$$\Delta\psi_i = \Delta\psi(x_i; x_s) = \arccos[\sin\delta_i \sin\delta_s + \cos\delta_i \cos\delta_s \cos(\alpha_i - \alpha_s)], \tag{9.2}$$

the great-circle angle between the reconstructed direction of event i and x_s.

For N observed events, the likelihood of the data given n_s signal events and $N - n_s$ background events is

$$\mathcal{L}(x; x_s, n_s, \gamma) = \prod_{i=1}^{N} \left[\frac{n_s}{N} \mathcal{S}_i(x_s, \gamma) + \frac{N - n_s}{N} \mathcal{B}_i \right], \tag{9.3}$$

where $\mathcal{S}_i(x_s, \gamma)$ and $\mathcal{B}_i$ are the signal and background probability densities for event i, the subscript i denoting evaluation at that event's data, so that $\mathcal{S}_i(x_s, \gamma) \equiv \mathcal{S}(x_i; x_s, \gamma)$. Each event is modeled as arising from either the signal or the background population with mixing fraction n_s/N. The signal density depends on the test location through the separation $\Delta\psi_i$ and on the assumed source spectrum through the spectral index γ. The background density depends only on the event observables under the null hypothesis. Of the likelihood's three arguments we fit two, the signal count n_s and the spectral index γ, and hold the test location x_s fixed. In principle x_s could be fitted as well, and some analyses do float it, but here we fix it to each candidate source position and scan it externally—the all-sky scan of Chapter 10 steps x_s across a grid of sky positions. That x_s is held rather than fitted is manifest in the likelihood ratio below, where it enters numerator and denominator identically, so the maximization over (n_s, γ) never moves it.

The spectral index appears because the test requires an assumed signal energy spectrum. Throughout this work that spectrum is a single unbroken power law in neutrino energy,

$$\frac{\mathrm{d}N}{\mathrm{d}E} \propto E^{-\gamma}, \tag{9.4}$$

with γ left free and fitted jointly with n_s (the all-sky scan restricts the fit to $\gamma \in [1, 4]$; Section 10.2). This single assumption sets the energy dependence of the signal density $\mathcal{S}_i(x_s, \gamma)$ everywhere in the analysis. The signal energy PDF is built from Equation (9.4) folded through the detector response (Section 9.5).

Two alternatives exist, though we adopt neither. First, γ could be fixed to a model-predicted value rather than fitted, which sharpens sensitivity when the assumed value is correct at the cost of robustness when it is not. Second, the power law could be replaced by an entirely different spectral model. For the obscured Seyfert galaxies of Chapter 13, the predicted neutrino spectrum is that of a corona model rather than a power law, and a recent IceCube search tested both as the signal hypothesis.[164]

[164] IceCube Collaboration 2026a.

For hypothesis testing we compare the likelihood to the background-only hypothesis $n_s = 0$, for which the spectral index drops out, through the likelihood

ratio

$$\frac{\mathcal{L}(x; x_s, n_s, \gamma)}{\mathcal{L}(x; x_s, 0)} = \prod_{i=1}^{N} \left[\frac{n_s}{N} \left(\frac{\mathcal{S}_i(x_s, \gamma)}{\mathcal{B}_i} - 1 \right) + 1 \right]. \tag{9.5}$$

Taking the logarithm turns the product over events into a numerically stable sum,

$$\ln \frac{\mathcal{L}(x; x_s, n_s, \gamma)}{\mathcal{L}(x; x_s, 0)} = \sum_{i=1}^{N} \ln \left[\frac{n_s}{N} \left(\frac{\mathcal{S}_i(x_s, \gamma)}{\mathcal{B}_i} - 1 \right) + 1 \right]. \tag{9.6}$$

The test statistic, denoted T throughout and often abbreviated TS in IceCube prose, is

$$T(x; x_s) = -2 \ln \frac{\mathcal{L}(x; x_s, 0)}{\mathcal{L}(x; x_s, \hat{n}_s, \hat{\gamma})} = 2 \ln \frac{\mathcal{L}(x; x_s, \hat{n}_s, \hat{\gamma})}{\mathcal{L}(x; x_s, 0)}, \tag{9.7}$$

where $\hat{n}_s$ and $\hat{\gamma}$ maximize the likelihood subject to the constraint $n_s \geq 0$, at the fixed test location x_s.

Both signal and background PDFs factorize into spatial and energy terms:

$$\begin{aligned} \mathcal{S}(x_i; x_s, \gamma) &= \mathcal{S}_{\text{space}}(\Delta\psi_i; \hat{\sigma}_i) \times \mathcal{S}_{\text{energy}}(\hat{E}_i; \gamma), \\ \mathcal{B}_i(\delta_i, \hat{E}_i) &= \mathcal{B}_{\text{space}}(\delta_i) \times \mathcal{B}_{\text{energy}}(\hat{E}_i \mid \delta_i). \end{aligned} \tag{9.8}$$

The signal spatial PDF $\mathcal{S}_{\text{space}}$ is the point-spread function (PSF, introduced in Chapter 8), evaluated at the angular separation $\Delta\psi_i$ between event i and the test location x_s and scaled by a per-event PSF scale estimate $\hat{\sigma}_i$. This is where the test-location dependence enters. The signal energy PDF $\mathcal{S}_{\text{energy}}$ reflects the assumed source spectrum (here a power law $E^{-\gamma}$) and is derived from MC simulation. This is where the γ dependence enters. The background spatial PDF $\mathcal{B}_{\text{space}}$ describes the declination-dependent event rate under the null hypothesis. The background energy PDF $\mathcal{B}_{\text{energy}}$ captures the energy distribution of background events as a function of declination.

The factorization in Equation (9.8) is an approximation: the spatial and energy responses are treated as independent, when a unified signal PDF that did not separate them could describe the signal more faithfully. The gain would likely be modest for the energy term, which contributes comparatively little to point-source power on its own (Section 9.12), but for the spatial model it could be significant: a spectrally aware signal PDF could sidestep the pull-correction problem (Section 8.4) entirely. A non-factorized signal PDF is left to future work, an idea due to Chiara Bellenghi.

In practice, the true background $\mathcal{B}$ is unknown and must be estimated from data. We denote the data-derived estimate $\mathcal{D}$; see Section 9.3 for details on how these PDFs are constructed and Section 9.7 for a derivation of the modified likelihood that accounts for signal contamination of $\mathcal{D}$. We store and evaluate the energy terms as a combined signal-to-data ratio $\mathcal{S}_{\text{energy}}(\gamma)/\mathcal{D}_{\text{energy}}$ rather than maintaining separate signal and data-derived energy PDFs.

Event samples and combination

When combining multiple samples in a single analysis, the individual sample likelihoods are multiplied together. The signal parameters n_s and γ are *shared* across all samples rather than fit independently for each. Each sample receives a fraction of the total signal determined by its relative acceptance:

$$\mathcal{L}_{\text{combined}}(x; x_s, n_s, \gamma) = \prod_k \mathcal{L}_k\big(x^{(k)};\, x_s,\, n_s \cdot f_k(\gamma),\, \gamma\big), \qquad f_k(\gamma) = \frac{A_k(\gamma)}{\sum_j A_j(\gamma)}, \tag{9.9}$$

where $x^{(k)}$ is the data of sample k, and $A_k(\gamma)$ is its signal acceptance (see Section 9.6). A sample with higher acceptance receives a proportionally larger share of the total signal. In practice, the combined log-likelihood ratio takes the same form as the single-sample case (a sum over all events), but each event uses the PDF terms from its own sample and is weighted by the effective signal count $n_s^{(k)} = f_k \cdot n_s$ for that sample:

$$\ln \frac{\mathcal{L}_{\text{combined}}}{\mathcal{L}_0} = \sum_k \sum_{i \in \mathcal{I}_k} \ln \left[\frac{n_s \cdot f_k}{N^{(k)}} \left(\frac{S_i^{(k)}}{B_i^{(k)}} - 1 \right) + 1 \right], \tag{9.10}$$

where $\mathcal{I}_k$ denotes the set of event indices in the sample indexed by k.

The per-sample PDFs matter asymmetrically. Through-going tracks dominate the combined statistics, so PDFs built from a single pooled sample would essentially describe the through-going population. In sensitivity tests that treated the samples as one, the combined signal-to-background PDFs were essentially those of the through-going population. Conversely, pooling the remaining samples into the through-going statistics changes the latter's PDFs by little more than a rounding error. Separation therefore protects the lower-statistics samples from being described by a pool dominated by through-going tracks—while the dominant sample is indifferent. For the two track topologies, the analogous asymmetry in the angular-error calibration has been measured directly (Section 8.4). Between starting tracks and cascades, whose statistics are of comparable scale, no equivalent measurement exists. We expect the protection to apply symmetrically there, with each sample benefiting from not being merged with the other.

Event overlap must be handled carefully. As Figure 9.1 shows, significant overlap exists between several sample pairs, particularly between track selections (PST, NT, ESTES, LT) and between ESTES and DNNC. Overlapping events must be removed to avoid double-counting, which would bias the likelihood and inflate significance. For LT (SLT + TLT), overlap is prevented by construction during the selection process, since the samples are two branches of the same selection pipeline. For LT + DNNC, we process the same MC used for the DNNC sample, enabling event matching and thus exact overlap removal on both data and MC.

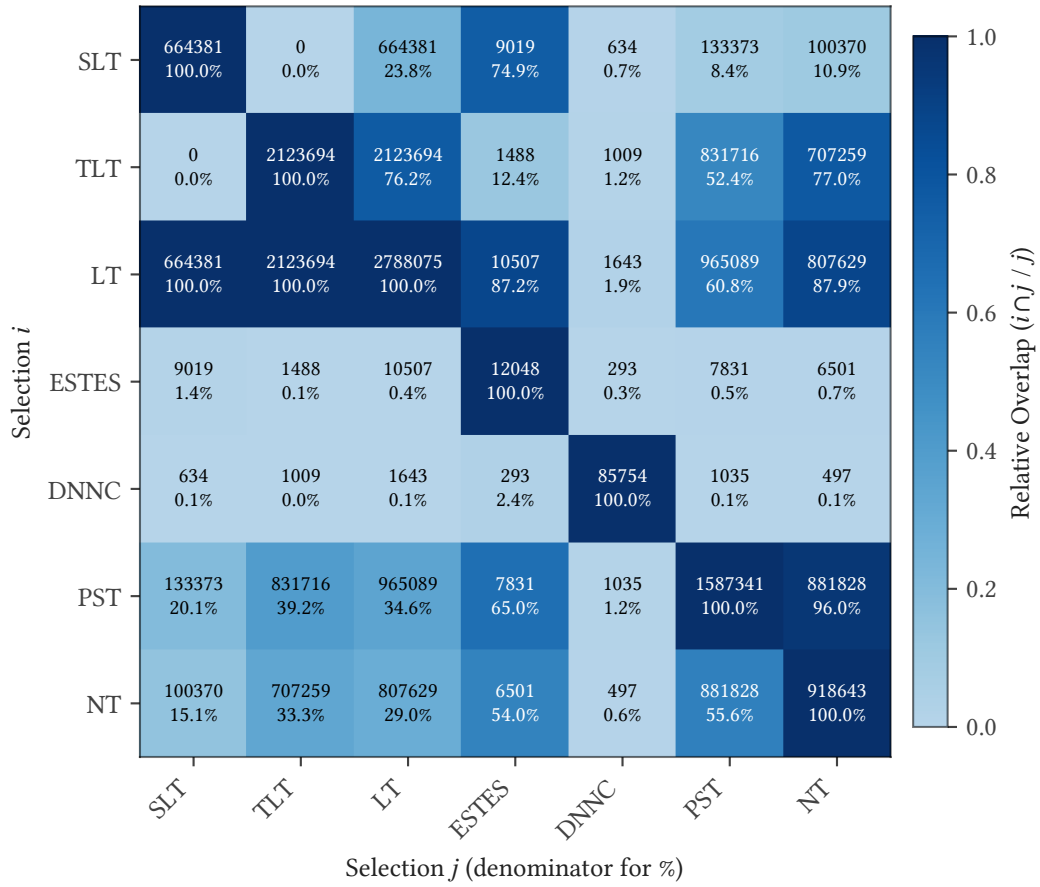

Figure 9.1: Overlap matrix for the major IceCube point-source samples. Each cell's absolute count is the number of data events shared by the corresponding pair of samples and is symmetric ($ij = ji$). The percentage below each count normalizes the overlap to the column sample j: for example, LT and NT share 807,629 events, 87.9% of NT's events but only 29% of LT's, since NT lacks LT's southern events.

9.2 Angular resolution

Although angular resolution is essential for point-source sensitivity, it is only predictive when other factors are comparable across selections. Angular resolution can even be artificially improved by applying tight cuts on the angular-error estimator. However, as long as the estimator is reasonably consistent with the PSF model used in the likelihood, such cuts *reduce* sensitivity instead of improving it.

A more fundamental limitation is that angular resolution cannot be compared meaningfully across reconstruction methods unless all methods are evaluated event-by-event on the same underlying sample. For this reason, comparing angular-resolution curves across selections provides limited insight.

The angular error distributions shown in Figure 9.2 should therefore be read alongside the coverage of the PSF model (Section 8.5), because better absolute reso-

lution does not by itself buy more analysis power. The energy selection makes this sharpest: a sample concentrated at high energies has excellent angular resolution but little low-energy effective area, and it can be less sensitive than a strict superset that also keeps the lower-energy events, provided those added events' angular errors remain reasonably covered by the PSF model. The same logic explains the tight-cut effect noted above: narrowing a selection to its best-reconstructed events sharpens the resolution while discarding usable information, and as long as that information is adequately modeled the net effect is a loss of sensitivity.

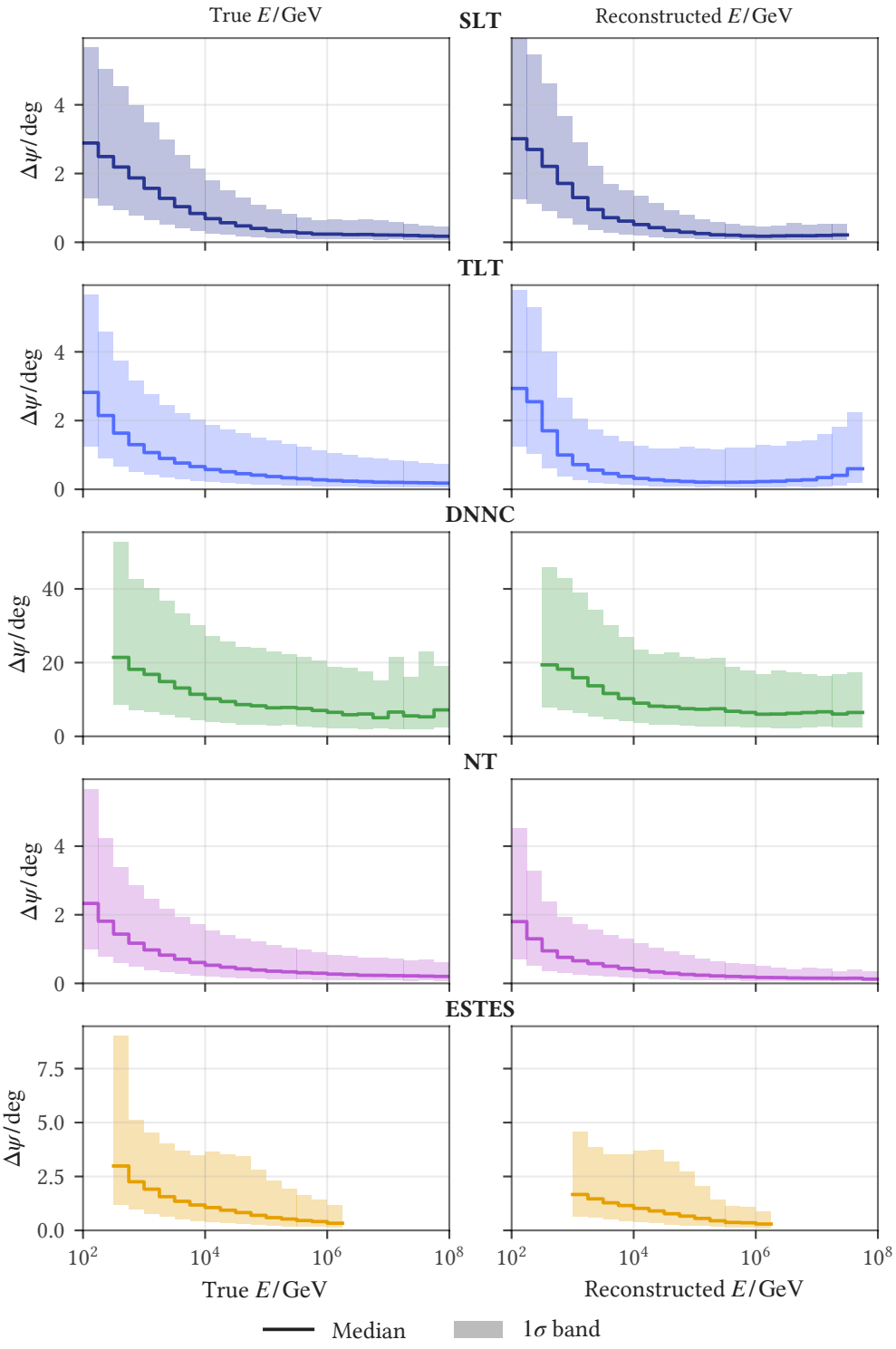

Figure 9.2: Angular resolution versus energy for each sample, shown against true neutrino energy (left column) and reconstructed energy (right column) for SLT, TLT, DNN Cascades, NT, and ESTES (rows). The line is the median angular error and the shaded band the central 16–84% interval (the $\pm 1\sigma$ range for a Gaussian). Bins with unreliable statistics (low effective sample size, including tail-dominated bins) are removed, and the shaded band is interpolated across the resulting gaps for visual continuity.

9.3 Data-driven background estimation

The background hypothesis assumes that all observed events arise from an isotropic background (atmospheric muons, atmospheric neutrinos, and any diffuse astrophysical neutrino flux), with no contribution from localized astrophysical point sources. Background modeling requires characterizing both the spatial distribution of events across the sky and their energy distribution as a function of declination.

Both background PDFs are constructed directly from the observed data rather than from Monte Carlo simulation. In principle, one could model atmospheric backgrounds using MC, but this approach faces severe practical limitations. Atmospheric neutrino flux predictions carry large systematic uncertainties (loosely estimated at the 10–30% level, driven by, e.g., the Barr flux parameters[165] and uncertainties in the hadronic interaction model), and producing atmospheric muon simulation with sufficient statistics across all energies is computationally prohibitive. MC-based background estimation is therefore only marginally feasible in the northern sky (where atmospheric muons are subdominant), and even there it requires careful treatment of systematics that can easily dominate over statistical uncertainties.

[165] Barr et al. 2006, "Uncertainties in Atmospheric Neutrino Fluxes".

The data-driven alternative exploits the fact that IceCube's background is isotropic in right ascension. Atmospheric neutrinos and muons arrive uniformly in azimuth (in local coordinates), and Earth's rotation smears any fixed local direction into a uniform ring in equatorial right ascension over the course of a sidereal day. Because the background is RA-independent, the background PDFs depend only on declination (and energy): the *spatial PDF* $\mathcal{D}_{\text{space}}(\delta)$ is simply the declination distribution of observed events, and the *energy PDF* $\mathcal{D}_{\text{energy}}(E|\delta)$ is the energy distribution within each declination band. We denote these data-derived PDFs with $\mathcal{D}$ rather than $\mathcal{B}$ to emphasize that they are empirical estimates constructed from the observed data, not the true (unknown) background distribution.

Remark 9.1. IceCube's location at the geographic South Pole is not required for data-driven background estimation: it simply makes the implementation more convenient. All neutrino detectors have zenith-dependent background rates in local coordinates (azimuth-independent, unless built next to a large mountain or similar obstruction). Seasonal variations do introduce time dependence, but for time-integrated searches we assume these average out over the observation period. Detector locations differ in how local coordinates map to equatorial coordinates. For detectors not at a pole, the transformation between equatorial coordinates (right ascension, declination) and local coordinates (azimuth, zenith) is time-dependent: a fixed point in right ascension and declination traces a path across different zenith angles throughout the day. This means the PDFs must be evaluated in local coordinates, and the source position (in local coordinates) is different for every event depending on its timestamp, adding complexity to the likelihood evaluation. At the South Pole, the declination-to-zenith mapping is effectively time-invariant (declination equals 90° minus zenith), and only right ascension rotates with local sidereal time. This allows us to work entirely in equatorial coordinates.

9.4 The background spatial PDF

The background spatial PDF $\mathcal{D}_{\text{space}}(\delta)$ describes the expected distribution of events across the sky under the null hypothesis. Because the background is uniform in azimuth, the PDF depends only on declination and reflects the detector's zenith-dependent acceptance.

Figure 9.3 shows the background spatial PDF for each sample. All samples peak near the horizon, where the ice overburden provides sufficient shielding to attenuate most atmospheric muons while remaining transparent to neutrinos: this region offers the best signal-to-noise ratio.

We refer to the transition from neutrino-dominated to muon-dominated backgrounds as the *muon horizon*: the declination where the atmospheric-muon rate crosses the atmospheric-neutrino rate and muons become the dominant background. Its location is both model- and selection-dependent, since the event-selection cuts themselves shift it. The event topologies behind this behavior are discussed in Chapter 2, and the bump cut that fixes the horizon placement for the throughgoing tracks is set in the sensitivity optimization (Section 5.2). North of the horizon the rate rises into territory dominated by atmospheric neutrinos; further south (more downgoing), atmospheric muons increasingly dominate.

Throughgoing track selections must apply progressively stricter energy cuts south of the muon horizon to achieve optimal signal-to-noise. Because throughgoing tracks lack topological information to distinguish single atmospheric muons from neutrino-induced muons (the two are topologically indistinguishable), the only available discriminant is the spectrum. Without the bump cut, the TLT sample exhibited a sharp spike in the background spatial PDF near the muon horizon due to this effect. Just south of the muon horizon, the ice overburden is thin enough to transmit high-energy neutrino-induced muons that are shielded by the thicker ice (and bedrock) at more northern declinations, but still thick enough to filter most low- and medium-energy atmospheric muons. These high-energy neutrinos interacted with the ice far from the detector, so their muon energy is right-censored: they appear as low reconstructed energy events. The result is a signal-to-background ratio above unity at low reconstructed energies in this region, and the data rate temporarily spikes before plummeting as cuts become stricter further south. The bump cut (Section 5.2) suppresses this feature by raising the MLP score threshold at the muon horizon, producing the smoother profile visible in Figure 9.3.

The Starting Lightning Tracks sample follows a different pattern. Starting tracks can use veto cuts based on event topology to reject atmospheric muons, but these cuts lose discriminating power in the southern sky. As a result, the optimal energy cut is loosened rather than tightened, yielding better overall sensitivity. These declination-dependent changes in the optimal cut produce the characteristic double-peak structure visible in Figure 9.3.

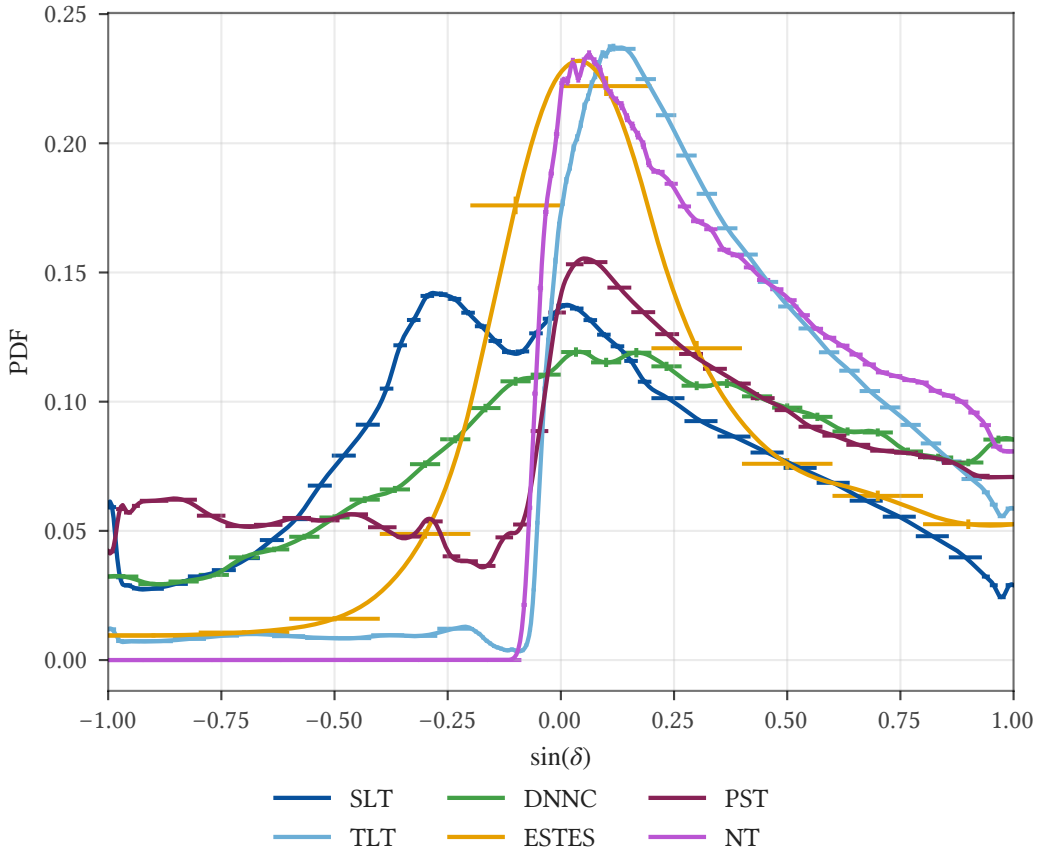

Figure 9.3: Background spatial PDF, the null-hypothesis declination distribution, for each sample.

9.5 The energy signal-to-background ratio

Recall that the full signal-to-data ratio in the likelihood factorizes as $\mathcal{S}_i/\mathcal{D}_i = (\mathcal{S}_{\text{space}}/\mathcal{D}_{\text{space}}) \times (\mathcal{S}_{\text{energy}}/\mathcal{D}_{\text{energy}})$. The spatial ratio is computed at runtime from the PSF and background spatial PDF (discussed in Section 9.4). This section focuses on the *energy ratio* $\mathcal{S}_{\text{energy}}(E;\gamma)/\mathcal{D}_{\text{energy}}(E|\delta)$, which is precomputed as 2D histograms and looked up at each event's reconstructed $(\sin\delta_i, \log E_i)$ coordinates during likelihood evaluation.

The energy ratio quantifies the relative likelihood that an event at energy E originates from a power-law source rather than atmospheric backgrounds. This ratio provides discrimination between signal and background based on the reconstructed energy proxy: for astrophysical sources with hard spectra, signal events tend toward higher energies than the softer atmospheric background.

Remark 9.2. A common misconception is that overestimating background rates or angular errors is *conservative*, and therefore harmless. Under the empirical calibration used here this is false, though not for the usual reason. Because the test

statistic is calibrated directly from background trials (Section 9.8), the significance of an excess and the coverage of the confidence regions stay valid however poorly the likelihood model matches reality. A systematic mismatch does not render the results invalid. What it costs is power: a less faithful model is less efficient, so the Feldman–Cousins construction overcovers, returning wider confidence regions than a well-specified model would. That extra width is not a free safety margin. It is constraint thrown away, so deliberately erring toward overcoverage because it feels conservative is itself a mistake. And where the mismatch also biases the mapping from the fitted (n_s, γ) to physical flux (Section 9.10), a theoretical model checked against the published region may be wrongly excluded or wrongly retained—an error the nominal confidence level conceals. Only explicit calibration against data, read with an honest account of the model's limits, keeps such conclusions trustworthy.

Construction

The $\mathcal{S}/\mathcal{D}$ ratio is constructed as a 2D histogram in $(\sin\delta, \log E_{\text{reco}})$ for each spectral index γ in a predefined grid. The data-derived histogram is built from observed events (with equal weight per event), while the signal histogram is built from MC weighted by $w_{\text{one}} \times E_\nu^{-\gamma}$ for the assumed spectral index. These raw histograms are converted to probability densities and interpolated with cubic splines to provide smooth $\mathcal{S}/\mathcal{D}$ values at arbitrary $(\sin\delta, \log E)$ positions within each histogram. During likelihood evaluation, the spectral index γ is a fit parameter that can take any value, not just the predefined grid points. We handle this by parabolic interpolation in log-space, identifying the three nearest grid points, fitting a parabola through the corresponding $\log(\mathcal{S}/\mathcal{D})$ values, and exponentiating the result.

Because we do not randomize declination or energy in background trials (for track selections), we can never encounter a bin during background trials or unblinding that has zero data events: every observed event falls into a bin that, by definition, contains at least that event. However, *empty bins* can arise in two ways: (1) regions of phase space where no data events were observed during the livetime, and (2) regions where no MC events exist. In Csky this is handled by filling empty data bins with the global minimum data density and setting signal values in empty bins equal to the corresponding data value. This yields $\mathcal{S}/\mathcal{D} = 1$ for any bin with no data, effectively treating it as uninformative. Bins with neither data nor MC are irrelevant even during signal injection (since no MC events can land there), but finite values must still be assigned to prevent 0/0 in the $\mathcal{S}/\mathcal{D}$ calculation (these fill values break the normalization of the energy PDF within each $\sin\delta$ slice, but this is acceptable since the affected bins are never populated). This filling strategy is why many $\mathcal{S}/\mathcal{D}$ histograms (e.g., Figure 9.4) show large regions with constant $\mathcal{S}/\mathcal{D}$ ratios. In particular, at extremely high reconstructed energies ($E_{\text{reco}} \gtrsim 10^8$ GeV) where we have zero data events and either zero or one MC event per bin, almost all bins fall into one of two categories: $\mathcal{S}/\mathcal{D} = 1$ (no data, no MC) or $\mathcal{S}/\mathcal{D} \ll 1$ (no data, one MC event). The latter case yields an arbitrary, binning-dependent but

typically tiny value because the flux-weighted MC contribution (for any reasonable spectral assumption) is much smaller than the global minimum data density used to fill empty data bins.

More broadly, the $\mathcal{S}/\mathcal{D}$ ratio is poorly constrained at high energies wherever data and MC statistics are sparse, including bins with only one or a few data events, where the ratio is highly sensitive to binning choices and statistical fluctuations. Of the two failure modes described above, the zero-data/nonzero-MC case ($\mathcal{S}/\mathcal{D} \ll 1$) affects only sensitivity and discovery potential calculations, since those bins are populated only during signal injection and never when evaluating real data. They are therefore irrelevant for unblinding p-values. The low-statistics case, however, affects all analyses. Signal subtraction with the energy term exacerbates both modes: arbitrary bin-filling artifacts and poorly constrained ratios are amplified into extreme signal-subtraction weights that can distort the likelihood surface and bias the best-fit n_s.

Alternatives to fixed-binning energy PDFs

These issues are compounded by a fundamental limitation of csky: it enforces *fixed binning* in both $\sin\delta$ and $\log E$ independently: we cannot use statistically optimized adaptive binning as we do for the PSF calibration. This means we either lose detail in high-statistics regions or suffer large fluctuations in low-statistics regions. One potential remedy that addresses both problems is to replace the fixed-binning histogram approach entirely with adaptive-bandwidth kernel density estimates of the full 2D $(\sin\delta, \log E)$ densities for signal and background, with explicit regularization (e.g., Laplace smoothing) replacing the ad hoc bin-filling. This eliminates the fixed-binning trade-off and makes all smoothing assumptions transparent model choices rather than implicit consequences of binning. However, tuning the KDE hyperparameters (bandwidth selection, boundary handling, and regularization strength) is a nontrivial optimization problem in its own right, and the resulting PDFs are sensitive to these choices. For the present analysis we retain the standard histogram approach with the per-component signal-subtraction configuration described in Section 9.7, but KDE-based energy PDFs remain a viable direction for future work.

Implications for hard spectra

These limitations are most acute for hard spectra ($\gamma \lesssim 2$), where the signal is dominated by the highest-energy events. These probe precisely the regions of phase space where we have zero or very few data events, so the $\mathcal{S}/\mathcal{D}$ ratio is driven by bin-filling artifacts rather than physical considerations. Sensitivity and discovery potential estimates for hard spectra should therefore be interpreted with care: they are highly model-dependent and strictly valid only for the exact assumptions made during the corresponding signal injection trials. The issue is not that the results are wrong—but that they depend sensitively on how the analysis handles regions with insufficient data to constrain the background. Other analysis frameworks (e.g.,

SkyLLH) address this by allowing a configurable default value for empty data bins, making the prior background assumption explicit. In practice, this is unlikely to significantly affect results since regions with zero data events are also unlikely to see substantial numbers of injected signal events for any reasonable source hypothesis.

This is further exacerbated by a subtle inconsistency in the TS calibration (see Section 9.8). The background trials used to define the TS threshold under H_0 are computed from the original data, which may contain zero or very few high-energy events in these sparsely populated bins. Those events can therefore never cluster by chance during background trials, and the resulting TS distribution reflects this. During signal injection trials, however, injected MC events populate these previously empty bins. Although the background PDFs are updated for each injection trial, the TS threshold is still calibrated against the original null distribution, which never saw such events. A fully self-consistent treatment would require rerunning background trials for every individual signal injection trial, recalibrating the null TS distribution with the specific injected events from that trial included in the data. This is computationally infeasible. Sensitivity, discovery potential, and upper limit estimates for such extreme spectra are therefore fundamentally limited by these uncorrectable approximations, and there is a reasonable argument for not publishing them at all.

The 2D signal-to-background ratio

Figure 9.4 shows the 2D signal-to-background ratio as a function of reconstructed energy and declination, separately for the SLT and TLT samples. Values above 1 indicate regions where signal is favored, values below 1 favor background. Astrophysical signal should dominate at high energies, and the northern sky typically has lower atmospheric background, increasing $\mathcal{S}/\mathcal{D}$. Softer assumed spectra ($\gamma \gtrsim 2.5$) shift the signal distribution to lower energies, reducing the discrimination power of the energy term.

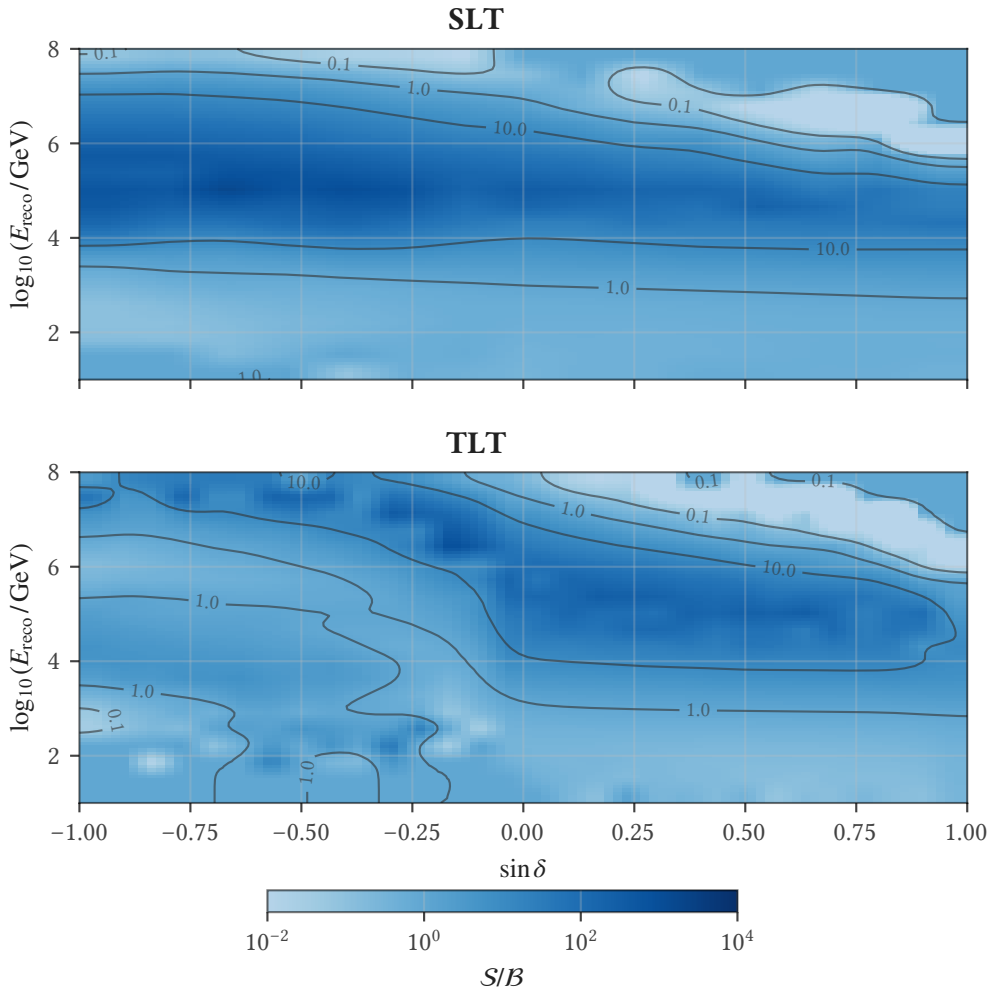

Figure 9.4: Two-dimensional signal-to-background energy ratio $\mathcal{S}/\mathcal{D}$ as a function of reconstructed energy and declination, shown separately for the SLT (top) and TLT (bottom) samples. Values above 1 favor signal, below 1 favor background. Computed at an assumed signal spectral index of $\gamma = 2.5$.

Energy PDF slices

Reading 2D color histograms can be difficult, so we also show 1D slices of the ratio. Figure 9.5 plots $\mathcal{S}/\mathcal{D}$ against reconstructed energy, one curve per fixed declination, isolating the energy discrimination at a given point on the sky. Figure 9.6 plots it against declination instead, one curve per fixed energy, showing how that discrimination varies across the sky. Both figures are split into the SLT and TLT samples. Figure 9.7 shows the same energy slices at $\sin\delta = 0$ with one curve per assumed spectral index, making the spectral dependence of the discrimination explicit.

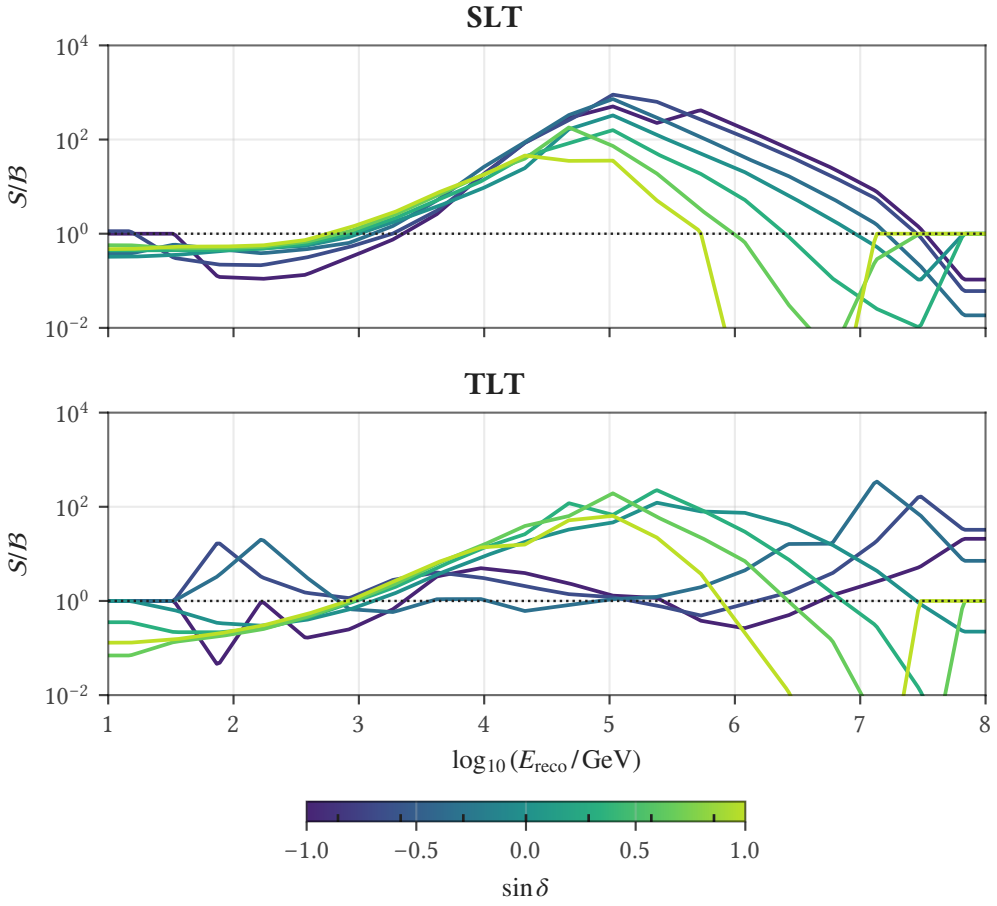

Figure 9.5: One-dimensional slices of the $\mathcal{S}/\mathcal{D}$ energy ratio versus reconstructed energy, each curve at a fixed declination, for the SLT (top) and TLT (bottom) samples. Computed at an assumed signal spectral index of $\gamma = 2.5$.

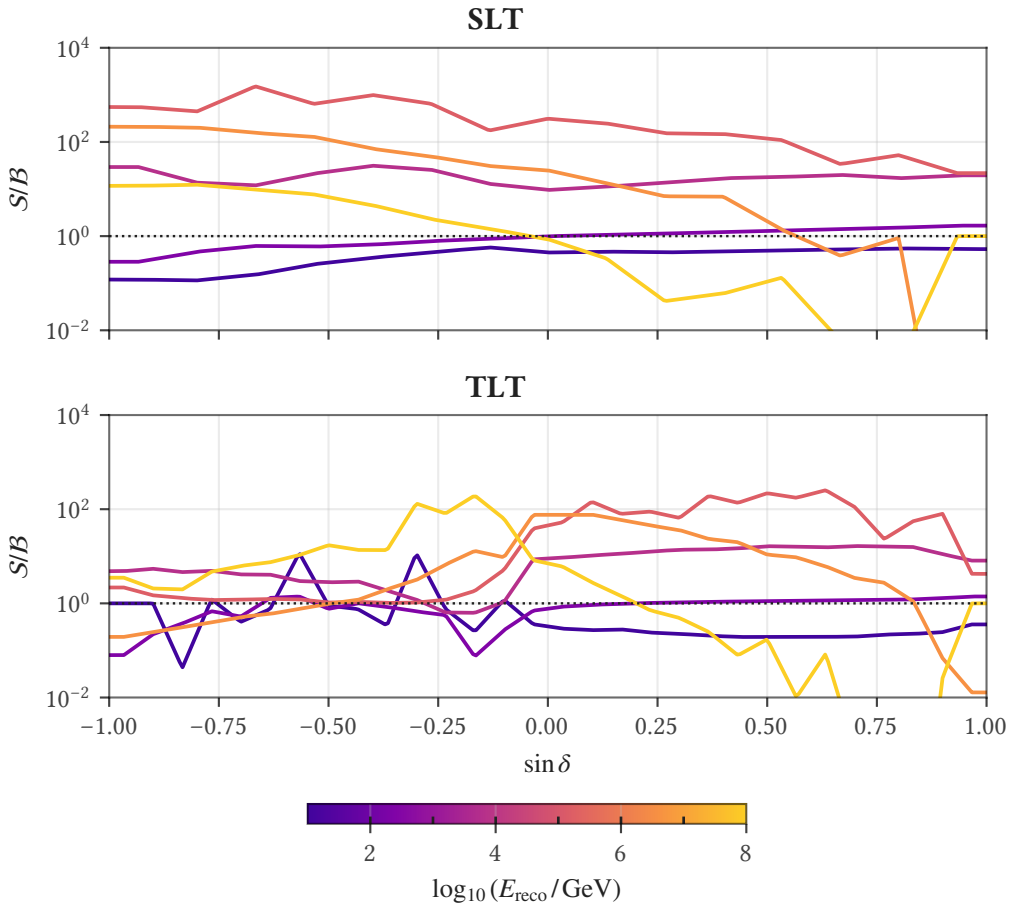

Figure 9.6: One-dimensional slices of the $\mathcal{S}/\mathcal{D}$ energy ratio versus declination, each curve at a fixed reconstructed energy, for the SLT (top) and TLT (bottom) samples. Computed at an assumed signal spectral index of $\gamma = 2.5$.

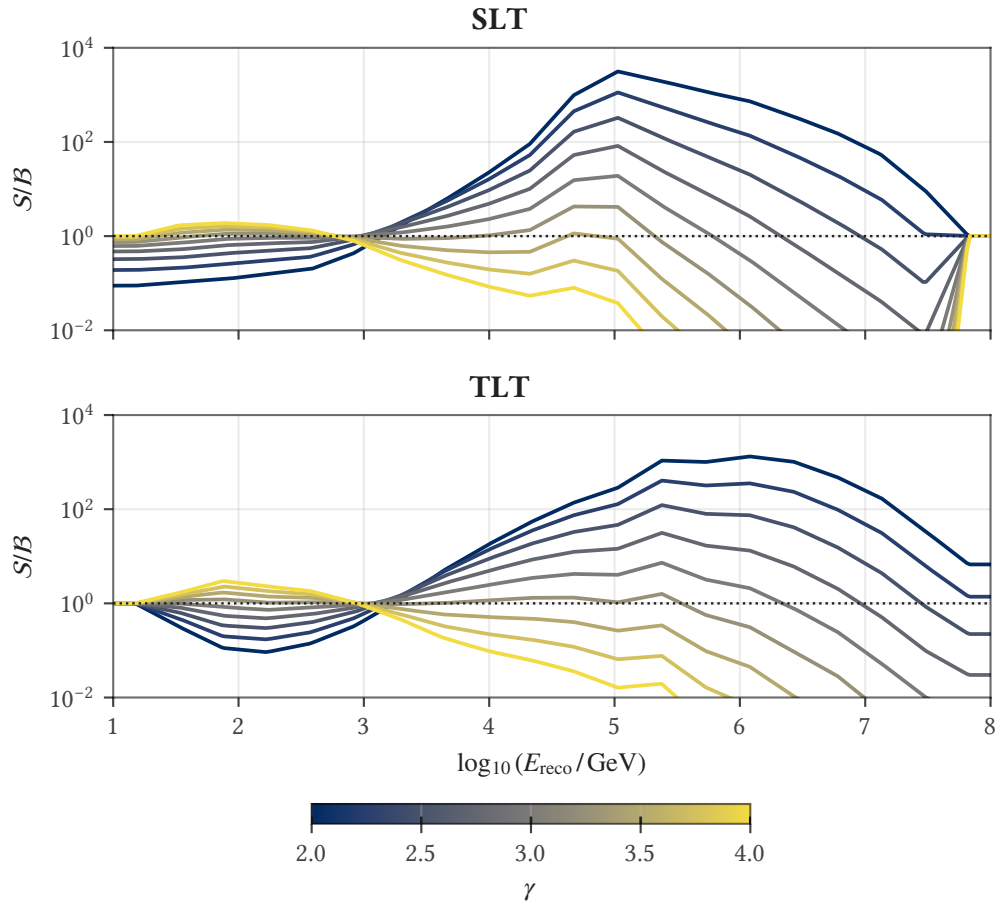

Figure 9.7: Spectral-index dependence of the energy ratio $\mathcal{S}/\mathcal{D}$ versus reconstructed energy at the celestial equator ($\sin\delta = 0$), one curve per assumed spectral index γ from 2.0 to 4.0 (colorbar), for the SLT (top) and TLT (bottom) samples.

9.6 Effective area and acceptance

The signal energy PDF $\mathcal{S}_{\mathrm{energy}}$ describes the expected energy distribution of signal events given the assumed source spectrum. Rather than appearing explicitly in the likelihood, this term is combined with the data-derived energy PDF into a signal-to-background ratio (see Section 9.5). Here we focus on the underlying quantities that determine the rate of signal events: effective area and acceptance.

The effective area A_{eff} is a standard measure of detection efficiency: it represents the equivalent cross-sectional area of an idealized detector that would observe the same number of events as the real instrument. Formally, it is defined such that the rate of detected events from a flux $\Phi(E)$ is

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \int \Phi(E)\, A_{\mathrm{eff}}(E, \Omega)\, \mathrm{d}E\, \mathrm{d}\Omega. \tag{9.11}$$

In general, A_{eff} depends on both energy and arrival direction (zenith and azimuth). For IceCube, located at the geographic South Pole, the detector's orientation relative to the celestial sky is nearly time-independent, and the hexagonal string layout introduces only weak azimuthal asymmetry. As a result, IceCube's effective area is well-approximated as a function of neutrino energy and declination alone: $A_{\text{eff}}(E_\nu, \delta)$. This effective area encapsulates the entire detection chain: the neutrino–nucleon interaction cross-section, the probability that the resulting lepton or hadronic shower produces detectable Cherenkov light, the geometric acceptance of the photomultiplier array, the trigger efficiency, and the survival probability through all selection cuts. Because a point source illuminates a single direction, the solid-angle integral in the rate formula above collapses: in the acceptance expressions that follow, $\Phi(E_\nu)$ denotes the point-source flux and $A_{\text{eff}}(E_\nu, \delta)$ the point-source effective area at the source declination.

In Monte Carlo simulation, the effective area is estimated from the ratio of detected to generated events:

$$A_{\text{eff}}(E_\nu, \delta) = A_{\text{gen}} \cdot \frac{N_{\text{det}}(E_\nu, \delta)}{N_{\text{gen}}(E_\nu, \delta)}, \tag{9.12}$$

where A_{gen} is the generation area (the cross-sectional area of the cylindrical or spherical volume over which neutrinos were injected) and the ratio reflects the fraction of generated events that pass all selection criteria. In IceCube's simulation framework, this information is encoded in the per-event *one-weight* w_{one}, which combines generation area, solid angle, and interaction probability into a single weight that allows efficient computation of expected event rates for arbitrary flux models. In the convention used here, w_{one} already includes division by the number of generated events, so no explicit $1/N_{\text{gen}}$ factor appears in the weighted sums below.

Figure 9.8 shows the effective areas for all samples. For PST and NT, the MC data contain only ν_μ and ν_τ. Neither selection has yet been processed on ν_e NuGen. Consequently, their effective-area curves do not exhibit the Glashow resonance[166] bump near $\approx$ 6.3 PeV, whereas it is clearly visible for Lightning Tracks and, as expected, for DNN Cascades. Presumably the resonance would also appear for NT and PST if they were processed on ν_e simulations, though their ν_e contamination may also simply be lower. Neither can be said for sure without checking, which is not possible here: those are not this work's samples, and ν_e is not included in the data they processed.

[166] Glashow 1960, "Resonant Scattering of Antineutrinos".

While effective area is a useful diagnostic quantity, it should not be interpreted as a direct indicator of point-source performance. Sensitivity is ultimately set by the *signal-to-noise ratio* ($S/\sqrt{B}$) and *angular resolution*, and effective area only correlates with sensitivity when background rates and angular errors are comparable across selections.

This can be seen in Figure 9.8, where NT has a slightly smaller effective area than PST. Yet, as shown in Figure 9.21, NT achieves better sensitivity. The underlying reason is difficult to attribute uniquely (lower background contamination,

improved angular reconstruction, or more accurate angular-error modeling may all contribute), but the conclusion remains—effective area alone is not a meaningful measure of point-source sensitivity.

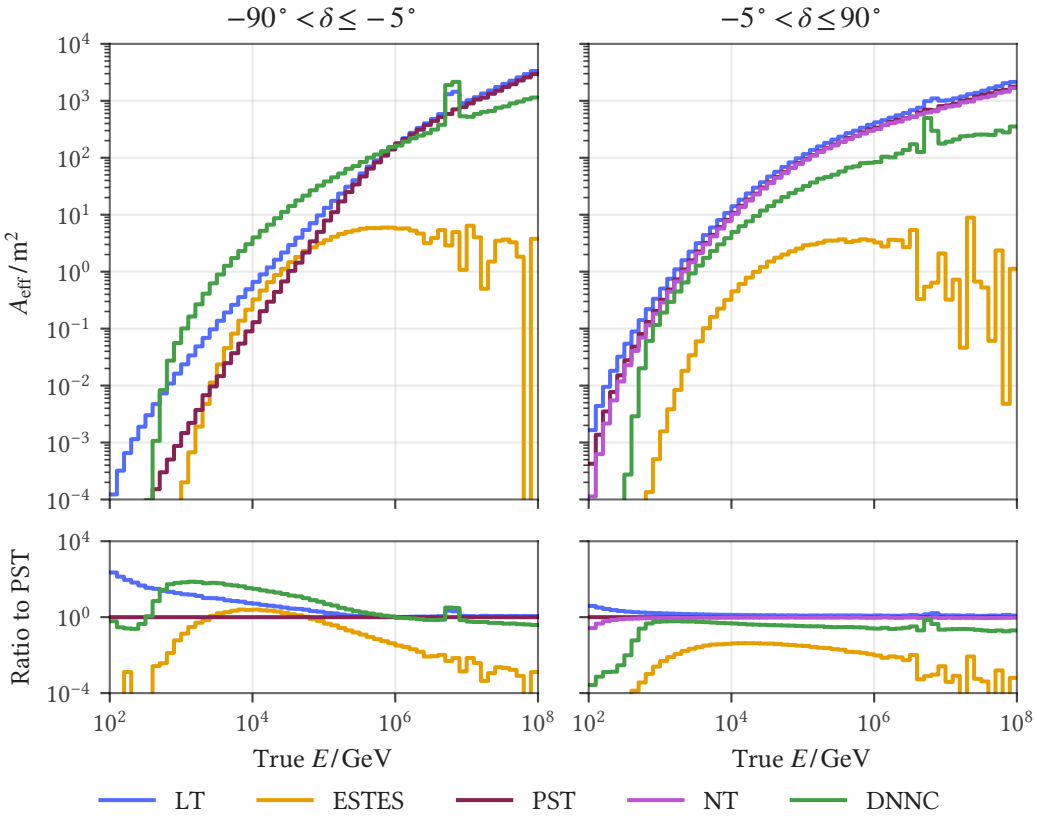

Figure 9.8: Effective area as a function of neutrino energy for each sample.

Figure 9.9 compares the Lightning Tracks and Northern Tracks effective areas in the northern sky. Lightning Tracks has a significantly larger low-energy effective area than Northern Tracks. Because muon energy is right-censored—a high-energy neutrino interacting far from the detector is reconstructed at lower energy than it carries—this larger low-energy acceptance also raises the Lightning Tracks effective area at high neutrino energies.

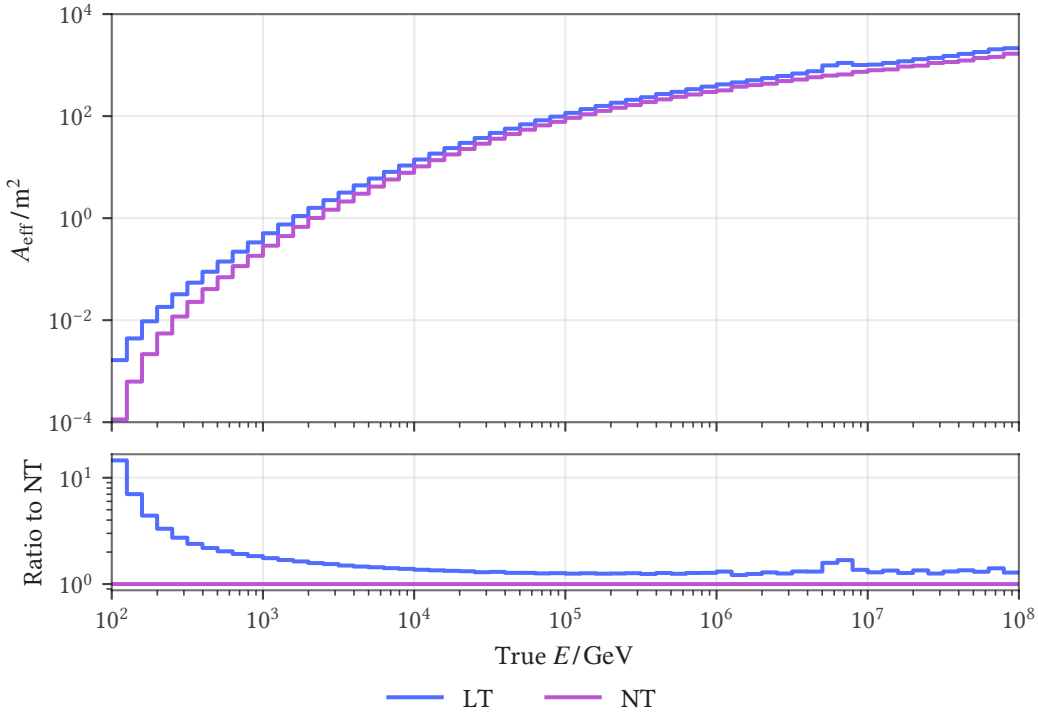

Figure 9.9: North-only effective area for Lightning Tracks and Northern Tracks, with the ratio to Northern Tracks below. Lightning Tracks holds a large low-energy effective-area advantage; because muon energy is right-censored, this advantage persists at high neutrino energies rather than falling to unity.

Acceptance

The *acceptance* $A(\gamma, \delta)$ integrates the effective area over the assumed source spectrum and observation time to yield an expected event count. For a source with differential flux $\Phi(E_\nu) = \Phi_0\, E^{-\gamma}$, the expected number of detected events is

$$\langle N \rangle = \tau \int \Phi(E_\nu)\, A_{\mathrm{eff}}(E_\nu, \delta)\, \mathrm{d}E_\nu = \Phi_0\, \tau \int E^{-\gamma}\, A_{\mathrm{eff}}(E, \delta)\, \mathrm{d}E \equiv \Phi_0\, A(\gamma, \delta), \quad (9.13)$$

where τ is the detector livetime. Thus the acceptance is the livetime-weighted spectral integral of the effective area:

$$A(\gamma, \delta) = \tau \int E^{-\gamma}\, A_{\mathrm{eff}}(E, \delta)\, \mathrm{d}E. \quad (9.14)$$

In IceCube's simulation framework, this integral is computed via a weighted sum over MC events:

$$A(\gamma, \delta) = \frac{\tau}{\Omega} \sum_i w_{\mathrm{one},i}\, E_{\nu,i}^{-\gamma}, \quad (9.15)$$

where the sum runs over MC events in a declination band of solid angle Ω around the source. The factor $1/\Omega$ converts this band-integrated sum into the point-source

acceptance: w_{one} carries the solid angle over which events were generated, so dividing by the band solid angle Ω recovers the per-source quantity that the continuous spectral integral above expresses directly.

The acceptance provides the crucial link between detected event counts and physical flux. Given a measured or fitted signal strength n_s, the corresponding flux normalization is $\Phi_0 = n_s / A(\gamma, \delta)$. This relationship underlies the conversion from sensitivity (expressed in expected signal events) to sensitivity (expressed as a flux limit), as discussed in Section 9.11.

Integrated over the northern sky, the Lightning Tracks acceptance advantage over Northern Tracks grows toward softer spectra, where the low-energy effective-area gain weighs most heavily. The northern-sky acceptance of Lightning Tracks exceeds that of Northern Tracks by a factor of 1.36 at $\gamma = 2$, 1.95 at $\gamma = 3$, and 2.48 at $\gamma = 3.5$.

Figure 9.10 shows the signal acceptance of the nominal analysis sample as a function of declination, with stacked contributions from its SLT, TLT, and DNN Cascades components.

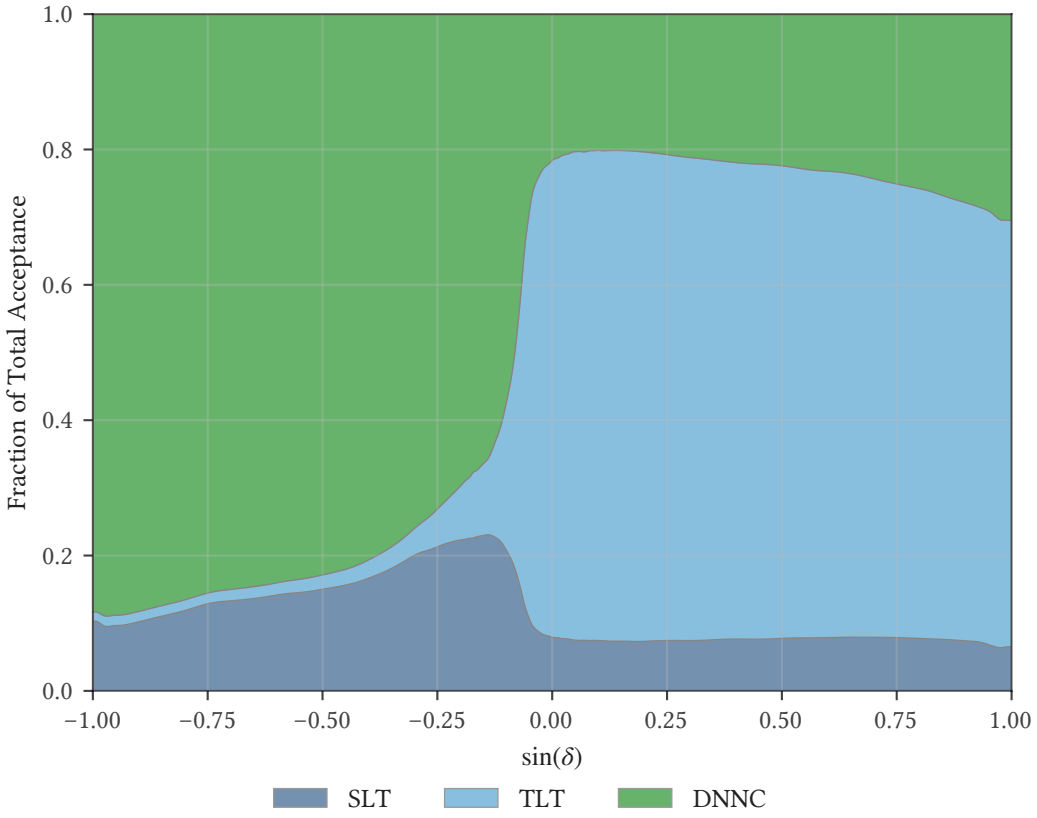

Figure 9.10: Signal acceptance $A(\gamma, \delta)$ versus declination for the nominal analysis sample, shown as stacked SLT, TLT, and DNN Cascades contributions.

9.7 Signal subtraction

As noted in Section 9.3, the background PDFs are constructed from the observed data. This raises a subtle issue: if the data contain a real astrophysical signal, that signal contaminates the background estimate. The background spatial PDF is constructed by averaging the data over right ascension. If the data contain a point source at position (α_s, δ_s), the RA-averaging smears the localized signal over all right ascensions while preserving its declination structure. This smeared contribution, $\mathcal{S}^{\text{sigsub}}$, is the RA-integrated signal PDF—the same spatial PSF and energy weighting as the full signal, but integrated over right ascension so that only the declination dependence remains.

The *signal subtraction* method corrects for this contamination. The RA-averaged data PDF $\mathcal{D}$ is fixed: it is the empirical distribution from the observed data. However, $\mathcal{D}$ is itself a mixture: if the data contain n_s signal events out of N total, then

$$\mathcal{D} = \frac{n_s}{N}\mathcal{S}^{\text{sigsub}} + \left(1 - \frac{n_s}{N}\right)\mathcal{B}, \tag{9.16}$$

where $\mathcal{B}$ is the assumed background. As n_s varies during fitting, our assumption of how $\mathcal{D}$ decomposes into signal versus background changes dynamically: $\mathcal{B}$ effectively depends on n_s. Solving for $\mathcal{B}$ and substituting into the standard per-event likelihood contribution $\frac{n_s}{N}\mathcal{S} + (1 - \frac{n_s}{N})\mathcal{B}$:

$$\frac{n_s}{N}\mathcal{S} + \left(1 - \frac{n_s}{N}\right) \cdot \frac{\mathcal{D} - \frac{n_s}{N}\mathcal{S}^{\text{sigsub}}}{1 - \frac{n_s}{N}} = \frac{n_s}{N}\mathcal{S} + \mathcal{D} - \frac{n_s}{N}\mathcal{S}^{\text{sigsub}}. \tag{9.17}$$

Rearranging:

$$\mathcal{L} = \prod_i \left[\mathcal{D}_i + \frac{n_s}{N}(\mathcal{S}_i - \mathcal{S}_i^{\text{sigsub}})\right]. \tag{9.18}$$

Factoring out $\mathcal{D}_i$:

$$\mathcal{L} = \prod_i \mathcal{D}_i \left[1 + \frac{n_s}{N} \cdot \frac{\mathcal{S}_i - \mathcal{S}_i^{\text{sigsub}}}{\mathcal{D}_i}\right]. \tag{9.19}$$

The background-only hypothesis corresponds to $n_s = 0$, giving $\mathcal{L}_0 = \prod_i \mathcal{D}_i$. The log-likelihood ratio is then:

$$\ln \frac{\mathcal{L}}{\mathcal{L}_0} = \sum_i \ln \left[1 + \frac{n_s}{N} \cdot \frac{\mathcal{S}_i - \mathcal{S}_i^{\text{sigsub}}}{\mathcal{D}_i}\right]. \tag{9.20}$$

In practice, we factorize the event weights as $W_i = W_i^{\text{space}} \times W_i^{\text{energy}}$, where $W_i^{\text{space}} = \mathcal{S}_{\text{space},i}/\mathcal{D}_{\text{space},i}$ is the spatial signal-to-data ratio (PSF divided by data-derived spatial PDF at the event position) and W_i^{energy} is the energy ratio (looked up from precomputed 2D histograms, as described in Section 9.5). The signal-subtraction weight differs from the signal weight only in the spatial factor: the RA-marginalized signal $\mathcal{S}_i^{\text{sigsub}}$ is computed at runtime by integrating the PSF over

right ascension, then divided by the data-derived spatial PDF to give $W_i^{\text{space,sigsub}}$. The energy factor W_i^{energy} is common to both. In the default *full signal subtraction* configuration, the signal-subtraction weight is $W_i^{\text{sigsub}} = W_i^{\text{space,sigsub}} \times W_i^{\text{energy}}$. An alternative *spatial-only* configuration sets $W_i^{\text{sigsub}} = W_i^{\text{space,sigsub}}$ by dropping the energy factor from the correction term entirely, which eliminates certain pathologies associated with poorly constrained energy PDFs at high energies. The log-likelihood ratio is then:

$$X_i = \frac{W_i - W_i^{\text{sigsub}}}{N}, \qquad \ln \frac{\mathcal{L}}{\mathcal{L}_0} = \sum_i \ln(1 + n_s X_i). \tag{9.21}$$

Whether it is ideal to include it or not is an analysis- and sample-specific choice. Solving Equation (9.16) for the assumed background density gives $\mathcal{B}_i = (\mathcal{D}_i - \frac{n_s}{N}\mathcal{S}_i^{\text{sigsub}})/(1 - \frac{n_s}{N})$, which turns negative once $\frac{n_s}{N}\mathcal{S}_i^{\text{sigsub}} > \mathcal{D}_i$. A negative implied background is not in itself a defect: it corresponds to a negative observed flux, which is unphysical, so any (n_s, γ) at which the signal density exceeds the data density must carry zero likelihood. Restricting the fit to the physical regime is the mathematically correct response, exactly as we enforce $n_s \geq 0$. The spatial-only correction keeps $\mathcal{S}_i^{\text{sigsub}}$ a bounded ratio of point-spread functions, so the constraint binds only where it physically should; including the energy factor does not. Both the signal and background energy PDFs are poorly sampled at high energies, where data and MC statistics are thin, and the resulting $\mathcal{S}/\mathcal{D}$ ratios are wrong: they can violate the constraint where the true densities would not. Starting tracks are the worst case, because at high energies they carry both large energy weight and large spatial weight, whereas cascades retain more moderate spatial weights. A few events with wildly inflated $\mathcal{S}/\mathcal{D}$ then distort the likelihood surface enough to push the true parameters into the forbidden region, where the implied background is negative, and the maximum-likelihood estimate stalls on the cliff between the forbidden and allowed regions, biasing $\hat{n}_s$. Regulating the PDFs directly, for instance with kernel density estimates, was tried at length, but nothing outperformed simply dropping the energy information from the correction term for SLT and TLT. For DNNC the spatial term appears to regulate the pathology well enough that keeping the energy term remains the better choice.

9.8 Null-hypothesis calibration

With the likelihood components defined, we now examine how the test statistic behaves under the null hypothesis, when no signal is present. Significance is assessed by comparing an observed TS against its background-only distribution (Section 7.2), so that distribution must be known. We obtain it empirically by generating *background trials*: pseudo-experiments that are, by construction, signal-free realizations of the background. Each trial applies *RA randomization*—drawing every event's right ascension uniformly from $\mathrm{U}(0, 2\pi)$ while preserving declinations and all other observables—and the full likelihood analysis is run on many such trials to build the empirical TS distribution.

For unbinned point-source searches, this distribution is a mixture: an atom (a point mass) at zero, when the best-fit $n_s = 0$, together with a continuous part on the positive axis, when $n_s > 0$. The mixture weight η is the observed fraction of trials with $n_s > 0$.

Asymptotic expectation

[167] Wilks 1938, "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses".

Wilks' theorem[167] (Section 7.5) predicts that, under regularity conditions, the TS asymptotically follows a χ^2 distribution with degrees of freedom equal to the number of free parameters. In our case the free parameters are n_s and γ, so the expected degrees of freedom are 2. The $n_s \geq 0$ boundary constraint produces the mixture structure: the atom at zero absorbs the trials where the unconstrained optimum would have $n_s < 0$, and the continuous part should follow χ^2_2 for the remaining trials. In practice, however, the conditions required by Wilks' theorem are not generally satisfied. Even restricting attention to the positive-TS tail (where the boundary constraint is not active and so the $n_s \geq 0$ issue does not apply), the χ^2 approximation fails for several reasons.

First, the nuisance parameter γ is not identified under the null hypothesis. When $n_s = 0$, the spectral index drops out of the likelihood entirely: it has no effect on the test statistic regardless of its value. This means the Fisher information matrix is singular at the null—violating a regularity condition of Wilks' theorem. This is a well-studied problem in statistics known as the Davies problem:[168] testing a hypothesis when a nuisance parameter is present only under the alternative. The consequence is that the effective number of degrees of freedom is not constant across the positive-TS tail. For trials with small fitted $\hat{n}_s$ (just barely above zero), the likelihood surface is nearly flat in γ. The data contain too few signal-like events to constrain a spectral index: γ is essentially noise, and only n_s carries information. In this regime the effective dimensionality is closer to 1. As $\hat{n}_s$ increases and more signal-like events contribute to the fit, γ becomes progressively better constrained, and the effective dimensionality approaches 2. Because the bulk of the positive-TS trials have small $\hat{n}_s$ (most background fluctuations produce only marginal excesses), the TS distribution is dominated by the ∼1-dof regime. This is directly visible in the fitted χ^2 degrees of freedom (Figure 9.14), which cluster around $n_{\text{dof}} \approx 1.0$–$1.2$ across declination, reflecting the core of the distribution, not the tail. The fitted n_{dof} is directly correlated with the local cut strength: in declination regions where harder cuts produce more signal-like background event distributions, n_{dof} rises (peaking at ∼1.3 near the muon horizon, where throughgoing track selections apply their most aggressive energy cuts for optimal sensitivity), while it drops closer to 1.0 in the northern sky where cuts are much less restrictive. In contrast, the truncated gamma fit to the deep tail (Figure 9.14) yields shape parameters α consistently close to 1.0. Since a χ^2_k distribution is a gamma distribution with $\alpha = k/2$ and $\theta = 2$, a tail shape of $\alpha \approx 1$ corresponds to effective dof ≈ 2, exactly what Wilks' theorem predicts when both parameters are well-constrained. The chi-squared fit therefore measures the effective degrees of freedom of the wrong part of the distribution (the core, where dof ≈ 1) and extrapolates it into the tail (where

[168] Davies 1977, "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative", Sec. 1, Davies 1987, "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternatives", Sec. 1.

dof ≈ 2)—guaranteeing a systematic mismatch. More importantly, the direction of the bias in the deep tail is controlled by the scale parameter, not the shape. A χ^2_k distribution is Gamma$(k/2, 2)$, so a χ^2 fit holds the scale fixed at $\theta = 2$ and adjusts only the shape (the dof). The empirical deep tail, however, falls off slightly faster: the truncated gamma fit recovers a scale $\theta \approx 1.9$. Forced to $\theta = 2$, the χ^2 tail decays as $e^{-T/2}$ where the data decay like $e^{-T/1.9}$, and since the exponential rate dominates the deep tail, the slower χ^2 decay places systematically more mass there. The χ^2 extrapolation therefore overestimates p-values and under-reports significance. The fitted thresholds make this concrete: for the LT + DNNC ring of Figure 9.13, the χ^2 extrapolation places the 5σ threshold at TS ≈ 32.2 while the truncated gamma places it at TS ≈ 27.5, so for a given observed TS the χ^2 assigns the larger p-value. Given the magnitude and consistency of this effect, there is strong reason to believe that previous csky-based all-sky searches that relied on χ^2 extrapolation for the TS-to-p conversion have reported systematically inflated pre-trial (per-ring) p-values.

Second, PDF mismodeling distorts the likelihood surface. Wilks' theorem requires that the likelihood be a proper likelihood, that is, that the per-event densities satisfy the axioms of probability distributions. Signal subtraction violates this: as discussed in Section 9.7, the modified per-event contribution can imply negative background densities, meaning the function being optimized is no longer a valid likelihood. The resulting log-likelihood ratio is therefore not a valid log-likelihood ratio in the sense that Wilks' theorem requires, and the χ^2 asymptotic guarantee does not apply.

Third, individual high-leverage events can produce discrete features in the TS distribution. Events with both high reconstructed energy and small angular error carry disproportionate weight in the likelihood because they simultaneously have large $\mathcal{S}/\mathcal{D}$ energy ratios and narrow spatial PSFs. When such an event lies close in declination to the tested source position, any RA randomization that places it near the test location produces a large TS contribution from that single event alone. Since the energy weight is constant for a given event and only the RA distance to the source varies across trials, the resulting feature in the TS distribution is a direct projection of the event's PSF: a bump centered at the TS value corresponding to perfect spatial alignment, with a shape determined by the logarithm of the PSF at fixed declination offset as a function of RA separation, localized to declinations near the high-weight event. The effect is most pronounced for starting samples (SLT, ESTES, DNNC), where the high-energy tail is very sparsely populated and individual events carry enormous relative weight (see Section 9.5)—as can be seen, e.g., in Figure 9.11—but it occurs to varying degrees in all samples, including throughgoing tracks (e.g., a mild bump around TS $\approx$ 6.5 for NT, also visible in Figure 9.11). Declination randomization in background trials attempts to circumvent this by smearing events across declination bands, diluting the influence of individual high-weight events. For DNNC, we use declination randomization, which is why the DNNC TS distribution plots in this chapter do not exhibit visible bumps. However, for track selections declination randomization should not be used: it introduces a significant risk of mismatch between the randomized background trials and the

unrandomized real data, invalidating the null calibration.

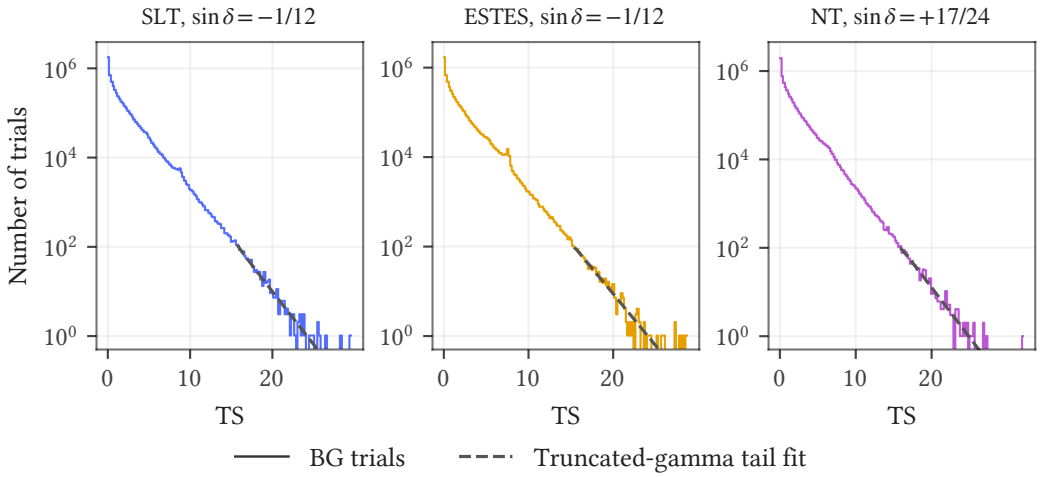

Figure 9.11: Per-sample high-leverage event bumps: fine-binned background TS distributions with the truncated-gamma tail fit, for SLT and ESTES at $\sin\delta = -1/12$ and NT at $\sin\delta = +17/24$.

These deviations are clearly visible in the background TS distributions shown below. For the purpose of determining significance thresholds, we therefore rely on empirical quantiles wherever feasible and use parametric fits only for extrapolation into the deep tail where empirical statistics are insufficient.

Tail extrapolation

Since direct empirical estimation of the 5σ threshold ($p \approx 2.87 \times 10^{-7}$) would require $\mathcal{O}(10^9)$ trials (even 10^9 trials yield only ~300 events above threshold, a ~6% statistical uncertainty on the quantile)—a parametric extrapolation is unavoidable in most cases. Two approaches are implemented.

The χ^2 *fit* fits a χ^2 distribution to all TS > 0 values. In addition to the fundamental issues discussed above, this approach has a practical problem: because the vast majority of positive-TS trials have small TS values near the core, the fit is dominated by the bulk of the distribution and has little sensitivity to the tail. As a result, χ^2 extrapolation systematically mismodels the tail, overshooting the 5σ threshold by a variable but often substantial margin. Individual-event bumps in the medium-TS range (around the 3σ threshold) further degrade the fit, since the χ^2 model has no capacity to accommodate discrete features. A χ^2 fit that appears adequate at $\mathcal{O}(10^4)$ trials may fail badly at $\mathcal{O}(10^7)$, because the tails where the fit diverges are not populated until the trial count is high enough to probe them.

The *truncated gamma fit* addresses these limitations by fitting only the tail of the distribution. Given M nonzero TS values sorted as $t_1 \le t_2 \le \cdots \le t_M$, a threshold $\tau = t_{M-k}$ is set such that exactly k trials lie above it (with a safety floor to prevent the fit from extending into the near-zero region). A gamma distribution

is then fitted via maximum likelihood to the shifted tail values $(t_i - \tau)$ for all $t_i \geq \tau$:

$$f(x;\alpha,\theta) = \frac{x^{\alpha-1}\, e^{-x/\theta}}{\theta^{\alpha}\,\Gamma(\alpha)}, \qquad x = T - \tau \geq 0, \tag{9.22}$$

where α (shape) and θ (scale) are the fit parameters. Note that a χ^2_k distribution is a special case: $\chi^2_k = \text{Gamma}(k/2, 2)$. In particular, $\chi^2_2 = \text{Gamma}(1, 2) = \text{Exp}(1/2)$, a pure exponential. The truncated gamma fit confirms this: the fitted $\alpha \approx 1.0$ across declinations recovers the expected dof $= 2$. The fitted $\theta \approx 1.9$ is close to but slightly below the χ^2_2 prediction of $\theta = 2$, reflecting the practical accommodation of the fit boundary: the top k trials include some from the transition region where the effective dof has not fully converged to 2, and the slightly reduced θ compensates (e.g. Figure 9.14). The survival function is then constructed piecewise: below τ, the empirical survival function is used directly; above τ, the fitted gamma tail is used:

$$\text{SF}(T) = \begin{cases} \text{empirical SF} & T < \tau, \\ \xi \cdot S_\gamma(T - \tau; \alpha, \theta) & T \geq \tau, \end{cases} \tag{9.23}$$

where $\xi = k/N_{\text{total}}$ is the fraction of trials above the threshold and S_γ is the gamma survival function. Continuity at τ is guaranteed because $S_\gamma(0) = 1$, so the fitted branch evaluates to ξ at the boundary, exactly matching the empirical count.

In practice, using only the top $k = 1{,}000$ nonzero TS values for the fit works well: this is deep enough in the tail that individual-event bumps at moderate TS values do not contaminate the fit, yet provides sufficient statistics for a stable two-parameter MLE. The gamma distribution models the same exponential tail decay expected from a χ^2 distribution but achieves a much better fit because it is not forced to simultaneously accommodate the core.

We adopt the truncated-gamma tail fit, introduced in a previous IceCube analysis,[169] but implement it differently from the original prescription. The original fixes a TS threshold and fits the gamma to every trial above it; we instead fix the number of tail trials k and take the threshold $\tau = t_{M-k}$ accordingly (above). At a fixed trial count the two are equivalent, but the analysis runs a variable number of trials per declination—topping up the deepest rings during unblinding until the empirical survival function reaches the observed TS—and fixing k holds the fitted region at a controlled tail depth as the trial count grows, where a fixed TS threshold would instead drift in coverage.

[169] IceCube Collaboration 2026a.

The tail fit is cross-checked against a *thinned* background TS distribution. Storing every trial from the billions generated for the deepest rings is prohibitively expensive, so the distribution is thinned where doing so is cheap: the deep tail is kept in full, while the densely populated core is dynamically subsampled, retaining only enough trials that the relative error on the survival function stays below a fixed tolerance. Because the tail—where the fit and the extrapolation operate—is left untouched, the thinned distribution reproduces the full one in the regime that matters at a small fraction of the storage.

Despite these improvements, parametric extrapolation remains an approximation. The empirical survival function is always preferred where statistics permit:

we use empirical quantiles for the median and 3σ thresholds, and the truncated gamma extrapolation only for the 5σ discovery potential. Even for the 5σ threshold, the extrapolation is only reliable when the trial count is high enough that the fitted region (k top trials) lies well within the empirical distribution: fitting to a tail that is itself poorly sampled simply moves the extrapolation problem rather than solving it. All background TS distributions in this chapter are based on at least 10^7 trials per declination.

The reliability of the tail extrapolation is particularly important for the all-sky scan (Chapter 10), where per-pixel p-values are computed at every declination and a single miscalibrated declination ring can corrupt the post-trial $-\log_{10}(p_{\max})$ distribution. For this reason, the all-sky scan aims to avoid parametric extrapolation entirely for any TS values encountered during unblinding, using exact per-ring empirical survival functions for the TS-to-p conversion (see Section 10.2). However, it is not feasible to generate enough trials to also cover the most extreme TS values encountered across the thousands of background sky scans used for the trial correction: the maximum TS over all scans can far exceed what any single ring's empirical survival function resolves. There, the truncated gamma fit is used where the empirical survival function runs out of statistics.

Remark 9.3. We report significances in the $n\sigma$ convention introduced in Section 7.2, even though the TS distribution is not Gaussian. A deficit would simply yield TS $= 0$.

Figure 9.12 shows the TS distribution at nine declinations for the combined LT + DNNC sample. The histogram shows the empirical distribution from $\mathcal{O}(10^7)$ background trials, with the fitted truncated gamma model overlaid. Vertical lines mark the median TS and the $3\sigma/5\sigma$ significance thresholds. The goodness-of-fit statistics reported for the fit, the Kolmogorov–Smirnov statistic D[170] (maximum CDF deviation) and the Anderson–Darling statistic A^2[171] (tail-weighted integrated CDF deviation), quantify the agreement between the parametric model and the empirical distribution.

[170] Smirnov 1948, "Table for Estimating the Goodness of Fit of Empirical Distributions".

[171] Anderson and Darling 1952, "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes".

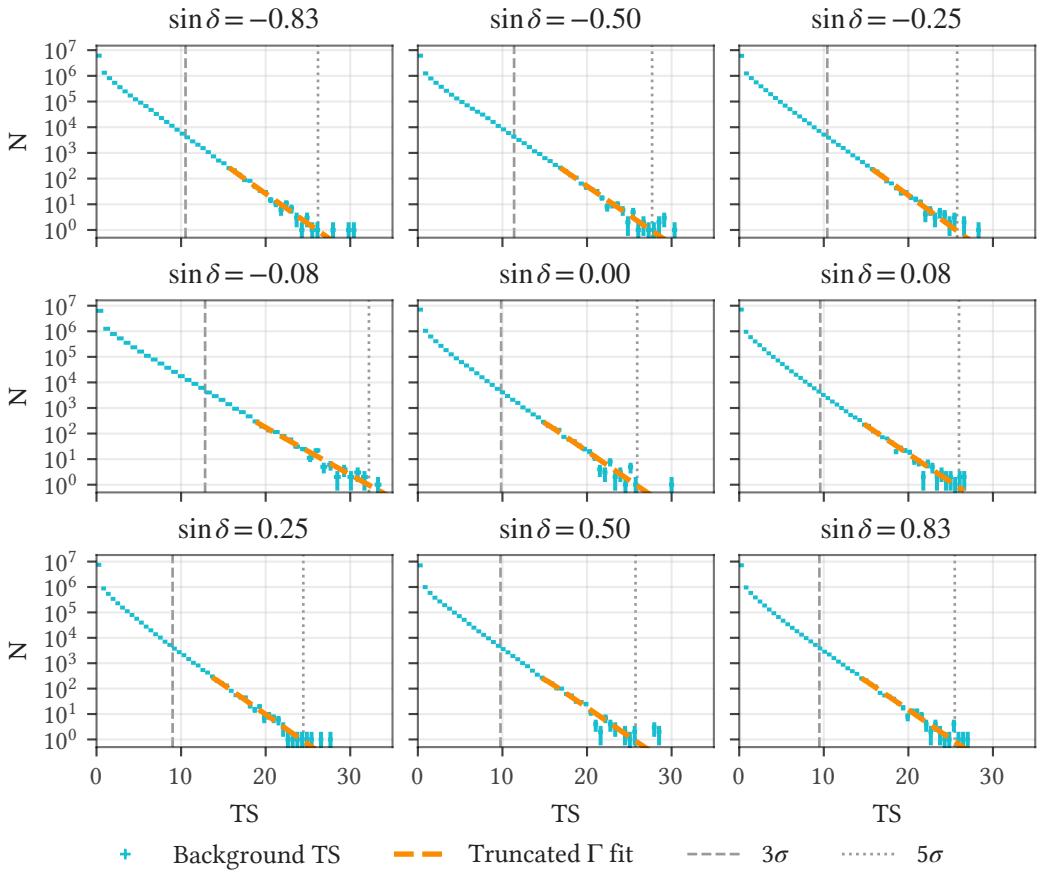

Figure 9.12: Background test-statistic distributions at nine declinations for LT + DNNC, in a 3×3 grid, with the fitted truncated-gamma model overlaid. Vertical lines mark the median and the $3\sigma/5\sigma$ thresholds.

Figure 9.13 shows the comparison directly for a single declination: both fit families on the same background-TS distribution, with the χ^2 failure visible against the truncated-gamma tail.

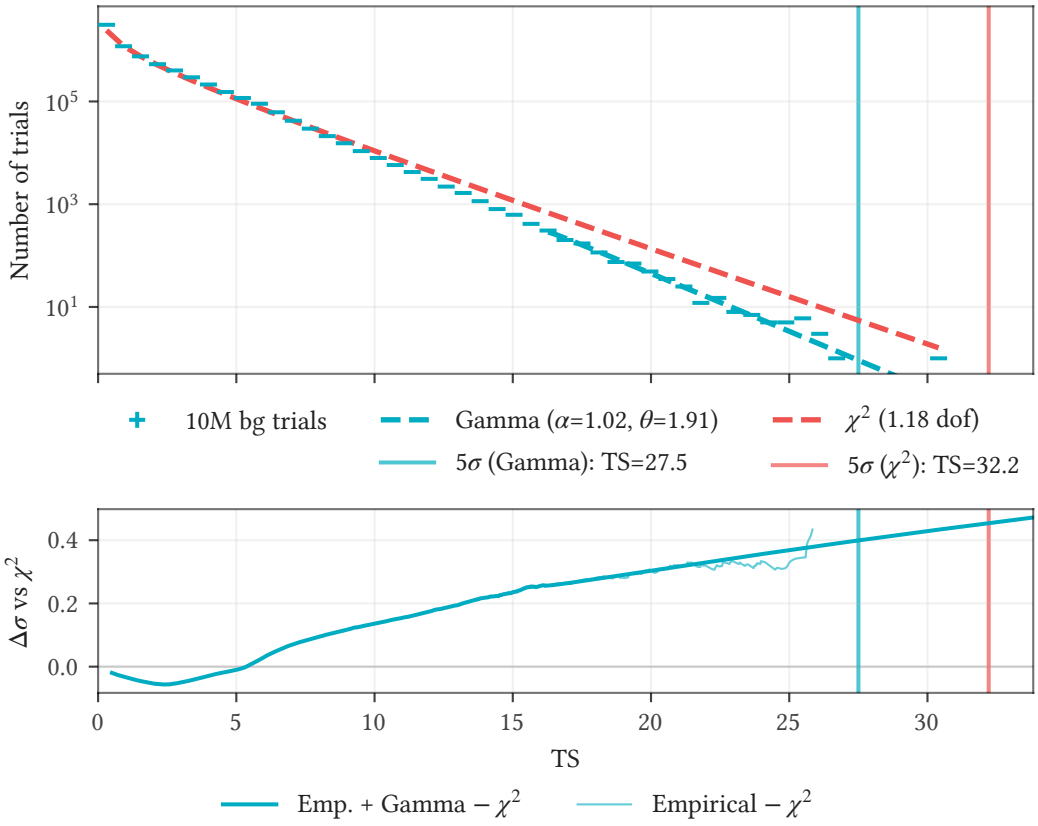

Figure 9.13: Comparison of the χ^2 and truncated-gamma tail fits for LT + DNNC at $\sin\delta = -1/3$. Top: the background TS distribution with both fits overlaid and the 5σ thresholds marked. The χ^2 extrapolation overshoots the tail while the truncated gamma tracks it. Bottom: significance difference relative to the χ^2 reference, showing the growing under-reporting of significance by the χ^2 extrapolation. The empirical curve confirms the piecewise empirical-plus-gamma calibration.

Fit parameters

Figure 9.14 through Figure 9.17 show the fitted distribution parameters as a function of declination: the effective degrees of freedom n_{dof} of the χ^2 fit together with the shape (α) and scale (θ) of the truncated-gamma tail fit, the mixture fraction η, the 3σ and 5σ thresholds, and the goodness-of-fit statistics D and A^2. Note that the 3σ *empirical vs. fit* comparison shown in Figure 9.16 is not meaningful for the truncated gamma: with $k = 1{,}000$ fit events and the trial counts used here, the fit threshold τ lies well above the 3σ TS value at all declinations, so the 3σ threshold falls entirely within the empirical region of the piecewise survival function and the *fitted* value is identically the empirical one.

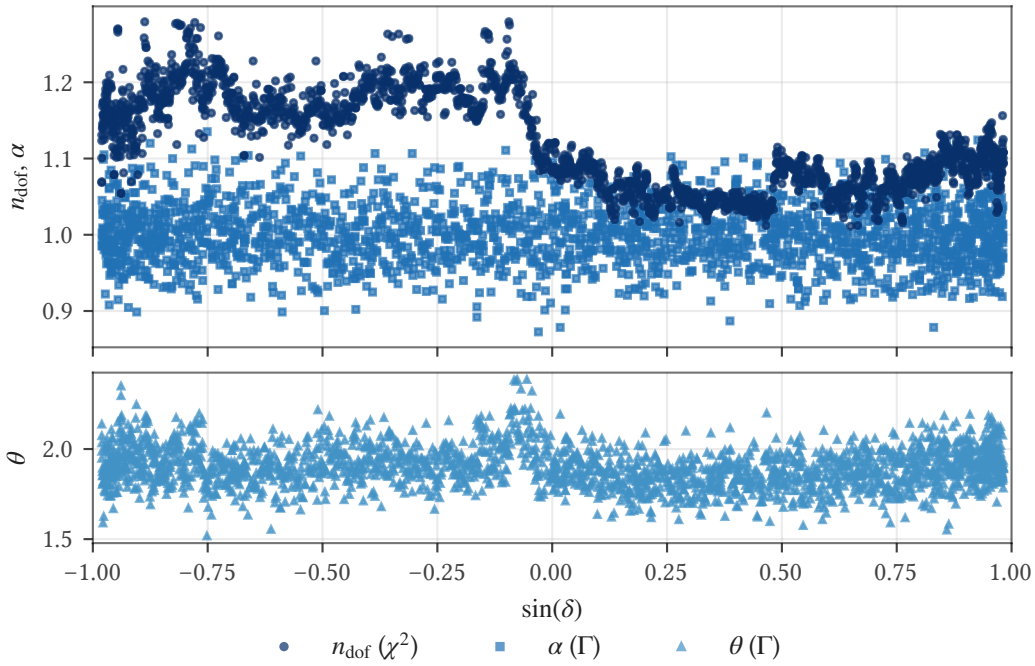

Figure 9.14: Fitted shape parameters of the background TS distributions as a function of declination for LT + DNNC. Top: the effective degrees of freedom n_{dof} of the χ^2 fit and the shape α of the truncated-gamma tail fit. Bottom: the truncated-gamma scale θ.

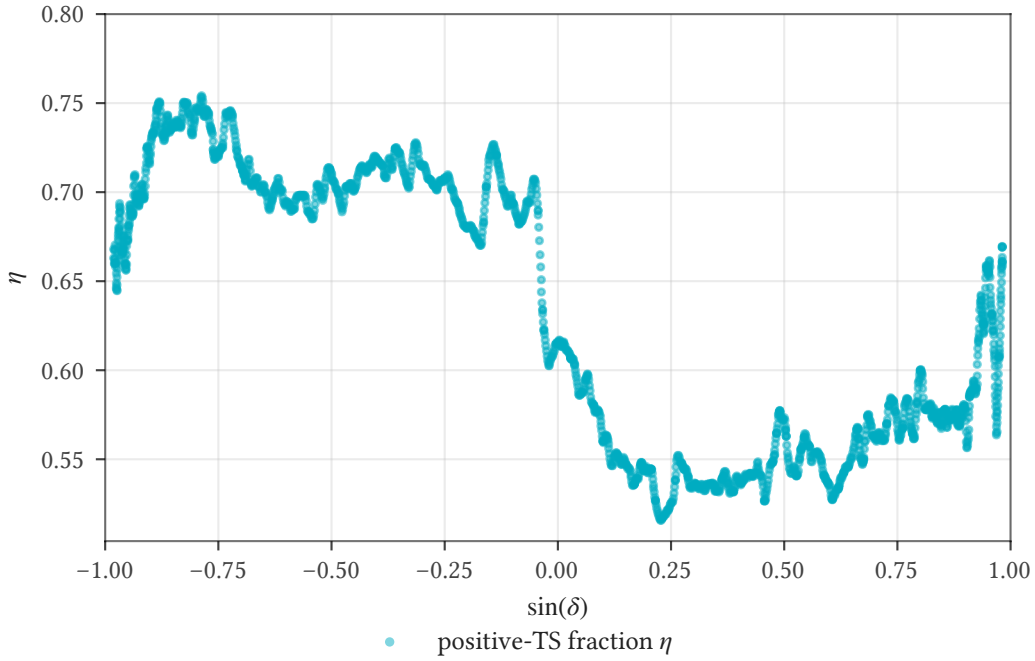

Figure 9.15: Fitted mixture fraction η of the background TS model—the fraction of background trials with best-fit $n_s > 0$, the weight of the continuous component beside the delta function at TS $= 0$—as a function of declination for LT + DNNC.

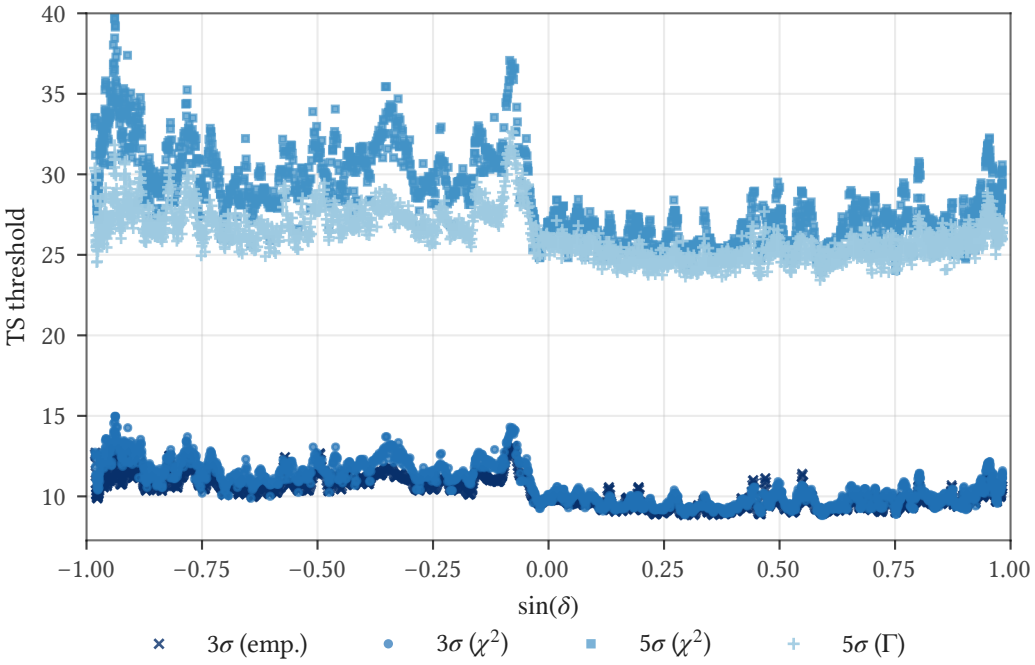

Figure 9.16: Significance thresholds as a function of declination for LT + DNNC: the 3σ TS threshold (empirical and χ^2-extrapolated) and the 5σ threshold (the χ^2- and truncated-gamma-extrapolated curves).

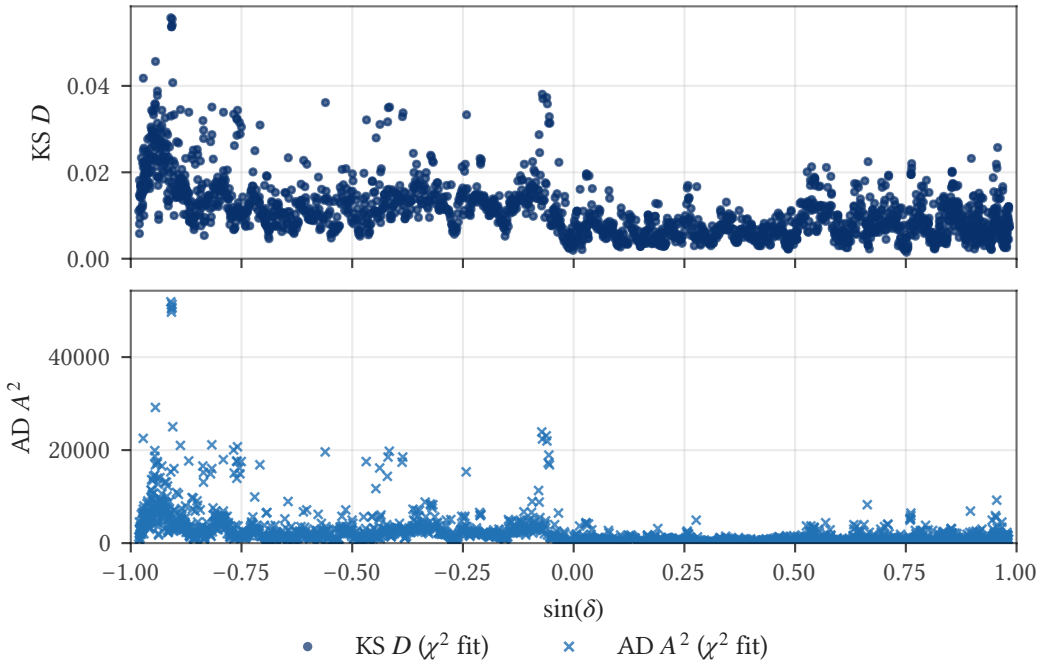

Figure 9.17: Goodness of fit of the background TS models as a function of declination for LT + DNNC. Top: the Kolmogorov–Smirnov statistic D. Bottom: the Anderson–Darling statistic A^2.

Fitted parameters from background trials

The spectral index γ and signal count n_s are free parameters in each likelihood fit, with γ constrained to the interval $[1, 4]$ and $n_s \geq 0$. In background-only trials, there is no injected signal and hence no *true* γ or n_s. The fitted values instead reflect how background fluctuations project onto the signal template. Figure 9.18 shows the median and 20–80% quantile range of the fitted γ and n_s distributions across declination, computed only from trials where the fitter found signal ($n_s > 0$). The sharp cutoff at $\gamma = 4$ is an artifact of the fit constraint, not a physical feature.

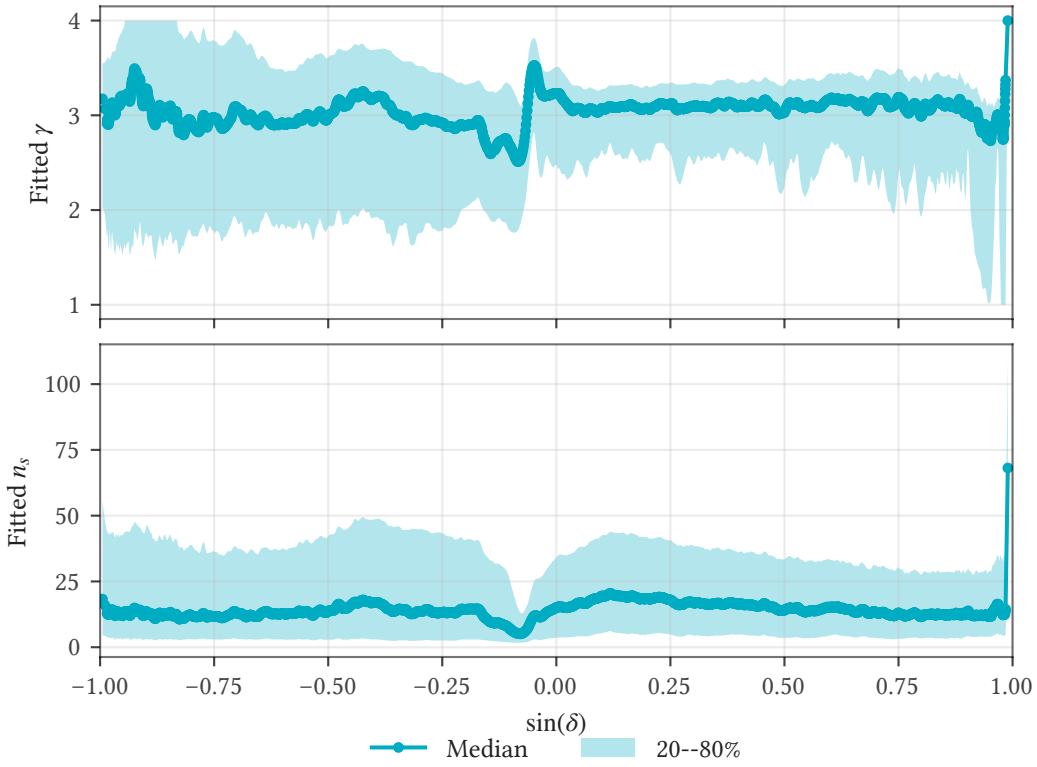

Figure 9.18: Median and 20–80% quantile range of the fitted γ and n_s across declination for LT + DNNC, from background trials with $n_s > 0$.

9.9 Validity of the empirical null calibration

The empirically sampled TS distributions introduced in Section 9.8 are used to compute the p-values for our analyses. A natural question is: under what conditions can we trust the final post-trial p-value? This section provides the rigorous answer. The proof is short, since it follows directly from the probability integral transform. The result is nevertheless important enough to warrant an explicit statement, because it establishes precisely what can and cannot affect the validity of the significance claim.

Empirical calibration guarantees valid p-values

We established the probability integral transform in Section 7.2. Here we prove it rigorously and apply it to the case at hand.

Claim 9.1. Let $x = \{(\hat{\alpha}_i, \hat{\delta}_i, \hat{E}_i, \hat{\sigma}_i, \ldots)\}_{i=1}^{N}$ be a dataset of N observed events, where each event carries a reconstructed right ascension $\hat{\alpha}_i$, reconstructed declination $\hat{\delta}_i$, reconstructed energy $\hat{E}_i$, angular error estimate $\hat{\sigma}_i$, and any number of additional observables. Let $T = T(x)$ be any scalar-valued function of these data: a likelihood

ratio, the number of events within 1° of a test position, the sum of all reconstructed energies, or the MD_5 hash of the event list modulo 10^6. Let F_0 denote the true cumulative distribution function of T under the null hypothesis H_0. Define the p-value as

$$p = 1 - F_0(T(x)). \tag{9.24}$$

Then p is uniformly distributed on $[0, 1]$ under H_0, regardless of the choice of T.

Proof. This is the probability integral transform.[172] Let $U = F_0(T)$. For any $u \in [0, 1]$:

$$P(U \leq u) = P(F_0(T) \leq u) = P(T \leq F_0^{-1}(u)) = F_0(F_0^{-1}(u)) = u, \tag{9.25}$$

which is the CDF of a Uniform(0, 1) random variable. Since $p = 1 - U$, it follows that $p \sim \text{Uniform}(0, 1)$ under H_0. □

[172] Rosenblatt 1952, "Remarks on a Multivariate Transformation", Casella and Berger 2002, *Statistical Inference*, Thm. 2.1.10, p. 54.

Remark 9.4. The probability integral transform requires F_0 to be continuous; the step $F_0(F_0^{-1}(u)) = u$ in the proof uses it. When T is discrete (for instance an event count, or the hash example above), F_0 is a step function and the transform gives $P(p \leq u) \leq u$ rather than equality, so the p-value is conservative (super-uniform) rather than exactly uniform. This is weaker but not a problem here: a conservative p-value still controls the type I error rate at any threshold and never overstates significance. The test statistics used in this analysis are continuous, so the exact-uniformity statement applies directly.

The critical observation is that this result is *entirely independent of the choice of T*. The function $T(x)$ can be a correctly specified likelihood ratio, a grossly misspecified one, or an arbitrary function of the data. The proof does not reference what T computes internally: no aspect of the likelihood, the PDFs, the reconstruction, or the rest of the analysis pipeline enters. It is a purely mathematical statement about CDFs: if we know the true null distribution of the test statistic and compute p-values from it, those p-values are uniform under the null.

Empirical realization with pseudo-experiments

In practice, F_0 is unknown and must be estimated. We do this by generating pseudo-experiments: for each pseudo-experiment, we construct a dataset $x^* = \{(\hat{\alpha}_i^*, \hat{\delta}_i, \hat{E}_i, \hat{\sigma}_i, \ldots)\}_{i=1}^N$ by redrawing each event's right ascension independently from $\hat{\alpha}_i^* \sim \text{Uniform}(0, 2\pi)$ while retaining all other observables exactly. Under the background-only hypothesis, the event rate is uniform in right ascension (see Section 9.3), so this resampling preserves the true marginal distributions of all observables while destroying any spatial clustering in right ascension, the only signature a localized point source can produce. Each x^* is therefore a valid realization of the data under H_0. The test statistic $T(x^*)$ is then computed under identical conditions (same likelihood, same PDFs, same reconstruction), as it would be for the real data.

The spatial clustering of a point source is two-dimensional (right ascension and declination), but the signal hypothesis probes this clustering via the PSF—a

two-dimensional function centered on the hypothesized source position. Breaking the clustering in one coordinate is sufficient to fully break the signal structure, because the point source signal hypothesis tests for coincidence in both coordinates via the PSF, unlike a diffuse signal hypothesis, which is itself assumed to be uniform in right ascension and therefore cannot be broken by RA resampling. What RA resampling does not break is the declination structure: if a strong source is present in the data at declination $\hat{\delta}_s$, the events from that source remain concentrated in a narrow band around $\hat{\delta}_s$ even after RA randomization. This declination-localized signal contamination is present in both the real data and the pseudo-experiments, so it is correctly captured by the empirical null calibration, but it reduces analysis power by inflating the background rate at the source declination (see Section 9.7 for the likelihood-level correction). It would be better for analysis power if this residual declination contamination were also removed, which is what MC-based background modeling achieves (see Section 9.9) and what signal subtraction attempts to partially counteract within the data-driven framework.

Let $T_1, T_2, \ldots, T_n$ denote the test statistic values from n pseudo-experiments. The empirical CDF

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(T_i \leq t) \tag{9.26}$$

converges uniformly to F_0 as $n \to \infty$ by the Glivenko–Cantelli theorem,[173] also known as the *fundamental theorem of statistics*:

[173] Vaart 1998, *Asymptotic Statistics*.

$$\sup_t |\hat{F}_n(t) - F_0(t)| \xrightarrow{\text{a.s.}} 0. \tag{9.27}$$

The p-value computed from $\hat{F}_n$ therefore converges to the true p-value, and the resulting significance estimate is valid in the sense that the type I error rate is correctly controlled at any pre-specified α. The convergence rate is $\mathcal{O}(1/\sqrt{n})$: the statistical uncertainty on the calibration shrinks with the square root of the number of pseudo-experiments.

What cannot affect the significance

The modeling imperfections below cannot bias the post-trial p-value when the null calibration is performed empirically, because they affect the pseudo-experiments and the real data in exactly the same way:

- **Energy reconstruction errors:** The energy proxy can systematically over- or underestimate the true neutrino energy, have arbitrary resolution, or bear no physical relationship to the true energy whatsoever. The null distribution of the resulting test statistic is still correctly sampled by the pseudo-experiments, and the p-values are still uniform under H_0. Whether such choices cost analysis power is a separate question. The PIT applies to the test statistic as constructed. Its proof is independent of how that construction was made.

- **Angular reconstruction and PSF mismodeling:** If the reconstructed direction is systematically biased, if the per-event angular error estimates are miscalibrated, or if the PSF functional form is wrong (whether it is modeled as von Mises–Fisher, Kent, Rayleigh, a Gaussian, a KDE, or any other function), the spatial signal PDF $\mathcal{S}_{\text{space}}$ does not match the true PSF. This changes the value of T, but pseudo-experiments use the same reconstruction and the same PSF model, so the null distribution of this changed T is still correctly sampled.

- **Energy PDF mismodeling:** The signal-to-background energy ratio $\mathcal{S}/\mathcal{D}$ may be poorly estimated due to sparse statistics, binning artifacts, or incorrect spectral assumptions (see Section 9.5). Whatever ratio is used, the pseudo-experiments use the same ratio and produce the matching null distribution: the TS landscape changes, but the null calibration remains valid.

- **Background spatial PDF errors:** Setting $\mathcal{D}_{\text{space}}$ to a constant (giving every event the same spatial weight regardless of declination) does not invalidate the p-values. As a limiting case, setting $\mathcal{D}_{\text{space}}$ to zero everywhere drives the signal-to-background ratio of every event to infinity and the TS to infinity for both the pseudo-experiments and the real data. The p-values remain formally valid (every observation ties with every pseudo-experiment) at the cost of all analysis power.

- **Signal subtraction errors:** Signal subtraction can cause the effective per-event likelihood contribution to imply negative background densities (see Section 9.7), violating the regularity conditions of Wilks' theorem.[174] This means the χ^2 asymptotic approximation fails, but it does not affect empirical calibration, which makes no assumptions about the functional form of F_0.

[174] Wilks 1938.

- **Incorrect effective area or acceptance:** These quantities enter the flux conversion directly, and they also affect the test statistic indirectly: the signal energy PDF $\mathcal{S}_{\text{energy}}$ is constructed from MC events weighted by acceptance, so biased acceptance shifts the $\mathcal{S}/\mathcal{D}$ ratio and changes the TS landscape. But once again, the pseudo-experiments use the same $\mathcal{S}/\mathcal{D}$ histograms, so the null distribution of the resulting TS is correctly sampled regardless of whether the acceptance is accurate.

- **Data-MC disagreement:** As long as the pseudo-experiments are generated from data (not from MC), any mismatch between data and simulation is irrelevant to the null calibration. The pseudo-experiments inherit the true data distribution automatically.

- **Other modeling decisions:** The list above is illustrative, not exhaustive. As established above, no choice of PDFs, reconstruction, weighting, or likelihood enters the proof of p-value uniformity under H_0, provided the null distribution is correctly estimated. Empirical calibration from pseudo-experiments that match the real data under H_0provides exactly that estimate.

The structure of the argument is uniform across these cases: the probability integral transform guarantees uniformity of p-values under H_0 for any test statistic, and empirical calibration estimates the null distribution of the test statistic actually computed, including one derived from a misspecified model. No property of the likelihood, the reconstruction, or the PDFs enters the proof.

For the all-sky scan (introduced in Chapter 10), this guarantee applies at two levels independently. The per-ring empirical survival functions calibrate the TS → pre-trial p-value conversion at each declination. But even if this first-level calibration were imperfect (even if the pre-trial p-values were not exactly uniform), the post-trial p-value would still be valid, because the trial correction is itself empirically calibrated. The background sky scan maxima distribution is built from scans that pass through the same (possibly imperfect) per-ring conversion. From the perspective of the trial correction, $\max(-\log_{10}(p_{\text{pre}}))$ is simply another test statistic, and the PIT applies to it in exactly the same way it applies to the raw TS at the per-ring level. The per-ring calibration exists to maximize analysis power, by equalizing the contribution of each declination to the global maximum (see Section 10.5), not to guarantee the post-trial p-value. That guarantee comes from the trial correction level. This two-level robustness holds only when the trial correction itself is empirical. An analytic correction such as Sidak or Bonferroni assumes that the pre-trial p-values are exactly uniform and divides the significance threshold α by the number of tests. If the pre-trial p-values are not uniform, $P(p \leq \alpha/m; H_0) \neq \alpha/m$ and the FWER is no longer controlled at the intended level. This is another advantage of the empirical approach over analytic trial correction. It also means that previous IceCube all-sky analyses that used χ^2 extrapolation[175] and declination interpolation for the per-ring TS-to-p conversion produced valid post-trial p-values, since their trial correction was empirical. The χ^2 miscalibration at the first level reduced analysis power by suppressing the pre-trial significance at most declinations (see Section 10.3), but the empirical trial correction at the second level remained valid regardless.

[175] IceCube Collaboration 2017a, "All-sky Search for Time-integrated Neutrino Emission from Astrophysical Sources with 7 yr of IceCube Data".

This assumes that the test statistic is constructed without knowledge of the real (unblinded) data. If an analyzer were to examine the unblinded data and then manipulate the likelihood, the PDFs, or the reconstruction to amplify a specific fluctuation, the resulting test statistic would no longer be independent of the data, and the pseudo-experiments (built before any such manipulation) would not sample the correct null distribution. This is the standard requirement that the analysis be defined before unblinding, and it is the reason IceCube enforces a formal unblinding procedure.

A related concern is whether modeling imperfections can produce a *false positive*: a high test statistic from a background fluctuation that would not have been high with a better model. This can happen: a misspecified likelihood reshapes the TS landscape, and some background fluctuations may project more strongly onto the misspecified signal template than they would onto a correctly specified one. This does not invalidate the p-value, however: the p-value is the probability of exactly such a fluctuation under the null. The empirical null calibration measures how often the (possibly misspecified) test statistic exceeds any given threshold

under H_0, which captures the ways that background fluctuations can masquerade as signal in the TS landscape created by whatever modeling choices were made. A background fluctuation producing $T = 20$ in a misspecified analysis is no more or less a false positive than a background fluctuation producing $T = 20$ in a perfectly specified analysis. The p-value quantifies how often each occurs under H_0, and the 5σ threshold ensures that both are sufficiently rare. What modeling imperfections *do* affect is how often a real signal produces a high test statistic (the analysis power), which is a separate question from whether the p-value is valid.

What can invalidate the calibration

The guarantee breaks down when the pseudo-experiments do not sample from the same distribution as the real data under H_0. Any systematic difference between the two means $\hat{F}_n$ converges to the wrong distribution, and the resulting p-values are no longer uniform under the true null distribution (they are still uniform under whatever distribution $\hat{F}_n$ happens to converge to, but that is not the distribution we *intend* to calibrate against):

- Parametric tail extrapolation with the wrong model: Replacing the empirical $\hat{F}_n$ with a parametric approximation (e.g., χ^2) that does not match the true tail of F_0 introduces a systematic error in the p-value that does not vanish with more pseudo-experiments. This is precisely the failure mode of the χ^2 extrapolation discussed in Section 9.8: the χ^2 fit is dominated by the core of the TS distribution (where the effective dof is ≈ 1 because of the Davies problem[176]) yet holds the deep-tail scale fixed at the χ^2 value $\theta = 2$, slightly above the empirical tail scale, so it places systematically too much mass in the tail and produces inflated pre-trial p-values. The truncated gamma fit mitigates this by fitting only the deep tail, where Wilks' theorem does apply, but any parametric extrapolation remains an approximation, which is why we use the empirical SF wherever statistics permit and validate the parametric fit against it (see Section 10.3).

[176] Davies 1977, Sec. 1, Davies 1987, Sec. 1.

- Declination randomization for tracks: Gaussian declination randomization changes the distribution of events in the pseudo-experiments relative to the real (unrandomized) data. For cascades with degree-scale angular resolution, the mismatch is small relative to the PSF and the effect is negligible; for tracks, it is not (see the declination randomization warning, Section 9.3). Within the data-driven RA-resampling scheme used to generate pseudo-experiments, this is the *only* operation that can systematically bias the empirical null calibration for a single-point test: the other failure modes listed in this section (parametric tail extrapolation and MC-based pseudo-experiments) come from replacing or augmenting the data-driven scheme rather than from a step inside it.

- MC-based pseudo-experiment generation: Using Monte Carlo simulation to generate pseudo-experiments (rather than RA-randomizing real data)

introduces data-MC disagreement as a direct source of calibration error. The pseudo-experiments then reflect the MC's approximation of the true data distribution rather than the data distribution itself. In the southern sky, where atmospheric muon simulation carries large systematic uncertainties (see Section 9.3), this can be severe. In the northern sky, where atmospheric neutrino predictions are more reliable, the effect is smaller but still present. In principle, MC-based background estimation can improve analysis power if the data-MC agreement is excellent, both because the background model captures the true declination and energy structure more smoothly than the finite data sample and because it avoids signal contamination of the background estimate entirely, unlike data-driven backgrounds, which require signal subtraction to correct for the presence of real sources in the data. In practice, the only IceCube analysis framework that uses this approach is SkyLLH with Northern Tracks,[177] where the Munich group has performed extensive validation of the data-MC agreement. For Lightning Tracks, data-MC agreement is adequate (see Chapter 6) but we would not trust it for background modeling, particularly in the southern sky where the systematic uncertainties are largest.

[177] IceCube Collaboration 2026a.

The common thread is that any operation applied asymmetrically (to the pseudo-experiments but not the real data, or vice versa) violates the premise of the probability integral transform. The calibration is valid if and only if, under H_0, the real data and the pseudo-experiments are drawn from the same distribution. For IceCube, the only way to guarantee this unconditionally is uniform RA resampling, which is equivalent to uniformly resampling the time of each event: every event retains all of its properties (energy, declination, angular error, reconstruction quality) except its right ascension, which destroys any spatial clustering while preserving the marginal distributions of all observables exactly. This is why it is often called *scrambling*.

Summary

In summary, empirical null calibration guarantees that the p-value is correct. It does not guarantee that the analysis is optimal or that the astrophysical interpretation is correct. The p-value answers the question: how often would background alone produce a test statistic this extreme or more? This question is answered by construction when the null distribution is estimated from pseudo-experiments that faithfully reproduce the real data under H_0. No property of the likelihood model enters this guarantee: not the PSF, not the energy PDFs, not the reconstruction, not the signal subtraction. What the likelihood model *does* affect is how extreme the test statistic becomes in the presence of a real signal (the analysis power) and how the fitted parameters map to physical quantities (the astrophysical interpretation). These are important, but they are entirely separate from the validity of the significance claim.

This is why the Lightning Tracks analysis uses purely data-driven background estimation, RA-only randomization for tracks, and, for the TS-to-p conversion,

empirical sampling of the TS distribution at exact declinations only (avoiding the declination interpolation used in earlier all-sky scans; see Section 10.2). Every component of the pipeline is designed to ensure that the pseudo-experiments are statistically indistinguishable from the real data under the null hypothesis. Where parametric extrapolation is unavoidable (in the deep tail of the per-ring TS distributions where empirical statistics are insufficient), it is fitted to the regime where the theoretical expectation holds (see Section 9.8) and validated against the empirical SF wherever they overlap (see Section 9.8 and Section 10.3). The sheer volume of calibration data ($\mathcal{O}(10^{10})$ total pseudo-experiments across all declination rings and $\sim 10^5$ background sky scans) exists precisely to minimize the statistical uncertainty on the null distribution and ensure that the empirical calibration is as close to exact as computationally feasible.

9.10 Signal recovery and bias

A correctly specified likelihood should recover the injected signal strength at the median: the fitted $\hat{n}_s$ should equal the true n_{inj}. Systematic deviations indicate model misspecification that can affect both detection power and flux estimation. To test signal recovery, we inject a known number of signal events n_{inj} into RA-randomized data and fit the likelihood. Repeating this many times for each n_{inj} value yields a distribution of fitted $\hat{n}_s$ values. If the likelihood is correctly specified, the median fitted $\hat{n}_s$ should lie on the diagonal $\hat{n}_s = n_{\text{inj}}$, with scatter consistent with statistical fluctuations.

Figure 9.20 shows the signal recovery diagnostic: fitted n_s (y-axis) vs. injected n_{inj} (x-axis). The solid line shows the median fitted value. Shaded bands indicate the 68% and 95% intervals across trials. The diagonal dashed line marks perfect recovery.

The fit recovers both $\hat{n}_s$ and $\hat{\gamma}$ simultaneously. At low injected signal strengths, the limited number of signal events provides little spectral information, making the soft degeneracy between n_s and γ particularly difficult to break.

As Figure 9.19 makes clear, signal recovery is far from perfect across all selections. The primary culprit is PDF mismodeling: both the spatial PSF and the energy $\mathcal{S}/\mathcal{D}$ ratio (see Section 9.5) contribute to discrepancies between the assumed and true signal-to-background distributions.

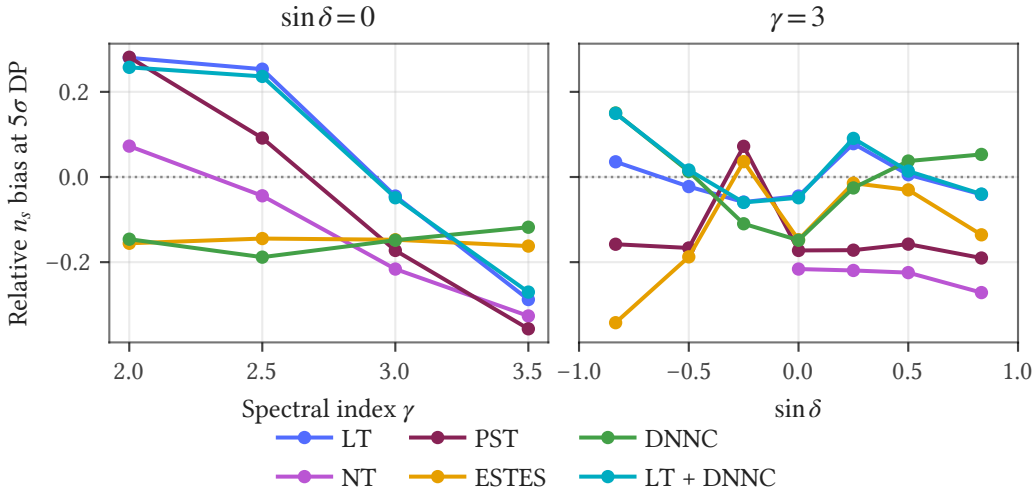

Figure 9.19: Relative n_s recovery bias at the 5σ discovery-potential strength, $\hat{n}_s/n_{\text{inj}}-1$, across samples. Left: versus spectral index γ at $\sin\delta = 0$. Right: versus $\sin\delta$ at $\gamma = 3$. Zero (dotted) is unbiased; negative is under-recovery.

The dominant signature in the recovery plots is a *spectral asymmetry*: at hard γ the median fit tracks the truth closely, while at soft γ ($\gamma \gtrsim 3$) the fit systematically under-recovers n_s (and to a lesser extent γ) across essentially all selections. This trend has a structural origin in the spatial PSF model. csky uses a von Mises–Fisher (vMF) PSF for tracks, which has lighter tails at large $\Delta\psi/\hat{\sigma}$ than the true angular-error distribution, particularly the tail contribution from the energy-dependent neutrino–lepton kinematic opening angle, which the vMF captures only as part of the calibration mixture. Because csky does not currently support a γ-dependent pull correction, a single pull-correction surface calibrated at an effective $\gamma_{\text{cal}} = 2.5$ is used as a forced compromise. At soft γ, signal events shift toward lower energies with broader true angular distributions, and live preferentially in the angular tail where the vMF model underpredicts. With model density $\mathcal{S}^{\text{model}}/\mathcal{B} \approx 0$ in that tail, the MLE treats real tail signal events as background-like, and $\hat{n}_s$ under-recovers accordingly. At hard γ, signal is dominated by high-energy events with small $\hat{\sigma}$, the tail mismodeling matters less, and the bias is correspondingly smaller, hence the spectral asymmetry.

This diagnosis is supported by an external comparison. The published Northern Tracks point-source analyses[178] are run with the SkyLLH framework, using γ-conditioned KDE-based PSFs[179] that fold the spectral dependence of the angular distribution into the spatial term explicitly. These analyses do not show the soft-γ under-recovery trend, even though our csky-based treatment of the same Northern Tracks data does: the recovery bias is visible across all csky-based selections here, NT included. The contrast strongly supports the conclusion that the bias is driven by the single-γ vMF PSF model rather than by reconstruction or simulation issues, which would persist across frameworks.

Ultimately, signal recovery is primarily a performance diagnostic, not a validity

[178] IceCube Collaboration 2022a.

[179] IceCube Collaboration 2026a, Sec. 3 & Fig. 7.

diagnostic. As long as the TS is calibrated empirically from background trials, accurate significance estimates do not require perfect PDF agreement, and the bias visible in these plots does not invalidate the actual parameter inference. Source flux and spectral-index inference are not performed by reading off the fitted MLE values directly: we use a Feldman–Cousins construction on a denser simulation grid built with the same pseudo-experiment procedure (see Chapter 12), which produces a bias-corrected point estimate and a frequency-correct 2D confidence region in (n_s, γ) without requiring the MLE to be unbiased. The recovery plots show what the MLE does at each truth point. The FC framework inverts that knowledge to recover the truth from an observation. Better PDF agreement narrows the resulting confidence region (better statistical power) but is not a prerequisite for inference validity. Where signal recovery and PDF mismatch do leak into the inference is in the *physical interpretation* of the recovered (n_s, γ) as an astrophysical flux: the FC region has guaranteed coverage of θ under the simulated signal model, but data-MC mismatches on signal acceptance shift the mapping between that θ and the true astrophysical flux. This separation between statistical validity (construction-level, robust to PDF mismatch) and physical interpretation (model-dependent) is discussed in detail in Section 12.6.

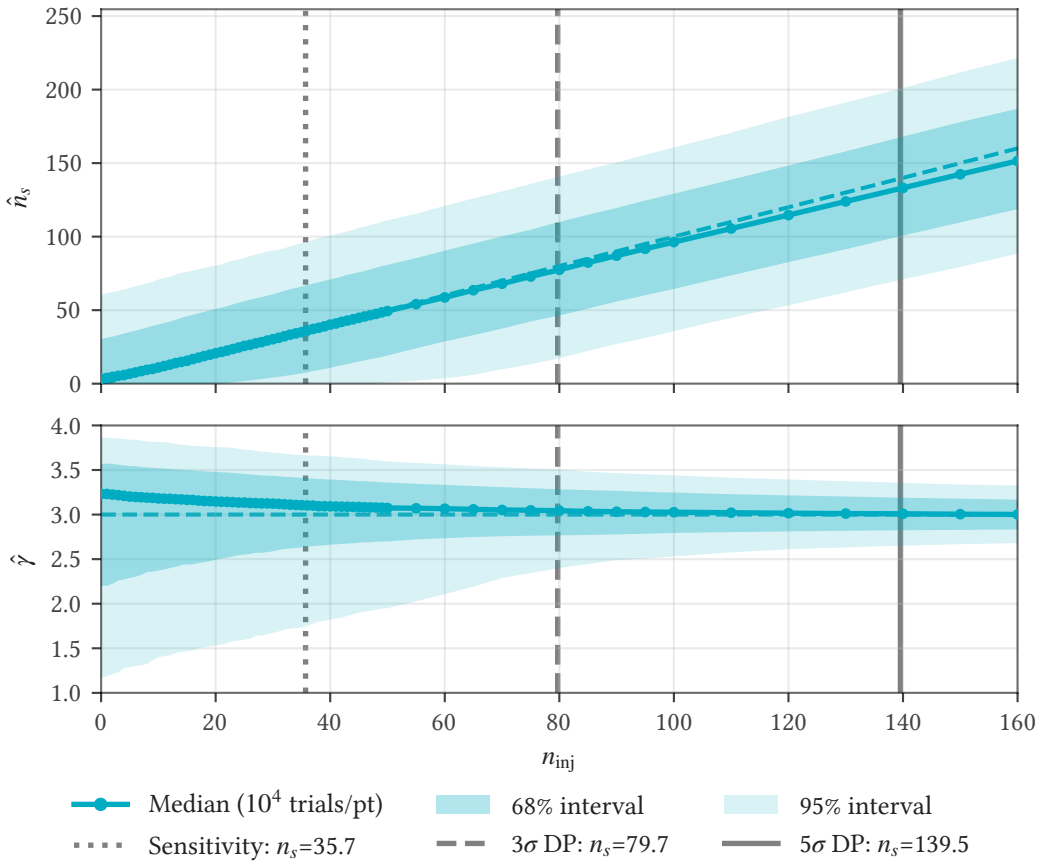

Figure 9.20: Signal-recovery diagnostic at true $\gamma = 3$, $\sin\delta = 0$, for LT + DNNC: fitted n_s versus injected n_{inj}, with the recovered spectral index $\hat{\gamma}$ in the companion panel. Solid line, median fitted value. Markers, the simulated truth points. Shaded bands, 68% and 95% intervals across trials. Dashed diagonal, perfect recovery. This is one example; the diagnostic can be produced for any truth γ and declination.

9.11 Sensitivity and discovery potential

The primary figures of merit for a point-source selection are *sensitivity* (the minimum source flux detectable at 90% confidence) and *discovery potential* (the flux required to achieve a given significance threshold, 3σ or 5σ, in 50% of experiments). These metrics integrate all aspects of the analysis: effective area, angular resolution, energy resolution, background rate, and the accuracy of the likelihood model. In both cases, lower flux values indicate superior performance. For caveats on the reliability of these estimates at hard spectral indices ($\gamma \lesssim 2$), see Section 9.5.

Integrated sensitivity

The standard sensitivity and discovery potential assume a steady point source emitting neutrinos with a power-law energy spectrum $E^{-\gamma}$ extending over all energies. The calculation proceeds by injecting simulated signal events matching the assumed spectrum on top of RA-randomized data and finding the signal strength at which a specified fraction of trials exceeds a TS threshold:

- **Sensitivity (90% CL):** the signal strength at which 90% of trials exceed the median background TS. The median is determined empirically from background-only trials.
- 3σ **discovery potential (50% CL):** the signal strength at which 50% of trials exceed the TS corresponding to a p-value of 1.35×10^{-3}. This threshold is determined empirically from background trials.
- 5σ **discovery potential (50% CL):** the signal strength at which 50% of trials exceed the TS corresponding to $p = 2.87 \times 10^{-7}$. Because this threshold lies deep in the tail, it is extrapolated from the fitted truncated-gamma tail model rather than determined empirically.

Signal injection proceeds as follows: MC events are pre-filtered to a declination band around the source and assigned weights proportional to $w_{\text{one}} \times E_\nu^{-\gamma}$ (see Section 9.6 for the definition of w_{one}). The injection uses only the spectral shape: no flux normalization is assumed at this stage. To inject n_{inj} signal events, we normalize the weights to probabilities and draw n_{inj} events by weighted rejection sampling. The drawn events are then rotated to the source position and added to the RA-randomized data. The declination band has a fixed full width of ≈ 0.035 in $\sin\delta$, which corresponds to very different angular extents depending on declination: approximately $\pm 1.0°$ at the horizon, $\pm 1.2°$ at $\delta = -30°$, and $\pm 5.8°$ at $\delta = -80°$ due to the $1/\cos\delta$ Jacobian. While the rotation places each injected event at the correct sky position, it retains the detector response (effective area, angular resolution, reconstruction bias) from its original simulated declination. The injected signal therefore does not perfectly represent the detector's response to a true source at the target position. It is an approximation that relies on detector properties varying slowly across the band. Near the poles, where the band spans more than 10° in declination, this assumption is considerably less justified than near the equator, adding a systematic uncertainty to all injection-based estimates that is difficult to quantify.

For combined samples such as LT (which combines SLT and TLT), we compute the acceptance of each component at the source declination and distribute the requested n_{inj} signal events among components by a multinomial draw weighted by relative acceptance. A component with three times the acceptance of another receives, on average, three times as many injected events.

Once the threshold n_{inj} is found, it is converted to a *physical flux* using the acceptance. Since $A(\gamma)$ gives the expected event count per unit flux normalization,

the flux that produces n_{inj} events is simply $\Phi_0 = n_{\mathrm{inj}}/A(\gamma)$. Evaluated at pivot energy E_0:

$$\left.\frac{\mathrm{d}N}{\mathrm{d}E}\right|_{E_0} = \frac{n_{\mathrm{inj}}}{A(\gamma)}\, E_0^{-\gamma}. \tag{9.28}$$

IceCube conventionally reports the *energy-weighted flux* $E^2\,\mathrm{d}N/\mathrm{d}E$ rather than the differential flux itself:

$$E_0^2 \left.\frac{\mathrm{d}N}{\mathrm{d}E}\right|_{E_0} = \frac{n_{\mathrm{inj}}}{A(\gamma)}\, E_0^{2-\gamma}. \tag{9.29}$$

For combined samples, $A(\gamma)$ is summed over all components.

Remark 9.5. The $E^2\,\mathrm{d}N/\mathrm{d}E$ convention is widespread in astronomy but somewhat arbitrary. For $\gamma = 2$, the pivot-energy dependence vanishes: $E_0^{2-\gamma} = 1$, so the reported value is simply n_{inj}/A regardless of E_0. For $\gamma \neq 2$, changing the pivot energy shifts the numerical value by a factor of $E_0^{2-\gamma}$. In log-space, this is a constant vertical offset of $(2-\gamma)\log E_0$. The choice of pivot energy is therefore meaningful only when comparing fluxes at different spectral indices, and one should always verify that the same E_0 was used before comparing numerical values across analyses. A more natural choice would be to report $E^\gamma\,\mathrm{d}N/\mathrm{d}E$, which equals n_{inj}/A for any spectral index and eliminates the pivot-energy ambiguity entirely—but this is not the convention that has taken hold.

Figure 9.21 shows the 5σ discovery potential as a function of source declination for the individual selections (SLT, TLT, LT, PST, NT, ESTES, and DNNC) together with the LT + DNNC combination, for a hard spectrum ($\gamma = 2$, upper panel) and a soft spectrum ($\gamma = 3.5$, lower panel). In the Northern sky ($\sin\delta > 0$), propagation through the Earth suppresses atmospheric muons while simultaneously introducing energy-dependent attenuation of the neutrino flux at the highest energies. Southern-sky performance depends more heavily on starting-track vetoes and self-veto effects. Softer assumed spectra ($\gamma \gtrsim 2.5$) shift the signal distribution to lower energies, which have poorer angular resolution but higher statistics. They also reduce the signal-to-background ratio because the atmospheric background spectrum is softer than a typical astrophysical signal.

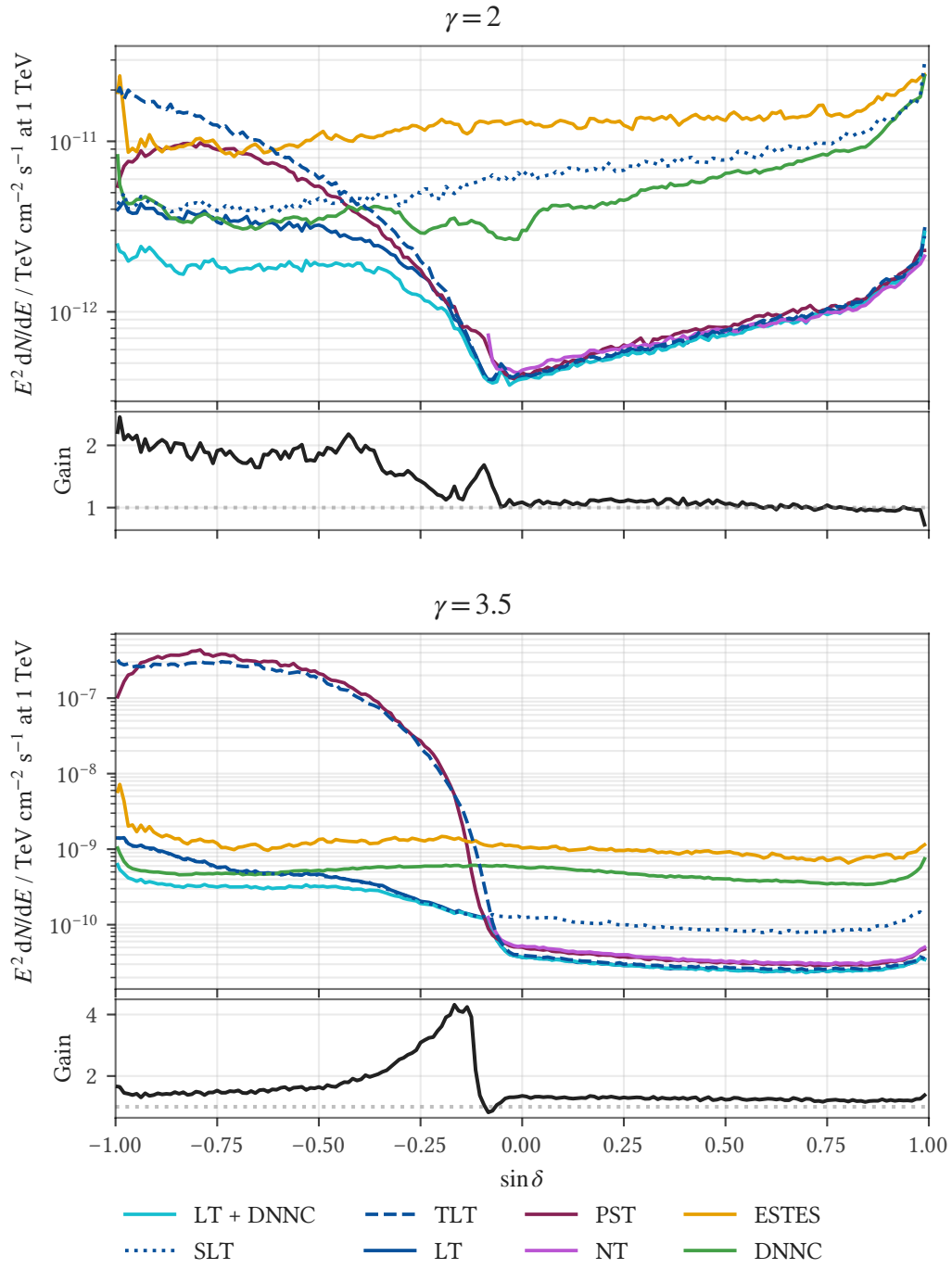

Figure 9.21: Discovery potential (5σ) flux as a function of source declination, comparing the LT + DNNC combination against the individual selections SLT, TLT, LT, PST, NT, ESTES, and DNNC, for a hard spectrum ($\gamma = 2$, top) and a soft spectrum ($\gamma = 3.5$, bottom), with the flux quoted at a 1 TeV pivot. The sub-panel under each block shows the ratio of the most sensitive of PST, NT, ESTES, and DNNC to LT + DNNC; above 1 means LT + DNNC is the most sensitive sample.

Figure 9.22 restricts the comparison to the northern sky, where NT has long

been the standard point-source instrument, and adds LT alone next to LT + DNNC, for a hard ($\gamma = 2$) and a soft ($\gamma = 3.5$) spectrum. In both, the combined curve sits on top of the LT curve: the cascade component adds nothing in the north, where the through-going tracks dominate.

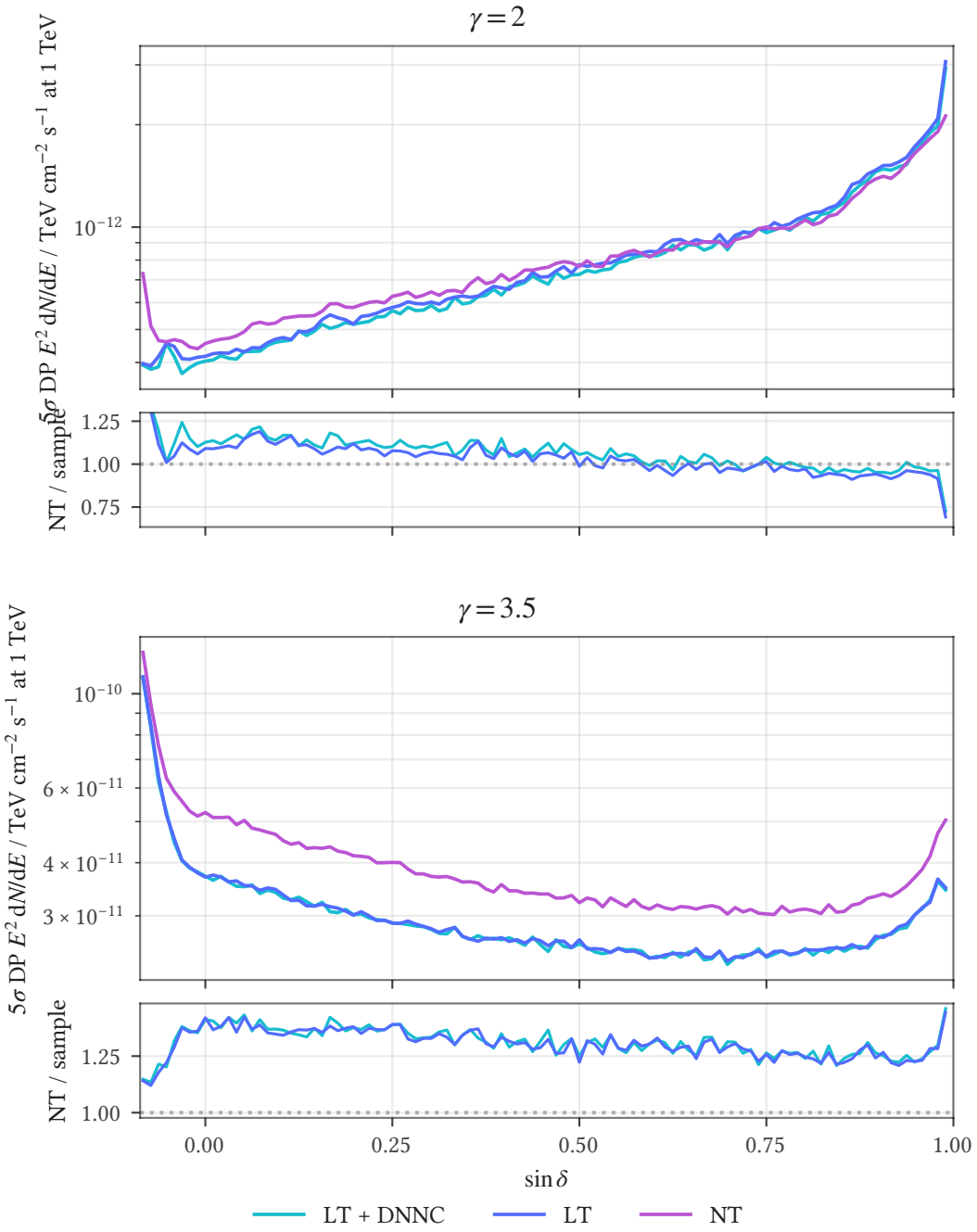

Figure 9.22: Discovery potential (5σ) flux as a function of source declination in the northern sky ($\sin\delta$ from $\sin(-5°)$, the LT/NT zenith cutoff at 85°, to 1) for LT + DNNC, LT alone, and NT, for a hard spectrum ($\gamma = 2$, top) and a soft spectrum ($\gamma = 3.5$, bottom), with the flux quoted at 1 TeV. The ratio panel beneath each shows LT + DNNC and LT relative to NT; values above 1 indicate improvement over NT. The LT + DNNC curve lies on top of the LT curve: the cascade component contributes nothing in the northern sky.

These comparisons are worth dwelling on, because in many ways they are

the central methodological contribution of this dissertation. Combining LT, the strongest track selection, with DNNC, the strongest cascade selection, gives the best point-source power available in every declination regime: the tracks carry the northern sky and the equatorial band, the cascades add reach in the deep south, and the combination is never worse than the better of its two parts.

Reading the size of these improvements requires care with the ratio convention. Each ratio is the flux a competing selection needs to reach a given significance divided by the flux LT + DNNC needs, so a ratio above one means LT + DNNC reaches the same significance with *less* flux. In the southern sky the improvements are large: at $\gamma = 3.5$ the ratio peaks at 4.3 near the muon horizon ($\sin\delta \approx -0.17$), where the combination reaches a given significance with under a quarter of the best single selection's flux, and even in the deep south at $\gamma = 2$ the ratio is 2.4 at $\sin\delta \approx -0.99$, less than half the flux. This is the regime Lightning Tracks was built for.

The northern improvements look smaller, but they may be the more consequential. At the horizon LT + DNNC reaches 5σ with about 71% of the flux Northern Tracks requires at $\gamma = 3.5$—Northern Tracks needs roughly 1.4× as much—narrowing to about 89% at $\gamma = 2$. The absolute sensitivity is far better in the north than in the south, so the same fractional improvement buys more real reach there: trimming the required flux by a third against the established northern-sky instrument moves the achievable limits more, in absolute terms, than the larger southern factors applied to a much weaker baseline.

It is worth being precise about where the combination actually helps. The combination improves on the best single selection only in the declination range where LT and DNNC have comparable power. Elsewhere the LT + DNNC curve simply follows its stronger component. The large southern improvement quoted above, the factor-of-4.3 peak near the muon horizon, is carried by LT alone, with no measurable cascade contribution there. DNNC adds reach only where its power approaches LT's, and only in that range is the joint selection genuinely better than either part on its own.

Figure 9.23 compares the three scan configurations used in the all-sky analysis; its ratio panel shows the improvement from combining LT and DNNC over the better of the two individual samples, which is meaningful only in the southern sky. Figure 9.24 quantifies the impact of overlap removal on the LT sample: the dashed curves show LT with DNNC-overlapping events removed, as used in the analysis (see the samples discussion, Section 9.1). The ratio panel shows no measurable impact of the overlap removal on LT sensitivity: the fluctuations in the sensitivity ratio are consistent with the 2.5% statistical uncertainty of the sensitivity estimate itself.

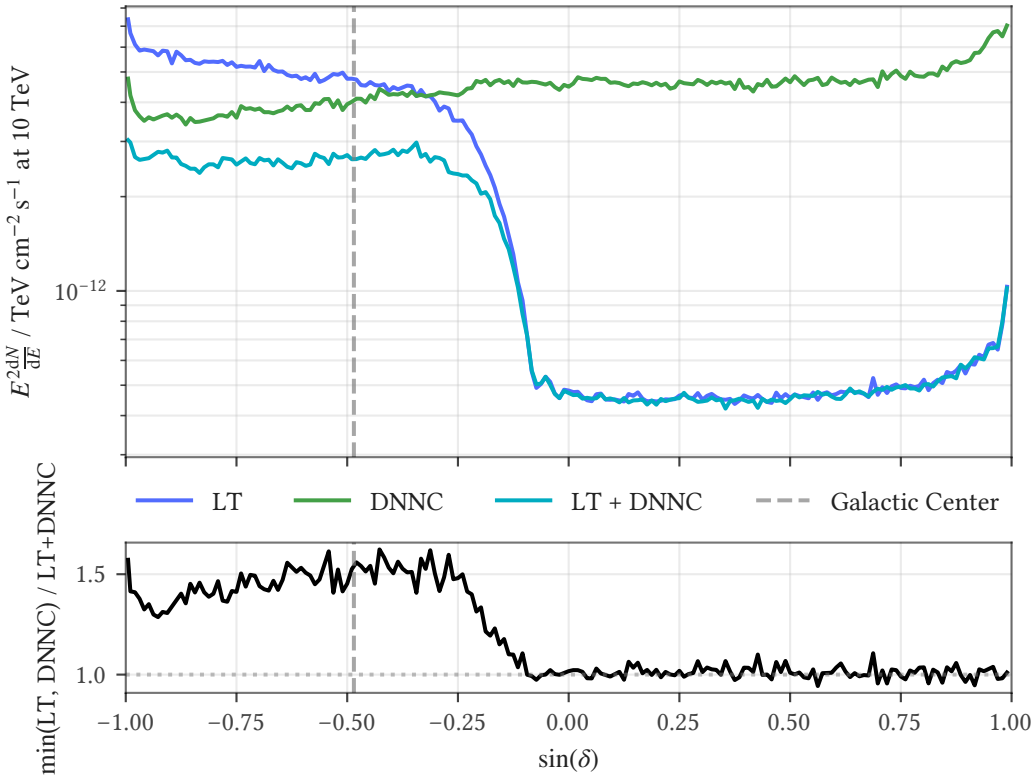

Figure 9.23: Sensitivity for the all-sky scan samples (LT, DNNC, LT + DNNC) assuming a point source with an $E^{-\gamma}$ spectrum. The ratio panel shows min(LT, DNNC)/(LT + DNNC).

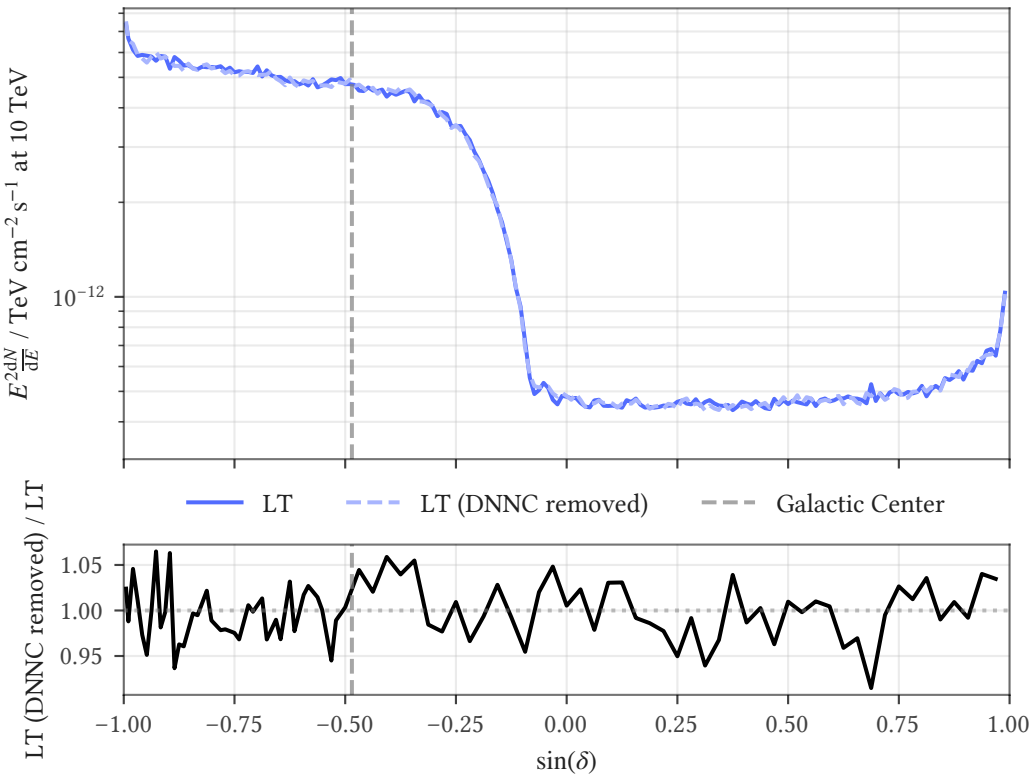

Figure 9.24: Overlap-removal impact on sensitivity: LT (full) vs. LT (DNNC overlap removed). The ratio panel shows LT (DNNC removed) / LT. Dashed lines indicate the overlap-removed version.

Differential sensitivity

Integrated sensitivity assumes a power-law spectrum extending over all energies, which is convenient for comparing selections but does not reveal how sensitivity varies with energy. *Differential sensitivity* attempts to address this by computing the sensitivity within restricted energy bins, yielding a spectrum-like curve that shows the minimum detectable flux as a function of energy.

The procedure mirrors integrated sensitivity, but signal injection is restricted to a specific energy range. For each bin $[E_{\min}, E_{\max}]$, we construct a flux hypothesis that is nonzero only within that bin:

$$\Phi(E) = \begin{cases} \Phi_0 E^{-\gamma} & \text{if } E \in [E_{\min}, E_{\max}] \\ 0 & \text{otherwise} \end{cases} \tag{9.30}$$

Signal injection then draws only from MC events whose true neutrino energy falls within the bin, and the bin-restricted acceptance is

$$A(\gamma; E_{\min}, E_{\max}) = \frac{\tau}{\Omega} \sum_{E_{\min} \le E_{\nu,i} \le E_{\max}} w_{\text{one},i}\, E_{\nu,i}^{-\gamma}. \tag{9.31}$$

The likelihood function itself is unchanged: it still uses the full energy PDF spanning all energies, exactly as in integrated sensitivity. Only the signal injection is restricted. The analysis itself remains agnostic to the energy range of the injected signal. Events from any energy can contribute to the test statistic, regardless of which bin is being tested. This means background trials can be reused across all energy bins, since the TS distribution under the null hypothesis depends only on the RA-randomized data and the likelihood model, neither of which changes between bins.

For each bin, the threshold n_{inj} is converted to flux using the bin-restricted acceptance:

$$E_0^2 \left.\frac{\mathrm{d}N}{\mathrm{d}E}\right|_{E_0} = \frac{n_{\mathrm{inj}}}{A(\gamma; E_{\mathrm{min}}, E_{\mathrm{max}})} E_0^{2-\gamma}. \tag{9.32}$$

As in the integrated case (Section 9.11), the *pivot energy* E_0 is the reference energy at which the flux is evaluated. For $\gamma = 2$, the factor $E_0^{2-\gamma} = 1$, so the pivot energy drops out entirely and the reported value is simply n_{inj}/A. For $\gamma \neq 2$, the choice of pivot energy affects the numerical value (though not the physical content: it is just a vertical shift in log-space).

We adopt the convention of evaluating the flux at the *lower bin edge*: $E_0 = E_{\mathrm{min}}$. This is a common choice in IceCube analyses and aligns with the interpretation that the reported value is an upper limit on the flux starting from that energy. For narrow bins (quarter-decade, spanning a factor of ~ 1.78 in energy), the choice of pivot energy within the bin makes little practical difference.

Figure 9.25 shows the differential sensitivity and discovery potential, comparing selections within each group at a fixed declination. We use quarter-decade (0.25 dex) energy bins from 10 GeV to 10^8 GeV, assuming a fixed spectral index γ within each bin.

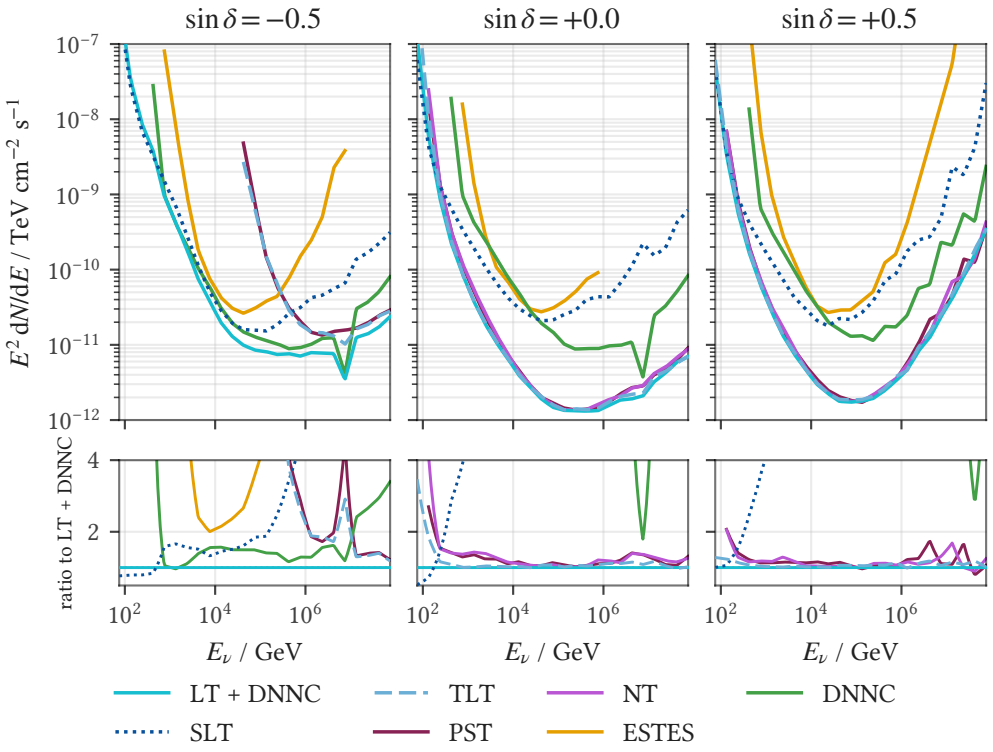

Figure 9.25: Differential sensitivity, comparing selections within each group at a fixed declination. Quarter-decade energy bins, with a fixed spectral index γ assumed within each bin. Each curve is linearly interpolated in log-log across sparsely populated or empty energy bins; the interpolation is internal only, and no curve is extended beyond the populated energy range of its selection.

9.12 Decomposing point-source power: spatial versus energy

The point-source likelihood draws on two kinds of information about each event: where it points (the spatial term) and how much energy it deposited (the energy term). A natural question is how much each contributes to the analysis's discriminating power, and in particular whether the energy reconstruction needs to be precise. The tests below use the primary analysis sample, the LT + DNNC combination; because they are run at the celestial equator ($\sin\delta = 0$), where the cascade (DNNC) component contributes negligibly (Figure 9.22), they effectively probe LT alone. They show that point-source power is dominated by the spatial term across most of the spectral range of interest. This is why the energy reconstruction in this work was kept deliberately simple: a more elaborate energy estimate would buy little point-source power. Throughout, it is power and parameter estimation that are at stake, never validity. The empirical background estimate (Section 9.3) absorbs any energy mismodeling, so p-values remain correct regardless of how well the energy is reconstructed.

Removing the energy term

The cleanest way to isolate the spatial contribution is to remove the energy term from the likelihood entirely, by fixing the signal-to-background energy ratio to unity ($\mathcal{S}_E/\mathcal{B}_E = 1$) so that the energy carries no weight. Comparing the analysis power with and without the energy term, as a function of the assumed signal spectral index γ at the celestial equator, gives the decomposition directly (Figure 9.26).

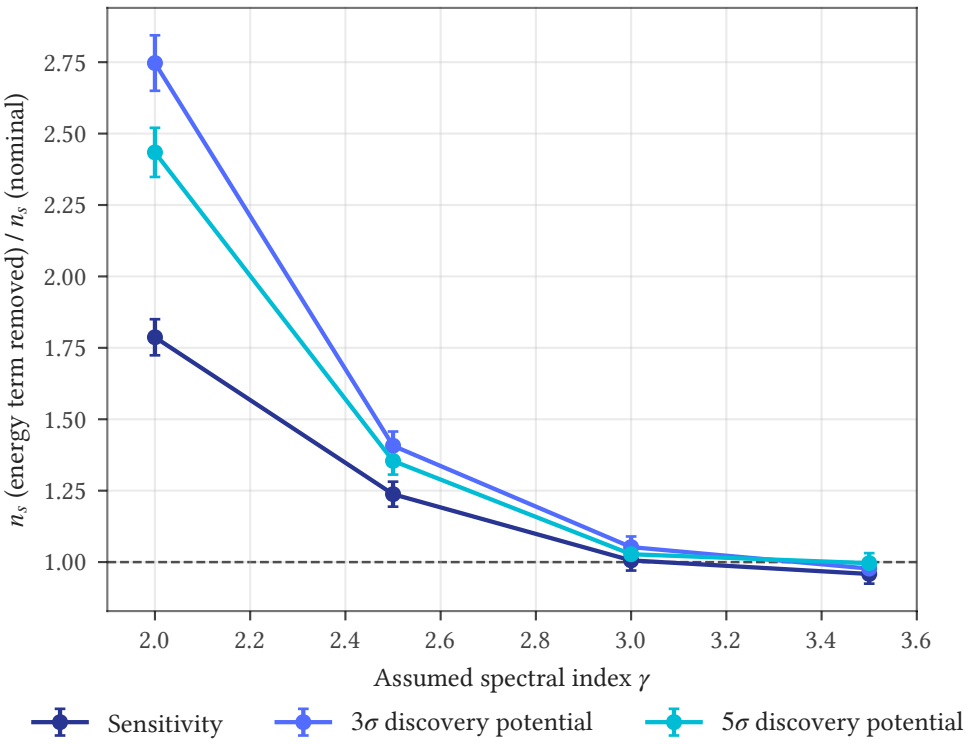

Figure 9.26: Cost of removing the energy term at $\delta = 0$: the ratio of the signal strength required without the energy term ($\mathcal{S}_E/\mathcal{B}_E = 1$) to the value with it, as a function of the assumed spectral index γ, for the sensitivity and the 3σ and 5σ discovery potentials. A ratio near unity means the energy term adds little; the ratio rises toward harder spectra. Error bars are propagated from the 2.5% statistical-error convergence threshold on each n_s, added in quadrature for the ratio (numerator and denominator treated as independent): $\sqrt{2} \times 2.5\% \approx 3.5\%$.

For very soft spectra the energy term contributes essentially nothing, and the power is entirely spatial. Near the atmospheric-background spectral index, removing the energy term actually *improves* power: at that index the signal and background energy distributions are nearly identical, so the energy term mostly adds an unidentified degree of freedom, and fixing it is equivalent to fixing γ to the truth. The energy term earns its place only toward harder spectra, and even there its contribution is modest. At $\gamma = 2.5$, removing the energy information costs only about ∼30% of the power, so roughly ∼70% of the point-source power

is spatial. The spatial and energy contributions approach parity only at the very hardest spectra.

This behavior has a simple origin in the Neyman–Pearson structure (Section 7.3). A PDF that is identical under the null and the alternative enters the likelihood ratio as unity and therefore carries no discriminating power. When the source spectral index matches the background index, the signal and background energy PDFs coincide and the energy term is exactly this powerless factor. The spatial analog is the isotropic alternative: a spatially uniform signal makes the point-spread function powerless, which is precisely the diffuse case. With $\mathcal{S}_E/\mathcal{B}_E$ fixed to unity the likelihood is flat in γ, so the spectral index is unidentified, with no gradient to fit; the spatial term nonetheless delivers power on clustering alone, and both the fitted signal strength $\hat{n}_s$ and the test statistic can remain large.

Removing low-energy events

A complementary test removes events below a threshold on the reconstructed energy, raising the MuEX energy threshold and tracking the effect on effective area (Figure 9.27) and sensitivity (Figure 9.28). This probes how much of the point-source power is driven by low-energy events.

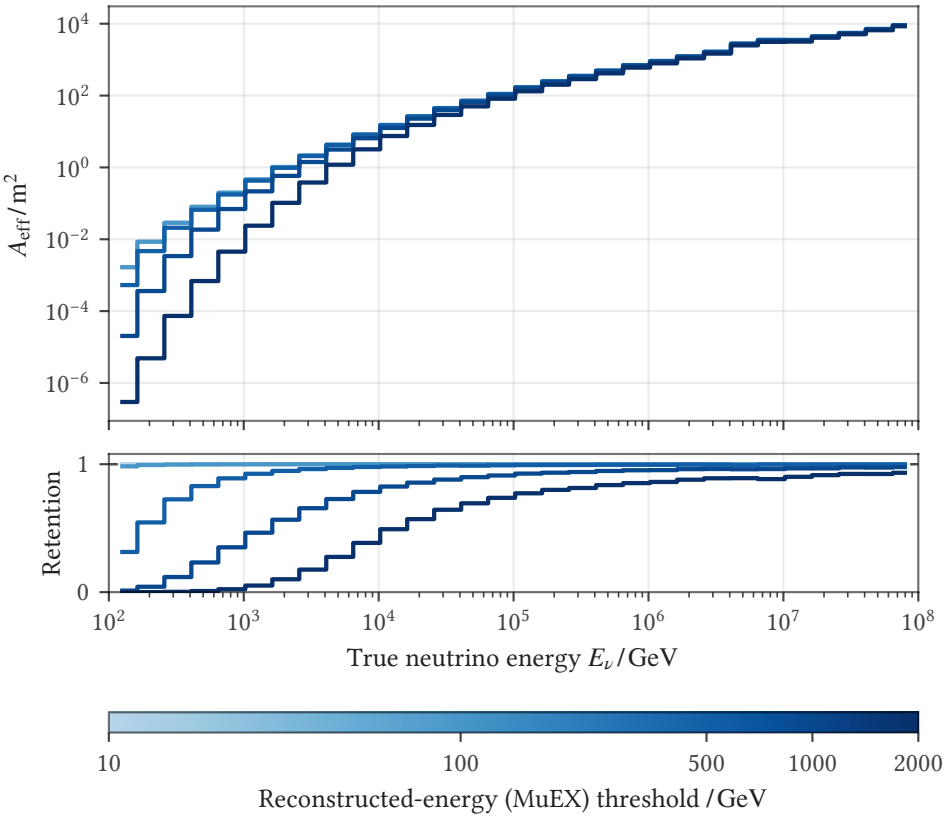

Figure 9.27: Effective area versus true neutrino energy at $\delta = 0$ for a set of reconstructed-energy (MuEX) thresholds, with a panel showing the fraction of events retained relative to no cut. Events below the threshold are removed.

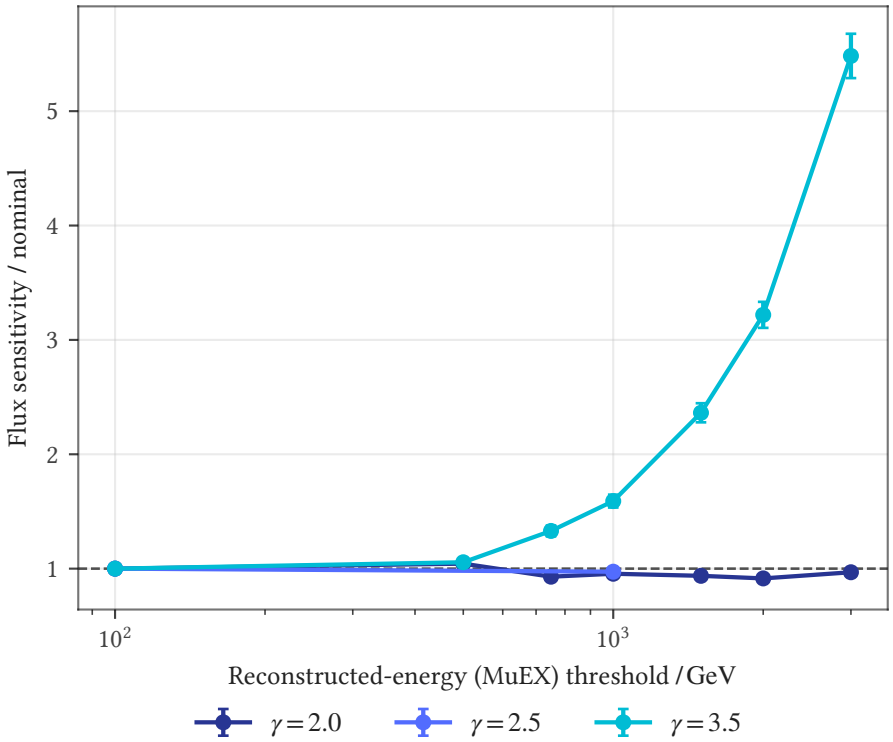

Figure 9.28: Point-source flux sensitivity at $\delta = 0$ relative to the no-cut (100 GeV) baseline (dashed line at unity), versus the reconstructed-energy (MuEX) threshold, one curve per assumed spectral index γ; events below the threshold are removed. Soft-source sensitivity worsens (the ratio climbs above unity) as the threshold increases, while the hardest spectra are little affected. Error bars are propagated as in Figure 9.26 ($\sqrt{2} \times 2.5\% \approx 3.5\%$).

Tightening the energy cut hurts sensitivity to soft sources, because the discarded low-energy events can no longer contribute spatially. For the very hardest spectra, removing low-energy events instead helps slightly, since those events carry little signal and mostly add background. The live sample applies only a loose quality threshold at 100 GeV, well below the regime where the cut would cost soft-source sensitivity.

The same energy dependence appears in the overlap with the Northern Tracks selection (Figure 9.29): the improvement in events relative to Northern Tracks is strongly energy-dependent and heavily skewed toward low energies. This overlap is measured on data alone, so the corresponding power improvement cannot be read off directly from matched simulation. It does, however, identify its origin: Lightning Tracks' improvement in soft-source discovery potential over Northern Tracks comes from the additional low-energy events admitted by the upgoing LCSC filter, which still contribute through their spatial clustering despite their poorer angular resolution—confirming the Lightning Tracks design principle of removing as few events as necessary.

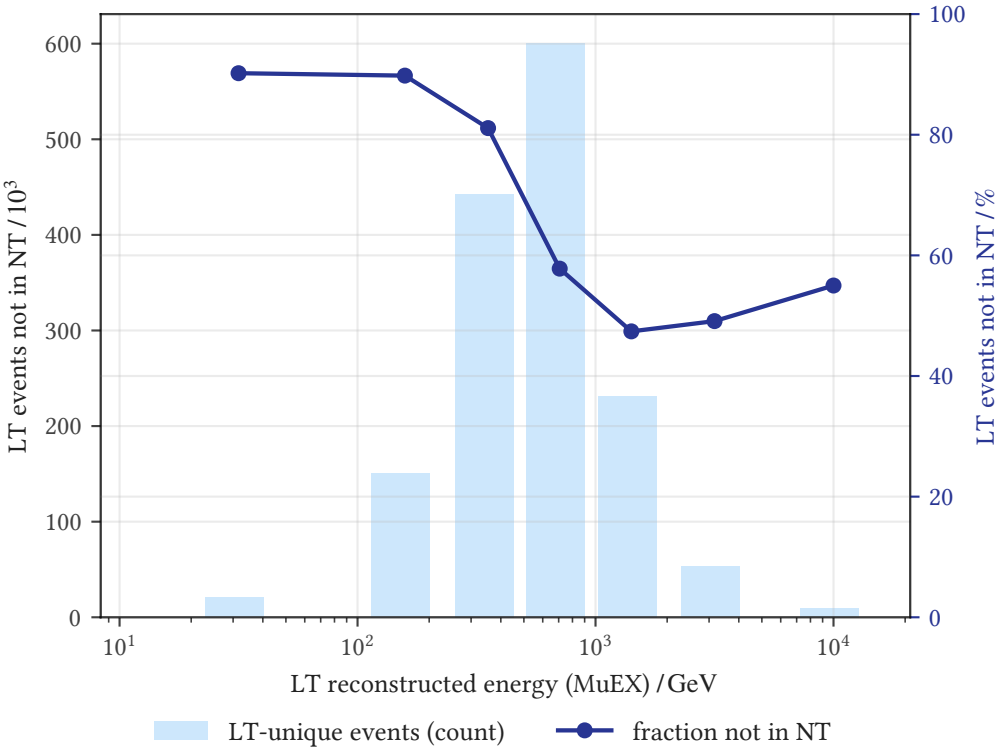

Figure 9.29: Energy dependence of the event overlap between Lightning Tracks and Northern Tracks, measured on data. The improvement relative to Northern Tracks is concentrated at low energies.

9.13 A high-performance C++ fitter

The empirical calibration program of this analysis—the per-ring null distributions of Section 9.8 together with the all-sky trial correction developed in Chapter 10—sets a computational demand that the standard fitter cannot meet. A purely empirical background description at $N_{\mathrm{side}} = 512$ requires at least 10^7 background trials at each of the 1,829 ring declinations, 10^8 at each catalog source declination, and 10^5 full-sky background scans for the trial correction itself, every one of which fits all 3,145,728 pixels. Each of these is an independent maximization of the point-source likelihood. This workload is infeasible in any practical wall time on csky's standard fitter. That fitter already evaluates the likelihood and the spatial term in C++, but it drives them from CPython and runs the outer optimization over γ—an L-BFGS-B loop with finite-difference gradients—in Python, so the per-evaluation interpreter overhead leaves it Amdahl-bound no matter how fast the compiled inner pieces are. Its parallelism compounds the problem: independent fits are farmed out to forked Python workers that cannot properly share the large read-only buffers. The resolution actually reached by recent all-sky analyses is set by exactly this cost: it is cheaper to scan a coarser grid, or to replace the deep tail of each declination's test-

statistic distribution with an analytic extrapolation, than to generate the empirical statistics a full-resolution scan would need.

We remove that constraint by carrying the entire optimization into C++, the outer γ loop included, with analytic envelope-theorem gradients in place of finite differences and OpenMP threads that share the read-only data rather than forked workers that copy it. The new fitter covers the single-source, stacking, and all-sky cases; configuration, dataset bookkeeping, and the construction of the spatial and energy probability density functions remain with csky's Python layer, and any case the C++ path does not implement falls back to the Python fitter automatically. The reimplementation changes how the fit is carried out, not what it computes: it solves the same maximization, with the same optimizer family, so that the resulting test statistic is the standard one and the speed is bought without any change to the statistical result.

Fit architecture

A single fit runs in three phases. Phase 1 is spatial filtering: for each event near the hypothesized source, the signal-over-background probability ratio is computed, and events beyond an angular-distance threshold are cut. Phase 2 is a seed scan over the spectral index γ, evaluating the likelihood on a strided subset of the γ grid; at each seed γ, the inner number-of-signal-events parameter n_s is found by Newton's method. Phase 3 refines the best seeds. The refinement is a nested one-dimensional profile likelihood: the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B) optimizes the single outer parameter γ, while at each γ the inner n_s is profiled out by Newton's method. The envelope theorem supplies the outer gradient analytically, so no finite differencing is needed. This is the same optimization architecture as the standard fitter, moved in its entirety into C++.

The inner n_s solve is cheap, typically about 3 Newton iterations, and the outer surface in γ is well behaved in practice: for any genuine signal the profiled likelihood is smooth and singly peaked, increasingly so at stronger signals. In principle L-BFGS-B could settle into a spurious local optimum that the nested scheme then misses, but this worry is overly cautious—such structure appears only at very low TS, where the surface is nearly flat and the trial is pure background noise, so the outcome is immaterial. We guard against it regardless by pre-scanning γ on a fixed grid before refinement: if every seed returns a null result, meaning no γ admits a positive n_s, refinement is skipped entirely. Roughly 40–50% of background trials are null in this sense and exit at the seed stage, and when seeds do survive, the strided pre-scan localizes the outer optimum well enough that L-BFGS-B is started from a good bracket.

Decomposing the signal-subtraction likelihood

The dominant cost in a background trial comes from signal subtraction with large event counts. The starting-track sample alone carries on the order of two million

events, and the signal-subtraction term sums a per-event contribution over all of them. Re-summing that term at every optimizer iteration, for every trial, is the bottleneck.

The reformulation that removes it splits the events into a far field and a near field relative to the source. Once the data are fixed, the far-field contribution depends only on (n_s, γ) and not on which trial is being fit, because the far-field spatial weights are marginalized over right ascension and the energy weights come from a fixed lookup table. We therefore precompute, on a grid in (n_s, γ), the term that assumes *every* event is far field,

$$F(n_s, \gamma) \;=\; \sum_{i=1}^{N} \log\!\Big(1 + \frac{n_s}{N}\, X_i^{\mathrm{ss}}(\gamma)\Big),$$

and then correct only the comparatively few near-field events back out, event by event, inside the optimizer loop:

$$L(n_s, \gamma) \;=\; F(n_s, \gamma) \;+\; \sum_{i \in \mathrm{near}} \Big[\log\!\Big(1 + \frac{n_s}{N} X_i\Big) - \log\!\Big(1 + \frac{n_s}{N} X_i^{\mathrm{ss}}\Big)\Big].$$

The $\mathcal{O}(10^5)$–$\mathcal{O}(10^6)$ far-field events thus become a single table lookup during fitting, evaluated by cubic Hermite interpolation in n_s and a Catmull–Rom spline in γ, with a conservative per-γ bound on n_s that keeps the interpolant clear of the regime where the implied background would go negative. The decomposition is exact for batched trials that share a fixed far field, which is precisely the background-trial workload, and it is the single largest source of the signal-subtraction speedup.

Lower-level optimization

Beneath the algorithmic reformulation, the fitter is tuned for the hardware it runs on. The hot inner loop is built around inlined, thread-safe replacements for the two most expensive math calls. The first is an exponentially scaled modified Bessel function from a Chebyshev approximation, which avoids the process-global state in the standard-library logarithmic-gamma routine that makes the special-function evaluation path unsafe under threaded execution.[180] The second is a textbook approximation to the exponential: in the CPython extension build the C library's exponential is an external call the compiler cannot inline, so an in-house version is supplied to inline and vectorize across events. The compiled extension ships in four Single Instruction, Multiple Data (SIMD) variants spanning successive x86-64 microarchitecture levels, and the matching one is selected at import time from the host's reported instruction-set support. Event records use an Array of Structs (AoS) layout, reversing the usual guidance that a Struct of Arrays (SoA) vectorizes better. Here each event is largely a set of indices into shared-memory spline tables, so the binding constraint is the number of concurrent memory streams; across the many parallel workers an SoA layout did not supply enough streams to keep the prefetchers busy, whereas AoS keeps each event's fields on a single stream. Events are additionally sorted into access order to cut cache misses. On the largest sample

[180] The hazard in the evaluation path is the standard logarithmic-gamma routine `lgamma`: its POSIX specification records the sign of the result in the static external variable `signgam`, so concurrent evaluation from several threads can race on that shared variable and silently corrupt the result. See `en.cppreference.com/w/cpp/numeric/math/lgamma`; the same hazard is reported at `github.com/stan-dev/math/issues/1250`.

(LT + DNNC) this lifted the retired instructions per cycle from roughly 0.95 to 1.87, taking the fit out of the memory-bandwidth-bound regime. The all-sky scan precomputes every event-intrinsic quantity once per scan rather than once per ring, which at $N_{\text{side}} = 512$ removes on the order of 36 million redundant exponential evaluations for the cascade sample alone, and finds each ring's candidate events by binary search on a declination-sorted array rather than a linear scan. Pixels are fit in parallel with a ring-affine work-stealing scheme so that the large equatorial rings are shared across all threads while the small polar rings are handled by single threads. Because the parallelism is over OpenMP threads sharing one process image, the large read-only event data is read in place from a single shared copy, sidestepping the copy-on-write page faults that the standard fitter triggers when its forked Python workers each touch a shared analysis object.

Performance

The combined effect is large. For the primary LT + DNNC sample, a full $N_{\text{side}} = 512$ sky scan—the end-to-end workload of the actual analysis—runs about 18× faster with the optimized fitter than with vanilla csky. This scan-level factor combines the per-trial fitter speedup with the scan-level optimizations below. Measured per trial, on the wall time to generate background trials, the optimized fitter is faster by a factor that grows with the per-trial event count: about 12× for the primary LT + DNNC background trials, and a smaller factor of a few for signal-injection trials. The per-trial gain is larger for samples with heavier per-event cost, reaching about 100× for Northern Tracks, the most favorable benchmarked configuration. The switch from finite-difference to analytic spectral gradients alone accounts for about 30% of the per-trial speedup. In concrete terms, a full $N_{\text{side}} = 512$ scan completes in under an hour on a large allocation—about 45 minutes measured on 156 cores, and faster still on a full 192-core AGX node—scaling to a few hours on the smaller scavenger allocations. The same scan is a multi-day proposition with the Python fitter. This is what made the fully empirical $N_{\text{side}} = 512$ program practical, with the analytic tail extrapolation used only in the deepest tail, where the empirical statistics run out, rather than across most declinations (Section 10.2).

Validation

Because the fitter is a from-scratch reimplementation, it is validated against csky directly rather than trusted on construction. The random-number generator is reimplemented to reproduce NumPy's stream bit for bit, so that a given seed produces the identical scrambled trial in both fitters. Each validation job then runs the same trial through the standard fitter and the C++ fitter back to back on the same node, and records the difference in the resulting test statistic. The two are not bit-identical at the level of the final test statistic, since the optimizer paths differ, but they are statistically identical: the likelihood evaluations themselves agree to the bit, and the residual test-statistic differences are tiny. Typical differences are $\mathcal{O}(10^{-2})$ or smaller, arising in the flat, near-zero regime where n_s and γ are

unconstrained and the surface is nearly degenerate, so the two optimizer paths settle at slightly different points. The largest differences are bigger—about 0.05 for the primary LT + DNNC sample, and below 0.06 across all selections—but they are not a fitter effect: those cells occur only at extreme declinations, where finite numerical precision in the transcendental NumPy evaluations used for the declination precuts shifts which events pass, rather than any disagreement in the fit itself. Even this absolute worst is negligible at the level of the test statistic and has no bearing on the result. This comparison was carried out across all samples, all signal-subtraction configurations, roughly 17 declinations, signal injection at spectral indices 2 and 3, and pixel-by-pixel for full sky scans, over tens of billions of trials in total. The validation is entirely on scrambled and simulated trials and touches no unblinded data.

Optimizer choices considered and rejected

The nested one-dimensional scheme was not the first design. A full two-dimensional optimizer over (n_s, γ) was implemented, using Levenberg–Marquardt (LM) with analytic gradients and the full Hessian, and was ultimately set aside. It performed no better than the nested one-dimensional profile while being substantially more complex, and it was less stable: at saddle points the Hessian can be indefinite, which forces a diagonal damping that, once applied, effectively reduces the joint step to the nested scheme anyway. The underlying reason is that n_s and γ have very different character. The n_s direction is trivial for Newton and the γ direction carries all the difficulty, so a method that profiles n_s out and works the one hard direction fits the problem better than a joint step that damps the two together. A two-dimensional L-BFGS-B variant was also tried and gave worse results. Finite-difference spectral gradients were used in an early version and replaced by the analytic envelope-theorem gradient, both for accuracy and because the analytic form removes a large number of redundant likelihood evaluations.

10 All-Sky Search Methods

The all-sky scan is the most assumption-free point source search a wide field-of-view experiment like IceCube can perform. By testing every direction on the sky for evidence of a localized neutrino excess, it requires no prior assumptions about source catalogs or candidate positions—and is therefore immune to experimenter bias in source selection. The source catalog search (Section 13.2) complements the all-sky scan by testing a pre-defined list of 110 astrophysical source candidates at a much smaller trial factor, extracted directly from the same sky scan data. Both searches use the unbinned likelihood framework described in Section 9.1.

The hypothesis testing pipeline can be summarized in four stages:

1. **Per-pixel hypothesis testing:** The sky is discretized into a HEALPix grid. At each pixel, the point-source likelihood is maximized over n_s and γ, yielding a test statistic T that quantifies the evidence for a localized excess.

2. **Local p-value conversion:** Each pixel's TS is converted to a *pre-trial* (local) p-value using the exact per-ring background TS distribution at that declination, with no interpolation between declination nodes.

3. **Trial correction:** The pre-trial p-values are corrected for the multiplicity of testing $\sim 3 \times 10^6$ pixels. This is done empirically by repeating the full scan on many RA-randomized pseudo-experiments and recording the most extreme pre-trial p-value in each (Section 10.3).

4. **Hotspot identification:** The most significant pixel in each hemisphere is reported with a post-trial p-value. Significance is declared when this post-trial p-value falls below the FWER-corrected threshold defined in Section 10.3.

10.1 Sky discretization

The sky is discretized with HEALPix,[181] a Hierarchical Equal Area isoLatitude Pixelization of the sphere. Two of its properties matter here: it is isolatitude, placing pixel centers on rings of constant declination (the rings the per-ring calibration of Section 10.2 relies on), and it is equal-area, so every pixel subtends the same solid angle. We use resolution $N_{\text{side}} = 512$, corresponding to 3,145,728 equal-area pixels with a characteristic pixel spacing of approximately 0.11° (6.9′). This is well below the angular resolution of all event samples. The angular offset between any

[181] Gorski et al. 2005, "HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere".

position on the sky and the nearest pixel center has a mean of 0.045° and a worst case of ∼0.12°, with the larger offsets occurring in the HEALPix polar cap regions ($|\delta| \gtrsim 41°$) where the pixel geometry transitions from equatorial rhombuses to triangular caps. In the equatorial belt, the worst-case offset is ∼0.09°. Even the polar cap worst case of 0.12° is at the tail of the best-resolved tracks, and angular resolution itself deteriorates toward the poles, so the relative pixelization error is smallest where the absolute offset is largest. Pixelization effects on sensitivity are negligible in both regimes. The likelihood is evaluated at the center of each pixel, with the hypothesized source position x_s set to the pixel center coordinates $(\alpha_{\text{pix}}, \delta_{\text{pix}})$.

The scan covers the declination range $-80° < \delta < 80°$, yielding 3,097,768 pixels across 1,829 unique HEALPix rings. The polar regions are excluded because the background TS distribution becomes poorly defined near the poles (vanishing solid angle per declination band, minimal event statistics). The hemisphere boundary is placed at $\delta = -5°$, in the muon-horizon transition region (see Section 5.2), close to, though not identical with, the through-going track bump-cut center at $\delta \approx -4.4°$ and the atmospheric-muon rate peak further south. This is a natural boundary between the physically distinct northern (neutrino-dominated) and southern (muon-dominated) sky regions. This yields 1,685,076 northern pixels (981 rings, $\delta > -5°$) and 1,412,692 southern pixels (848 rings, $\delta \leq -5°$). Hotspot identification, trial correction, and post-trial p-values are computed separately for each hemisphere.

10.2 Local p-value conversion

At every pixel of the sky grid, the all-sky scan runs the point-source hypothesis test developed in Chapter 9. The background-only hypothesis H_0 is the data-driven null—right ascension uniform, with the declination and energy distributions taken from the data themselves (Section 9.3)—and the alternative $H_1(x_s)$ adds a point source at the pixel position x_s contributing $n_s \geq 0$ signal events, with the spatial distribution given by the point-spread function (PSF) and an unbroken power-law spectrum $E^{-\gamma}$, $\gamma \in [1, 4]$ (Section 9.1). At each pixel the likelihood is maximized over (n_s, γ) subject to $n_s \geq 0$, and the test statistic is the resulting maximum-likelihood ratio against H_0 (Equation (9.7)).

The same event list is reanalyzed at every pixel. The likelihood evaluates all events at each tested position, but the spatial term weights each event by its angular separation from x_s, so events far outside the PSF contribute negligibly. Two pixels separated by more than a PSF width are therefore effectively independent, sharing no events with non-negligible weight, while pixels within a width are strongly correlated, driven by the same nearby events. This correlation structure is exactly what the empirical trial correction (Section 10.3) measures, and it is why the grid can be made fine enough that no true source falls between tested positions without paying a trial-factor penalty beyond the genuine multiplicity the correction resolves.

A raw test statistic is not comparable across declinations. The background TS distribution varies with declination, because the event density, energy spectrum,

and angular resolution all do (Section 9.8): a TS of 15 at the celestial equator does not carry the significance of a TS of 15 near the pole. To place every pixel on a common scale, each pixel's TS is converted to a pre-trial p-value against the background TS distribution at that pixel's declination,

$$p_{\text{pre}}(\delta, T) = P(T' \geq T; H_0, \delta) = 1 - F_T(T; \delta), \tag{10.1}$$

where $F_T(\cdot\,; \delta)$ is the cumulative background TS distribution at declination δ, sampled empirically from the per-ring trials of Section 9.8. Each ring's survival function is the construction introduced in Section 9.8: the empirical survival function through the bulk, with a truncated-gamma fit to the top 1,000 TS values of the ring carrying the deep tail, where too few trials lie above a given TS to read the survival function off the empirical distribution.

Per-ring matching versus interpolation

The standard approach converts a TS to a p-value by interpolating between a set of pre-computed declination nodes. For track selections this is inadequate, independently of how well the tail is modeled. The background TS distribution carries fine declination-dependent structure that varies even between adjacent rings: the 3σ TS threshold for the track samples ranges from roughly 9 to 14 across declination (Figure 9.16), and this structure is real rather than statistical noise, since each ring rests on at least 10^7 background trials. The sharpest feature sits in the narrow band at the muon horizon (Section 5.2), where the atmospheric muon-to-neutrino transition leaves systematically heavier TS tails. Interpolating through that feature would test horizon TS values against a reference averaged over the lighter-tailed neighboring rings, so the resulting p-values would not be uniform under H_0 and would therefore not be valid p-values at all: the per-pixel type-I error rate would no longer be controlled at the intended α. Those artificially small horizon p-values would also dominate the hemisphere maximum, inflating the trial factor and suppressing sensitivity at every other declination.

We therefore convert each pixel against the exact empirical survival function of its own HEALPix ring, with no interpolation between declinations. Each ring's background trials are generated at that ring's exact declination; a pixel is mapped to its ring by integer index and its TS read directly against that ring's empirical survival function. Because every unblinded TS is topped up with further trials until its ring's empirical survival function covers it (Section 9.8), the conversion at any evaluated point is purely empirical, and the tail extrapolation enters only the background-maxima distribution at the trial-correction stage, whose extreme-value treatment we develop in detail later (Section 10.5).

How deep the per-ring sampling reaches is itself a diagnostic: Figure 10.1 shows, for each HEALPix ring, the largest TS its background trials actually cover, with a running median across declination. The rings are sampled deep enough that the empirical survival function, not the gamma extrapolation, carries the conversion at the significances the scan reaches.

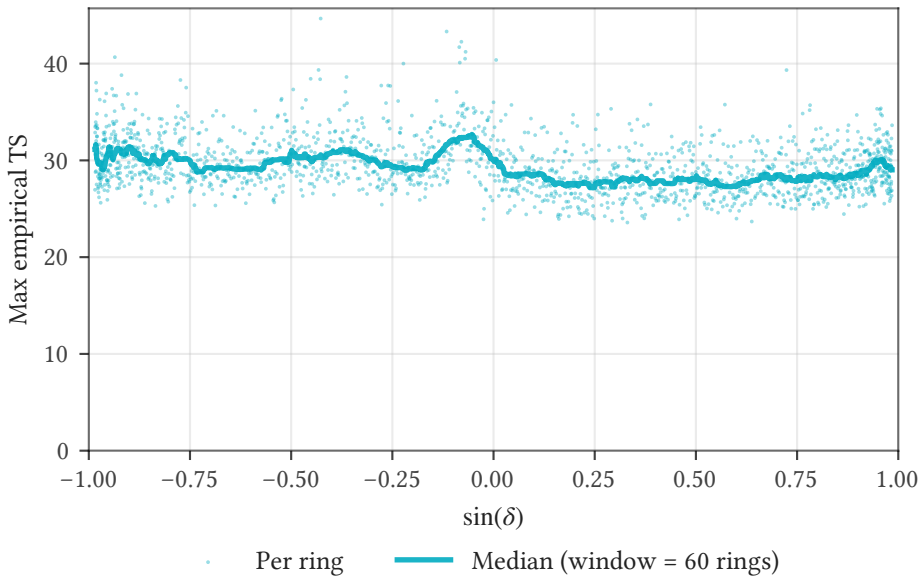

Figure 10.1: Background trial statistics per HEALPix ring for LT + DNNC: the maximum TS value empirically covered by the trial data at each ring, with running median overlay.

The same background scans also show where the hottest pixel of each scan lands. Figure 10.2 gives the distribution of those background hotspot positions: uniform in right ascension, as the RA resampling guarantees, and non-uniform in $\sin\delta$, clustering toward the horizon where the higher event density and shorter pixel-to-pixel correlation length make the effective number of independent tests largest. Declinations with fewer effective independent tests are correspondingly less likely to host a hemisphere's hottest pixel.

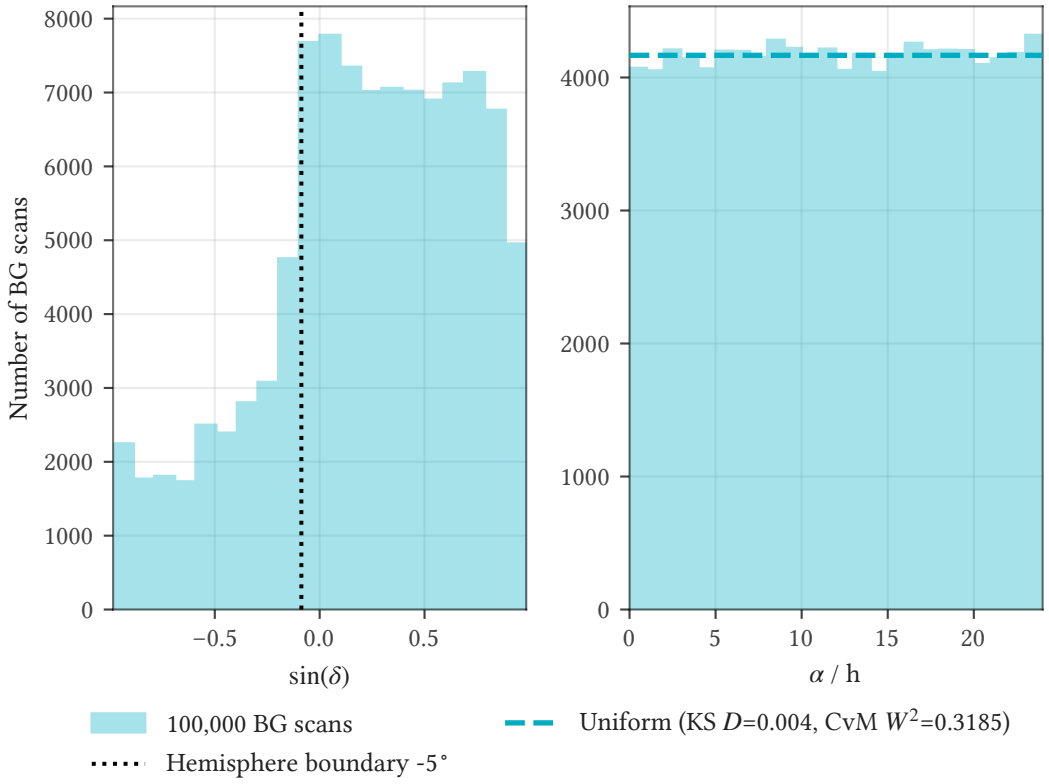

Figure 10.2: Hotspot location distributions for LT + DNNC. Left panel: $\sin(\delta)$ distribution. Right panel: right ascension distribution (expected uniform).

10.3 Trial correction

The all-sky scan tests $\sim 3 \times 10^6$ pixels. Even under the background-only hypothesis, the most significant pixel will have a small pre-trial p-value simply by chance. *Trial correction* quantifies how extreme the observed hottest pixel is relative to what is expected from pure background fluctuations across the entire sky.

The trial correction follows the standard IceCube approach:[182] an empirical null calibration of the scan-level test statistic $T_{\text{hem}} = \max_{\text{pix}} \left[-\log_{10}(p_{\text{pre}})\right]$ (the *scan maximum*), built from its empirical distribution across many background-only sky scans.

The procedure is:

1. Generate many background sky scans, each using independently RA-randomized LT + DNNC data.
2. For each scan, record the maximum $-\log_{10}(p_{\text{pre}})$ across all pixels. This produces one entry in the trial-correction distribution per hemisphere.
3. The collection of maxima defines the distribution of the most extreme fluctuation expected under H_0, fully accounting for pixel-to-pixel correlations

[182] IceCube Collaboration 2017a, "All-sky Search for Time-integrated Neutrino Emission from Astrophysical Sources with 7 yr of IceCube Data".

induced by overlapping declination bands, shared events between neighboring pixels, and any other spatial structure in the scan.

The *post-trial p-value* of the observed hottest pixel is then

$$p_{\text{post}} = P\left(\max_{\text{pix}} [-\log_{10}(p_{\text{pre}})] \geq -\log_{10}(p_{\text{pre, obs}}); H_0\right), \tag{10.2}$$

which is estimated from a Gumbel distribution fitted to the background sky scan maxima (see Section 10.5). The Gumbel fit provides a more stable estimate in the deep tail than the purely empirical survival function, which has large statistical uncertainty when only a handful of background scans exceed the observed value.

Empirical vs. analytic trial correction

Analytic trial correction methods such as the Bonferroni correction ($p_{\text{post}} = N_{\text{trials}} \times p_{\text{pre}}$) or the Sidak correction[183] ($p_{\text{post}} = 1 - (1 - p_{\text{pre}})^{N_{\text{trials}}}$) assume independent tests. In the all-sky scan, neighboring pixels share events and have correlated TS values, so the effective number of independent tests is much smaller than the number of pixels. Applying these corrections with $N_{\text{trials}} = N_{\text{pix}}$ would be overly conservative. The empirical approach automatically accounts for all correlations without requiring an estimate of the effective number of independent tests. In fact, with empirical trial correction, it is optimal to choose a resolution fine enough that nearest-neighbor pixels become strongly correlated: the trial correction automatically absorbs the increased pixel count, while the finer grid eliminates sensitivity loss from a true source falling between tested positions. For the catalog search, by contrast, the 110 tested positions are well separated and effectively independent, so the analytic Sidak correction would suffice: it differs from the empirical Gumbel by only $< 0.01\sigma$. We adopt the fitted Gumbel for completeness and to establish precedent for future catalogs, whose denser or more clustered sources may make the analytic and empirical corrections diverge. Where the empirical approach does matter is in avoiding unnecessary trial factor costs: for example, the Bonferroni correction for a hemisphere split costs $\sim 0.2\,\sigma$ at the 3σ evidence threshold (see Chapter 11). A detailed discussion of the theoretical expectations for the trial correction distribution and the role of pre-trial p-value calibration is given in Section 10.5.

[183] Šidák 1967, "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions".

Hemisphere split

As described in Section 10.1, the scan is split into northern ($\delta > -5°$) and southern ($\delta \leq -5°$) hemispheres, each with its own trial-correction distribution. The pragmatic motivation is that a single whole-sky trial correction would be dominated by the northern sky, where the higher event density produces more effectively independent hypothesis tests. A genuine source in the southern sky would have to exceed the typical northern-sky fluctuation to register as the global hotspot—making the south effectively invisible. The hemisphere split avoids this by allowing each region to set its own fluctuation threshold.

The hemisphere split is standard convention in IceCube all-sky searches.[184] However, testing two hemispheres is two tests, and conventionally no correction is applied to the per-hemisphere p-values or thresholds to account for this multiplicity. To address this, we apply a Bonferroni correction to the per-hemisphere significance thresholds. Concretely, the standard 3σ and 5σ significance levels correspond to one-sided p-values of 1.35×10^{-3} and 2.87×10^{-7}, respectively. With two independent hemisphere tests, the per-hemisphere thresholds are halved to 6.75×10^{-4} and 1.44×10^{-7}, ensuring that the FWER, the probability of at least one false positive across both hemispheres, is controlled at the intended 3σ and 5σ levels. The per-hemisphere post-trial p-values themselves are unchanged; only the thresholds at which they are declared significant are adjusted. In practice, this shifts the effective discovery threshold from 5.0σ to 5.13σ, which is unlikely to affect any real outcome. At the evidence level, however, the shift from 3.0σ to 3.21σ is more substantial and could matter for marginal excesses. We include the correction because testing two hemispheres without it is not statistically justifiable. The Bonferroni correction assumes independence between the two tests, which holds to good approximation since the hemispheres use disjoint sets of pixels. The one exception is a hotspot at the hemisphere boundary: the hottest pixel in one hemisphere is the hotspot itself, while the hottest pixel in the other is a nearby correlated pixel from the same excess, making the two hemisphere maxima correlated. In such cases Bonferroni is slightly conservative, which is the safe direction.

[184] IceCube Collaboration 2017a.

Each background sky scan contributes two values to the trial-correction distributions: the maximum $-\log_{10}(p_{\text{pre}})$ among northern LT + DNNC pixels and the maximum among southern LT + DNNC pixels. The observed data scan is then compared against the appropriate hemisphere-specific distribution.

Correlated background scans

To enable direct comparison between the combined and component sky maps, and to support the trial factor analysis (Section 10.4) that informed the analysis design, all background sky scans use hash-based per-component seeding to ensure cross-sample consistency. Each component (SLT, TLT, DNNC) receives a deterministic seed derived from SHA-256 of the master scan seed and the component name, independent of which combined sample it belongs to. At the same master seed, the three scans use the exact same scrambled events per component: they are correlated views of the same underlying background fluctuation.

This seeding scheme is also why the standalone LT scan must use the overlap-removed data (see the samples discussion, Section 9.1): if the standalone LT scan included events that are removed in the LT + DNNC combination (due to DNNC overlap removal), the RA randomization sequences would diverge. The scrambling is performed sequentially from a single RNG stream per component, not based on event IDs, so even a single additional or missing event shifts the entire sequence for all subsequent events. Using identical event lists is therefore required for deterministic scramble consistency between the combined and standalone scans.

BG maxima distributions

Figure 10.3 and Figure 10.4 show the distribution and survival function of the maximum $-\log_{10}(p_{\text{pre}})$ across background sky scans for LT + DNNC, per hemisphere; the per-component and joint trial-correction cost is evaluated separately (Section 10.4). All are well-described by the fitted Gumbel distribution. This was not the case before switching from χ^2 to the truncated gamma fit for tail extrapolation in the per-ring TS-to-p conversion (see the background TS distributions, Section 9.8): with χ^2 extrapolation, the BG maxima distribution exhibited clear bimodality. The two modes correspond not simply to *empirical vs.* χ^2 *fallback* rings, but to the declination-dependent severity of the χ^2 error: one mode is dominated by rings where the χ^2 fit fails less catastrophically (including rings with sufficient empirical coverage), the other by rings where the χ^2 tail mismatch is most severe. The χ^2 distortion is particularly insidious because it is heterogeneous across declination: the fitted χ^2 degrees of freedom vary from ring to ring (see the asymptotic expectation discussion, Section 9.8), meaning some rings overestimate the local p-value while others underestimate it. A uniform bias across all declinations would partially cancel in the trial correction (both the BG scan maxima and the data scan maximum would be shifted by a similar amount), but a declination-dependent bias does not cancel. The BG scan maxima are dominated by the few declinations where the χ^2 error is smallest: these rings have the most accurate (and therefore most extreme) $-\log_{10}(p)$ values. The majority of declinations, at which χ^2 inflates the p-values more severely, produce systematically suppressed $-\log_{10}(p)$ values that rarely win the hemisphere maximum. Sources at those declinations are effectively invisible—their post-trial p-values are inflated because the trial correction threshold is set by the correctly calibrated rings. The truncated gamma fit resolves this at the source by accurately modeling the per-ring TS tail at every declination, producing a smooth, unimodal BG maxima distribution consistent with the Gumbel expectation. Even with accurate tail modeling, residual calibration imperfections could in principle affect analysis power (see the discussion of pre-trial calibration and power, Section 10.5, above), though in our case the effect is marginal (Section 10.5).

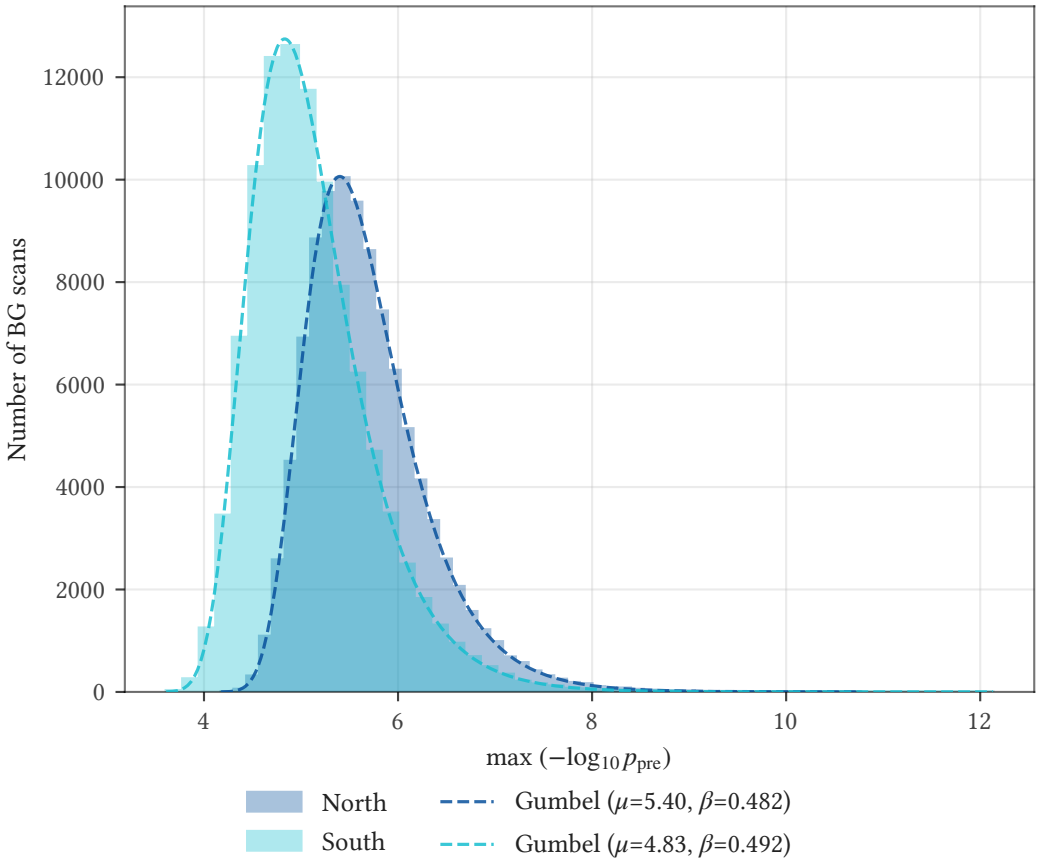

Figure 10.3: BG maxima distribution for LT + DNNC. Histograms of the maximum $-\log_{10}(p_{\text{pre}})$ across background sky scans for the northern (dark blue) and southern (cyan) hemispheres, each with its Gumbel fit overlay.

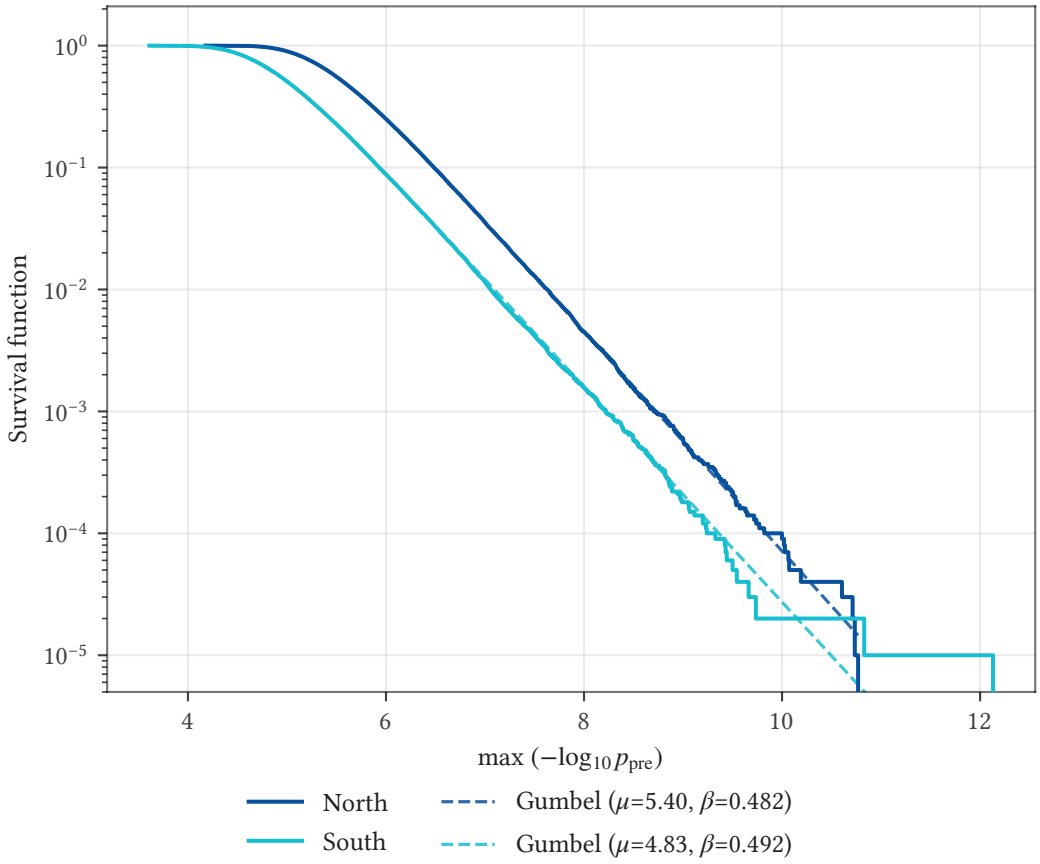

Figure 10.4: BG maxima survival function for LT + DNNC. Solid lines show the empirical SF of the maximum $-\log_{10}(p_{\mathrm{pre}})$ from background sky scans for the northern (dark blue) and southern (cyan) hemispheres; dashed lines show the Gumbel fits.

Figure 10.5 shows the absolute trial correction for LT + DNNC, per hemisphere: the top panel maps pre-trial to post-trial p-values (in sigma units; offset from the 1:1 diagonal is the p-value penalty from trial correction), and the bottom panel shows the corresponding effective number of independent trials as a function of pre-trial p-value. The effective trial factor is computed as the ratio $N_{\mathrm{eff}} = p_{\mathrm{post}}/p_{\mathrm{pre}}$ from the fitted Gumbel survival function. For the stationary independent case ($\beta = 1/\ln 10$), this ratio is constant and equal to 10^{μ}, the number of independent tests. When $\beta \neq 1/\ln 10$, N_{eff} depends on the pre-trial significance: it increases with significance when $\beta > 1/\ln 10$ (as in the all-sky scan) and decreases when $\beta < 1/\ln 10$. This is why the effective trial factor is shown as a curve rather than a single number.

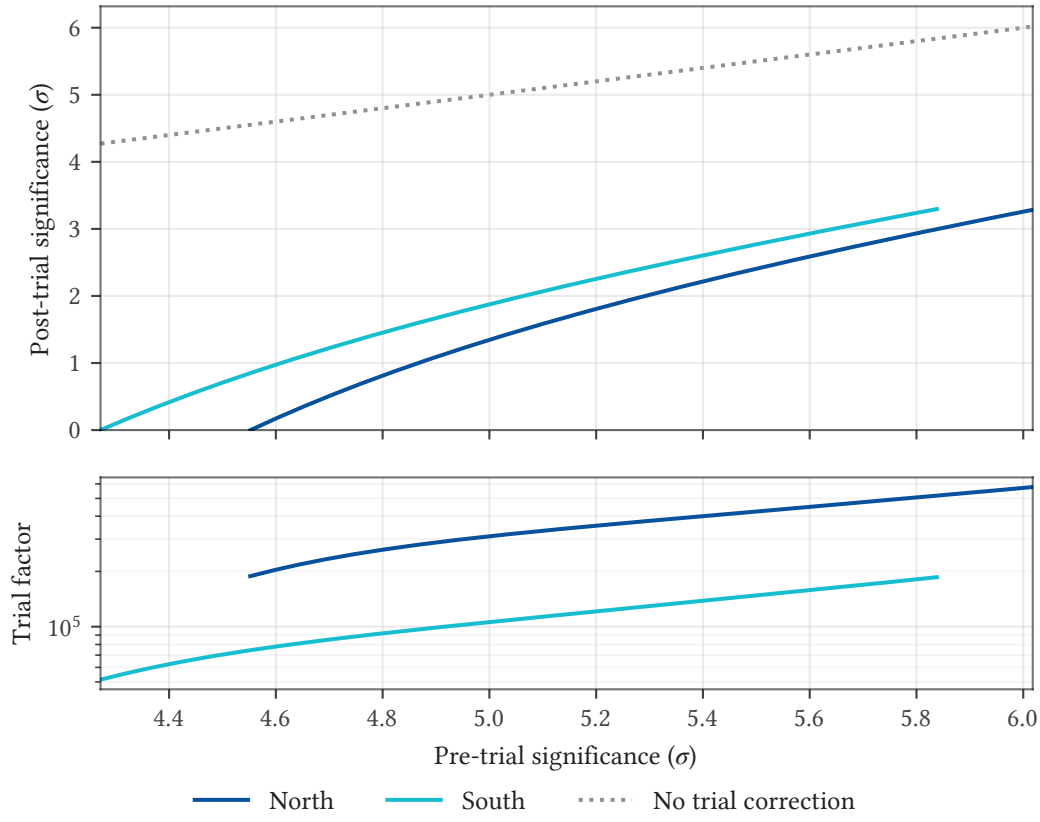

Figure 10.5: Trial correction summary for LT + DNNC, northern (dark blue) and southern (cyan) hemispheres. Top: pre-trial vs. post-trial p-value (in sigma units) with 1:1 reference line. Bottom: effective number of independent trials as a function of pre-trial p-value (log scale).

Goodness of fit

The Gumbel fit quality is assessed via the Anderson–Darling A^2 statistic,[185] which weights the tails more heavily than the Kolmogorov–Smirnov test[186] and is therefore more sensitive to the regime where the trial correction operates. Because the Gumbel parameters μ and β are estimated from the same data, standard critical value tables do not apply. The p-value is computed via parametric bootstrap (drawing from the fitted Gumbel, refitting, and recomputing A^2 for each replicate). The bootstrap confidence interval on the Gumbel fit is too narrow to be visible at plot resolution: all 10^5 scans constrain the Gumbel parameters so tightly that the fit uncertainty is small even where the empirical SF has only a handful of entries in the tail.

As a direct validation, the fitted Gumbel survival function agrees with the purely empirical SF to within $< 0.04\sigma$ in post-trial significance everywhere the empirical SF has coverage (down to $\mathcal{O}(10)$ scans above threshold, corresponding to $\sim 5.5\sigma$ pre-trial in the north and $\sim 6.0\sigma$ in the south). Beyond this, the Gumbel provides the

[185] Anderson and Darling 1952, "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes".

[186] Smirnov 1948, "Table for Estimating the Goodness of Fit of Empirical Distributions".

only estimate. For the catalog search, replacing the Gumbel entirely with a simple Sidak correction ($p_{\text{post}} = 1 - (1 - p_{\text{pre}})^{110}$) shifts the post-trial significance by only $< 0.01\sigma$, consistent with the well-separated catalog sources being effectively independent (see the empirical-vs-analytic discussion, Section 10.3). In effect, the catalog amounts to just 110 effectively uncorrelated tests, in contrast to the ∼3.1 million strongly correlated HEALPix pixels of the all-sky scan (Section 10.1); a Sidak factor is adequate for the catalog but would grossly over-correct the scan, which is why the empirically calibrated Gumbel is required there. This agreement is arguably the most relevant goodness-of-fit measure for the trial correction: it confirms that the Gumbel accurately describes the data in the regime where it overlaps with the empirical SF, validating the extrapolation into the deep tail.

10.4 Component scans and joint trial correction

Testing the component scans (LT and DNNC individually) as part of the analysis was considered. Reporting per-sample post-trial p-values without a joint trial correction would be statistically incorrect—looking at three sets of results is three tests, and any excess found in a component scan must be accounted for. To evaluate the trial factor cost, we implement the joint trial correction using correlated background sky scans with hash-based per-component seeding (Section 10.3) and measure the resulting penalty empirically.

Figure 10.6 overlays the BG maxima survival functions for all samples and the joint combination. The horizontal offset between the LT + DNNC curve and the joint curve directly reflects the trial factor cost of including the component scans. Figure 10.7 quantifies this: in the northern sky, the additional trial factor is ∼1.2× (corresponding to a ∼0.09 σ loss at a pre-trial 5σ anchor), which is negligible: the component scans are nearly redundant with LT + DNNC in the north because the effective number of independent tests in the component scans is largely captured by the combined scan. In the southern sky, however, the DNNC component contributes a substantial number of effectively independent tests, and the additional trial factor rises to ∼2.2×, a post-trial p-value penalty of ∼0.35 σ at a pre-trial 5σ anchor, and larger at lower significance: the penalty falls with increasing significance, as forced by the joint ensemble's smaller fitted Gumbel scale. This is not negligible—a penalty of this size near the evidence threshold can determine whether an excess is claimed or not.

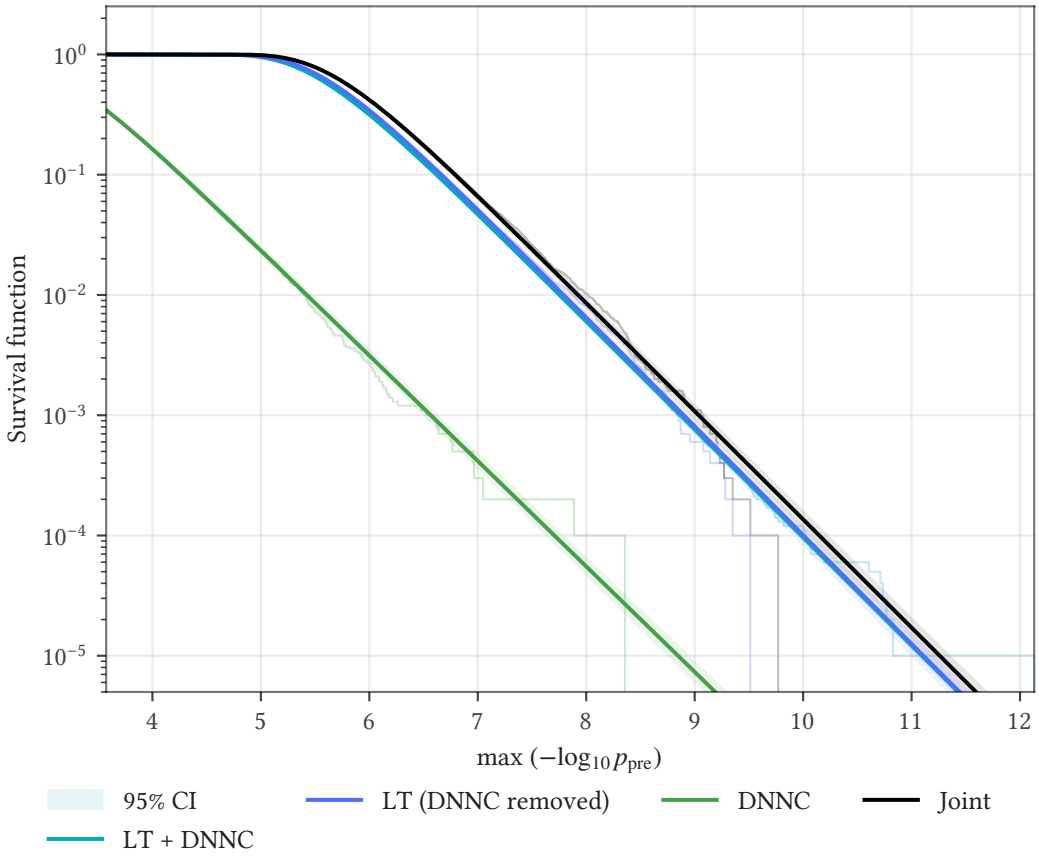

Figure 10.6: Overlaid BG maxima survival functions for all samples and the joint combination. The horizontal offset between curves at a given significance level reflects the trial factor cost of including additional selections.

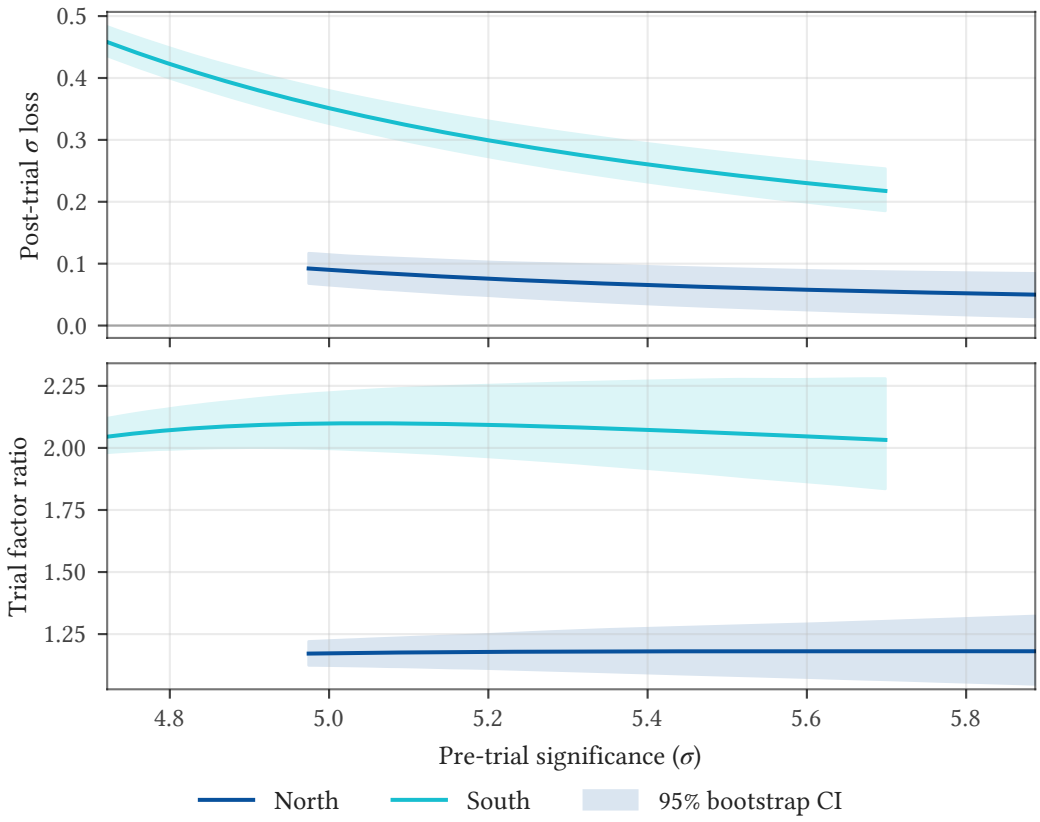

Figure 10.7: Joint trial correction penalty for the northern (dark blue) and southern (cyan) hemispheres as a function of pre-trial significance. Top: post-trial significance loss from including the component scans, $\sigma_{\mathrm{LT+DNNC}} - \sigma_{\mathrm{Joint}}$. Bottom: ratio of the joint effective trial factor to the LT + DNNC-only trial factor (values near 1 indicate negligible additional penalty). Shaded bands are 95% bootstrap confidence intervals.

Since LT + DNNC is at least as sensitive as either component at every declination (and meaningfully more sensitive in the south), the component scans cannot improve the rejection probability: they can only add trial factor. Including them in the trial correction would reduce analysis power by more than $\sim 0.35\,\sigma$ near the evidence threshold in the southern sky (the penalty falls with significance, so it exceeds its 5σ-anchor value there) with no compensating gain.

Apart from the trial-factor cost, the standalone component scans are also not informative as diagnostics. Per-component pre-trial p-values are p-values of distinct hypothesis tests against the same null, not a quantitative measure for comparing alternative hypotheses against each other. Standalone LT and DNNC scans on the unblinded data are therefore not part of the analysis, and per-component pre-trial p-value maps will not be produced or reported.

10.5 Extreme value theory and the Gumbel expectation

The distribution of the maximum $-\log_{10}(p_{\rm pre})$ across pixels in a background sky scan is an extreme value problem: it is the maximum of a large number of correlated random variables. We begin with the independent case, which establishes the theoretical baseline, and then discuss how correlation and non-stationarity modify the picture. The Fisher–Tippett–Gnedenko theorem (also known as the extreme value theorem or the extremal types theorem[187]) establishes that the maximum of n *independent* random variables, after suitable normalization, converges to one of three universal limit distributions, all of which are special cases of the *generalized extreme value* (GEV) distribution:

$$F(x;\mu,\sigma,\xi) = \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right), \tag{10.3}$$

where μ is the location, $\sigma > 0$ the scale, and ξ the shape parameter. The three families are: Gumbel ($\xi = 0$, unbounded right tail), Fréchet ($\xi > 0$, heavy tail), and reversed Weibull ($\xi < 0$, bounded right tail). The Gumbel limit ($\xi \to 0$) reduces to

$$F(x;\mu,\beta) = \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right), \tag{10.4}$$

where $\beta = \sigma$ in the Gumbel case (we use β rather than σ for the Gumbel scale to follow the standard convention and avoid confusion with the significance unit $n\sigma$).

Under the null hypothesis, each pixel's pre-trial p-value is uniform by construction (guaranteed by the per-ring empirical survival functions, Section 10.2), so $x_i = -\log_{10}(p_i)$ follows an exponential distribution with rate $\lambda = \ln 10$. This *marginal distribution* (the distribution of any single pixel's value, ignoring all other pixels) is identical across the entire sky, regardless of declination. The per-ring calibration guarantees this in principle: even though the underlying TS distributions vary dramatically with declination, the calibrated p-values should be uniform everywhere (minor deviations from exact uniformity due to finite calibration statistics and tail extrapolation are discussed in Section 10.5). What varies across the sky is not the marginals but the *dependence structure*, the spatial correlations between pixels, which depend on the local event density and angular resolution.

For n *independent* variables with identical Exp(λ) marginals, the maximum converges to a Gumbel. The standard result for Exp(1) marginals gives location $a_n = \ln n$ and scale $b_n = 1$ (Majumdar et al.[188]). Rescaling to Exp($\ln 10$) gives $\beta = 1/\ln 10 \approx 0.434$ and $\mu = \log_{10}(n)$. This is confirmed by the catalog search (Chapter 11), where 110 well-separated source positions are effectively independent: the fitted scale $\beta \approx 0.43$ matches the independent-test prediction, and the fitted location $\mu \approx 2.04$ corresponds to $10^{\mu} \approx 110$ effective independent tests, closely matching the actual source count of 110.

For the all-sky scan, the $\sim 3 \times 10^6$ pixels are strongly correlated: neighboring pixels share events and have overlapping declination bands. Correlation modifies the Gumbel parameters relative to the independent case:

[187] Fisher and Tippett 1928, "Limiting forms of the frequency distribution of the largest or smallest member of a sample", Gnedenko 1943, "Sur la distribution limite du terme maximum d'une série aléatoire".

[188] Majumdar, Pal, and Schehr 2020, "Extreme value statistics of correlated random variables: A pedagogical review", p. 5, Eq. (14).

Location and the effective number of independent tests

For weakly correlated variables with correlation length ξ_{corr}, the problem reduces to $n_{\text{eff}} = n/\xi_{\text{corr}}$ effectively independent variables (Majumdar et al.[189]), and the Gumbel location shifts to $\mu = \log_{10}(n_{\text{eff}})$. More generally, Leadbetter et al.[190] define the extremal index $\theta \in (0, 1]$ for stationary sequences, and Auld and Papastathopoulos[191] generalize this to non-stationary sequences via an average extremal index γ: in both cases, $\mu = \log_{10}(n \cdot \theta)$ or $\mu = \log_{10}(n \cdot \gamma)$, and the scale $\beta = 1/\ln 10$ is preserved. From the fitted μ, we can estimate the effective number of independent tests at the typical (modal) fluctuation level as $n_{\text{eff}} = 10^{\mu}$, which implicitly encodes the average correlation length across the scan. More generally, the effective trial factor $N_{\text{eff}} = p_{\text{post}}/p_{\text{pre}}$ depends on the pre-trial significance level (see Section 10.3). It ranges from $\sim 2 \times 10^5$ to $\sim 5 \times 10^5$ in the north (Figure 10.5) and $\sim 6 \times 10^4$ to $\sim 10^5$ in the south (Figure 10.5). The lower southern value reflects the interaction of three declination-dependent factors: the event density and angular resolution of each sample (which determine the spatial correlation length between pixels), and the HEALPix pixel density itself (which varies with declination due to the grid geometry). In the south, DNNC's $\mathcal{O}(10°)$ angular resolution contributes a comparable event rate to tracks, producing longer correlations and fewer effective independent tests. All three factors are automatically captured by the empirical approach.

[189] Majumdar, Pal, and Schehr 2020, p. 7, Sec. IV.A.

[190] Leadbetter, Lindgren, and Rootzén 1983, *Extremes and Related Properties of Random Sequences and Processes*, p. 67, Sec. 3.7, Thm. 3.7.2.

[191] Auld and Papastathopoulos 2021, "Extremal clustering in non-stationary random sequences", p. 6, Thm. 2.1.

Scale and the marginal tail rate

The theoretical prediction $\beta = 1/\ln 10 \approx 0.434$ is remarkably robust: as discussed above, it holds for independent, stationary correlated, and non-stationary correlated sequences with identical marginals. The catalog search, where 110 sources are effectively independent, confirms this: the fitted $\beta \approx 0.43$ matches the prediction. For the all-sky scan, however, the fitted $\beta \approx 0.48$–0.49 exceeds the prediction by 39 (north) and 45 (south) standard errors ($n = 10^5$ scans), computed as $(\beta - 1/\ln 10)/\text{SE}_\beta$ with $1/\ln 10 = 0.4343$ and SE_β the bootstrap scale error from refitting the background maxima. Per hemisphere, the fits give 0.482 (north) and 0.492 (south) for the combined LT + DNNC, 0.498 (north) and 0.474 (south) for DNNC alone, and 0.478 (north) and 0.472 (south) for tracks alone. Since the theory proves $\beta = 1/\lambda$ for any correlation structure as long as the marginals are identical, the deviation implies that the calibrated $-\log_{10}(p)$ values are not exactly Exp(ln 10) at every pixel. One candidate is the truncated gamma tail extrapolation used in the per-ring TS-to-p conversion (Section 10.2): with $\sim 10^7$ calibration trials per ring and only $\sim$50–70% having TS > 0, there are only $\mathcal{O}(10)$ empirical calibration events at the $p \sim 10^{-5}$ level where the all-sky maximum typically lands. The gamma extrapolation handles this regime well, vastly better than the χ^2 extrapolation it replaced (see Section 9.8), but small declination-dependent biases in the fitted gamma parameters could produce slightly different effective tail rates across rings. The catalog never probes these extreme quantiles ($\mu \approx 2.0$, well within empirical coverage), which explains why it recovers $\beta \approx 1/\ln 10$. Whether this marginal effect fully accounts

for the deviation, or whether value-dependent spatial correlations also contribute, is settled in the next subsection. The same scale inflation ($\beta \approx 0.47$) is observed in an independent all-sky scan[192] using different data (Northern Tracks vs. Lightning Tracks), a different likelihood framework (SkyLLH with KDE spatial PSFs vs. csky with von Mises–Fisher), and a different gamma tail fit threshold, confirming this is a general characteristic of IceCube all-sky scans rather than an artifact unique to this analysis. The authors note that "the parameters of the Gumbel distribution cannot be determined analytically due to non-trivial correlations between adjacent pixels and must be calibrated from simulations".[193] As shown above, the correlation structure does affect μ but should not affect β: the theoretical prediction $\beta = 1/\ln 10$ holds regardless of the correlation structure, provided the correlations are independent of the values themselves.

[192] IceCube Collaboration 2026a, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube".

[193] IceCube Collaboration, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube", p. App. D.1.

Value-dependent correlation and the copula diagonal

There is a subtlety specific to likelihood-based test statistics on shared-event grids: the spatial correlation between neighboring pixels may itself depend on the significance level. The events driving a high-TS fluctuation have specific angular errors and positions that affect the likelihood at neighboring pixels differently depending on how extreme the fluctuation is. This would violate the assumptions of the standard EVT results cited above, which all require the dependence structure to be independent of the variable values. A recent result from probability theory provides a relevant framework for understanding this. Herrmann et al.[194] show that the classical convergence of maxima to GEV relies on the copula diagonal (the probability that all pixels are simultaneously below the same significance threshold, which determines $P(\max(-\log_{10}(p_{\text{pre}})) \leq x)$ converging to a power function u^{θ}. When the copula diagonal converges to a non-power distortion function $D(u)$, the limiting distribution of the maximum is $D \circ H$ where $H \in \text{GEV}$, which is not itself GEV. Proposition 2.5 of that work further shows that *any* continuous distribution can arise as the limit of suitably normalized maxima under appropriate dependence: the GEV universality is a property of independence (or power-diagonal dependence), not a general law.[195] We do not claim expertise in this area, but the implication for our case seems clear: if value-dependent correlations produce a non-power copula diagonal, fitting a Gumbel to the resulting distribution would yield apparent parameters (including β) that differ from the marginal prediction of $1/\ln 10$.

[194] Herrmann, Hofert, and Neslehová 2024, "Limiting Behavior of Maxima under Dependence", Thm. 2.2.

[195] Herrmann, Hofert, and Neslehová 2024, Prop. 2.5.

To summarize: two plausible and not mutually exclusive mechanisms could explain the observed $\beta \neq 1/\ln 10$. The first, discussed above, is that small imperfections in the truncated gamma tail extrapolation cause the calibrated $-\log_{10}(p)$ marginals to deviate slightly from $\text{Exp}(\ln 10)$ at the extreme quantiles probed by the all-sky maximum. The second is that value-dependent spatial correlations between pixels produce a non-power copula diagonal, placing the problem outside the classical EVT framework where $\beta = 1/\ln 10$ is guaranteed. Which of the two dominates is an interesting question in its own right, but it is out of scope for this analysis: the trial correction does not depend on the answer, since the Gumbel

parameters are calibrated directly from the background simulations rather than predicted analytically. It is hard to move on while an interesting question like this remains unanswered, however, so we test it anyway in the next subsection.

Identifying the mechanism

To tell the two apart, we run an experiment directly on the background scans: we force the per-ring marginals to be exactly uniform while leaving the spatial correlations untouched. The tool is a ranking. Within each HEALPix ring, an iso-latitude band whose pixels all share the same TS-to-p-value calibration, we replace each pixel's calibrated $-\log_{10}(p)$ by its rank among all of that ring's values pooled over the background scans. Ranking only relabels values in increasing order, so it leaves untouched which pixels are large together, that is, the spatial dependence, or copula, while making each ring's marginal perfectly uniform by construction. This isolates the marginals: if the inflated β came from marginal miscalibration, uniformizing them returns it to $1/\ln 10 = 0.4343$.

It does not. Refitting the Gumbel to the maxima of these rank-uniformized scans leaves the scale high, $\beta = 0.4695 \pm 0.0011$ in the north and 0.4867 ± 0.0012 in the south, still far above $1/\ln 10$. The marginals are therefore not the cause. Provided no third effect is hiding that is neither the marginals nor the copula, the copula carries the inflation: about 74% of it in the north and 91% in the south. The marginal calibration is only a percent-level effect on its own, since a marginal-only model reaches just $\beta = 0.438$ (north) and 0.431 (south), both within a percent of 0.4343.

To confirm that the copula is in fact responsible, and not some unidentified residual, its physical origin can be measured directly: the angular size of a significant patch shrinks as the excess grows. The mean exceedance-cluster size falls from about 8 pixels at lower significance to about 5.6 at higher, so a more extreme fluctuation occupies a smaller patch of sky and the scan effectively holds more independent tests at high significance than at low. This is the value-dependent correlation, the non-power copula diagonal anticipated above, observed in the data, and it accounts for the bulk of the $\beta > 1/\ln 10$ deviation.

For the analysis the practical conclusion is unchanged: the trial correction is empirical and never relied on β taking its classical value. What the measurement settles is the interpretation: $\beta > 1/\ln 10$ reflects a real property of the sky's dependence structure, not a calibration defect. We conclude that the authors of the precedent analysis[196] are correct: there are indeed non-trivial correlations between adjacent pixels that prevent the Gumbel parameters from being determined analytically.

[196] IceCube Collaboration 2026a.

Pre-trial calibration and power

Imperfect pre-trial p-value calibration does not bias the post-trial result when the trial correction is empirical. The Gumbel is fitted to the actual distribution of BG scan maxima, and both background scans and the unblinded data pass through the

same calibration. As long as the Gumbel describes the actual BG max distribution well, which the Anderson–Darling goodness of fit (Section 10.3) confirms, the post-trial p-value is correct regardless of whether the underlying marginals are exactly Exp(ln 10). This robustness is specific to empirical trial correction. For analytic corrections such as Sidak or Bonferroni (where the pre-trial p-value is the post-trial p-value up to a multiplicative factor), accurate pre-trial calibration is essential, since any bias propagates directly into the final result.

What imperfect calibration can in principle affect is *analysis power*. If some declination bands have slightly inflated $-\log_{10}(p)$ tails (effective tail rate faster than ln 10), those bands contribute disproportionately to the BG scan maxima, inflating β and raising the trial correction threshold for all declinations, even correctly calibrated ones. Whether this measurably reduces power in practice depends on the magnitude of the miscalibration. In our case, the effect is not directly observable and may be negligible. The per-pixel p-value calibration exists precisely to equalize the contribution of each declination to the global maximum. Without it, declinations with the heaviest TS tails would dominate the trial correction and suppress sensitivity everywhere else. In the extreme case of using raw TS values without any p-value conversion, only the single declination with the highest typical TS would matter: a real source at any other declination would need an enormously high TS to compete, even if it is locally very significant. This is why experiments where the background is explicitly modeled, such as Fermi-LAT, which incorporates diffuse emission templates into the likelihood and applies a fixed TS > 25 threshold for source detection (the likelihood test statistic was introduced for EGRET by Mattox et al.[197] and later adopted by the Fermi-LAT Collaboration[198]), can work directly with TS values, while IceCube's strongly declination-dependent TS distributions require the intermediate pre-trial p-value calibration step. It is also why accurate TS tail modeling matters for power: χ^2 extrapolation introduces declination-dependent calibration errors (see Section 9.8) that suppress the $-\log_{10}(p)$ values at most declinations, leaving sensitivity effectively limited to the few rings where the χ^2 error is smallest (see Section 10.3). The truncated gamma fit minimizes this effect. For the trial correction itself, the Gumbel fit to the BG scan maxima agrees with the purely empirical survival function to within $< 0.04\sigma$ everywhere the empirical SF has coverage (down to $\mathcal{O}(10)$ scans above threshold), confirming that the Gumbel is an accurate description of the data rather than an imposed assumption. Beyond the empirical coverage (which ends at $\sim 5.5\sigma$ pre-trial with 10^5 scans), the Gumbel provides the only estimate.

[197] Mattox et al. 1996, "The Likelihood Analysis of EGRET Data".

[198] Fermi-LAT Collaboration 2020, "Fermi Large Area Telescope Fourth Source Catalog".

11 Source Catalog Testing

The all-sky scan of the previous chapter (Chapter 10) tests every pixel on the sky. The same machinery serves a second search, in which the tested positions are a pre-defined list of source candidates. For this chapter, that is all a catalog is: a list of 110 (δ, α) positions tested in place of the all-sky pixels. Why those positions are worth testing, and how the list is constructed, is Part III material (Section 13.2).

An all-sky scan is, in a sense, just a catalog search whose catalog is every pixel on the sky. The methodology is identical: test a set of positions, convert each to a pre-trial p-value, and apply an empirical trial correction built from the extremum p-value across the tested positions. The only difference is the test set: the all-sky scan tests the full $N_{\text{side}} = 512$ grid ($12N_{\text{side}}^2 \approx 3.1 \times 10^6$ pixels), while the catalog tests the 110 source positions. Everything else (the per-position likelihood fit and the trial correction alike) is identical. The conceptual difference is that a catalog search incorporates prior knowledge about which sky locations are worth testing (informed by observations at other wavelengths), while the all-sky scan makes no assumptions about source positions. Because it tests far fewer positions, the catalog trial factor is much smaller than the all-sky one, providing substantially higher post-trial sensitivity to sources at catalog positions.

11.1 Trial correction

The observed test statistic and pre-trial p-value at each catalog source are obtained by fitting the likelihood at the source position, exactly as for any tested position in the all-sky scan (Section 10.2). The trial correction is built the same way as the all-sky one (Section 10.3); the catalog takes nothing from the all-sky scan's pixel grid. For each background sky realization, the likelihood is evaluated at the 110 exact source positions, and the catalog test statistic for that realization is the global maximum over those sources, $T_{\text{cat}} = \max_{\text{src}} \left[-\log_{10}(p_{\text{pre}})\right]$. The distribution of T_{cat} over the background realizations defines the empirical trial correction.

The catalog trial correction is applied globally, without a hemisphere split, with significance thresholds at the standard 3σ (1.35×10^{-3}) and 5σ (2.87×10^{-7}) and no Bonferroni correction. The hemisphere split was originally considered, during the analysis design phase, for consistency with the all-sky scan, but it is unnecessary here and costs $\sim 0.2\,\sigma$ at the evidence threshold with no compensating benefit: the all-sky scan needs the split because the north has $\sim 10\times$ more effective independent tests and would otherwise suppress the south, whereas the catalog

tests all 110 sources under a single global correction. Because the catalog contains far fewer positions than the full sky, the trial factor is correspondingly smaller, and the background realizations provide ample statistics for a stable empirical trial correction. Figure 11.1 and Figure 11.2 show the resulting T_{cat} distribution and its survival function, both well-described by the fitted Gumbel distribution.

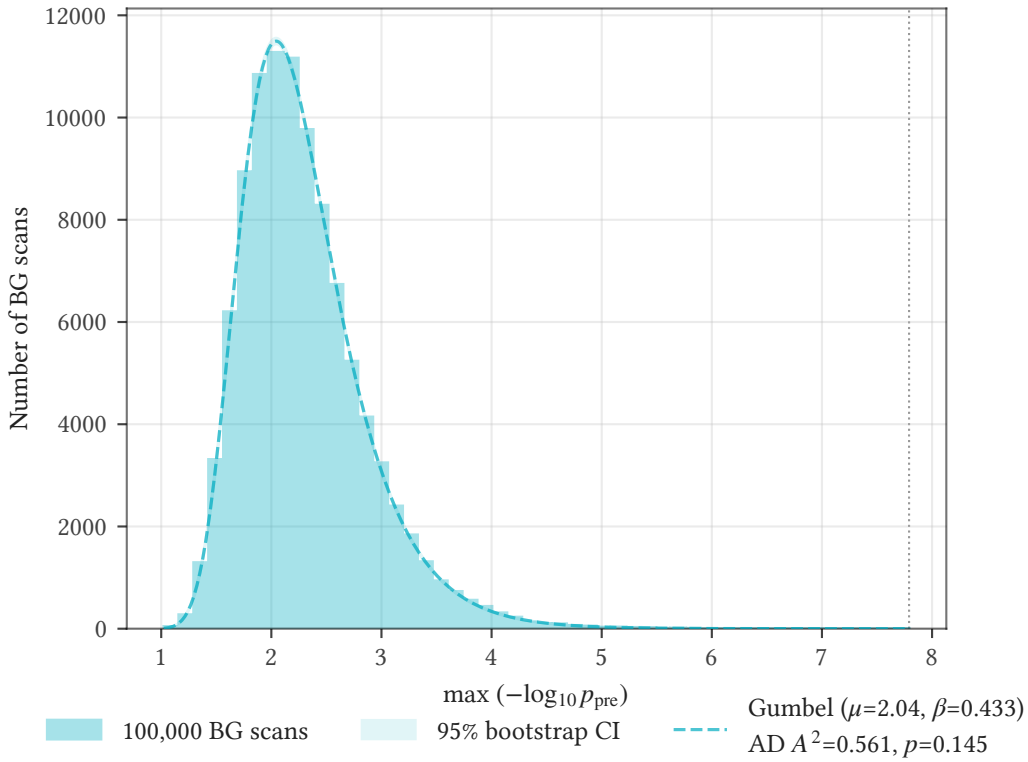

Figure 11.1: Catalog BG maxima distribution. Distribution of the minimum pre-trial p-value (as $-\log_{10}(p_{\mathrm{pre}})$) across all 110 catalog positions from background sky scans, with Gumbel fit overlay and 95% CL bootstrap confidence band.

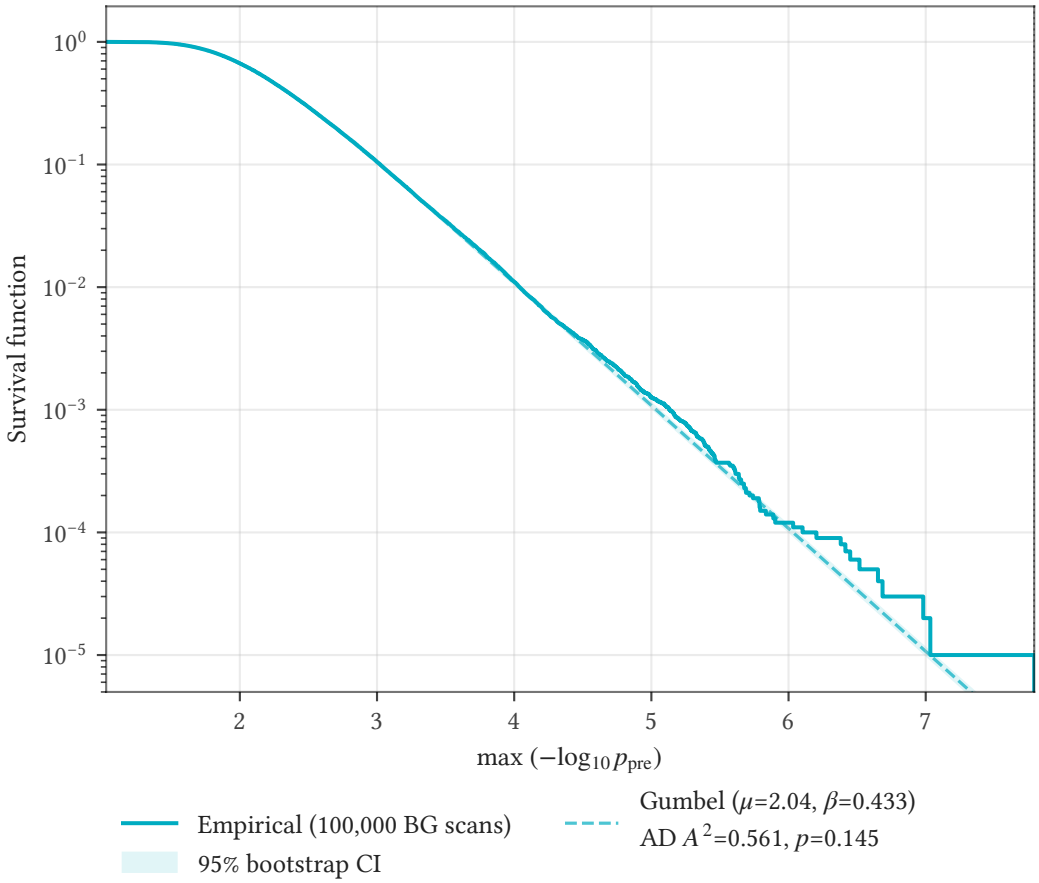

Figure 11.2: Catalog BG maxima survival function. Solid line: empirical SF. Dashed line: Gumbel fit with 95% CL bootstrap confidence band.

Figure 11.3 shows the pre-trial to post-trial p-value mapping and the effective number of independent trials for the catalog search. The effective trial factor peaks at ∼109 around 3.9σ pre-trial, then slowly declines at higher significance. This behavior reflects the fitted (point-estimate) β sitting slightly below the theoretical $1/\ln 10$: the Gumbel tail drops marginally faster than the exponential rate needed to maintain a constant trial factor. Unlike the all-sky case, the catalog β sits only $\sim 1.2\,\sigma$ below the theoretical $1/\ln 10$ (SE ≈ 0.0011 at 10^5 scans), statistically consistent with the prediction: where the 110 well-separated sources are effectively independent, the Gumbel scale takes its classical value, and the catalog is the well-behaved control against the all-sky deviation. The post-trial result is insensitive to this in any case: at 5σ pre-trial the empirical Gumbel and a simple Sidak correction[199] differ by only $< 0.01\sigma$, which indicates no significant correlations between the catalog sources. The all-sky effective trial factor (Figure 10.5) increases monotonically by contrast, consistent with the all-sky $\beta > 1/\ln 10$. At the same post-trial level, the all-sky trial factor is $\mathcal{O}(10^5)$: the catalog search reduces the trial penalty by a factor of $\sim 10^{3.8}$.

[199] Šidák 1967, "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions".

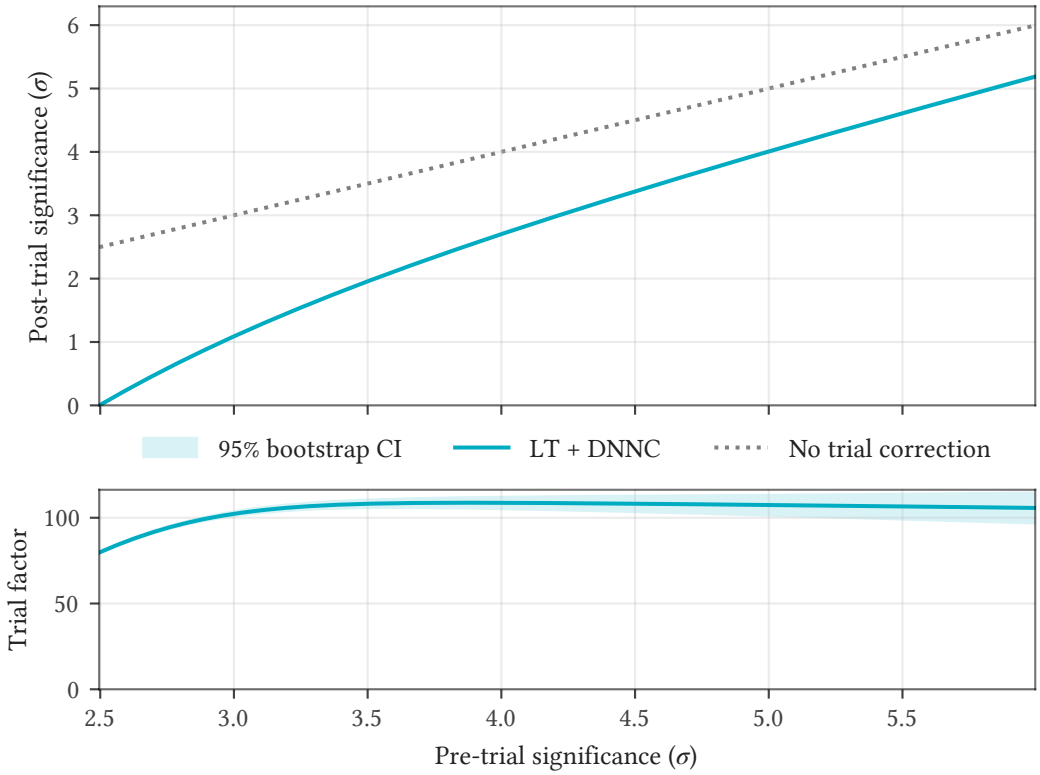

Figure 11.3: Catalog search trial correction summary. Top: pre-trial vs. post-trial p-value (in sigma units) with 1:1 reference line. Bottom: effective number of independent trials as a function of pre-trial p-value.

11.2 Step-down procedure for multiple significant sources

The standard IceCube catalog trial correction reports a post-trial p-value only for the single hottest source. This effectively treats the catalog as a single composite test asking only whether there is *at least one* neutrino source at any of the catalog positions, structurally identical to the all-sky scan's per-hemisphere composite test, with the per-source pre-trial p-values acting only as intermediate scalars feeding the catalog-level maximum. Unlike the all-sky scan, however, the catalog positions are pre-specified astrophysical objects where per-source inference is well-defined, so collapsing to a single composite test is an unnecessary limitation. In standard multiple hypothesis testing, every tested hypothesis is individually compared against a corrected significance threshold. Under Bonferroni, for example, the significance threshold α is divided by the number of tests: each hypothesis (here each catalog sky location) is individually compared against α/m, and any tested location passing the corrected threshold is significant, not just the hottest one. The IceCube convention works fine if there is no or only one significant excess, but it leaves no pre-defined procedure for the case where multiple sources have very small pre-trial p-values. Without such a procedure, any claim about the k-th

hottest source would be ad hoc regardless of how significant it appears. The cleanest approach is to pre-define a procedure, even if it is unlikely to be needed.

We therefore adopt the following step-down procedure, applied to the full catalog:

1. Apply the standard empirical trial correction (Gumbel fit to the distribution of BG scan maxima) to the full catalog. If H_0 is rejected at the hottest source (post-trial p-value crossing the evidence or discovery threshold), report the post-trial p-value as defined by the standard procedure.

2. Remove the rejected source from the catalog. Rebuild the BG maxima distribution from the same background sky scans, with the maximum now taken over the reduced set of source positions.

3. Apply the trial correction to the new hottest source. If H_0 is rejected, report it. Reported post-trial p-values are held non-decreasing down the ranked list, so that a locally less significant source is never assigned a more significant post-trial value (justified in full below).

4. Repeat until H_0 fails to be rejected on the reduced catalog.

Each step is the standard IceCube catalog trial correction on a progressively smaller catalog. The question asked at each step remains the same: can the background-only hypothesis be rejected at this specific source position? Unlike methods based on k-th order statistics, the post-trial p-value at each step is never boosted by the existence of more significant sources: it is always compared directly against background-only fluctuations on the reduced catalog. This is also distinct from false discovery rate (FDR) or population-level binomial methods, which ask fundamentally different questions about the source population as a whole rather than individual hypotheses; the catalog's own population-level test, the binomial population test (Section 11.3), addresses that complementary question. This procedure controls the FWER, equivalent to the Holm procedure[200] with empirical trial correction replacing the analytic Bonferroni factors. Since the empirical trial correction accounts for spatial correlations automatically, the result is equivalent to analytic Holm in principle, but more powerful because it does not require the independence assumption. This gain in power is, however, marginal in our regime: testing against $m - 1$ instead of m sources after H_0 is rejected at one of them gives a negligible improvement at $m \sim 100$. We adopt the procedure regardless because it is cleanly pre-defined, compatible with the IceCube convention, and computationally free given the background realizations already in hand: there is no good reason not to have it in place. The procedure is applied at unblinding regardless; what the data decide is only whether it uncovers more than one significant source, and the plan is fixed beforehand either way. Looking ahead, the procedure also sets a clean methodological precedent for an era in which next-generation neutrino observatories begin to resolve multiple catalog

[200] Holm 1979, "A Simple Sequentially Rejective Multiple Test Procedure".

sources simultaneously, a regime where the standard convention of reporting only the single hottest source would leave additional real findings unreported.

More precisely, the procedure is the Holm step-down when the trial correction is the analytic Bonferroni factor, and the resampling step-down of Westfall and Young[201] (their minimum-p-value construction) when the correction is empirical. Romano and Wolf[202] give the assumption-free generalization. It was developed here independently of these references, the correspondence identified only afterward. Strong control of the FWER rests on *subset pivotality*:[203] the joint null distribution of the surviving sources' test statistics must not change depending on whether a removed source actually carries signal. The per-source null here is generated by RA scrambling, which holds each surviving source's background fixed regardless of which source is removed, so subset pivotality can fail in only one way: positive spatial correlation, in which a strong real source leaks signal into a close neighbor's test statistic. The neighbor's scramble null treats that strong source as part of the background and therefore does not contain the leak. Once the strong source is removed, the residual leak could in principle inflate the neighbor's apparent significance. The exact remedy would rebuild the neighbor's null with the strong source's signal injected, which RA scrambling alone cannot do.

[201] Westfall and Young 1993, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Sec. 2.6.

[202] Romano and Wolf 2005, "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing", Sec. 2.

[203] Westfall and Young 1993, pp. 42–43.

The measured dependence structure of the sky bounds this failure mode. The copula-diagonal analysis of the background maxima (Section 10.5) finds that extreme excesses decorrelate: the more extreme a fluctuation, the shorter its angular correlation length. Signal contamination of a neighbor is point-spread function leakage, and a more significant excess is a higher-energy, better-reconstructed one with a tighter point-spread function, so the same scale governs both the contamination and the correlation length. A source bright enough to reject the background-only hypothesis therefore has its shortest reach precisely in the regime where the rejection occurs. We do not claim this as an identity (the copula result is measured on the background field, whereas contamination is a signal effect), but because the two share the point-spread function scale, the measured extremal decorrelation makes signal leakage beyond the tightest catalog separation implausible at the significances where the step-down rejects.

The single exception is the closest catalog pair (Section 13.2), whose two members lie within each other's point-spread function. If both were to cross the evidence threshold, per-source attribution would be ill-posed: two sources that close cannot be cleanly separated, a resolution limit shared by every method rather than a defect of this one. A reporting-level monotonicity floor handles this case automatically. The reported post-trial p-values are made monotone by a cumulative maximum down the ranked list (each source's reported value is set to the larger of its own and the running maximum above it), so the reported significances are non-increasing from the hottest source downward, exactly as the adjusted p-values of the Holm and Romano–Wolf constructions are monotone by definition. The floor changes no reject decision and applies only to the reported significances. The raw per-source values are retained in the analysis output. For the unresolvable pair it caps the cooler source at the hotter one's significance, so the two are reported at a common significance: one excess in a region the analysis cannot resolve, with no spurious

second detection.

11.3 Binomial population test

The step-down addresses individual hypotheses, source by source. A population-level test addresses the complementary question of whether some subset of the catalog harbors a collective excess of mildly significant sources, even when no single source crosses the discovery threshold. We additionally apply a binomial population test to the catalog's pre-trial p-values, a standard binomial-tail construction in statistical inference.[204] Applying it to a neutrino source catalog follows the precedent of the IceCube NGC 1068 searches.[205]

The N catalog sources (110 in this analysis; Section 13.2) are ranked by pre-trial p-value, $p_{(1)} \le p_{(2)} \le \cdots \le p_{(N)}$. For each rank k, the probability that the background-only hypothesis produces at least k sources with pre-trial p-value at or below the observed $p_{(k)}$ is

$$P_k = \sum_{j=k}^{N} \binom{N}{j} p_{(k)}^{j} \left(1 - p_{(k)}\right)^{N-j}, \tag{11.1}$$

and the population test statistic is the most significant such value, $P_{\mathrm{bin}} = \min_k P_k$, with $k^* = \arg\min_k P_k$ the size of the most significant sub-population. Each P_k is a binomial-tail probability: under the background-only hypothesis the per-source pre-trial p-values are uniform (Section 10.5), so the number falling at or below a fixed threshold is a binomial count. Taking the minimum over k has no closed form and introduces a look-elsewhere effect across the rank scan, so P_{bin} is calibrated empirically against background draws, which also absorbs the ways the real catalog departs from the independent-uniform idealization.

Calibrating the population significance

The minimum over k must be turned into a significance that accounts for the scan over the rank. The N rank statistics P_k are strongly correlated, since each is built from an order statistic of the same per-source p-values and adjacent ranks share almost all of theirs. The effective number of independent looks is therefore far below N, the trial factor relating P_{bin} to a global p-value is modest, and in the deep tail it approaches a constant.

That global p-value would, in principle, be obtained exactly as the single-source catalog trial correction (Section 11.1): the whole procedure—sort, scan k, take the minimum—run on each background sky realization, with the distribution of P_{bin} over the realizations giving the empirical calibration. The available number of background realizations, however, is too small to calibrate a significant population this way. The empirical p-value floors at roughly the reciprocal of the number of realizations, so the direct calibration yields only a one-sided Clopper–Pearson upper limit[206] on the global p-value, equivalently a lower limit on the global significance.

[204] Casella and Berger 2002, *Statistical Inference*, p. 89, Sec. 3.2.

[205] IceCube Collaboration 2022a, "Evidence for neutrino emission from the nearby active galaxy NGC 1068", IceCube Collaboration 2026a, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube".

[206] Clopper and Pearson 1934, "The use of confidence or fiducial limits illustrated in the case of the binomial".

The alternative to calibrating on real background is not reading P_{bin} off a closed-form binomial: the minimum over k has no closed form, so any calibration of it is itself a Monte Carlo. The idealized choice draws the per-source p-values independent and exactly uniform on $[0, 1]$ and runs the same procedure, the null a reader reconstructing the population p-value from the published per-source values would assume. The real catalog departs from that independent-uniform idealization in two ways, of opposite sign. The catalog sources are not perfectly independent: neighboring positions can share events. For a population test, ignoring such correlations is *anti-conservative* (correlated sources effectively double-count and inflate the joint significance), which is the opposite sign from the single-source trial correction, where positive correlation between pixels *reduces* the effective number of trials. The signs differ because the single-source statistic is a maximum, which correlation makes less extreme, whereas the population statistic is a count, which correlation makes more variable and so heavier-tailed. Taken on its own, including the correlations in the calibration makes the population significance slightly *more* conservative.

The larger departure is a point mass at zero test statistic. About 30% of the catalog sources return $\text{TS} = 0$ (no background fluctuation at all) and are tied at the maximal pre-trial p-value rather than spread over the upper unit interval. The real marginal is uniform only on $[0, \eta)$, with $\eta = \Pr[\text{TS} > 0]$ the positive-TS fraction, plus an atom at the top. Because the tied sources cannot fluctuate, the background almost never produces the broad excesses of many slightly-small p-values that would otherwise populate the extreme tail of P_{bin}. The empirical null is lighter-tailed than the independent-uniform null, and a given observed P_{bin} is correspondingly *more* significant under the empirical calibration. This atom effect outweighs the correlation effect, so on balance the independent-uniform significance is slightly conservative.

The two effects can be separated by column-shuffling the background matrix, which breaks the source-to-source correlations while preserving each source's true marginal, atom included. Evaluated at two reference values of the observed P_{bin}, chosen so that the independent-uniform null assigns them 2.33σ (its 99th percentile) and 3.00σ, the three nulls give the post-trial significances collected in Table 11.1.

Table 11.1: Post-trial significances for the full-catalog binomial under three null models, each evaluated at two reference values of the observed P_{bin}: the value the independent-uniform null places at its 99th percentile (2.33σ) and the value it places at 3.00σ. The shuffled null breaks the source-to-source correlations while preserving each source's marginal, atom included. The empirical null retains both.

null model	99th-percentile ref.	3σ ref.
independent uniform	2.33σ	3.00σ
atom only (shuffled)	2.46σ	3.11σ

null model	99th-percentile ref.	3σ ref.
empirical (atom + correlation)	2.42σ	3.05σ

The shuffled row isolates the atom, raising the significance by $\sim 0.1\sigma$ over the independent-uniform null, while the step from shuffled to empirical isolates the correlation, which lowers it by a few hundredths of a sigma. The net, empirical minus independent-uniform, is $\lesssim 0.1\sigma$ and shrinks deeper in the tail. The practical consequence is that a reader who reconstructs the population p-value from the published per-source p-values, treating them as independent and uniform, obtains a slightly conservative number. The analysis itself reports the empirical calibration.

When a strong source hides a population

Running the binomial on the full catalog, matching the precedent searches, is the simplest choice to state, but it has a conceptual weakness when a small, strongly significant subset of sources is present, whether one source or a few. Such a subset feeds its own significance into the test and forms a deep minimum at small k. If a weaker population of individually sub-threshold sources is also present, it forms a shallower minimum at a larger k, and the full-catalog test reports whichever minimum is deeper. Which one wins depends on the relative significance and size of the two populations: the global minimum may fall on the strong subset alone, or on the full population combining the strong and weak sources. When it falls on the strong subset, a genuine weaker population can be left hidden behind it (Figure 11.4). The full-catalog test runs this risk but does not force that outcome.

This failure becomes more likely as a catalog yields more individual detections. Each detection occupies a top rank and contributes its own deep small-k dip, and the weaker population can set the minimum only if its larger-k dip runs deeper still. As catalogs grow and individually resolved sources become routine, an unconditional full-catalog binomial therefore becomes progressively less informative about the unresolved population it was introduced to probe.

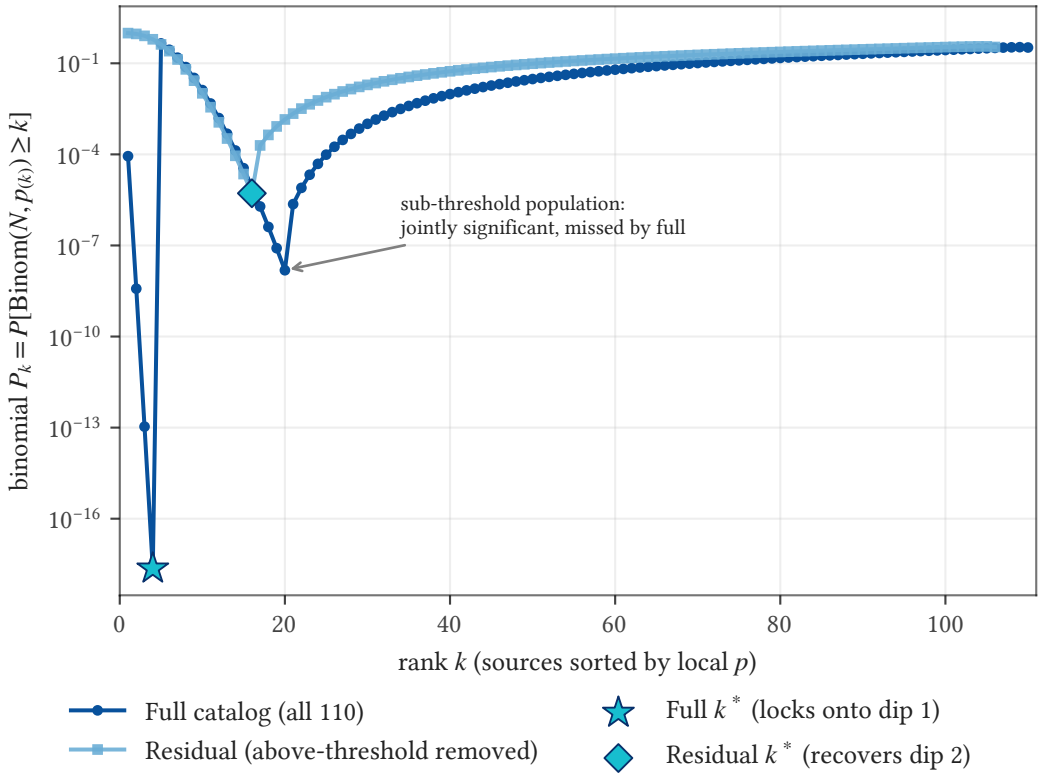

Figure 11.4: Constructed illustration of the full-catalog binomial's blind spot (independent-uniform p-values, $N = 110$; not real data). DIP 1: four sources individually above the 3σ threshold (4.8σ each). DIP 2: sixteen sources individually below the 3σ threshold (1.75σ each) but jointly significant. The full-catalog minimum over k locks onto DIP 1 ($k^* = 4$), leaving the sixteen-source population at DIP 2 invisible to it. The residual variant, run after the four detections are removed, surfaces DIP 2 at $k^* = 16$ ($\sim 4.4\sigma$ local).

The residual population test

The clean way to ask the population question is to set the individually significant sources aside first and run the binomial on what remains. That set is exactly what the step-down procedure (Section 11.2) rejects, so the two compose into a single procedure: the step-down handles the individual detections, and the binomial then asks, of the sources that fall below the individual-detection threshold, whether they are jointly significant. We call this the *residual* population test, as against the *full-catalog* test, and both are implemented.

The residual test is a composite procedure, and its null distribution is built to match. Each background pseudo-experiment has the step-down applied first, at the same fixed threshold used on the data, and the binomial is evaluated on the sources that survive. The threshold that defines an individual detection is therefore part of the test construction rather than a separate look at the data, and

the residual null is read from the same background scans that calibrate the catalog trial correction. With the resolved sources removed, the residual p-value speaks only to the population of individually sub-threshold sources, which is the quantity of interest once a catalog begins to yield individual detections.

The two tests are not redundant, and neither dominates. Removing a strong source before the binomial can split an aggregate excess across the two procedures: a catalog that the full-catalog test would carry over a threshold by counting the strong source among its population can fall short once that source is removed. The opposite failure, a genuine sub-threshold population buried under a strong source, is the one the residual test is built to avoid. Reporting both would carry a trial factor for the two looks. Although the residual test asks the cleaner sub-threshold population question, this analysis uses the *full-catalog* test as its primary population test, to follow the precedent searches and allow a direct comparison.[207]

[207] IceCube Collaboration 2022a, IceCube Collaboration 2026a.

12 Feldman-Cousins Parameter Estimation

As the field matures and strong sources such as NGC 1068 begin to emerge, a point-source search is no longer concerned only with the significance of a source but also with its properties. In this analysis those properties are the parameters of the assumed signal model: the flux Φ and the spectral index γ. For every catalog source position, and for the all-sky hotspot in each hemisphere, the analysis reports parameter estimates and confidence regions on the source signal parameters (n_s, γ) and the corresponding flux Φ. The reporting structure is unified across rejection status: every source receives the same FC interval on $\Phi(E_{\text{pivot}})$ and γ, with the rejection threshold determining only whether a rejection claim is made rather than how we construct and constrain the estimated parameters.

The primary operational motivation for the construction below is that published parameter values propagate to downstream uses: injection studies, multi-messenger followups, and —most importantly—the source models built from the published parameters, such as spectral-energy-distribution fits. Characterizing the nature of the sources is, ultimately, the point of the measurement. Downstream consumers often grab the headline best-fit number as the input rather than propagating the full confidence region. The only exact-coverage frequentist propagation is to push the entire 2D FC region through the downstream computation as a set: by the invariance property of confidence sets, the image of a $1 - \alpha$ FC region under any deterministic downstream map D is itself a $1 - \alpha$ confidence set on $D(\vartheta)$, where ϑ denotes the true parameter value. This is impractical in most realistic followup scenarios because few downstream pipelines are built to handle a 2D-set input rather than a point estimate with errors. In practice many downstream consumers plug in a single central value and propagate without uncertainty.

If the published central number is biased, the bias propagates indefinitely through every downstream consumer of the result, and the confidence interval's coverage guarantee, while statistically valid in isolation, does not break the propagation chain because few downstream users propagate the entire interval. Reporting a bias-corrected $\tilde{\theta}$ as the central value breaks the propagation chain at the source. The combination of bias-corrected central value plus coverage-correct confidence region gives both a usable point estimate and a defensible interval. Reporting the uncorrected MLE plus a valid interval gives a usable interval but a number that anyone using downstream propagates the bias into their own results.

In addition to this operational motivation, the construction's *statistical* validity is itself a substantive payoff (proved using Neyman duality, Section 12.6). The structure mirrors Section 9.9 and runs in parallel to it: both are construction-level

guarantees that decouple the validity of the inference from any property of the likelihood model or the maximum-likelihood estimator's behavior. In particular, the validity of the reported parameter estimates and confidence regions does *not* depend on the level of agreement between $\hat{\theta}$ and injected truth on the signal recovery diagnostic plots (Section 9.10), in the same sense that the validity of the post-trial p-value does not depend on the agreement between data and MC at the PDF level, or, equivalently, on the χ^2 agreement of the test statistic with its asymptotic Wilks form. Both cases are two sides of the same coin: known PDF mismatches violate Wilks' regularity conditions, and the observed bias of $\hat{\theta}$ is itself direct proof of that violation (the regularity conditions are sufficient for an asymptotically unbiased MLE, so observed bias implies they must be violated). Acknowledging this by avoiding Wilks for significance computation while still relying on Wilks to set confidence regions for the estimated parameters would be methodologically inconsistent (see Section 12.1). The FC construction we adopt is the parameter-estimation analog of the empirical null calibration that solves the problem for significance—recovery quality determines the *width* of the confidence region at fixed coverage; it does not determine whether the coverage holds.

One caveat distinguishes this from the null-calibration case: unlike the post-trial p-value, which inherits its validity from data-driven pseudo-experiments and is therefore shielded from signal-model imperfections, the FC construction *does* depend on the simulated signal model, and a mismodeled signal acceptance would propagate directly into the reported parameter estimate and confidence region. The coverage statement is therefore conditional on the assumed signal model, unavoidable for any simulation-based parameter-estimation construction, ours included, in which the sampling distribution at each truth point is defined by injecting simulated signal events. As long as the same signal model is used in the simulation grid and in the unblinded fit, the coverage guarantee holds under that signal assumption. Completely analogous to the validity of the empirical null calibration: the construction's statistical guarantee is correctly controlled given the stated hypothesis, but it does not protect against the risk of the hypothesis itself being poorly chosen for the underlying physics, or of the result being misinterpreted as pertaining to a different hypothesis than the one the construction actually answers.

12.1 Why Wilks-based intervals are inadequate

An alternative construction is to define (n_s, γ) confidence regions using Wilks' theorem (Section 7.5): $\{\theta : -2\log[L(x;\theta)/L(x;\hat{\theta})] \leq \chi^2_2(\alpha)\}$. This is the standing IceCube point-source precedent for sources where H_0 is rejected, set by the Northern Tracks analyses ("68% and 95% confidence levels derived from Wilks' Theorem";[208] "[t]he flux intervals, instead, are computed assuming Wilks' theorem"[209]). The same precedent papers, however, explicitly disclaim Wilks for *significance* computation due to a regularity violation that affects parameter estimation in exactly the same way[210]:

[208] IceCube Collaboration, "Evidence for neutrino emission from the nearby active galaxy NGC 1068", *Science* 378, no. 6619 (2022): p. Fig. 3.

[209] IceCube Collaboration, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube", *The Astrophysical Journal Letters* 1000, no. 1 (2026): p. App. D.3.

[210] IceCube Collaboration 2026a, App. C.

> "the signal-strength parameter lies on the boundary of the parameter space under the null hypothesis ($n_s = 0$), thereby violating the required regularity conditions. In this situations [sic], the asymptotic χ^2 approximation is formally invalid [...] no χ^2 distribution is used anywhere in this work to compute local or global significances [...]"
>
> — IceCube Collaboration, 2026, p. App. C

Using Wilks for parameter-estimation contour levels but not for significance on the same data is methodologically uneven, but the precedent's choice is defensible in their regime: First, in the deep-signal active-fit interior away from boundary atoms, where the precedent papers' published MLE sits, the regularity conditions appear to hold, provided the MLE is unbiased, as it mostly is for them (Figure 12.2), as the empirical Wilks closure check confirms (Section 12.1). Second, FC at this per-source simulation-grid scale has historically been computationally prohibitive—our low-level software optimization (Section 9.13) makes FC tractable here.

These are the closest prior work to the analysis presented here. They test a source list built by the same construction (Section 13.2), though the resulting catalogs aren't identical, with the same catalog-search method, and they include an all-sky scan component, in practice covering a single hemisphere with the Northern Tracks sample alone. The X-ray AGN analysis retests everything the NGC 1068 analysis did and adds the X-ray AGN catalog on top. Their methodological choices, and in particular their use of Wilks-based confidence regions, are therefore the natural benchmark for the construction developed here.

Wilks-asymptotic in our regime fails in both directions (see Figure 12.1): it over-covers at and near the $n_s = 0$ boundary and can under-cover under PDF misspecification. A natural-seeming hybrid is to apply marginal bias correction on top of Wilks: shift the central point estimate from $\hat{\theta}$ to the marginal-median estimate $\tilde{\theta}_{\mathrm{med}}$ (Equation (12.6)), chosen component-wise so that median$[\hat{\theta}\,;\tilde{\theta}_{\mathrm{med}}]$ matches the observed $\hat{\theta}_{\mathrm{obs}}$, while retaining the Wilks-derived width from the Hessian of the *model likelihood* at $\hat{\theta}$. This hybrid is internally inconsistent: the center is taken from a bias-corrected empirical distribution while the width is taken from the model likelihood's Hessian at the biased MLE, so the two halves reflect different distributions. Observable bias in the model-likelihood MLE is itself evidence that asymptotic normality has failed for that likelihood (a regular MLE is only asymptotically unbiased; a bias large enough to matter signals the finite-sample regime is far from asymptotic, so the Hessian width is unreliable), so the foundation of the width derived from the model Hessian has already broken down by the time the bias correction is needed. The internally consistent variant takes both center and width from the same per-truth empirical sampling distribution, with Wilks applied to the empirical density rather than to the misspecified model likelihood. Whether the per-truth empirical density is well-approximated by a Gaussian is itself an empirical question, addressed in the empirical Wilks closure check (Section 12.1) below.

FC replaces the asymptotic χ^2 with the empirical sampling distribution at each candidate θ and inherits no asymptotic regularity assumption. The inner

density model we use, the FFT KDE described in the bias-corrected point estimate (Section 12.3), accommodates whatever shape the empirical cloud takes without committing to Gaussianity at any step. The empirical-LR variant we use produces both the bias-corrected center $\tilde{\theta}$ and the CI width (through the truth-point thresholds $R_c(\theta)$) from one unified procedure: the same empirical density $\hat{p}(x;\theta)$ at each truth point defines both. This is the parameter-estimation analogue of the unified-construction property the validity theorem (Section 12.6) establishes for the p-value: validity is a structural guarantee of the construction.

The Wilks construction is computationally cheap (only a Hessian evaluation at the MLE is needed), while empirical FC requires pseudo-experiment grids at every truth point. We invested substantial effort in low-level software optimization to make per-declination FC simulation grids computationally tractable, removing the cost barrier that has historically discouraged FC for analyses at this scale. With FC tractable here, it is the natural choice. Wilks would commit the analysis to regularity assumptions that the section below shows are in tension with this likelihood at multiple identifiable points.

Regularity conditions and their failure modes

We have argued before, most fully in the context of the background test-statistic distribution (Section 9.8), that Wilks' theorem does not hold for this likelihood. The same regularity violations bear directly on confidence-interval construction; there the case was made largely in words, whereas here we make it with the formalism of the likelihood-ratio test's regularity conditions. The canonical modern statement of the LRT χ^2 limit is Theorem 16.7 of van der Vaart,[211] which subsumes both the classical Wilks[212] χ^2 result and the Chernoff[213] boundary half-χ^2 under a single set of conditions. Other textbook formulations[214] require three-times differentiability of $\log f(x;\theta)$ and a uniform third-derivative envelope, sufficient but not necessary, and explicitly note "a notable exception being if the MLE is on the boundary of the parameter space."[215] Theorem 16.7 replaces these with the minimum conditions under which the asymptotic χ^2 limit can be derived for a composite null:

(i) the model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at the true ϑ with nonsingular Fisher information matrix I_ϑ;

(ii) a local Lipschitz envelope on the log-density: there exists a measurable $\dot{\ell}(x)$ with $P_\vartheta \dot{\ell}^2 < \infty$ such that $|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \le \dot{\ell}(x) \cdot \|\theta_1 - \theta_2\|$ for all θ_1, θ_2 in a neighborhood of ϑ;

(iii) the maximum-likelihood estimators $\hat{\theta}_{n,0}$ (under the null) and $\hat{\theta}_n$ (unconstrained) are consistent under ϑ;

(iv) the local parameter sets $H_n = \sqrt{n}(\Theta - \vartheta)$ and $H_{n,0} = \sqrt{n}(\Theta_0 - \vartheta)$ converge (in the set-convergence sense of van der Vaart[216]) to limit sets H and H_0.

Under these conditions, $-2\log\Lambda_n$ converges in distribution to $\|I_\vartheta^{1/2}X - I_\vartheta^{1/2}H_0\|^2 - \|I_\vartheta^{1/2}X - I_\vartheta^{1/2}H\|^2$ for $X \sim \mathcal{N}(0, I_\vartheta^{-1})$. When ϑ is interior to Θ (so $H = \mathbb{R}^k$) and H_0 is an l-dimensional linear subspace, this reduces to the classical χ^2_{k-l}. When ϑ sits on the boundary of Θ so that H becomes a half-space rather

[211] Vaart 1998, *Asymptotic Statistics*, Theorem 16.7, p. 233.

[212] Wilks 1938, "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses".

[213] Chernoff 1954, "On the Distribution of the Likelihood Ratio".

[214] Lehmann and Casella 1998, *Theory of Point Estimation*, Sec. 6.5, pp. 462–463.

[215] Casella and Berger 2002, *Statistical Inference*, Sec. 10.6.2, p. 516.

[216] Vaart 1998, p. 232.

than the full $\mathbb{R}^k$, the same theorem produces $(Z \vee 0)^2$, the Chernoff half-χ^2 result, derived as a special case rather than as a separate phenomenon.[217]

The textbook chain underlying the classical χ^2_q result for the LRT goes through (1) sum of squared independent standard normals as χ^2_k;[218] (2) Mahalanobis distance $T(x) = (x-\mu)^\top \Sigma^{-1}(x-\mu)$ for $x \sim \mathcal{N}(\mu, \Sigma)$ as χ^2_k via the whitening transform $z = \Sigma^{-1/2}(x-\mu) \sim \mathcal{N}(0, I)$, so $T = z^\top z = \sum z_i^2$; (3) MLE asymptotic Gaussianity under regularity,[219] with the rigorous version due to van der Vaart;[220] (4) Wilks's Taylor expansion of $-2\log\Lambda$ around the MLE, giving $-2\log\Lambda \to \chi^2_q$ where q is the number of constrained parameters.[221]

The IceCube point-source likelihood is in tension with each of (i)–(iv). Table 12.1 maps the conditions to IceCube features that put the analysis outside the regular regime where Theorem 16.7[222] reduces to plain χ^2.

[217] Vaart 1998, Examples 16.9 and 16.10, p. 235.

[218] Casella and Berger 2002, Sec. 5.3, p. 218.

[219] Casella and Berger 2002, Sec. 10.1.2, p. 470.

[220] Vaart 1998, Theorem 5.39, p. 65.

[221] Casella and Berger 2002, Sec. 10.3.1, p. 488.

[222] Vaart 1998, Theorem 16.7, p. 233.

Table 12.1: Mapping of the van der Vaart Theorem 16.7 conditions to IceCube point-source likelihood features.

VDV 16.7 condition	IceCube tension
(i) DQM with nonsingular Fisher information I_ϑ	At the $n_s = 0$ boundary, γ is unidentified (the likelihood does not depend on γ when there is no signal), so I_ϑ is singular along the γ axis (strict Davies problem). At small $n_s > 0$, γ is only *weakly identified* and I_ϑ is ill-conditioned rather than singular. The Davies regime degrades continuously.
(ii) Local Lipschitz envelope on $\log p_\theta$	The signal-subtracted construction (Section 9.7) is not a proper probability density: for events with extreme weights, the assumed background goes negative, and at sufficiently large n_s the per-event factor $(1 + n_s X_i)$ reaches zero and then turns negative for those events, making $\log L$ undefined and removing any possibility of a Lipschitz envelope on the affected region. More fundamentally, VDV 16.7's framework presupposes a parametric family of densities (so that the score zero-mean property, derived by differentiating $\int p_\theta\, dx = 1$ under the integral sign, holds). A construction that takes negative values is outside that framework regardless of how the Lipschitz envelope question is posed.
(iii) MLE consistency under truth	Under model misspecification, the MLE converges to a pseudo-true parameter $\theta^* \neq \vartheta$ rather than to truth. The PSF tail-light approximation discussed in the bias-corrected point estimate (Section 12.3) and the related PDF mismatches are the dominant misspecification source for our likelihood. The signal-recovery diagnostics (Section 9.10) directly demonstrate $\hat{\theta} \neq \vartheta$ in the median.
(iv) Set convergence $H_n \to H$ and $H_{n,0} \to H_0$ to limit shapes that give plain χ^2 (full $\mathbb{R}^k$ for H, linear subspace for H_0)	$\Theta = [0, \infty) \times [1, 4]$ has three boundary faces: $\{n_s = 0\}$ (where (i) also co-fires since γ is undefined there), $\{\hat{\gamma} = 1\}$, and $\{\hat{\gamma} = 4\}$. Truth on any face puts the analysis in the boundary regime VDV's Examples 16.9–16.10 work through: H_n converges to a half-plane rather than $\mathbb{R}^2$ (a quadrant at a corner), with limit $(Z \vee 0)^2$ (Chernoff half-χ^2) rather than χ^2_1 and factor-of-2 over-coverage in the canonical 1D-boundary case. Empirically these manifest as the $\hat{n}_s = 0$ atom and the $\hat{\gamma} \in \{1, 4\}$ atoms in pseudo-experiments.

Conditions (i) and (iv) both originate at the $n_s = 0$ boundary, where they

act simultaneously. Condition (i) in particular is continuous rather than sharp. At $n_s = 0$ exactly, the score for γ vanishes identically and I_ϑ is singular along the γ axis. At small $n_s > 0$ the score is nonzero but small (proportional to n_s at leading order), so $I_{\gamma\gamma} \to 0$ as $n_s \to 0$ and the condition number of I_ϑ diverges from below: γ is *weakly identified*[223] in this regime rather than strictly unidentified. Standard Wilks regularity is recovered only once n_s is well above zero on the per-experiment scale that supports separating the spectral index. Each tension in isolation has a known correction: condition (iv) alone (boundary geometry with identifiable θ_0 and nonsingular Fisher) gives the Chernoff[224] half-χ^2 result, over-coverage by approximately a factor of 2 in the tail at the boundary in the canonical 1D case (modern HEP exposition by Cowan et al.[225]). Condition (i) alone (unidentifiable nuisance only under the alternative) gives the supremum-of-χ^2 correction of Davies.[226]

Empirically, the truncated-gamma tail fit (Section 9.8) recovers effective dof ≈ 2 in the deep tail of the per-pixel test statistic, while the body of the distribution sits at effective dof ≈ 1.1, and additional Chernoff-type atoms appear at the $\gamma \in [1, 4]$ endpoints from the bounded model-LR MLE. The conjunction of (i) and (iv) in the IceCube setup is consistent with both effects compounding, but we are not aware of a closed-form correction valid across the full distribution. Even a hypothetical correction combining Davies (for the strict γ-unidentified boundary) with Chernoff (for the half-space cone) would address $n_s = 0$ exactly but not the weak-identification regime at small $n_s > 0$ where I_ϑ remains ill-conditioned. The continuous Davies degradation discussed above is itself a structural argument against a closed-form fix. We do not attempt to attribute specific effective-dof regimes to specific named effects. Replacing analytic Wilks with empirical FC calibration is one of the original motivations for our construction.

For broader context on treating LRT regularity violations carefully, see Protassov et al.[227]. An empirical demonstration of Wilks coverage failure in toy oscillation problems analogous to ours is given by NOvA,[228] where Wilks systematically over-covers while the FC construction recovers nominal coverage.

Conditions (i) and (iv) recede deep in the active-fit interior, where non-zero signal pushes the MLE far above the $n_s = 0$ boundary and $\hat{\gamma}$ is well separated from the atoms at $\{1, 4\}$. Condition (iii) does not: under misspecification the MLE converges to a pseudo-true parameter rather than to truth,[229] and the PDF misspecification, chiefly the PSF tail-light approximation discussed in the bias-corrected point estimate (Section 12.3), persists across the parameter grid, and the signal-recovery diagnostics (Section 9.10) directly demonstrate $\hat{\theta} \neq \vartheta$ in the median. Wilks-based confidence regions on the model likelihood would therefore be centered on the wrong distribution anywhere on the grid, including the deep interior where (i) and (iv) are no longer consequential.

The four conditions touch different parts of the analysis: (i) is information-geometric (singular Fisher matrix from unidentifiable nuisance), (ii) is regularity of the log-density itself (no Lipschitz envelope), (iii) is model-data alignment (MLE not consistent for the parameter we want to infer), and (iv) is the geometry of the parameter-space tangent cone at ϑ. They are triggered by separate properties of

[223] Staiger and Stock 1997, "Instrumental Variables Regression with Weak Instruments", Andrews and Cheng 2012, "Estimation and Inference with Weak, Semi-Strong, and Strong Identification".

[224] Chernoff 1954.

[225] Cowan et al. 2011, "Asymptotic formulae for likelihood-based tests of new physics", Sec. 3.5.

[226] Davies 1977, "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative", Davies 1987, "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternatives".

[227] Protassov et al. 2002, "Statistics, Handle with Care: Detecting Multiple Model Components with the Likelihood Ratio Test".

[228] NOvA Collaboration 2025, "Monte Carlo method for constructing confidence intervals with unconstrained and constrained nuisance parameters in the NOvA experiment", Figs. 2 and 5.

[229] White 1982, "Maximum Likelihood Estimation of Misspecified Models", Vuong 1989, "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses".

the IceCube setup and do not collapse into a single underlying cause. The per-cell empirical sampling distributions in Figure 12.5 are visibly non-Gaussian: boundary atoms at $\hat{n}_s = 0$, ridge asymmetry, truth-dependent correlation orientation, and $\hat{\gamma}$-boundary atoms at $\gamma \in \{1, 4\}$. This non-Gaussianity is consistent with the regularity tensions tabulated above. The closure check below quantifies the deviation from χ^2_2 Mahalanobis directly.

Empirical Wilks closure check

The empirical-FC construction does not assume Wilks' theorem holds anywhere. The per-cell empirical α-quantile of the test statistic gives exact coverage by Neyman duality (Section 12.6) regardless of whether the asymptotic χ^2 approximation is accurate. Checking whether Wilks does in fact hold on the bias-corrected empirical density is nevertheless a useful supporting diagnostic: agreement places our construction in a familiar asymptotic-statistics regime; disagreement (especially near boundary atoms) shows where FC is necessary for valid confidence regions.

The closure check tests whether the asymptotic χ^2 conclusion of Theorem 16.7 (Section 12.1) is empirically reached on our per-cell sampling distributions. The test statistic is the Mahalanobis distance

$$T_i = (x_i - \theta_c)^\top \Sigma_\theta^{-1} (x_i - \theta_c), \tag{12.1}$$

computed for each active-fit pseudo-experiment's MLE pair $x_i = (\hat{n}_s, \hat{\gamma})_i$ at truth $\theta = (n_{\text{inj}}, \gamma_{\text{inj}})$ on the FC simulation grid, with Σ_θ the sample covariance of the active-fit interior trials at θ. Under asymptotic MLE Gaussianity around the centering θ_c with covariance Σ_θ, T is exactly χ^2_k (see the textbook chain in the regularity conditions, Section 12.1, above), so empirical deviation from χ^2_2 on our $k = 2$ parameter space is direct evidence by modus tollens that the Wilks regularity assumptions are not all met. The centering choice θ_c controls which failure modes the check can detect: the lenient version uses the sample mean, $\theta_c = \mu_\theta$, which absorbs MLE bias and can flag only shape non-Gaussianity; the complete test uses the truth, $\theta_c = \theta$, which additionally exposes MLE bias and $\hat{\gamma}$-bound atom behavior as miscoverage of truth. This excludes atom observations and probes only the interior distribution. The atoms themselves are already proof of non-Gaussianity, and no Wilks-based confidence region would be constructed at an atom observation in any case.

We report results at the diagnostically relevant $\chi^2_2(\text{erf}(1/\sqrt{2})) \approx 2.296$ Wilks threshold to match the confidence level at which both the 2D joint FC region and the 1D profile-FC intervals are reported as deliverables. Deviations from χ^2_2 Gaussianity at this contour translate directly into miscalibration of both under a Wilks construction. The closure logic itself does not depend on the threshold—χ^2_2 asymptotics predict a specific coverage value at every quantile, so empirical disagreement at *any* threshold suffices to prove regularity violation via modus tollens on Theorem 16.7.

Figure 12.1 shows the closure check at $\sin\delta = 0$. Reading the colormap: the cream band at ≈ 0.6827 is where Wilks-asymptotic and the empirical sampling distribution coincide. The magenta tail (empirical coverage below ≈ 0.6827) marks cells where Wilks under-covers: its contour is narrower than the data supports, giving false confidence in rejection. The truncated-Blues tail (empirical coverage above ≈ 0.6827) marks cells where Wilks over-covers: its contour is wider than the data supports, giving false confidence in compatibility.

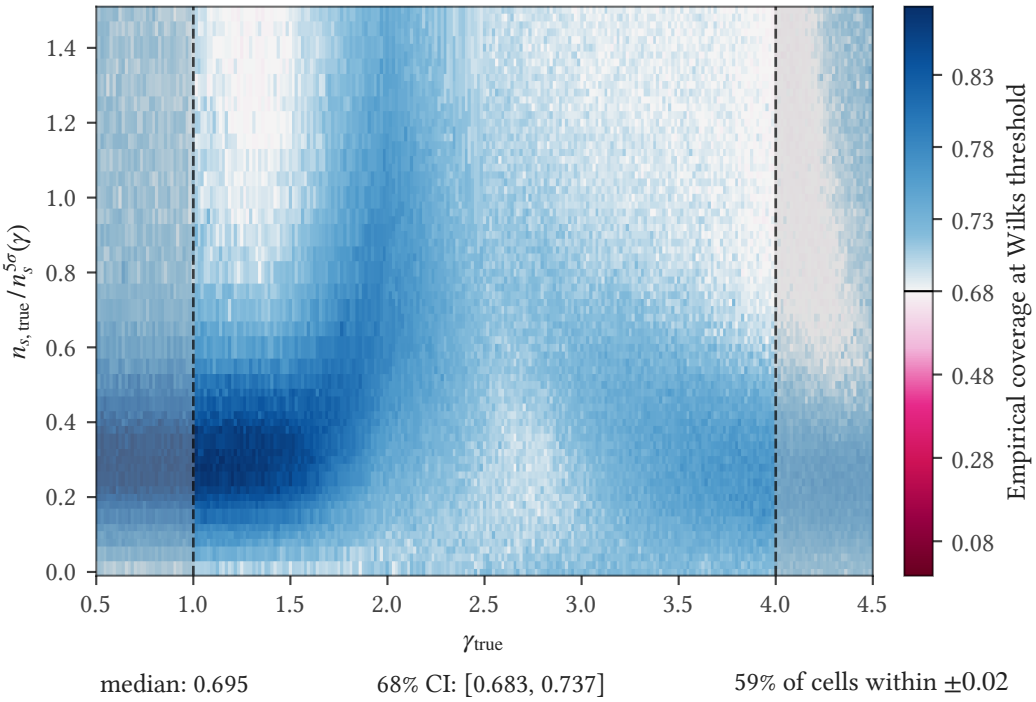

Figure 12.1: Per-cell empirical coverage at the $\chi^2_2(\mathrm{erf}(1/\sqrt{2})) \approx 2.296$ Wilks threshold on the FC simulation grid for LT + DNNC, at $\sin\delta = 0$. The vertical axis is the injected signal strength normalized by significance, $n_s/n_s^{5\sigma\mathrm{DP}}(\gamma)$, so that 1.0 marks the 5σ discovery-potential level at the corresponding γ. Per-trial Mahalanobis distance uses the sample covariance of the active-fit interior trials. Color encodes the fraction of trials below the threshold. Pure Gaussianity predicts ≈ 0.6827. Diverging colormap: magenta = under-covers, cream = nominal, Blues = over-covers.

Sample-mean centering discussion

We observe no significant undercoverage within the model-bound interior of the simulation grid (Figure 12.1): every interior cell is at nominal coverage or in the over-coverage half. The few cells below nominal sit at the grid edges (beyond the γ model-fit bound and at the outer n_s boundary), where the boundary-truncation caveat discussed in this section applies. Boundary atoms leave no visible footprint either, consistent with restricting the check to the active-fit interior.

What remains is a bimodal ridge structure of over-coverage. The two ridges sit symmetrically on either side of $\gamma = 2.5$ (the vMF pull correction calibration point and the location of best agreement) and are mirror images of each other in shape but not in magnitude: each runs along $\sim 0.3 \times 5\sigma$-DP at low signal and curves up vertically at $\gamma \approx 2$ on the hard side and $\gamma \approx 3$ on the soft side, but the hard arm is consistently stronger. At low signal (~ 0.2–$0.4 \times 5\sigma$-DP) the soft arm produces $\sim$78% over-coverage while the hard arm peaks near $\sim$90%. At higher signal ($\geq 1 \times 5\sigma$-DP) both sides recover toward nominal away from the ridge arms. The mid-hard band $\gamma \approx 1.5$–2.5 retains a milder $\sim$80% over-coverage.

This pattern is consistent, though we cannot prove it rigorously, with PSF mismodeling as the dominant remaining mechanism after bias absorption. The vMF pull correction is calibrated at a single $\gamma_{\text{cal}} = 2.5$ as a forced compromise (see the angular error calibration), so the angular distribution's coverage itself is expected to worsen at spectral-index extremes. The 0.2° angular-error floor compounds this on the hard-γ side, where high-energy events with the best intrinsic angular resolution get floored at exactly the regime in which the vMF approximation already over-covers the PSF core and under-covers the tail. PSF over-coverage propagating into Wilks over-coverage is a plausible mechanism for a ridge with this orientation. The pattern repeats qualitatively across declinations with one suggestive asymmetry: the hard-γ over-coverage is less severe at northern declinations than at southern ones (the south-to-north ratio of hard-side to soft-side over-coverage is more extreme in the south). This is consistent with the PSF interpretation, since the southern selection cuts harder on low-energy events and therefore retains a relatively higher fraction of the high-resolution floor-affected events, where the vMF mismatch is most pronounced. The picture is further complicated by the multi-component sample: the structure here appears tracks-driven, while DNN Cascades carry their own PDF modeling imperfections but plausibly contribute less to the angular-mismatch component of the over-coverage given their substantially wider PSF scale. The shape match is suggestive but speculative without a dedicated isolating test.

Truth centering discussion

The bimodal over-coverage structure of the sample-mean version persists in the interior of the truth-centered map (Figure 12.2), as expected. The truth-centered version additionally exposes severe under-coverage at all three boundary atoms—severe enough that "under-coverage" is almost a misnomer: at the atoms, empirical coverage bottoms out at effectively zero because a Wilks-CI built on a clipped MLE lands too far from truth to cover it.

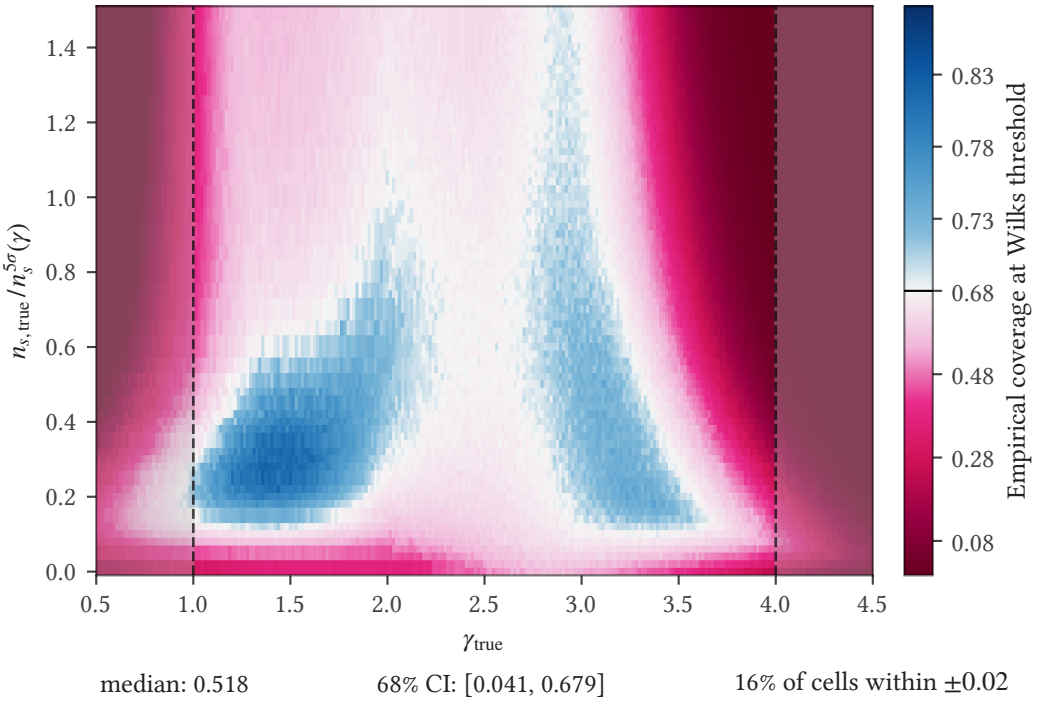

Figure 12.2: Per-cell empirical coverage at the $\chi^2_2(\mathrm{erf}(1/\sqrt{2})) \approx 2.296$ Wilks threshold on the FC simulation grid for LT + DNNC, at $\sin\delta = 0$, with the per-trial Mahalanobis distance centered on the truth $\theta_c = \theta$ (Wald-equivalent) rather than the sample mean. The vertical axis is the injected signal strength normalized by significance, $n_s/n_s^{5\sigma\mathrm{DP}}(\gamma)$, so that 1.0 marks the 5σ discovery-potential level at the corresponding γ. This centering additionally exposes MLE bias and the boundary-atom under-coverage. Color encodes the fraction of trials below the threshold; pure Gaussianity predicts ≈ 0.6827. Diverging colormap: magenta = under-covers, cream = nominal, Blues = over-covers.

The MLE bias documented in the signal-recovery diagnostics (Section 9.10) is also directly visible. At soft spectra, $\hat{\gamma}$ under-recovers, and the most extreme under-coverage zone gets pushed into the $\gamma \approx 3.5$ interior, where it would otherwise sit only outside the model fit bound at $\gamma > 4$. The hard side behaves differently: the extreme under-coverage stays outside the fit bound at $\gamma < 1$, and just above $\gamma = 1$ we see only a narrow band of severe under-recovery before it eases to mild under-recovery and approaches Wilks agreement deeper in. Even there, the active-fit interior still under-covers: empirical coverage runs around 58% at the 1σ Wilks threshold.

A separate band of severe under-recovery sits in the low-n_s regime, adjacent to the $\hat{n}_s = 0$ atom, with a gap of milder under-coverage (closer to the Wilks prediction) running through the center that moves with declination. The gap may again be related to the PSF calibration, here combined with distance from the γ atoms: the central $\gamma \approx 2.5$ region has the best PSF agreement and the greatest distance from both γ atoms, and the least severe coverage issues land there,

consistent with both mechanisms acting together.

This plot alone cannot disentangle which of the regularity tensions in Table 12.1 produce which observed features. The pattern is most likely a complex combination of all of them, and our interpretations of specific features above are largely speculative. What is not speculative is that Wilks does not apply anywhere on the simulation grid.

Modus tollens

The closure check is modus tollens applied to the textbook chain that delivers the asymptotic χ^2 result. The implication direction runs from Theorem 16.7's (Section 12.1) regularity conditions (i)–(iv) to the prediction that the per-cell Mahalanobis distance to the empirical centroid follows χ^2_2 on our 2D parameter space: under conditions (i)–(iii) the MLE is asymptotically Gaussian,[230] with the rigorous version due to van der Vaart,[231] and a sum of k squared standard normals is χ^2_k[232] (via the whitening transform $z = \Sigma_{\hat{\theta}}^{-1/2}(\hat{\theta} - \mu_{\hat{\theta}}) \sim \mathcal{N}(0, I)$). With 10^4 pseudo-experiments per truth cell the MC sample size is more than sufficient to resolve any genuine χ^2_2 limit, so the systematic deviations observed across substantial regions of the simulation grid imply that the χ^2_2 prediction fails, hence at least one of (i)–(iv) is non-trivial wherever the closure fails. We do not attempt to attribute specific deviations to specific conditions. The boundary atoms (a manifestation of (iv)) are excluded from the Mahalanobis cloud by construction, but boundary truncation can still distort the active-fit interior near parameter-space edges, so cells near boundaries can carry mixed (i)–(iv) contributions. The plot is *not* a prerequisite for our coverage (the empirical-FC construction we use gives exact coverage by Neyman duality, Section 12.6, regardless of whether T is asymptotically χ^2_2), but it documents the empirical inadequacy of Wilks for this analysis. Wilks-based confidence regions would be invalid across the parameter grid (under-covering near boundary atoms, over-covering at specific interior locations), making FC the necessary construction throughout for both regimes (sources where H_0 is rejected and where it is not), not merely a preferred one.

[230] Casella and Berger 2002, Sec. 10.1.2, p. 470.

[231] Vaart 1998, Theorem 5.39, p. 65.

[232] Casella and Berger 2002, Sec. 5.3, p. 218.

12.2 The empirical Feldman-Cousins construction

Our parameter estimation uses the unified prescription of Feldman and Cousins[233] over the joint parameter space $\theta = (n_s, \gamma)$, with the boundary $n_s \geq 0$ enforced as in the original construction.

[233] Feldman and Cousins 1998, "Unified approach to the classical statistical analysis of small signals".

Intuition: how the ranking yields a confidence region

Before the technical description below, an intuitive picture of what the construction does. For each candidate parameter θ, simulating many pseudo-experiments at θ gives a complete picture of *what observations look like under* θ: the sampling distribution $p(x; \theta)$. Within that sampling distribution, define an *acceptance region*

$A(\theta)$ as the subset of data space containing the observations most consistent with θ, totaling probability $1-\alpha$ under $p(x;\theta)$. What *most consistent* means is the ranking choice. We use the empirical likelihood ratio (see Section 12.2), but the coverage argument works for any choice. The *confidence region* $C(x_{\text{obs}})$ is then the set of θ values whose acceptance region contains the actual observation: "θ is in the CI" if and only if "the observed data is a not too unusual outcome under θ."

The frequentist coverage guarantee follows directly from the construction: when $\theta = \vartheta$ (the truth), the actual observation is drawn from $p(x;\vartheta)$ by definition; $A(\vartheta)$ contains $1-\alpha$ of that probability mass by construction; therefore $x_{\text{obs}} \in A(\vartheta)$ a fraction $1-\alpha$ of the time, equivalently $\vartheta \in C(X)$ a fraction $1-\alpha$ of the time. The coverage holds regardless of how the ranking was defined: this is the Neyman duality argument (Section 12.6), applied to a simulation-based estimate of the sampling distribution.

The technical specification of the construction follows.

Observable vector

$x = (\hat{n}_s, \hat{\gamma})$ from the same likelihood fit applied to the unblinded data. The empirical density $\hat{p}(x;\theta)$ on x is the input to both the calibrated MLE (Section 12.3) and the empirical-LR ranking (Section 12.2) below.

Conceptually, the model likelihood serves here as feature compression on the event-level data: it takes the $\mathcal{O}(N_{\text{events}})$-dimensional raw observables (per-event positions, energies, angular errors) and projects them onto the 2D summary $(\hat{n}_s, \hat{\gamma})$ that the empirical density operates on.

In the equations below, x should be read as the MLE pair $(\hat{n}_s, \hat{\gamma})$, not as event-level data, even though the classical FC reference equations use x in that more general sense. The two readings are operationally equivalent (the MLE compression step is implicit in $\hat{p}$, so $\hat{p}(x;\theta)$ reduces to $\hat{p}((\hat{n}_s, \hat{\gamma});\theta)$ either way), but throughout this section we work with x as the post-compression summary that the empirical density operates on.

The $;$ convention (Section 7.1) applies cleanly to $\hat{p}(x;\theta)$, but the two slots play roles that are easy to confuse—and the role each (n_s, γ) symbol plays depends on which level of the construction we are at. At the FC level, the frequentist parameter is whatever indexes the family of sampling distributions on the ranking function: here the truth $\theta = (n_s, \gamma)$ on the right of $\hat{p}(x;\theta)$. The MLE outputs $(\hat{n}_s, \hat{\gamma})$ on the left, which *would* be parameters at the model-likelihood level $L(x_{\text{events}};\theta)$, are random variables in this context: realizations of the MLE statistic $\hat{\theta}(X_{\text{events}})$ with their own sampling distribution under each candidate truth. The role flip is because at the FC level we use L purely as a *compression function* mapping event-level data to the summary $(\hat{n}_s, \hat{\gamma})$, not as a probabilistic model in its own right. The probabilistic model at this level is $\hat{p}(x;\theta)$. Estimator and estimand share the symbol family ($\hat{\theta}$ vs. θ, hat vs. no-hat) but live in opposite slots: the parameter θ indexes the family, and the random-variable realization $\hat{\theta}$ is the value the family assigns probability density to.

Parameter grid

Pseudo-experiments are run on a discrete truth grid $\Theta = \{\theta = (n_s, \gamma)\}$ at the hypothesized source declination. The grid covers $n_s \in [0, n_{\max}]$ (where $n_{\max}$ is set well above the largest plausible signal level) and $\gamma \in [0.5, 4.5]$ at uniform spacing $\Delta\gamma = 1/64 \approx 0.0156$ (257 points). The truth-grid extent in γ is intentionally wider than the model-fit bounds $\hat{\gamma} \in [1, 4]$. Per-cell coverage holds throughout the simulated grid under the simulated signal model (see Section 10.2).

Our construction is strict frequentist on Θ: θ is a parameter, not a random variable, and we attach no probability density to it. The Bayesian alternative would treat θ as a random variable with a posterior on Θ, with off-grid statements coming from evaluating the posterior density. We don't take that path. Frequentist-compatible smoothing alternatives also exist: simulation-based inference methods that learn a continuous surrogate for the likelihood ratio or the per-cell sampling density across nearby cells to absorb finite-N MC noise,[234] but those buy off-grid resolution at the cost of swapping exact per-cell Neyman coverage for asymptotic coverage in the simulation-budget limit. We prefer the exact guarantee—empirical acceptance regions $A(\theta)$ at the grid points we sampled, with Neyman duality (Section 12.6) giving exact coverage at those points alone and no claim off-grid.

This is not a claim that off-grid points behave pathologically: if the true likelihood ratio were known at every θ, Wilks-like smoothness would hold and off-grid points deep inside the region would qualitatively belong. What we lack is a *quantitative* coverage statement off-grid, and post-hoc interpolation in Θ does not provide one: linear interpolation contributes nothing because the argmax of a linearly-interpolated landscape still lands on grid points, and nonlinear interpolation creates artificial maxima between grid points that are interpolation artifacts rather than characteristics of the true sampling distribution. Our resolution in Θ is therefore set by the grid. The way to add resolution is to simulate at a finer grid—not to smooth what we have.

[234] Cranmer, Pavez, and Louppe 2015, "Approximating Likelihood Ratios with Calibrated Discriminative Classifiers", Dalmasso et al. 2024, "Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning for Reliable Simulator-Based Inference".

Pseudo-experiments

At each grid point $\theta = (n_{\text{inj}}, \gamma_{\text{inj}})$, where n_{inj} is the expected signal-event count ($= \Phi_0 \cdot A_{\text{eff}}(\gamma, \delta)$ using the per-declination acceptance), N_{trials} pseudo-experiments are generated. Each pseudo-experiment takes RA-randomized real data as the background substrate (identical to the empirical-null pseudo-experiments (Section 9.9) used for the significance result), then injects $n^* \sim \text{Poisson}(n_{\text{inj}})$ MC signal events at the hypothesized position with spectral index γ_{inj}, matching the original Feldman and Cousins[235] construction (Sec. III.B, Eq. 3.2: the acceptance belt at truth μ is built by summing $P(n; \mu) = \text{Poisson}(\mu + b)$ over the count n). (Sec. III.B, Eq. (3.2), p. 3876) The fit is applied with the same code path as the unblinded analysis, giving $(T, \hat{n}_s, \hat{\gamma})$ for each trial. The simulation grid is therefore a grid of empirical joint sampling distributions of the MLE $(\hat{n}_s, \hat{\gamma})$ under the injected truth $(n_{\text{inj}}, \gamma_{\text{inj}})$.

[235] Feldman and Cousins 1998.

Ranking function

[236] Feldman and Cousins 1998.

The classical Feldman and Cousins[236] prescription uses the model likelihood-ratio ordering, $R_{\text{model}}(x;\theta) = L(x;\theta)/L(x;\hat{\theta}(x))$, where $\hat{\theta}(x)$ is the constrained MLE under $n_s \geq 0$. This ranking has a clean operational interpretation: $R_{\text{model}} = 1$ at the MLE by construction, so the resulting CI naturally centers on $\hat{\theta}$, and $R_{\text{model}} \ll 1$ means the data strongly prefer some other parameter under the model. Sorting pseudo-experiments by R_{model} and taking the top $1-\alpha$ fraction picks out the observations for which θ is a defensible explanation relative to the best alternative the model can find. When L exactly equals the true sampling density, $\hat{\theta}$ is unbiased, the CI is correctly anchored, and the LR ordering is the most powerful by Neyman–Pearson (Section 7.3) reasoning.[237]

[237] Neyman and Pearson 1933, "On the Problem of the Most Efficient Tests of Statistical Hypotheses".

In general we cannot assume $\hat{\theta}$ is unbiased: the $n_s \geq 0$ boundary, the n_s–γ degeneracy along the flux-acceptance ridge, and any residual PDF mismatches (chiefly the vMF PSF tail mismatch detailed below) all contribute, and the actual bias depends strongly on the spectral index. At least part of the cause is almost certainly in the PSF calibration: csky's vMF PSF has lighter tails than the true angular-error distribution and is calibrated at a single effective $\gamma_{\text{cal}} = 2.5$ as a forced compromise (γ-dependent pull correction is currently not implemented and likely computationally infeasible for our analysis, see Section 8.1 for the methodology limitation), so soft-spectrum events, which sit preferentially in the angular tail, are treated as background-like by the model because they receive lower spatial signal weights than they would under the true PSF, and the result is n_s under-recovery. The same trend is observed for all data samples (including NT) when using the von Mises–Fisher (or, equivalently in the small-angle limit, Rayleigh) distribution for PSF modeling, while published SkyLLH-based Northern Tracks analyses,[238] which use γ-conditioned KDE PSFs,[239] do not exhibit it, supporting the diagnosis that the bias is driven by the single-γ vMF PSF model rather than by selection or simulation issues. The full discussion of the mechanism is in Section 9.10. Coverage of ϑ is still exact regardless: by Neyman duality, the model-LR CI does cover the truth at the nominal rate. What's biased is the *CI center*: classical FC anchors on $\hat{\theta}$, which is systematically off whenever the underlying MLE is. The CI is then a confidence statement built around the biased best-fit value. The empirical-LR construction below corrects this without compromising the coverage guarantee. The relative width compared to the classical model-LR FC is set out in Section 12.6.

[238] IceCube Collaboration 2022a.

[239] IceCube Collaboration 2026a, Sec. 3 & Fig. 7.

We replace the model likelihood ratio with its *empirical likelihood-ratio* (empirical-LR) counterpart,

$$R(x;\theta) = \frac{\hat{p}(x;\theta)}{\hat{p}(x;\tilde{\theta}(x))}, \tag{12.2}$$

where $\hat{p}$ is the empirical sampling density estimated from the same pseudo-experiment grid that builds the FC acceptance regions (but from an independent draw; see Section 12.3), and $\tilde{\theta}(x) = \arg\max_\theta \hat{p}(x;\theta)$ is the calibrated MLE on the physically allowed grid (the mode of the empirical density at the observation; see Section 12.3). The model likelihood L in both numerator and denominator is

replaced by the empirically estimated sampling density $\hat{p}$, and the model MLE $\hat{\theta}$ is replaced by the calibrated MLE $\tilde{\theta}$. The operational meaning of the ratio is the same as in the classical case ($R \approx 1$ at the calibrated MLE; $R \ll 1$ for implausible candidates), but with the bias-correction baked in.

The structural ingredients of this construction are well-established in the recent simulation-based inference and likelihood-free inference (LFI) literature. The *Likelihood-Free Frequentist Inference (LF2I)* framework of Dalmasso et al.[240] covers the LR-based test statistic in Sec. 3.2.1, the empirical-quantile critical-value construction in Sec. 3.3, and the coverage guarantee independent of the test-statistic choice in Theorem 1 (Secs. 3.2.1, 3.3 & Thm. 1, p. 5056). The ACORE test statistic of Dalmasso et al.[241] is a precursor. The broader simulation-based inference lineage is reviewed in Cranmer et al.[242]. What that literature typically does *on top* of these ingredients is substantial ML machinery: neural classifiers for odds-ratio estimation, neural density estimators, learned off-grid surrogates. Our construction is much simpler—classical Monte Carlo pseudo-experiments at the discrete truth grid points $\theta \in \Theta$, with the per-cell empirical sampling density estimated by KDE on the per-cell MC samples (covered in Section 12.3). No neural networks, no learned surrogates, no parameter-space density estimation. The references are cited for the structural overlap, not for any ML methodology we adopt.

[240] Dalmasso et al. 2024.

[241] Dalmasso, Izbicki, and Lee 2020, "Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting".

[242] Cranmer, Brehmer, and Louppe 2020, "The frontier of simulation-based inference".

Two principled advantages over the classical model-LR follow:

1. The CI naturally centers on the bias-corrected estimate $\tilde{\theta}$ rather than on the biased model MLE $\hat{\theta}$, since $R \approx 1$ at $\tilde{\theta}$ by construction.

2. No model likelihood enters the ranking, so PDF mismatches that bias L have no influence on R. In the limit of infinite simulation statistics, $\hat{p} \to p_{\text{true}}$ and the empirical-LR converges to the likelihood ratio built from the *true* sampling density of the MLE summary $(\hat{n}_s, \hat{\gamma})$, which is the most powerful ranking on summary-statistic space by Neyman–Pearson reasoning. The relationship to the classical model-LR FC is made precise in Section 12.6: the two constructions yield identical confidence regions when $\hat{\theta}$ is sufficient for θ (in particular, asymptotically when the likelihood is correctly specified) and diverge in centroid and shape under misspecification. Other rankings give correct coverage too (the Neyman duality is independent of the choice of ordering), but the empirical-LR is the cleanest principled choice in our regime. The transition from an upper limit to a two-sided interval happens automatically as the data shifts from "consistent with no signal" to "supporting a non-zero signal," exactly as in the classical FC construction.

The non-negativity constraint $n_s \geq 0$ is built into the candidate grid Θ and inherited by $\tilde{\theta}(x)$, preserving the unified-construction property of the original FC prescription that "yields intervals which automatically change over from upper limits to two-sided intervals as the 'signal' becomes more statistically significant. This eliminates undercoverage caused by basing this choice on the data ('flip-flopping')."[243]

[243] Feldman and Cousins, "Unified approach to the classical statistical analysis of small signals", p. 16.

LR throughout this section refers specifically to this empirical-LR ranking function on candidate parameter points θ, not to the log-likelihood ratio $T = 2\ln[L(x;\hat{\theta})/L(x;0)]$ that defines our significance test statistic (Section 10.2). The two are conceptually related (both are ratios of probabilities at x across two parameter points) but operationally distinct: T uses the model likelihood L with the MLE in the numerator, the FC ranking uses the empirical density $\hat{p}$ with the calibrated MLE in the denominator.

Acceptance and confidence regions

At each truth grid point $\theta \in \Theta$, sort the pseudo-experiments by $R(x;\theta)$ and define the acceptance region $A(\theta;\alpha)$ to contain the top $1-\alpha$ fraction. The confidence region for an observation is then

$$C(x_{\text{obs}};\alpha) = \{\theta \in \Theta \,:\, x_{\text{obs}} \in A(\theta;\alpha)\}. \tag{12.3}$$

This is the standard Neyman construction read in two directions: A partitions data space at fixed θ; C collects parameter points whose acceptance regions contain the observation. We build both at the 1σ Gaussian level (the headline) and the 2σ Gaussian level (supplementary outer contour). Per HEP convention, these are the exact 1D Gaussian probability masses $\text{erf}(1/\sqrt{2}) \approx 0.6827$ and $\text{erf}(\sqrt{2}) \approx 0.9545$ up to float precision.

Reparameterization between (n_s, γ) and $(\Phi(E), \gamma)$

The construction is native to (n_s, γ) space. This native parameterization is in bijection with $(\Phi(E), \gamma)$ only when the signal model assumed in the likelihood matches the one for which a coverage statement is sought. To cover systematic uncertainties we also construct intervals for mismatched signal models (Section 12.7); there the bijection breaks—the acceptance differs trial by trial on the truth side—and the simulation grid must then be built directly in Φ as well. The reparameterization to $(\Phi(E), \gamma)$ through the per-declination signal acceptance ($\Phi_0 = n_s/A_{\text{eff}}(\gamma,\delta)$ at $E = 1$ TeV, or $\Phi(E) = \Phi_0 \cdot (E/1\,\text{TeV})^{-\gamma}$ at any other reference) is deterministic and lossless: every truth grid point maps bijectively to a corresponding $(\Phi(E), \gamma)$ point, and the per-cell $A(\theta;\alpha)$ thresholds are exactly preserved. Only the *shape* of the 2D region changes: the $(\Phi(E), \gamma)$ region is a sheared and scaled version of the (n_s, γ) region because $\Phi(E)$ depends on γ.

The 1D profile-FC construction depends on which axis is being profiled, and the profile lines do not transform trivially under reparameterization. For the γ POI, the inner-profile slice at fixed γ is the same physical set of grid cells in either parameterization (since γ is on the same axis in both). For the n_s POI vs the $\Phi(E)$ POI, the slicing directions are genuinely different: a slice at fixed n_s in (n_s, γ) is a curve at varying $\Phi(E)$ in $(\Phi(E), \gamma)$ (because $\Phi(E)$ depends on γ), and a slice at fixed $\Phi(E)$ in $(\Phi(E), \gamma)$ is a curve at varying n_s in (n_s, γ). The resulting profile lines for the n_s POI vs the $\Phi(E)$ POI are therefore different physical curves through 2D

parameter space, not just relabellings of each other. Profile-FC intervals must be computed natively in the parameterization in which they are reported. They do not transform trivially between parameterizations.

Pivot energy selection

The *pivot energy* E_{pivot} (also commonly called the *decorrelation energy*) is the per-source energy at which the local correlation between $\Phi(E)$ and γ within the FC region is at its minimum: equivalently, the energy at which the 1D profile-FC interval on $\Phi(E)$ has the smallest residual dependence on γ. It is determined per source from the 2D FC region itself, by sweeping the test energy E across a grid and choosing the value that minimizes the projection width of the 2D region onto $\log_{10}\Phi(E)$. Reporting fluxes at the per-source pivot energy is standard practice in gamma-ray astronomy.[244]

Figure 12.3 shows an example: as the test energy is swept, the projection width of the 2D FC region onto $\log_{10}\Phi(E)$ falls to a minimum at the pivot energy and rises on either side.

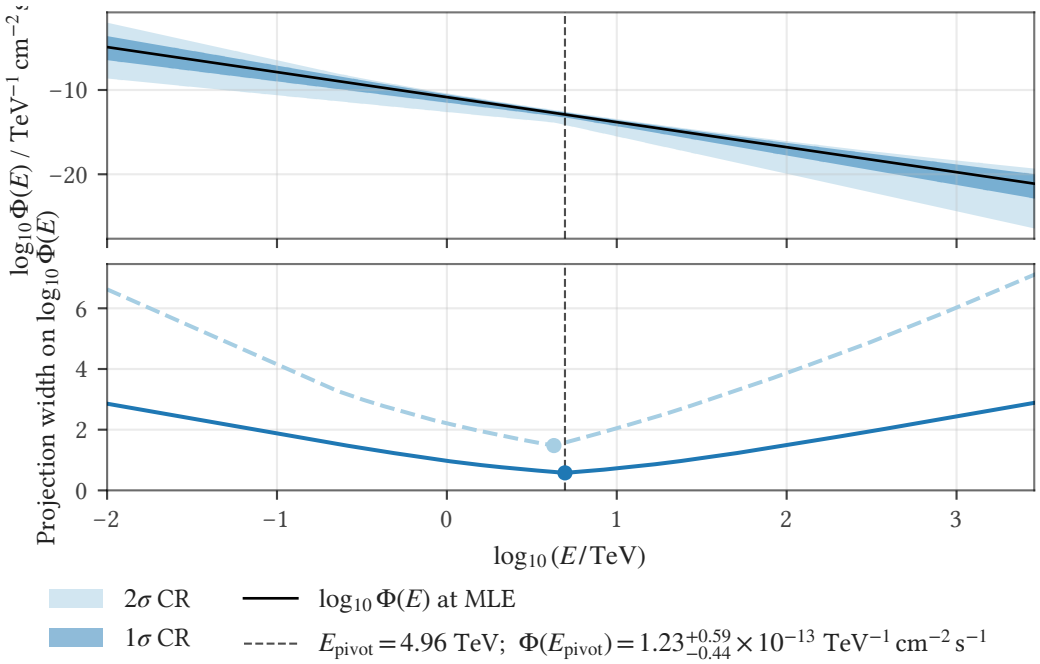

Figure 12.3: Example pivot-energy selection. The projection width of the 2D Feldman–Cousins region onto $\log_{10}\Phi(E)$, swept across test energies, is minimized at the pivot energy E_{pivot}; reporting Φ there gives the flux its smallest residual dependence on γ.

The geometric consequence of the decorrelation is that the 2D region in $(\Phi(E_{\text{pivot}}), \gamma)$ coordinates is approximately axis-aligned, and the profile lines for both axes approach orthogonals to those axes. The 1D profile-FC interval on

[244] Fermi-LAT Collaboration, "Fermi Large Area Telescope Fourth Source Catalog", *The Astrophysical Journal Supplement Series* 247, no. 1 (2020): Sec. 3.3, H.E.S.S. Collaboration, "Gamma-ray blazar spectra with H.E.S.S. II mono analysis: the case of PKS 2155-304 and PG 1553+113", *Astronomy & Astrophysics* 600 (2017): footnote 1.

$\Phi(E_{\text{pivot}})$ is therefore approximately decorrelated from γ, and the joint $(\Phi(E_{\text{pivot}}), \gamma)$ region is approximately the product of its two 1D profile-FC intervals.

At any other reference (the native n_s axis itself, or Φ at a non-pivot energy), the profile lines through the 2D region are curved and the 1D interval carries non-trivial γ-correlation residue. The 2D region is correspondingly sheared. The pivot-energy reporting choice is what makes the 1D Φ number admit a clean per-axis decorrelation interpretation. For any other reference, the 1D Φ number alone cannot reproduce the joint structure.

1D profile-FC intervals

Headline 1D intervals on γ and on $\Phi(E_{\text{pivot}})$ are constructed by applying the POI-profiling procedure (Section 12.4) with the empirical-LR ranking. The 1D interval is the range of POI values whose observed profile-LR exceeds the per-cell critical value. The strict-frequentist discrete-grid version we use is the NOvA "Profiled Feldman-Cousins approach".[245]

[245] NOvA Collaboration, "Monte Carlo method for constructing confidence intervals with unconstrained and constrained nuisance parameters in the NOvA experiment", "Profiled Feldman-Cousins approach".

The γ interval uses slices at fixed γ (the same slice in either parameterization). The $\Phi(E_{\text{pivot}})$ interval uses slices at fixed $\Phi(E_{\text{pivot}})$, which is a column slice in the $(\Phi(E_{\text{pivot}}), \gamma)$ parameterization (each cell's $\Phi(E_{\text{pivot}})$ is computed from its native (n_s, γ) through the deterministic reparameterization above). Both have exact 1D Neyman coverage.

A subtlety worth being explicit about: the unique test statistic values belonging to each pseudo-experiment used in the 1D profile-FC construction are *not* the same as those used to build the 2D FC region. In the 2D construction at truth cell θ, the test statistic is the empirical-LR $R(x;\theta)$ evaluated at that specific cell. In the 1D profile-FC construction at POI value θ_{POI}, the test statistic is the profile-LR

$$R_{\text{prof}}(x;\theta_{\text{POI}}) = \max_{\theta_{\text{nuis}}} R(x;\theta_{\text{POI}},\theta_{\text{nuis}}), \tag{12.4}$$

where the inner argmax is taken independently for each pseudo-experiment in the slice. The test statistic, computed for each pseudo-experiment, is therefore evaluated at the conditional-MLE nuisance for *that* pseudo-experiment, which is generally a different cell than the conditional-MLE nuisance for any other pseudo-experiment. The per-cell α-quantile of this new test statistic distribution defines the profile-FC threshold separately from the 2D threshold. This is the source of the power gain over a marginal 2D projection: the 1D construction directly tests the 1D POI hypothesis with the conditional MLE absorbing nuisance variation per pseudo-experiment, whereas the marginal projection of the 2D region tests the joint 2D hypothesis and then collapses the result, losing the per-experiment refit information. The marginal projection therefore over-covers relative to the nominal 1D level. See the NOvA "Conservative" discussion[246] for an explicit demonstration of this failure mode and the case for proper profile-FC as the principled alternative.

[246] NOvA Collaboration, "Monte Carlo method for constructing confidence intervals with unconstrained and constrained nuisance parameters in the NOvA experiment", "Conservative".

The $\Phi(E_{\text{pivot}})$ profile-FC carries a small discretization detail not present in the γ profile. The native truth grid is in (n_s, γ) coordinates with n_s sampled on a hybrid grid: $\Delta n_s = 0.5$ steps from $n_s = 0.5$ to 10, and $\Delta n_s = 1$ above. The finer

low-n_s sampling is deliberate: in extreme cases the 5σ discovery potential sits below $n_s = 10$, and in most cases at least the 3σ discovery potential for $\gamma = 2$ is around $n_s = 10$, so sampling the Poisson μ more finely at low n_s keeps the relative flux error in a roughly similar regime across the grid. Candidate $\Phi(E_{\text{pivot}})$ slices generally do not align with grid cells regardless: a curve of constant $\Phi(E_{\text{pivot}})$ in (n_s, γ) specifies a real-valued n_s for each γ, and the finite Δn_s means the exact target rarely sits on a cell. The implementation handles this by taking each native (n_s, γ) cell as its own candidate POI value ($\sim 5 \times 10^4$ candidates in the production grid, irregularly spaced in $\Phi(E_{\text{pivot}})$ by construction), and for each candidate cell θ_{self} assembling the slice by snapping to the nearest n_s on each γ-row (one cell per γ-column, $\sim$193 cells per slice). The per-column snap error is well below 1% relative in $\Phi(E_{\text{pivot}})$ across the populated grid. One simplification in the grid construction above will need revisiting once per-trial systematic marginalization is in place. Proper per-trial Cousins–Highland draws make the detector acceptance depend on the systematic parameters drawn for each trial, so the truth-side $n_s \leftrightarrow \Phi$ conversion is no longer a fixed mapping. The treatment we adopt in principle is to use, on the truth side, the acceptance of each trial's own draw (it is part of the simulation), while the measured side always converts $n_s \to \Phi$ through the single fixed unblinding acceptance (part of the estimator, which must not depend on the truth). A further complication arises in two dimensions: because the n_s and Φ profiles differ across γ, interval endpoints cannot simply be transformed between the two parameterizations, which suggests sampling the grid independently in n_s and Φ. Per-trial marginalization is not implemented yet; the baseline-equivalent construction actually used, and the systematics treatment that accompanies it, are described in Section 12.7.

This snap to the nearest cell is *not* in tension with the lossless per-cell $(n_s, \gamma) \to (\Phi(E), \gamma)$ mapping discussed earlier (Section 12.2): the forward map (each cell to its exact Φ value) is exact. Only the inverse operation of finding cells on a constant-Φ slice has to live with our finite n_s resolution. Interpolating $\hat{p}$ between n_s cells would violate our strict-grid no-interpolation policy, and the snap error is far below the per-cell MC noise on $\hat{p}$ in any case.

The discrete-grid argmax also avoids numerical pathologies (NaN-discontinuous landscape near grid edges, spurious convergence on partial calibration grids) that an off-grid continuous-profile optimizer would face. If finer resolution is needed, we simulate at a finer grid.

12.3 Bias-corrected point estimate

The maximum-likelihood estimator $\hat{\theta}(x)$ is biased in general (see Section 9.10). The bias arises from a combination of the $n_s \geq 0$ boundary, the n_s–γ degeneracy along the flux-acceptance ridge, and the irreducible PDF mismatches that any analysis carries: here chiefly the tail-light vMF PSF and the energy-PDF imperfections (Section 9.10). None of this affects the validity of the construction (Section 12.6) below, but $\hat{\theta}$ is not the right value to report as the best-fit parameter.

We report instead a point estimate that inverts the empirical joint sampling distribution from the same simulation grid that builds the FC acceptance regions:

$$\tilde{\theta} = \arg\max_{\theta} \ \hat{p}(x_{\text{obs}}; \theta). \tag{12.5}$$

This is the empirical analog of the classical maximum-likelihood estimator: the same argmax-in-θ construction, with the model likelihood $L(x;\theta)$ replaced by the empirically estimated sampling density $\hat{p}(x;\theta)$. In the asymptotic limit of correct model specification and infinite simulation statistics, the calibration reduces to the identity transform and $\tilde{\theta} = \hat{\theta}_{\text{obs}}$ (formal statement: see Section 12.6). Under model misspecification, the transform is non-trivial and absorbs the model's recovery bias by construction. Equivalently: $\tilde{\theta}$ is the parameter value for which the observed $(\hat{n}_s, \hat{\gamma})$ is most typical of the actual sampling distribution at θ, in contrast to $\hat{\theta}$ which optimizes the original likelihood model and inherits whatever bias the model carries.

As a baseline alternative, the *marginal-median* estimator finds the truth value that produces observation-matched marginal medians:

$$\tilde{\theta}_{\text{med}} \ : \ \text{median}[\hat{n}_s \,;\, \tilde{\theta}_{\text{med}}] = \hat{n}_s^{\text{obs}}, \qquad \text{median}[\hat{\gamma} \,;\, \tilde{\theta}_{\text{med}}] = \hat{\gamma}^{\text{obs}}. \tag{12.6}$$

By construction, the marginal-median estimator is *median-roundtrip consistent*: at the chosen $\tilde{\theta}_{\text{med}}$, the marginal medians of pseudo-experiments at that truth match the observation exactly, so the recovery diagnostic below would show the marginal-median curve sitting on the diagonal by definition. But median-roundtrip and empirical-MLE answer different questions: the calibrated MLE asks *which truth makes the observation most likely under the empirical sampling density?* (the natural empirical-MLE question), while the marginal median asks *which truth, if injected, would typically produce the observation in the marginal-median sense?* For symmetric distributions where mode and marginal median coincide, the two estimators give identical results; for asymmetric distributions, they differ—and the difference is a property of the cloud asymmetry, not a defect of either estimator.

The empirical sampling distribution in our regime is asymmetric in ways that make the two estimators land at different points. The joint structure is non-trivial: $(\hat{n}_s, \hat{\gamma})$ are correlated along the flux-acceptance ridge. The Pearson correlation $\rho(\hat{n}_s, \hat{\gamma})$, measured directly on 10^4 pseudo-experiments per truth point at $\sin\delta = 0$, varies systematically with the injected spectral index. At the 3σ-discovery-potential-equivalent injection levels we measure $\rho = 0.71$ at $\gamma_{\text{inj}} = 2.0$, $\rho = 0.63$ at $\gamma_{\text{inj}} = 2.5$, $\rho = 0.39$ at $\gamma_{\text{inj}} = 3.0$, and $\rho = 0.21$ at $\gamma_{\text{inj}} = 3.5$. The trend has a clean physical interpretation: hard spectra produce high-energy events that simultaneously constrain n_s and γ, so a fluctuation in one correlates with a fluctuation in the other; soft spectra are dominated by low-energy events that resemble background and constrain γ less tightly. The ridge is therefore most prominent at hard spectra and weakens toward soft spectra ($\gamma \sim 3.4$). On top of the correlation, the per-cell shapes are non-Gaussian: boundary-driven skewness near $\hat{n}_s = 0$ and asymmetric tails along the ridge at hard injection both decouple the joint mode from the marginal

medians. The calibrated MLE inverts the full joint $\hat{p}(x;\theta)$ and inherits both the correlation and the asymmetry automatically.

We adopt the calibrated MLE and do not report the marginal-median estimator—it uses the same per-cell density model as the FC inversion machinery downstream (so the centroid and the contour come from one unified procedure) and centers the FC contour on itself by construction. The empirical-LR ranking gives $R(x;\tilde{\theta}) \approx 1$ at the calibrated MLE, approximate rather than exact because finite-stats sampling noise on the per-cell density estimate $\hat{p}$ can displace its maximum slightly from the true mode.

Boundary mixture model

We model the empirical sampling distribution at each truth point as a four-component mixture: three atoms induced by the fit's parameter-space boundaries plus an active-fit continuous interior. The fit constrains $\hat{n}_s \geq 0$ and $\hat{\gamma} \in [1,4]$. Pseudo-experiments whose unconstrained MLE lands outside the allowed region collapse onto the corresponding boundary. Each atom is modeled as an explicit component of $\hat{p}$ with its own per-truth empirical weight and (where applicable) its own conditional density.

- $\hat{n}_s = 0$ **point mass:** When the model-likelihood maximum sits on the $n_s = 0$ boundary, the spectral index is unidentified: the likelihood is independent of γ when there is no signal contribution, so the maximum is the entire line $\{(0,\gamma) : \gamma \in [1,4]\}$ rather than a single $(\hat{n}_s, \hat{\gamma})$ pair. We treat $\hat{\gamma}$ on this atom as statistically undefined and exclude it from any γ-related estimation. This is the Davies problem[247] operating at the boundary, here treated by leaving $\hat{\gamma}$ unassigned. The atom carries empirical mass $w_{n_s=0}(\theta)$, the fraction of pseudo-experiments at θ that fit to $\hat{n}_s = 0$.

[247] Davies 1977, Davies 1987.

- $\hat{\gamma} = 1$ **and** $\hat{\gamma} = 4$ **boundary lines:** Trials whose unconstrained $\hat{\gamma}$ MLE would land outside $[1,4]$ get clipped to the corresponding bound, producing 1D atoms along $\{\hat{\gamma} = 1\}$ and $\{\hat{\gamma} = 4\}$ in observation space. Clipping mostly affects low-TS background fluctuations whose unconstrained spectral fit is poorly determined, but the $\hat{\gamma}$-boundary atoms can be load-bearing at higher TS too: a single Poisson realization can clip the $\hat{\gamma}$-fit boundary at the 1σ level, as shown later in Figure 12.12. This is one of the regimes where Wilks-asymptotic intervals fail and the FC construction still gives correct coverage: accurate modeling of the atoms in the mixture is methodologically necessary, not cosmetic. Unlike the $\hat{n}_s = 0$ atom, these atoms have meaningful 1D structure: γ is the *constrained* parameter at its boundary, so $\hat{n}_s$ takes a continuous distribution along the atom line. Empirical mass is $w_{\gamma=1}(\theta)$ and $w_{\gamma=4}(\theta)$. Conditional densities $p_{\hat{n}_s|\hat{\gamma}=1}(\hat{n}_s;\theta)$ and $p_{\hat{n}_s|\hat{\gamma}=4}(\hat{n}_s;\theta)$ describe the $\hat{n}_s$ distribution along each line.

- **Active-fit continuous interior:** Trials with $\hat{n}_s > 0$ and $\hat{\gamma} \in (1,4)$ form the continuous 2D component $p_{\text{cont}}(x;\theta)$ on the interior support $(0,\infty) \times (1,4)$.

We partition the atoms disjointly by precedence on $\hat{n}_s$: trials with $\hat{n}_s = 0$ go to the $\hat{n}_s = 0$ atom (with $\hat{\gamma}$ treated as undefined), and the $\hat{\gamma}$-boundary atoms collect only the *active-fit* trials ($\hat{n}_s > 0$) on each $\hat{\gamma}$ boundary. The full mixture is

$$\begin{aligned}\hat{p}(x;\theta) = {} & w_{n_s=0}(\theta)\,\delta_{\hat{n}_s=0}(x) \\ & + w_{\gamma=1}(\theta)\,\delta_{\hat{\gamma}=1}(x)\,p_{\hat{n}_s|\hat{\gamma}=1}(\hat{n}_s;\theta) \\ & + w_{\gamma=4}(\theta)\,\delta_{\hat{\gamma}=4}(x)\,p_{\hat{n}_s|\hat{\gamma}=4}(\hat{n}_s;\theta) \\ & + \left(1 - w_{n_s=0} - w_{\gamma=1} - w_{\gamma=4}\right) p_{\text{cont}}(x;\theta),\end{aligned} \tag{12.7}$$

with δ's denoting indicators of the boundary sets. We estimate all weights and densities separately at each truth point.

In practice, the atoms are significant across most of the operating regime, not just at edge cells. As one example point: at $(n_{\text{inj}}, \gamma_{\text{inj}}) = (20, 3.0)$, $\sin\delta = 0$ (a weak-signal cell), ∼15% of trials land in the $\hat{n}_s = 0$ atom, and that fraction grows as truth n_s decreases toward zero. The $\hat{\gamma} = 1$ atom is a few-percent effect at hard truth ($\gamma_{\text{inj}} \lesssim 2.5$); the $\hat{\gamma} = 4$ atom is its mirror at soft truth, load-bearing for very soft sources where FC contours can clip at the $\hat{\gamma} = 4$ boundary at the 1σ level (as already noted in the bullet above). Atom magnitudes recede only where the truth sits jointly far from all three boundaries: high enough n_s to thin the $\hat{n}_s = 0$ atom and γ far enough from both $\{1, 4\}$ to thin the $\hat{\gamma}$-boundary atoms. Anywhere else, at least one atom is non-negligible. This pattern is one of the principal reasons we cannot defer to Wilks asymptotics for parameter estimation—a Gaussian-asymptotic construction has no representation for boundary atoms at all, so it can be approximately right only in a deep-signal regime where all three atoms are far from the contour, and badly wrong wherever an atom is near. The Wilks coverage closure check (Section 12.1) below documents the resulting over- and under-coverage directly across the truth grid.

Figure 12.4 traces this atom-versus-interior balance directly: walking the truth grid along the per-γ discovery-potential curves, the atom and interior mass weights cross over at the saturation point past which the calibrated MLE pins to the $\hat{\gamma}$ boundary and further γ-grid extension carries no information at that signal level.

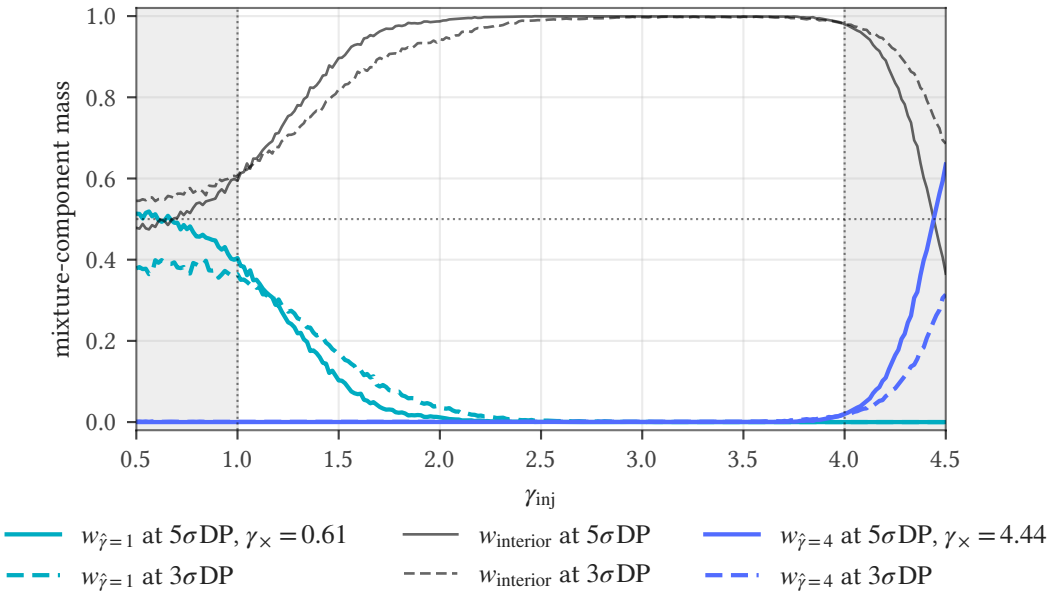

Figure 12.4: Atom-mass saturation diagnostic at $\sin\delta = 0$. Walking the $(\gamma_{\mathrm{inj}}, n_{s,\mathrm{inj}})$ truth grid along the per-γ 5σ and 3σ discovery-potential curves, the three boundary-mixture mass weights (w_{interior}, $w_{\gamma=1}$, $w_{\gamma=4}$) are plotted at each step. The crossover between the atom and the interior is the operational saturation point past which the calibrated MLE pins to the $\hat{\gamma}$ boundary and further γ-grid extension carries no information at that signal level.

Continuous-component KDE

The three continuous components p_{cont}, $p_{\hat{n}_s|\hat{\gamma}=1}$, and $p_{\hat{n}_s|\hat{\gamma}=4}$ are estimated nonparametrically by Gaussian-kernel KDE at each truth point. The interior uses a 2D KDE on the active-interior subset of pseudo-experiments. The atom-side conditionals use 1D KDEs on the $\hat{n}_s$ marginal of trials with $\hat{\gamma}$ clipped to the respective bound. Both use bandwidths from the Normal reference rule ($h \propto N_{\mathrm{trials}}^{-1/(d+4)}$ in d dimensions;[248]) and are evaluated on regular grids with linear-binning Fast Fourier Transform (FFT) convolution.[249]

A nonparametric KDE, rather than a parametric multivariate-Gaussian fit on the empirical $(\mu_\theta, \Sigma_\theta)$, is the right choice for the interior: the per-cell empirical density visible in Figure 12.5 carries shape features (asymmetry along the n_s–γ ridge, non-Gaussian skew near the $\hat{n}_s = 0$ boundary-atom shoulder, and a truth-dependent $(\hat{n}_s, \hat{\gamma})$ correlation orientation that varies with γ_{inj}) that the FFT KDE captures and a 2-parameter Gaussian ellipse cannot. We compared the two backends directly during methodology development and adopted the FFT KDE. Coverage of the truth is preserved either way by Neyman duality (Section 12.6): the per-cell empirical α-quantile is exact regardless of the inner density model, but the *interpretation* of the FC region (its centroid, marginal-error widths, and joint shape) is meaningful only

[248] Wasserman 2006, *All of Nonparametric Statistics*, Sec. 6.3.

[249] Wand and Jones 1995, *Kernel Smoothing*, Sec. D.2 (1D, pp. 183–187) & Sec. D.5 (p. 191, Fig. D.3).

under a density model that fits the per-cell cloud, which the parametric Gaussian does not.

A subtlety attaches to the KDEs at the support boundaries. The interior support is $(0, \infty) \times (1, 4)$ for p_{cont}, and $(0, \infty)$ for the atom-side conditionals: hard boundaries at $\hat{n}_s = 0$ for both, plus $\hat{\gamma} \in \{1, 4\}$ for the interior. The FFT-binned Gaussian KDE leaks mass past these hard boundaries. Each per-cell continuous component is under-normalized over its support by an amount that varies cell-to-cell with the proximity of its pseudo-experiments to a boundary. The textbook fix is a boundary-aware kernel estimator: Jones[250] for boundary-corrected symmetric kernels achieving $\mathcal{O}(h^2)$ bias on bounded-support densities (an order-of-magnitude improvement on the $\mathcal{O}(h)$ boundary bias of the uncorrected KDE), Chen[251] for beta kernels designed natively for $[0, 1]$ support (rescalable to $[1, 4]$ for $\hat{\gamma}$), and Chen[252] for gamma kernels designed natively for $[0, \infty)$ support (the $\hat{n}_s$ axis). These approaches share the property that the kernel shape varies with the evaluation point: Chen explicitly notes that "an eminent feature of the beta kernels is that the kernel shape changes according to the value of x",[253] which makes them incompatible with the FFT-binned convolution structure that powers our density estimation. Using a kernel that is not translation-invariant returns evaluation to $\mathcal{O}(N \cdot G)$ per cell and puts density estimation out of computational reach for our dense truth grid. Reflective KDE preserves the FFT structure but is physically wrong here: the reflected mass would represent trials whose unconstrained MLE clips to the $\hat{\gamma}$-boundary atoms, which the mixture already parameterizes separately through $w_{\gamma=1}$ and $w_{\gamma=4}$, so reflecting would double-count.

[250] Jones 1993, "Simple boundary correction for kernel density estimation".

[251] Chen 1999, "Beta kernel estimators for density functions".

[252] Chen 2000, "Probability density function estimation using gamma kernels".

[253] Chen, "Beta kernel estimators for density functions", p. 133.

We instead apply per-cell renormalization. The structural justification is that pseudo-experiments are first classified into one of the four mixture categories, and each KDE is fit only on the subset belonging to its own component, which lies strictly inside the component's support by definition. Each KDE is therefore empirically a model of the conditional density on its support. The kernel-smoothed mass placed past the support boundary is a smoothing artifact rather than a distributional claim, since the conditional has zero density past the boundary by construction and there are no pseudo-experiments there. Renormalizing by the in-support integral is the standard truncated-density treatment: at each cell, the integral Z_θ is computed numerically at fit time, and $\log Z_\theta$ is subtracted from the per-cell log-density grid as a constant offset. Each per-component continuous density then integrates to one over its own support, while the empirical category fractions $w_{n_s=0}, w_{\gamma=1}, w_{\gamma=4}, w_{\text{int}}$, which are measured quantities, are untouched. A third option, projecting the spillover mass into the atom weights, was rejected on related grounds: the weights are empirical counts and should not be perturbed by the kernel-smoothing geometry of a separate component.

Renormalization fixes the *integral* bias the boundary-aware kernels would also fix, but not the *shape* bias: Jones[254] shows that the residual bias of renormalization remains $\mathcal{O}(h)$ near the boundary, an order of magnitude worse than the $\mathcal{O}(h^2)$ interior bias. Coverage is unaffected by the choice between the textbook boundary-kernel fix and our per-cell renormalization: the Neyman-duality (Section 12.6) argument absorbs any per-cell multiplicative shift on the density symmetrically

[254] Jones 1993.

into the trial-side critical-value calibration and the unblinded R evaluation. Power should also be unaffected: a uniform multiplicative shift on the density inside a cell does not reshuffle the cell-wise ranking for any single observation, though we have not directly measured this. What does change is the calibrated MLE for boundary-near observations and the FC region shape near the boundaries. Both are reported deliverables and are now produced from a properly renormalized per-cell density model.

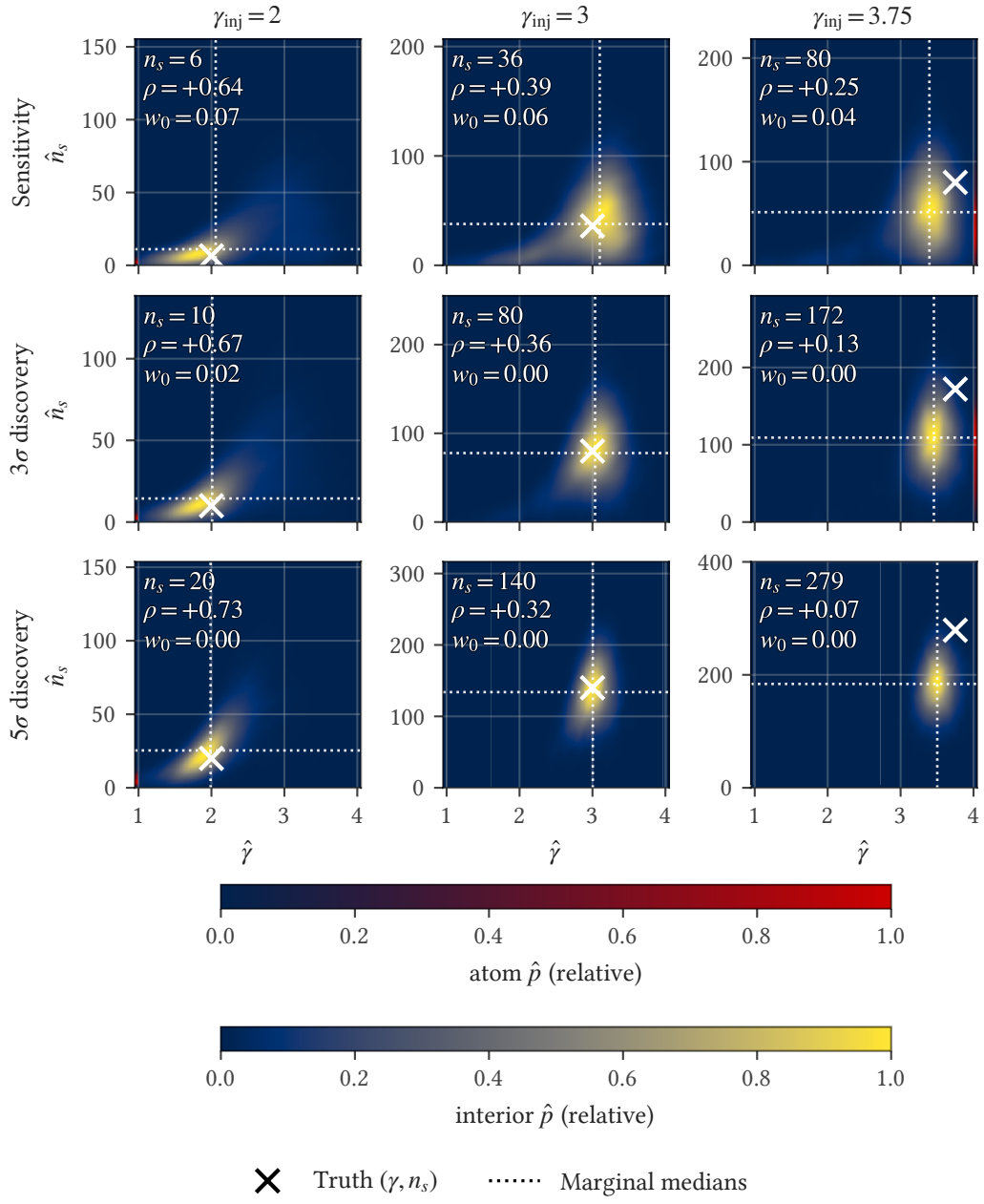

Figure 12.5: Joint density model of the MLE $(\hat{n}_s, \hat{\gamma})$ across the simulation grid at $\sin\delta = 0$ for LT + DNNC. Columns are the injected spectral index $\gamma_{inj} \in \{2.0, 3.0, 3.5\}$; rows are the injected signal strength (sensitivity, 3σ discovery potential, and 5σ discovery potential). Each panel shows the FFT-KDE interior density of $(\hat{n}_s, \hat{\gamma})$ with normal reference rule bandwidth as a heatmap, plus the two $\hat{\gamma}$-boundary atom strips ($\hat{\gamma} = 1$ and $\hat{\gamma} = 4$) along the panel edges. The cross marks the truth $(\gamma_{inj}, n_{s,inj})$; the dotted lines mark the marginal medians of $\hat{n}_s$ and $\hat{\gamma}$, with their intersection the joint median. Each panel is annotated with its injected n_s, the Pearson correlation $\rho(\hat{n}_s, \hat{\gamma})$, and the $\hat{n}_s = 0$ atom mass $w_0 = P(\hat{n}_s = 0; \theta)$. The interior and atom colorbars are normalized per panel, so the colors compare density shape across the grid rather than absolute mass.

Calibrated MLE computation

At observation time, the per-truth log-densities $\log \hat{p}(x_{\text{obs}}; \theta)$ are evaluated from the per-cell mixture at every truth grid point $\theta \in \Theta$, and the calibrated MLE is the strict argmax over the discrete grid Θ. We do not interpolate between grid points: the empirical density is sampled at the grid only, so smooth interpolation through MC-noisy grid samples would add off-grid maxima driven by the noise pattern of a specific simulation realization rather than by the underlying physics. With 10^4 pseudo-experiments per truth point the per-knot MC noise is small, but the methodologically clean position is to admit no off-grid resolution that the data does not directly support. Higher resolution comes from simulating at a finer grid, not from interpolation.

For an $\hat{n}_s = 0$ observation, the empirical density collapses to its atom component: $\hat{p}(x; \theta) = w_{n_s=0}(\theta)$, a function of θ alone. The argmax-in-θ of $\hat{p}(x; \theta)$, which is what defines the calibrated MLE $\tilde{\theta}(x)$ and which then enters the R-denominator $\hat{p}(x; \tilde{\theta}(x))$, is ill-conditioned at finite MC for this observation type. In expectation, $w_{n_s=0}$ is constant along the entire $\{n_s = 0\}$ slice of the truth grid: the pseudo-experiments at $n_s = 0$ truth are pure RA-randomized data backgrounds with no signal injected, flat across γ-cells by construction, and any signal injection at $n_s > 0$ can only thin the $\hat{n}_s = 0$ atom fraction. The asymptotic argmax is therefore the boundary $\{n_s = 0\}$ slice itself, with γ a non-existent parameter on it: no γ exists at $n_s = 0$ in the model likelihood (the same Davies degeneracy noted in the boundary-mixture bullet for $\hat{n}_s = 0$), and correspondingly no $\hat{\gamma}$ exists on the observation. An apparent finite-MC argmax at $n_{s,\text{truth}} > 0$ (Poisson noise on $w_{n_s=0}$ of order $\sqrt{w(1-w)/N_{\text{trials}}} \sim 0.5\%$ at our trial count, randomly lifting some non-zero cell above the no-signal baseline) is unphysical. We therefore define the calibrated MLE piecewise:

$$\tilde{\theta}(x) = \begin{cases} (n_s = 0,\ \gamma \text{ undefined}) & \text{if } \hat{n}_s = 0, \\ \arg\max_\theta \hat{p}(x; \theta) & \text{otherwise,} \end{cases} \tag{12.8}$$

the $\hat{n}_s = 0$ branch being the *asymptotic boundary closure* of the empirical-density argmax: the value the unrestricted $\arg\max_\theta \hat{p}(x; \theta)$ would take in the $N_{\text{trials}} \to \infty$ limit, with γ undefined. For an $\hat{n}_s = 0$ observation, $R(x; \theta) = w_{n_s=0}(\theta)/w_{n_s=0}(0, \cdot)$, and the observation contributes to acceptance regions through this atom-mass ratio alone. FC coverage holds.

FC at the $n_s = 0$ truth cell uses the same pseudo-experiments as the empirical null calibration (Section 9.9) of the significance test, and is therefore the case of a single truth cell (Section 12.9) of the same Neyman-duality machinery, with the empirical-LR ranking on $w_{n_s=0}(\theta)$ alone playing the role that the standard model-LR test statistic plays for the per-pixel hypothesis test elsewhere.

Figure 12.6 visualizes the calibrated MLE's median recovery, with shaded bands spanning the 16th to 84th percentile of the recovery distribution at each truth point. The trials used for this diagnostic are the same as those used for the bias diagnostic in Section 9.10 (Figure 9.20, signal recovery for LT + DNNC). Importantly, they do not include Poisson variance (each pseudo-experiment injects exactly n_{inj} signal

events), giving a pure diagnostic for recovery of the signal events actually present in each trial.

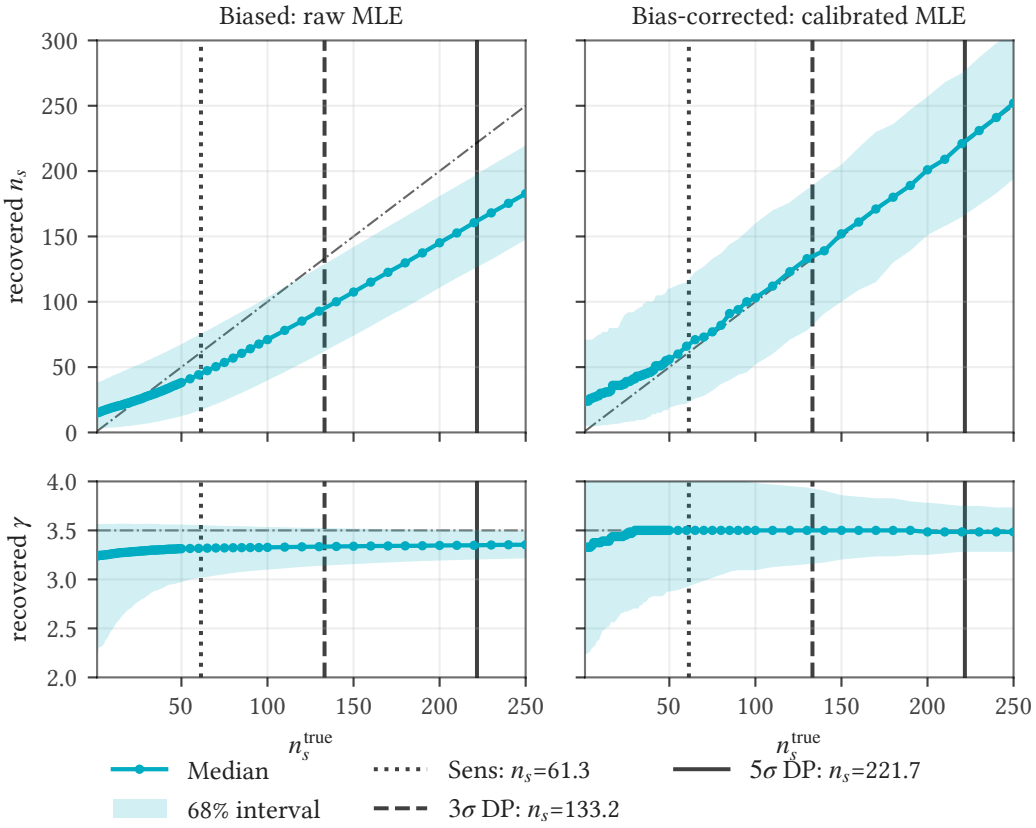

Figure 12.6: Signal recovery at the soft-source cell $\gamma_{\text{inj}} = 3.5$, $\sin\delta = 0$ (LT + DNNC), where the raw-MLE bias is largest. Left column: the raw maximum-likelihood estimate $\hat{\theta}$. Right column: the bias-corrected calibrated MLE $\tilde{\theta}$. Top row: recovered n_s versus injected n_{inj}; bottom row: recovered γ. Solid lines are the median recovery and shaded bands the 16th–84th percentile of the recovery distribution at each truth (the spread of the estimator, not Poisson scatter: each pseudo-experiment injects exactly n_{inj}). Dash-dot lines mark perfect recovery ($\tilde{n}_s = n_{\text{inj}}$ on top, $\gamma = 3.5$ on the bottom); vertical lines mark the sensitivity, 3σ, and 5σ discovery-potential injection levels. The biased median sits below perfect recovery in both parameters, while the calibrated median lies on it.

[255] Dalmasso et al. 2024, Sec. 1, Fig. 1.

[256] Masserano et al. 2023, "Simulator-Based Inference with Waldo: Confidence Regions by Leveraging Prediction Algorithms and Posterior Estimators for Inverse Problems", Sec. 3, Fig. 1, Alg. 1.

Pseudo-experiment sample split

The per-cell MC pseudo-experiment sample is split into two disjoint subsets following the LF2I three-sample prescription of Dalmasso et al.[255] and Masserano et al. (WALDO)[256]. Their T is a sample-naming convention, not to be confused with the test statistic T used elsewhere in the hypothesis-testing context (Section 10.2). The T branch (10^4 trials per cell) fits the per-cell KDEs that define $\hat{p}(x;\theta)$, and therefore

the empirical-LR ranking $R(x;\theta)$ and the calibrated MLE $\tilde{\theta}(x)$. The T' branch (10^4 trials per cell, disjoint from T) supplies the per-cell empirical α-quantile $R_c(\theta)$ that defines the FC acceptance threshold (Section 12.2). The T'' diagnostics branch uses the existing signal recovery sample (Section 9.10), an independent draw at each truth point used for empirical-coverage validation against the constructed regions.

The disjointness of T and T' eliminates a finite-N in-sample bias that would otherwise arise. At each in-sample point $x_i \in T$, the KDE carries a self-contribution from the kernel centered at x_i itself, identified explicitly as the diagonal term $n^{-1}L_g^{(r)}(0)$ in the density-functional bias decomposition of Wand and Jones[257] (pp. 67–70: "the first term being independent of the data"[258]). Tail x_i have $\tilde{\theta}(x_i) \neq \theta$, so the denominator $\hat{p}(x_i;\tilde{\theta}(x_i))$ is built from a different per-cell sample that does *not* contain x_i and carries no such self-contribution. The in-sample $R(x_i;\theta)$ values are biased upward in the tails, lifting the empirical α-quantile threshold by the same order and producing a small under-coverage at finite N. T/T' disjointness avoids this exactly: each $R(x_i;\theta)$ for $x_i \in T'$ is evaluated against a KDE built without x_i, identical to how a fresh observation x_{obs} is treated. Per-cell coverage is then exact up to the unbiased empirical-quantile MC noise of $\sqrt{\alpha(1-\alpha)/|T'|} \sim 0.47\%$ at $|T'| = 10^4$ (see Section 12.6 for the rank-fraction vs R-value distinction), which shrinks as we extend the per-cell pseudo-experiment count, the same knob that controls boundary raggedness, see Section 14.3.

[257] Wand and Jones 1995, Sec. 3.5.

[258] Wand and Jones, *Kernel Smoothing*, p. 68.

Two alternatives to the T/T' split were considered and rejected. Leave-one-out KDE (rebuilding $\hat{p}$ on T omitting each $x_i \in T'$ in turn) is infeasible at FFT KDE cost: each LOO evaluation requires its own full FFT pass, so $|T'|$ rebuilds per cell across the full grid is prohibitive. The per-sample binning-aware LOO correction that would replace it (subtracting the self-contribution analytically, accounting for the linear-binning weights of x_i on its 2^d corner bins) carries a residual approximation specific to the FFT linear-binning step. K-fold cross-validation with $K > 2$ does not give an exact construction either: each fold-out KDE differs slightly from the full KDE, breaking the consistency between the calibration R-function and the observation R-function that the Neyman-duality coverage statement requires. The T/T' split matches the LF2I prescription verbatim.

12.4 Confidence intervals for a single parameter of interest

The 1D profile-likelihood construction in classical statistics holds the parameter of interest (POI) θ fixed and maximizes the joint likelihood over the nuisance ϕ:

$$L_{\text{profile}}(x;\theta) = \max_{\phi} L(x;\theta,\phi). \tag{12.9}$$

With an analytical model likelihood, this maximization is just numerical optimization over the nuisance subspace at fixed POI. The resulting profile-LR test statistic has the standard $-2\ln\Lambda \to \chi^2_1$ Wilks limit asymptotically.[259]

[259] Casella and Berger 2002, Theorem 10.3.3, p. 490.

The random profile-LR test statistic at candidate POI θ is $T = R_{\text{prof}}(X;\theta)$, with observed value $t = R_{\text{prof}}(x_{\text{obs}};\theta)$. The p-value at observed t under POI θ

and nuisance ϕ is $p(t;\theta,\phi) = P_{\theta,\phi}(T \geq t)$. In the literature the data argument is conventionally suppressed in this notation, written as $p(\theta,\phi)$ or $p(\theta)$[260]. NOvA further mark the conditional best-fit nuisance with a double hat $\hat{\hat{\phi}}$ to distinguish it from the globally optimal $\hat{\phi}$. We use the explicit form $p(t;\theta,\phi)$ and a single hat in our own statements; the quotations below keep each source's notation.

[260] E.g., NOvA Collaboration 2025, Berger and Boos 1994, "P Values Maximized Over a Confidence Set for the Nuisance Parameter".

The profile-LR is the *test statistic*. How its threshold is computed is a separate question. In standard statistical usage "profile" refers to the test statistic alone (the max over nuisance). NOvA[261] appropriate the term to a specific threshold method, which they call the "Profiled Feldman-Cousins approach," setting the threshold from the plug-in p-value $p_{FC}(t;\theta) = p(t;\theta,\hat{\phi})$:

[261] NOvA Collaboration 2025, App. B, Eq. B.2.

> "effectively assuming that the nuisance parameters which give the largest likelihood value (and thus the largest p-value under Wilks' theorem) will also have the largest p-value with the pseudoexperiment-calculated critical values."
>
> — NOvA Collaboration, 2025, p. 23

This is the *plug-in* approach (a standard term in statistics for substituting an estimate where the formula calls for the truth value): calibrate the empirical critical value at the single truth point (θ,ϕ), where ϕ is the conditional best-fit nuisance.

The plug-in is not a valid frequentist p-value in general. Berger and Boos[262] state this directly:

[262] Berger and Boos 1994, Sec. 2.

> "Storer and Kim (1990) and others have used this idea to propose as a p value $p(\hat{\theta})$, where $\hat{\theta}$ is an estimate of θ (usually the maximum likelihood estimate). But p values defined in this way may not be valid; see the computations of Storer and Kim (1990)."
>
> — Berger and Boos, 1994, p. 1013

Recovering strict frequentist validity classically requires the *Conservative* threshold, $p_{\text{cons}}(t;\theta) = \sup_{\phi} p(t;\theta,\phi)$, which takes the supremum over the entire nuisance space. This is strictly valid but admits values of the nuisance the data clearly excludes—producing substantial avoidable over-coverage. The Berger-Boos threshold sits between the two extremes: rather than the unrestricted supremum, take the supremum over a $(1-\beta_C)$ confidence set C_{β_C} for the nuisance restricted to the slice, and add β_C to the resulting p-value,

$$p_{\beta_C}(t;\theta) = \sup_{\phi \in C_{\beta_C}} p(t;\theta,\phi) + \beta_C. \tag{12.10}$$

The construction uses the data to constrain the nuisance (unlike the unrestricted Conservative supremum) but puts an explicit β_C bound on the confidence of that constraint (unlike the plug-in, which assumes the conditional MLE is exactly right). The Berger-Boos Lemma (Sec. 2, p. 1013) establishes that p_{β_C} is a valid frequentist p-value at any level α, regardless of the data distribution and without any Wilks regularity. We adopt this construction directly with $\beta_C = 10^{-3}$.

[263] NOvA Collaboration 2025, Sec. 2.3.

NOvA acknowledge Berger-Boos as the strict-frequentist alternative[263]:

> "Berger-Boos. This method is philosophically similar to the conservative method, but introduces a limiting principle for which values of nuisance parameter to consider. At each point in parameter space, θ_i, determine the range of nuisance parameters consistent with the data at significance level β, and then calculate p-values empirically (i.e. using pseudoexperiments) for all values of the nuisance parameters within that range. The overall p-value for point θ_i is based on the largest p-value within that set, $p = \max_\phi p(\theta_i, \phi) + \beta$. [...] Since the nuisance parameters in the likelihood and the pseudoexperiments are moved together, this method does not have the same problem of over-coverage as the Conservative method, but it is still computationally infeasible for making confidence intervals or for a large number of nuisance parameters. Appendix B shows the use of this method to cross-check the significance in a single hypothesis test."
>
> — NOvA Collaboration, 2025, p. 5

We strongly disagree with NOvA's "philosophically similar to the conservative method" framing. The three threshold constructions sit on a continuum parameterized by the confidence one places in the data's localization of the nuisance. Berger-Boos places $(1 - \beta_C)$ confidence in C_{β_C} and pays a $+\beta_C$ correction in exchange for strict frequentist validity by the Berger-Boos Lemma. The unrestricted Conservative supremum is the $\beta_C \to 0$ limit (no localization trusted, no correction needed). The plug-in is the opposite limit in spirit: full confidence placed in the conditional MLE alone, without the additive correction that the Berger-Boos Lemma requires for strict validity. The Lemma's explicit constraint $\beta_C < \alpha$ for a non-trivial test rules out approaching the plug-in limit at any reportable confidence level. The plug-in's standing "validity" is asymptotic-Wilks, not strict-frequentist.

Under Wilks regularity, the distinction between the three would not matter: the asymptotic χ^2_1 distribution of the profile-LR is independent of the nuisance, so any supremum over nuisance reduces to the same constant and plug-in, Berger-Boos, and Conservative would coincide up to the small β_C floor on the Berger-Boos p-value. The methods diverge only when Wilks regularity fails. Our empirical Wilks closure check (Section 12.1) below demonstrates that it does fail across substantial parts of the parameter grid in our setup.

Adopting the plug-in on the implicit assumption that the threshold is nuisance-independent would amount to invoking the very Wilks-asymptotic argument we have rejected at the test-statistic calibration level, methodologically one step short of using a Wilks-based CI directly, which we have already rejected on the same grounds. Calling Berger-Boos the "more conservative" cousin of the plug-in obscures this. In our reading, Berger-Boos is the rigorous version that the plug-in approximates only when Wilks regularity holds—and our setup is demonstrably outside that regime.

NOvA's own results show that the power cost of moving from plug-in to Berger-Boos is small. Their App. B cross-check on the mass-ordering significance, run with $\beta = 5 \times 10^{-3}$ relative to a $p \approx 0.30$ operating point, found that "[Berger-Boos] did

not uncover a larger p-value than the one reported from the profile construction, and so is consistent with that result".[264] The implicit objection that Berger-Boos is too conservative to use as primary is therefore quantitatively unsupported in the regimes either they or we operate in.

[264] NOvA Collaboration, "Monte Carlo method for constructing confidence intervals with unconstrained and constrained nuisance parameters in the NOvA experiment", p. 20.

The defensible reason to prefer plug-in over Berger-Boos is computational. Brute-force sampling of C_{β_C} scales poorly in the nuisance dimension, and NOvA's setup with on the order of 50 nuisance parameters made Berger-Boos infeasible to apply at every grid point. Even there, however, the brute-force approach is not the only option: an iterative search for the Berger-Boos threshold after unblinding, on a pre-specified protocol, locates the relevant supremum without precomputing the entire C_{β_C} slice. We acknowledge that this comes with substantial implementation complexity, and we understand NOvA's choice on those grounds. In our setup, the question does not arise: single nuisance dimension at a time, with the per-cell calibration data already extractable from the same pseudo-experiment ensemble that builds the 2D FC region. Berger-Boos is the principled choice rather than a too-expensive ideal.

We follow Berger and Boos's literal recommendation, $\beta_C = 10^{-3}$ (Sec. 2: "rather small, such as .001 or .0001"[265]). For our 1σ intervals ($\alpha \approx 0.3173$) the relative threshold shift $\beta_C/\alpha \approx 0.3\%$ is well below the per-cell MC quantile noise ($\approx 0.47\%$ at 10^4 pseudo-experiments per cell). NOvA's mass-ordering Berger-Boos cross-check used $\beta = 5 \times 10^{-3}$ at a $p \approx 0.30$ operating point.[266] Our tighter choice is appropriate because we adopt Berger-Boos as the primary construction rather than as a cross-check.

[265] Berger and Boos, "P Values Maximized Over a Confidence Set for the Nuisance Parameter", p. 1013.

[266] NOvA Collaboration 2025, App. B.

Boundary behavior

The 1D profile-FC interval on $\Phi(E_{\text{pivot}})$ constructed in Section 12.2 above transitions automatically between one-sided (effective upper limit, $[0, \Phi^{1\sigma}_{\text{upper}}]$) when the data does not reject zero at the chosen CL, and two-sided (lower bound > 0) when it does. This is the original Feldman–Cousins unified-construction property.[267]

[267] Feldman and Cousins 1998, Sec. VII.

We rely on this property to avoid the practice of switching CL between regimes (e.g., 1σ two-sided / 90% one-sided), which is itself flip-flopping on the confidence level and contradicts FC Sec. VII's automatic-transition framing.

For sources whose unblinded fit collapses to the boundary $\hat{n}_s = 0$, the FC interval automatically takes the form $[0, \Phi^{1\sigma}_{\text{upper}}]$. The bias-corrected point estimate $\tilde{\theta}$ is undefined on the boundary atom (no spectral information to anchor it; see Section 12.3), and E_{pivot} is undefined as well: the 2D FC region for a boundary-fit observation includes $\Phi = 0$ at every γ, which makes the projection-width minimization that defines E_{pivot} unbounded below. For these sources we report the FC interval on Φ at the standard fixed reference energy of 1 TeV instead.

The shape of the FC region in (n_s, γ), and therefore whether the 1D interval on $\Phi(E_{\text{pivot}})$ is one-sided or two-sided, is determined by whether the data rejects the null at the chosen CL ($\sim 1\sigma$ for the 1σ interval, $\sim 2\sigma$ for the 2σ), not by the 3σ evidence threshold of the significance test. Sources at marginal pre-trial significance can therefore have two-sided 1σ FC intervals with non-zero lower bounds on Φ,

even though they fall below the evidence threshold. The interval is reported as-is in either case. Rejection of H_0 is read off the post-trial p-value column where one is reported (see the step-down procedure, Section 11.2, for which sources receive one).

12.5 Noise-free examples

The figures in this section show noise-free FC regions for representative cells of the simulation grid. "Noise-free" refers to the observation we invert around: at each truth point $\theta = (n_{\mathrm{inj}}, \gamma_{\mathrm{inj}})$ we take the mode of the per-cell sampling density, $x_{\mathrm{nf}}(\theta) = \arg\max_x \hat{p}(x;\theta)$ (the most likely $(\hat{n}_s, \hat{\gamma})$ observation at that truth), and invert the FC construction around it. This is the reverse of the calibrated MLE lookup $\tilde{\theta}(x) = \arg\max_\theta \hat{p}(x;\theta)$. It is distinct from setting $\tilde{\theta} = \theta_{\mathrm{truth}}$, because the calibration is itself imperfect. The region construction is otherwise unaffected: the pseudo-experiments behind each acceptance region carry the full Poisson scatter, so each region covers the truth at its nominal probability mass exactly as for any random realization. Each figure therefore shows the central diagnostic region for its truth point.

Every example is shown at $\sin\delta = 0$; the construction is the same at every declination, and the declination dependence carries no trend that changes the behavior relevant to the confidence-interval construction discussed here. Each panel shows the 1σ and 2σ Feldman–Cousins confidence regions built with the Berger-Boos threshold, with the 1σ profile-FC intervals drawn as error bars on the MLE. Because the examples are noise-free, the calibrated MLE sits essentially at the truth in each one and the coverage is correct by construction. The grey shaded bands ($\gamma \in [0.5, 1]$ and $\gamma \in [4, 4.5]$) mark the extended truth grid beyond the model-fit bounds $\hat{\gamma} \in [1, 4]$: pseudo-experiments are injected there and the construction keeps valid per-cell coverage even though the fit clips the recovered $\hat{\gamma}$ to the bound and cannot return an unconstrained estimate (Section 12.2). The gallery walks five representative cases. Figure 12.7 is the $\hat{n}_s = 0$ (TS = 0) cell, where the interval collapses to a one-sided upper limit on the flux and γ is undefined on the boundary atom. Figure 12.8 is a hard ($\gamma = 2.0$) source at the sensitivity level, a weak-signal case whose region is broad and skewed and whose MLE is pulled toward lower γ (a harder spectrum), with the 2σ region extending below the $\hat{\gamma} = 1$ fit bound into the shaded sub-bound range. Figure 12.9 ($\gamma = 3.5$ at the 5σ discovery potential) is a strong-signal cell where the recovered MLE sits on the truth. The final pair reports the same $\gamma = 3.0$, 3σ-discovery-potential cell at two energies: Figure 12.10 at the per-cell pivot energy, where the flux and γ nearly decorrelate and the region is compact, and Figure 12.11 at a fixed 1 TeV, where the residual flux–γ correlation tilts the region into an elongated band. That contrast is why fluxes are reported at the pivot energy. A single Poisson realization shows how the boundary atoms manifest in a real fit: Figure 12.12 clips the $\hat{\gamma} = 4$ bound at 1σ.

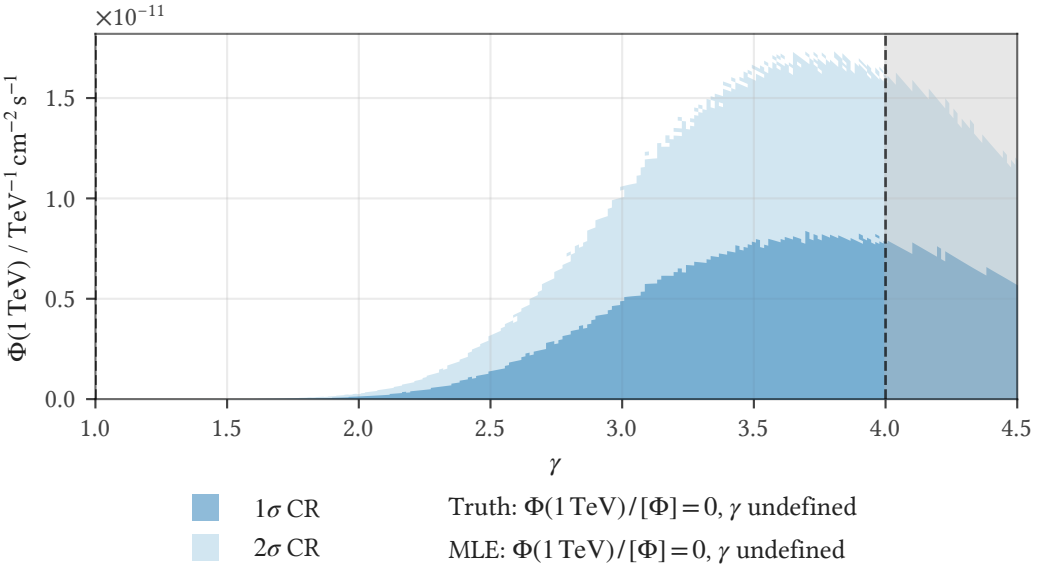

Figure 12.7: Noise-free FC region for an $\hat{n}_s = 0$ (TS $= 0$) observation at $\sin\delta = 0$. The intervals automatically take the form of one-sided upper limits on $\Phi(1\,\mathrm{TeV})$; γ is undefined on the boundary atom.

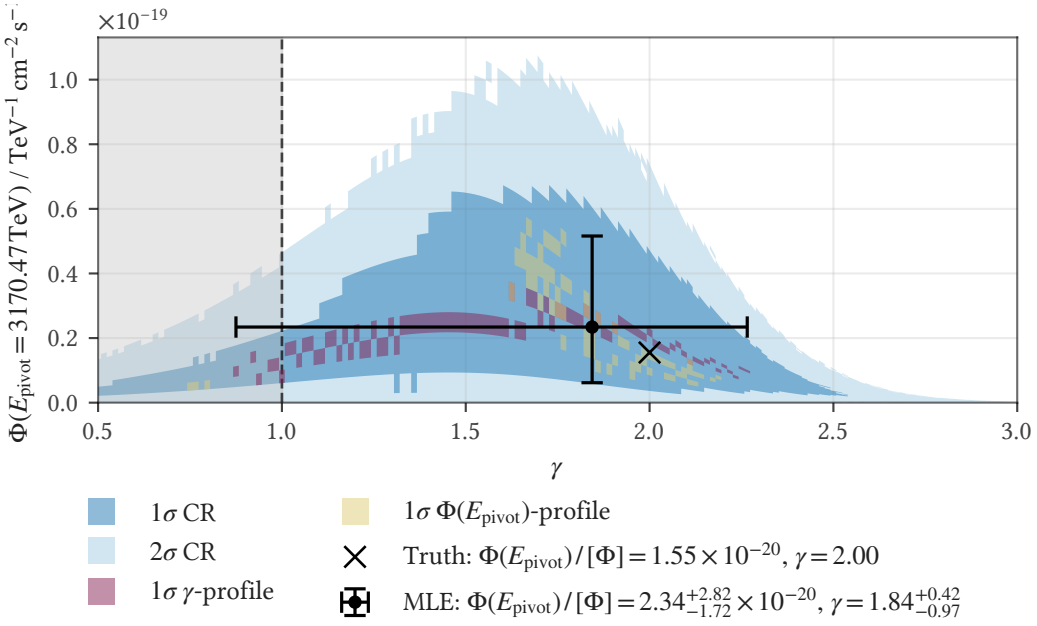

Figure 12.8: Noise-free FC region at truth $\gamma = 2.0$, sensitivity-level injection, $\sin\delta = 0$, with flux at the per-cell pivot energy. The weak signal gives a broad, skewed region with the MLE pulled toward lower γ (a harder spectrum); the 2σ region extends below the $\hat{\gamma} = 1$ fit bound.

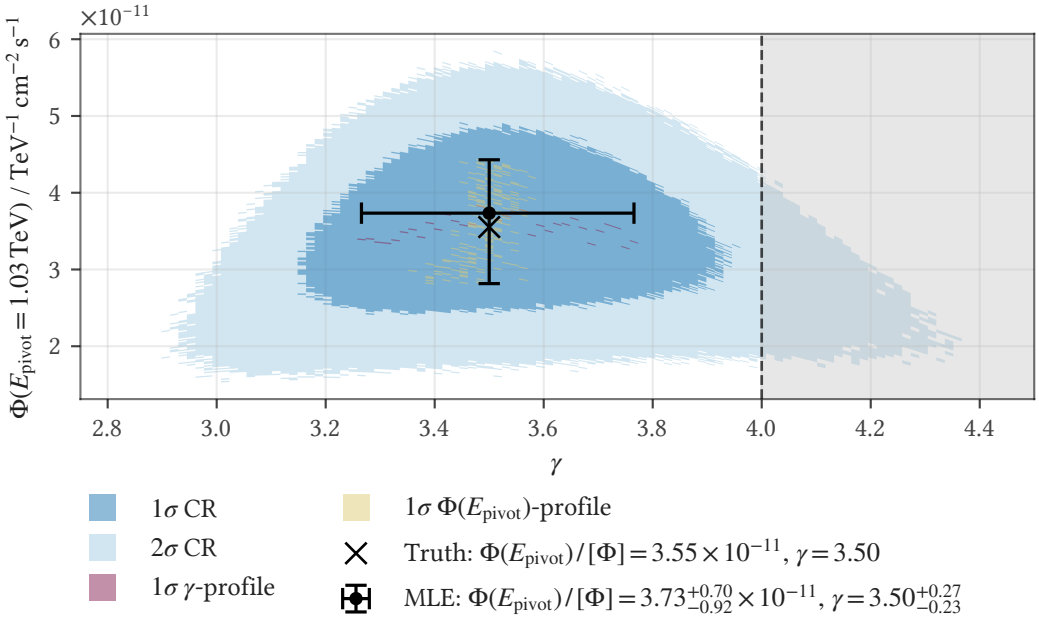

Figure 12.9: Noise-free FC region at truth $\gamma = 3.5$, 5σ-discovery-potential injection, $\sin\delta = 0$, with flux at the per-cell pivot energy. The recovered MLE lies on the truth.

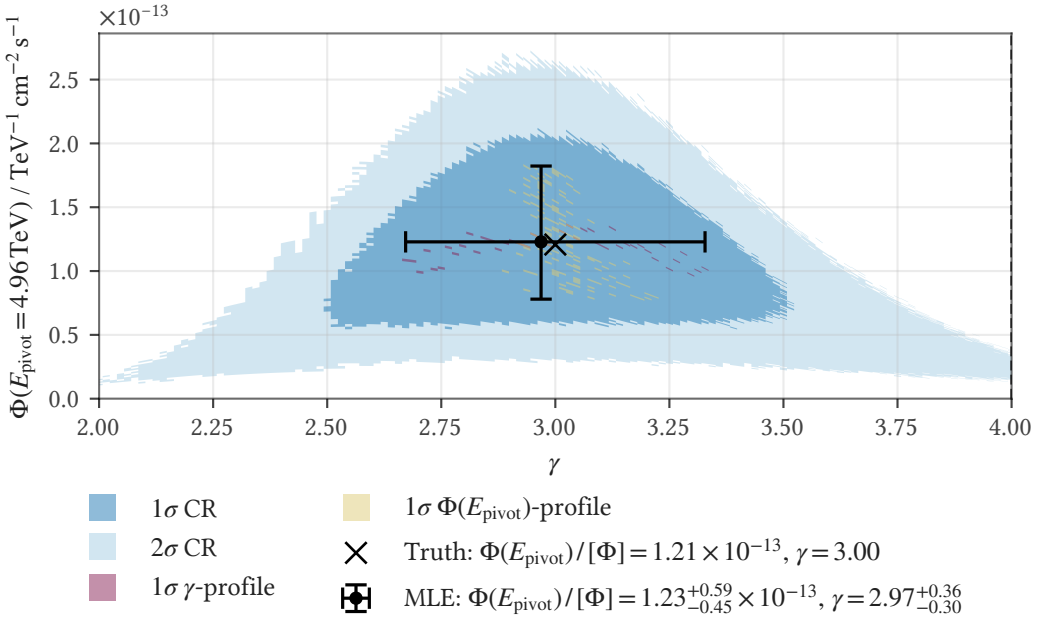

Figure 12.10: Noise-free FC region at truth $\gamma = 3.0$, 3σ-discovery-potential injection, $\sin\delta = 0$, with flux at the per-cell pivot energy. Compare Figure 12.11, the same cell reported at 1 TeV.

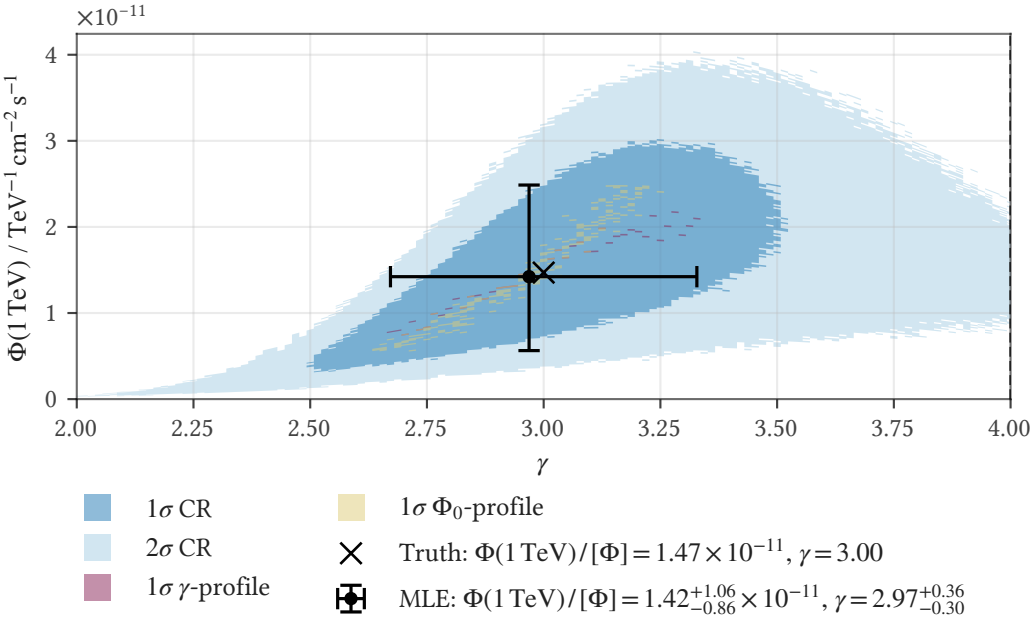

Figure 12.11: The same $\gamma = 3.0$, 3σ-discovery-potential cell as Figure 12.10, but with flux at a fixed 1 TeV: the residual flux–γ correlation tilts the region into an elongated band.

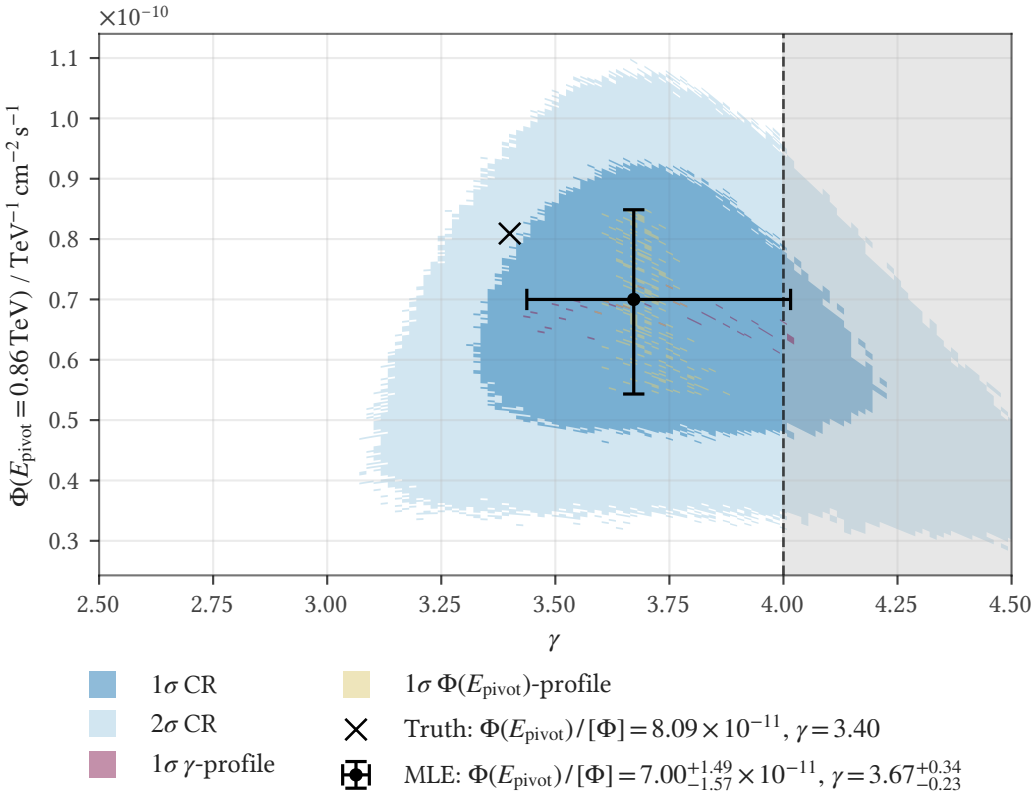

Figure 12.12: Feldman–Cousins 1σ and 2σ confidence regions for a single Poisson realization of NGC 1068 at $\sin\delta = 0$, illustrating a $\hat{\gamma}$-boundary clip. The calibrated MLE recovers $\gamma = 3.67$, and the 1σ region clips the $\hat{\gamma} = 4$ fit boundary; the injected truth ($\gamma = 3.4$) lies inside the 2σ region.

12.6 Validity of the construction

Claim 12.1. Let $x = (T, \hat{n}_s, \hat{\gamma}, \ldots)$ be any vector of summary statistics extracted from a single experiment, and let $\theta \in \Theta$ index a family of signal hypotheses, here $\theta = (n_s, \gamma)$. For each candidate $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level-α test of $H_\theta : \theta = \theta_0$, defined by simulating x under θ_0 and including the highest-ranked $1 - \alpha$ fraction of pseudo-experiments under any chosen ranking function $R(x; \theta_0)$. Define the confidence set

$$C(x) = \{\theta_0 \in \Theta : x \in A(\theta_0)\}. \tag{12.11}$$

Then $P_\theta(\theta \in C(X)) \geq 1 - \alpha$ for every $\theta \in \Theta_{\text{grid}}$, *regardless of the choice of ranking function* R, where Θ_{grid} is the simulation-grid extent on which acceptance regions $A(\theta)$ have been constructed through pseudo-experiments.

Proof. This is Theorem 9.2.2 of Casella and Berger,[268] the duality between hypothesis-test acceptance regions and confidence sets. By construction of a

[268] Casella and Berger 2002, p. 421.

level-α test, $P_{\theta_0}(X \notin A(\theta_0)) \leq \alpha$, so $P_{\theta_0}(X \in A(\theta_0)) \geq 1-\alpha$. Since θ_0 was arbitrary within Θ_{grid}, for every $\theta \in \Theta_{\text{grid}}$,

$$P_\theta(\theta \in C(X)) = P_\theta(X \in A(\theta)) \geq 1-\alpha, \tag{12.12}$$

establishing $C(X)$ as a $1-\alpha$ confidence set on Θ_{grid}. □

The conditionality on Θ_{grid} is essential—for truth values outside the simulation-grid extent, no acceptance region has been constructed. The confidence statement is therefore undefined, and the coverage guarantee does not apply. The operational rule that ensures the simulation grid is large enough at the time of reporting is that the 1σ contour must lie fully in the grid interior; see also the treatment of what can invalidate the coverage (Section 12.6) below for the broader treatment of grid-coverage requirements.

The $\geq$ in the coverage statement is structural: Neyman-construction confidence regions generally *over-cover* rather than achieve $1-\alpha$ exactly. Two sources of this:

- **Discrete acceptance:** The acceptance region $A(\theta)$ is built by including all x with rank above some empirical-quantile threshold $R_c(\theta)$. With finite-trial empirical quantiles or atom-induced ties in the rank distribution, the threshold gets set so that $A(\theta)$ contains at least $1-\alpha$ of the probability mass, often slightly more.
- **Boundary atom at $\hat{n}_s = 0$:** Our boundary mixture (Section 12.3) places non-negligible discrete mass at the $\hat{n}_s = 0$ point. $R(0;\theta)$ takes a single value at each θ, so the atom is included in $A(\theta)$ in full or excluded in full. The inclusion/exclusion flip as θ varies produces a discrete step in coverage of size $P(\hat{n}_s = 0 \mid \theta)$. The conservative-rule convention (include the atom when $R(0;\theta) \geq R_c(\theta)$) lands the step at or above $1-\alpha$ rather than below (the $\hat{\gamma} \in \{1, 4\}$ boundary mass is a *line* in $\hat{n}_s$ along which R varies continuously, so the threshold sweeps it smoothly and it contributes like the continuous bulk).

Practically, in a finite-N sample of pseudo-experiments at any truth θ, the fraction f_{enclose} whose CI encloses the truth satisfies $f_{\text{enclose}} \geq 1-\alpha$ in expectation, with binomial fluctuation $\sigma_f = \sqrt{(1-\alpha)\alpha/N}$ around the lower bound and a small systematic offset above it from the discrete-acceptance and atom effects. The construction guarantees a lower bound on coverage, not equality.

The original definition of confidence sets in Neyman[269] requires exactly this property. The confidence-coefficient-α functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ are the data-dependent lower and upper endpoints of the confidence interval $[\underline{\theta}(E), \bar{\theta}(E)]$ on the parameter of interest θ_1, with E denoting the experimental result (Neyman's original notation, *not* energy). They are required to satisfy $P\{\underline{\theta}(E) \leq \theta_1^0 \leq \bar{\theta}(E)\} = \alpha$ for every value of θ_1 and of any nuisance parameters $\theta_2, \theta_3, \ldots$. The independence of this guarantee from the choice of ordering principle is made explicit by Feldman and Cousins: "we use the freedom inherent in Neyman's construction in a novel way to obtain a unified set of classical confidence intervals…The new element is

[269] Neyman 1937, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability", Eq. 20.

a particular choice of ordering, based on likelihood ratios, which we substitute for more common choices in Neyman's construction."[270](Sec. I) Their coverage statement is asserted in Sec. II *before* the LR ordering of Sec. IV is introduced ("By construction, Eq. (2.3) is satisfied for all μ"[271]), so it is independent of any specific ordering choice. The choice of ranking function determines the *shape* and *width* of the resulting confidence region (and hence the analysis power), but the coverage guarantee is independent of that choice.

[270] Feldman and Cousins, "Unified approach to the classical statistical analysis of small signals", p. 1.

[271] Feldman and Cousins, "Unified approach to the classical statistical analysis of small signals", p. 3.

Parallel to the empirical null section

This is the parameter-estimation analogue of the empirical null calibration result (Section 9.9) above. Both are construction-level guarantees:

- the PIT (Section 9.9) makes the post-trial p-value valid regardless of how the test statistic is constructed, so long as the pseudo-experiments correctly sample the data under H_0;
- the Neyman duality (Section 12.6) makes the confidence region valid regardless of how the maximum-likelihood estimator behaves, so long as the pseudo-experiments correctly sample the data under each candidate θ.

The parallel runs deeper than mere analogy. In both constructions, the background component of the pseudo-experiments is identical: RA-randomized real data, with no Monte Carlo backgrounds entering at any stage. CORSIKA mismatches, atmospheric muon mismodeling, atmospheric flux uncertainties, and any other concern about background simulation fidelity are equally irrelevant to the validity of both inferences. The only piece of the FC pseudo-experiments that depends on Monte Carlo simulation is the *signal injection* at each candidate θ. The validity statement of the FC construction is therefore conditional on the simulated signal model (see what affects only the physical interpretation, Section 12.6, below)—but not on background fidelity in any sense.

What cannot affect the coverage

The following properties of the maximum-likelihood fit do not affect the coverage of the FC confidence region or the bias-corrected point estimate:

- **Bias in $\hat{\theta}$:** Even when the sampling distribution of $\hat{\theta}$ under θ does not center on θ (whether measured by $E[\hat{\theta}]$, the median, or the mode), the construction holds: the acceptance region $A(\theta)$ is defined by the *empirical sampling distribution* of x under θ, and that distribution is correctly captured regardless of how the bias is measured. The bias-corrected point estimate is obtained by inverting the same empirical density. *The signal recovery agreement plots* (Section 9.10) *document the bias of $\hat{\theta}$; they are not validity diagnostics for parameter estimation.* They motivate the calibrated-MLE point estimator we

report instead, but a worse-recovering analysis would still produce frequency-correct confidence regions and a frequency-correct calibrated MLE, just with broader regions and noisier point estimates.

- **Reconstruction quality:** Energy resolution, angular resolution, and any other per-event reconstruction property can be arbitrarily poor without affecting coverage. The calibration pseudo-experiments and the unblinded data pass through the same reconstruction code path, so per-event smearing or bias absorbs identically into both halves of the construction. The interesting separate question is whether the *simulated* reconstruction performance applied to injected signal MC matches the *real* reconstruction performance applied to the unblinded data: signal-side data-MC agreement on reconstruction. That is a signal MC fidelity question and falls under the next section: it can shift the physical interpretation of the recovered θ, but not coverage of θ under the simulated signal model.

- **Background-side data-MC agreement and atmospheric systematics:** The background component of the pseudo-experiments contains no MC at any stage: backgrounds are modeled entirely from data through RA randomization (Section 9.9) of the observed event list. Atmospheric flux models, cosmic-ray hadronic interaction models, atmospheric muon simulation fidelity, and any data-MC question on background-event reconstruction performance are therefore irrelevant to coverage end to end: there is no background MC for them to disagree with. Only signal MC fidelity can enter the inference at all, and even there only the physical interpretation, not the coverage statement (see next section).

- **Variance and precision of $\hat{\theta}$:** Whether $\hat{\theta}$ is precise or noisy at a given truth, the construction characterizes its full sampling distribution and inverts. The width of the FC region scales with the spread of $\hat{\theta}$; its coverage does not.

- **Failure of Wilks' theorem:** The Davies problem[272] of unidentified nuisance parameters under the null and more general regularity violations all invalidate Wilks-based confidence intervals. They do not affect the FC construction, which makes no asymptotic assumption.

- **Choice of ranking function:** As proven above, coverage is independent of R. Any choice (LR, T alone, $\hat{n}_s$ alone, an arbitrary scalar of x) gives valid coverage. We use the empirical-LR ranking $R = \hat{p}(x;\theta)/\hat{p}(x;\tilde{\theta}(x))$ (see Section 12.2) because in the limit of infinite simulation statistics it converges to the likelihood ratio built from the *true* sampling density of the MLE summary, the most powerful ranking on summary-statistic space by Neyman–Pearson (Section 7.3)[273] reasoning. The relationship to the classical model-LR is set out in the equivalence to standard FC under sufficiency (Section 12.8) below: under correct specification with sufficient MLE the two rankings coincide asymptotically, while under misspecification they diverge in centroid and

[272] Davies 1977, Davies 1987.

[273] Neyman and Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses".

shape. In our finite-sample regime, all reasonable orderings give comparable widths, but the empirical-LR is the cleanest choice on principle and is the only one that produces a CI naturally centered on the calibrated MLE.

- **Choice of the observable vector:** Whatever the choice of x, coverage is preserved by construction. The choice affects analysis power, not validity.
- **Correlations among observables:** The construction uses the joint sampling distribution and is unaffected by any correlation structure among the components of x.

In every case, the argument is the same: the FC construction estimates the joint sampling distribution of x under each candidate θ from pseudo-experiments and inverts the resulting acceptance regions. No property of the likelihood functional form, the estimator unbiasedness, the asymptotic regime, or the recovery agreement quality enters this guarantee. The bias-corrected point estimate inherits the same shielding for the same reason: it is constructed precisely to compensate for whatever sampling-distribution shape the simulation grid reveals.

This is the parameter-estimation analogue of the PIT-theorem statement that "no property of the likelihood model enters the p-value validity." The concern that "the analysis must reproduce injected signal accurately to be trusted" is misplaced for both inferences. *Recovery quality determines power (the width of the confidence region at fixed coverage), not validity.* Genuine model improvements (refining the energy PDF, fixing a known reconstruction systematic) can simultaneously reduce recovery bias and improve power. Cut-based "improvements" (removing events whose properties happen to make the recovery look worse on a diagnostic plot) shrink the sample and harm both the main hypothesis test and parameter estimation. Validity itself is shielded against either by construction.

What affects only the physical interpretation, not the coverage

Statistical validity is established by the construction. What can shift is the *physical interpretation* of the parameter θ that the construction reports. External critique of an FC interval must therefore be framed as a critique of the simulated signal model (a physics question) rather than the construction itself.

Signal MC fidelity

The pseudo-experiments inject signal events drawn from MC with assumed reconstruction, energy response, and angular response. If any of these differs from how real neutrinos are reconstructed (mismodeled detector response at the signal end, neutrino cross-section uncertainties at energies poorly constrained by data, or anything else), then the parameter θ that "explains" the observed data under the MC is not the same parameter that would explain it under reality. The FC region has guaranteed coverage of θ *under the simulated signal model.* The physical interpretation as a true astrophysical flux depends on the MC's signal fidelity. This

is not specific to FC: any analysis that quotes a flux number inherits the dependence on signal-MC fidelity.

A concrete example is the reconstruction-settings mismatch documented for this sample (Section 6.3): the modern simulation was processed with unfolding settings that differ from those applied to the data, which shifts the reconstructed angular error. The FC coverage statement is unaffected, because it holds under whatever signal model the pseudo-experiments inject. A large *known* mismatch would still be problematic in practice—the construction would give the correct answer to the wrong question—but here the size of the effect was evaluated directly: the legacy-versus-modern comparison bounds it at a ∼2% upward bias in recovered signal flux, far below the width of the confidence intervals reported here. It can moreover be folded into the coverage statement itself through the detector-systematics sampling (Section 12.7), discussed in its own section below.

Source hypothesis

The signal hypothesis underlying the FC construction makes three modeling assumptions, each of which scopes the interpretation of the reported $\hat{\gamma}$ and $\hat{\Phi}(E_{\text{pivot}})$:

- **Spectral form:** signal pseudo-experiments are injected with a single unbroken power law $\Phi(E) \propto E^{-\gamma}$ at each truth point on the grid. The truth grid extends over $\gamma \in [0.5, 4.5]$, wider than the model-fit bounds $\hat{\gamma} \in [1, 4]$. The fit bounds do not constrain coverage: the FC construction operates on the per-cell empirical sampling density rather than on the model likelihood, so per-cell Neyman coverage holds at every truth point on the simulated grid regardless of whether the fit can recover an unconstrained MLE there. What the fit bounds do affect is power: at truth γ outside $[1, 4]$, the active-fit interior thins out as more pseudo-experiments clip into the $\hat{\gamma}$-boundary atom, and the FC interval at those truths broadens correspondingly. Cutoffs, broken power laws, and other deviations from the single power-law family are not represented in the truth-grid signal injection at all, so the FC interval makes no claim about them.

- **Source position:** signal pseudo-experiments inject point sources exactly at the hypothesized declination and right ascension. If the actual source has a small angular offset from the hypothesized position, the FC interval applies to the assumed-position fit, not to the offset source.

- **Source extension:** signal pseudo-experiments inject exact point sources with no spatial-extension parameter on the truth grid. Spatially extended sources are therefore outside the truth-grid signal-injection family entirely, and the FC coverage statement makes no claim about them.

What can invalidate the coverage

The coverage guarantee fails only when the FC per-cell threshold calibration uses a different construction than the one applied to the unblinded data, i.e., when the calibration pipeline and the unblinding pipeline are internally inconsistent. As with the empirical null calibration (Section 9.9) above, this is a statement about *internal* consistency, not about how well the simulated signal matches reality. Signal MC fidelity (whether the injected MC events match real neutrinos) is *not* a coverage issue. It falls under what affects only the physical interpretation (Section 12.6) above. Statistical coverage requires only that the calibration and unblinding go through the same code path:

- **Pipeline inconsistency between calibration and unblinding:** What matters for coverage is that the *analysis* pipeline applied to each calibration pseudo-experiment matches the analysis pipeline applied to the unblinded data: same model likelihood, same fitter (literally the same code, executed against the same per-event PDFs), same per-cell empirical density model $\hat{p}(x;\theta)$, same empirical-LR ranking R. Any code-path difference in any of these between the two halves means the simulated $\hat{p}(x;\theta)$ does not represent what the unblinded pipeline would produce under θ, and the per-cell acceptance regions are calibrated against the wrong distribution. We enforce identity by construction: the analysis code is the same in both regimes wherever it can be, and the only difference between the two paths, the signal-injection code that builds each calibration pseudo-experiment, has no analog in the unblinded analysis (the unblinded data is real data, no injection). Anything specific to the signal-injection code itself belongs to the signal-MC fidelity discussion above and does not affect coverage.

- **In-sample density-model evaluation against the per-cell α-quantile sample:** A subtle but real version of the same pipeline-inconsistency failure mode arises if the per-cell empirical density $\hat{p}(x;\theta)$ is fit to the *same* sample of pseudo-experiments that is then used to compute the per-cell empirical α-quantile of the empirical-LR $R(x;\theta)$. The density model evaluated at its own fitting points (in-sample) carries a small finite-N bias relative to the same model evaluated at fresh draws (out-of-sample), so the calibration α-quantile of R would be computed on a distribution slightly shifted from the one a fresh unblinded observation produces, a mismatch between the calibration and the unblinding R-distributions. The bias is small at our per-cell trial budget, but we eliminate it by construction with the disjoint T/T' split (see the bias-corrected point estimate, Section 12.3): T fits the density, T' calibrates the α-quantile, and the unblinded x_{obs} being real data (never in T) follows the same out-of-sample regime as T'.

- **Inadequate parameter grid coverage:** If the grid does not contain (or interpolate finely enough around) the true θ, no candidate's acceptance region is calibrated correctly at that θ, and the coverage guarantee is vacuous in that

region of parameter space. The grid must span the physically plausible signal range and resolve γ finely enough that the bias correction and acceptance-region boundary are well-defined.

- **Insufficient pseudo-experiments per grid point:** Finite Monte Carlo statistics yield per-cell empirical α-quantile thresholds $R_c(\theta)$ with binomial sampling noise of $\sqrt{\alpha(1-\alpha)/|T'|}$ on the rank fraction (the fraction of T' samples below R_c). At our default $|T'| = 10^4$ per cell this is small: ∼0.47% at the 1σ quantile, ∼0.21% at the 2σ quantile. The 2σ number is *smaller* because the binomial $p(1-p)$ is maximized at $p = 0.5$ and shrinks toward the tails: this is the noise on the *coverage probability* $P_\theta(R \geq R_c)$, which is what matters for whether the construction hits its nominal $1-\alpha$ rate. The corresponding noise on the *threshold value* R_c in R-units is the same binomial divided by the local density of the R distribution at R_c ($\sigma_{R_c} = \sqrt{\alpha(1-\alpha)/|T'|}/f_R(R_c)$), and grows in the tails where f_R is small. A separate $|T|$-driven noise on the R values themselves comes from finite-$|T|$ KDE smoothing of $\hat{p}$ (relative error $\propto 1/\sqrt{|T|\, h^d\, \hat{p}(x;\theta)}$). The two noise sources share an amplification geometry (tail of the x-distribution at θ maps to tail of the R-distribution at θ, so wherever $\hat{p}$ is small both f_R and the KDE relative error are large), but they are mechanistically independent and separately controllable: extending $|T|$ alone shrinks the KDE noise, extending $|T'|$ alone shrinks the threshold noise. The visible boundary raggedness in the FC region figures is expected to be dominated by the $|T|$-driven component at our $|T| = |T'| = 10^4$ scale: the R values jitter cell-to-cell with the finite-$|T|$ KDE landscape more than the threshold does, most visibly in the tails where $\hat{p}$ is small and the region boundary sits. The disjoint T/T' split makes the threshold extensible post-unblinding: T' can be enlarged arbitrarily to shrink both threshold-noise components without touching T, and therefore without changing the empirical density $\hat{p}(x;\theta)$ or the empirical-LR ranking R that define the unblinded result. This refinement converges the construction more precisely to its own asymptotic limit but does not increase analysis power: more T' samples give the density model no new information, so the ranking landscape is unchanged. Re-fitting the KDEs (changing T) would change R and the unblinded result with it.

In contrast to the empirical-null case, where the failure modes are operations applied asymmetrically to pseudo-experiments vs real data on the *background* side, FC coverage is fundamentally tied to internal consistency between the *signal-side* calibration code path and the unblinded analysis code path. This is the structural difference between p-value validity (anchored in real data via scrambling) and confidence-region validity (anchored in simulated signal via injection). Both rely on the self-consistency of their respective constructions, not on physics-level fidelity to reality.

12.7 Treatment of detector systematic uncertainties

For LT events we use the SnowStorm MC ensemble,[274] which marginalizes the simulation over a multivariate prior on the detector systematic parameters rather than fixing them at central values; the parameters, their priors, and their measured effect on the observables are described in Section 6.4 (Section 6.4).

Methodologically, this implements the *Cousins–Highland 1992 hybrid Bayesian–frequentist approach* to systematic-parameter handling[275] (generalized to the FC ordering scheme with coverage analysis under Gaussian, log-normal, and flat nuisance priors in[276]). The classical Cousins–Highland prescription marginalizes the per-event PDF over the systematic prior, $f(x;s) = \int f(x;s,\epsilon')\,P(\epsilon'|\epsilon)\,d\epsilon'$, where x is the per-event observable and s is the signal hypothesis. Our model likelihood is built from SnowStorm MC and inherits this per-event PDF marginalization automatically, but that is not what carries the FC coverage statement: the Neyman-duality coverage argument (Section 12.6) operates directly on the per-cell empirical sampling density $\hat{p}(x;\theta)$ defined in the bias-corrected point estimate (Section 12.3), with the model likelihood entering only as the MLE-summary feature compression. The relevant marginalization is therefore the one that lands in $\hat{p}(x;\theta)$. The Cousins–Highland method is used here only for systematic nuisance parameters; for the parameter of physics interest, we use empirical-LR FC inversion (Section 12.2 above), explicitly *not* a Bayesian-prior marginalization on the POI.

In the full Cousins–Highland construction, each pseudo-experiment draws a single ϵ' realization from the prior $P(\epsilon'|\epsilon)$ and is evaluated entirely under that draw. Across the per-cell ensemble the realizations span the prior support, and the per-cell sample of MLE pairs $(\hat{n}_s, \hat{\gamma})_i$ inherits prior-marginalized statistics by construction. This is the standard HEP convention for systematic-parameter handling in confidence-interval construction. The over-coverage property discussed in the literature is a property of this full construction. Cousins–Highland is known to over-cover relative to the (unknown) fixed-truth coverage when the systematic parameter has a fixed true value across experiments rather than varying experiment-to-experiment[277]—but this is the design intent, not a flaw. When the true systematic value is unknown, marginalizing over the prior is exactly how the construction accounts for that uncertainty in the resulting confidence statement. Over-coverage at any specific fixed truth value is the correct consequence of integrating over our prior uncertainty about which truth value applies. The "over-coverage" framing in the cited references measures coverage at a single fixed truth value rather than the prior-marginalized coverage the construction is built to provide. The prior-marginalized target is the inferential statement we actually want when the truth is unknown. One way to keep this fully frequentist-consistent in spirit: rather than treating the data as a draw from a fixed but unknown true ϵ^*, treat it as a draw from a hypothetical ensemble of possible detector realizations indexed by ϵ, with the prior $P(\epsilon)$ encoding our uncertainty about which detector we actually have. The integration over the prior is then a frequentist statement over that ensemble: the construction's coverage is exact across realizations, and the over-coverage at any specific fixed-ϵ^* reality is the correct consequence of integrating over our

[274] IceCube Collaboration 2019, "Efficient propagation of systematic uncertainties from calibration to analysis with the SnowStorm method in IceCube".

[275] Cousins and Highland 1992, "Incorporating systematic uncertainties into an upper limit".

[276] Tegenfeldt and Conrad 2005, "On Bayesian Treatment of Systematic Uncertainties in Confidence Interval Calculation".

[277] Tegenfeldt and Conrad 2005.

uncertainty about which one to condition on. This is exactly the Cousins–Highland hybrid framing as they describe it themselves[278]: "Our treatment of the Poisson parameter is classical, the type of statistics we generally prefer. Because we average over a probability distribution for the experimental sensitivity, our treatment of that quantity is necessarily Bayesian."[279]

[278] Cousins and Highland 1992, Sec. 1.

[279] Cousins and Highland, "Incorporating systematic uncertainties into an upper limit", p. 331.

The current implementation delivers the marginalization only at the per-event level: the pseudo-experiment event sample is the full-systematics SnowStorm ensemble, so each event carries an independent systematic draw (the per-event PDFs and the injected events are marginalized event-by-event), but a given pseudo-experiment mixes events from many ϵ' realizations rather than being simulated under a single draw. The over-coverage discussion above therefore describes the full construction, not what is currently implemented. The relationship between the two treatments is one of bias versus variance. The per-event mixture under-covers compared to proper per-trial draws: mixing realizations within a trial suppresses the coherent trial-to-trial systematic variation that the full construction is designed to propagate, so the resulting intervals are expected to behave like intervals built on the baseline simulation. Per-trial draws over-cover with respect to any one fixed truth, which may or may not coincide with the best-fit baseline. If the baseline is unbiased, the tighter baseline-like coverage would be correct. If the baseline is actually biased, which we cannot know for sure, the baseline-like intervals would under-cover, and the per-trial systematics treatment is more likely to cover the truth while over-covering with respect to any fixed truth. That is exactly the point of the construction. As a cross-check of the current per-event setup, we compared results from the full-systematics ensemble against the baseline simulation and observed agreement (Section 6.4), consistent with the systematic effects canceling at the per-event level.

A proper per-trial Cousins–Highland marginalization (each pseudo-experiment drawing a single ϵ' realization from $P(\epsilon'|\epsilon)$ and evaluating the entire trial under it) is *not yet implemented*. It is planned for after unblinding, for the publication. It is not a drop-in change: per-trial draws make the detector acceptance systematics-dependent, so the truth-side $n_s \leftrightarrow \Phi$ conversion stops being a fixed mapping. The guiding principle for the implementation is that the truth-side $\Phi \to n_s$ conversion uses that trial's own draw acceptance (it is part of the simulation), while the measured-side $n_s \to \Phi$ conversion always uses the exact, fixed unblinding function (it is part of the estimator and may not depend on truth). In 2D there is a further complication: the n_s/Φ profiles differ across γ, so interval endpoints cannot simply be transformed. A possible clean approach under consideration is to sample the truth grid independently in n_s and Φ. The prior choice itself is also open. The full uniform ensemble priors are almost certainly overly conservative, and Gaussian priors centered on the best-fit baseline are under consideration. The effects of the perturbed parameters are moreover correlated. For tracks the two most impactful parameters are bulk-ice absorption and DOM efficiency, both of which affect the total light yield: perturbing one of them upward moves the other's conditional best fit downward, so perturbing both upward together would unrealistically over-cover. Past applications used likewise uniform, uncorrelated ensemble priors, the

most conservative prior shape.[280] Compared to that, the plain baseline seems more likely to give closer-to-truth coverage. Per-trial sampling itself has precedent: the Galactic-plane analysis sampled a realization of the systematic parameters for every pseudo-experiment.[281] Partial perturbation also has precedent: prior IceCube FC applications perturbed up to three SnowStorm parameters per pseudo-experiment, which strictly under-covers relative to the full parameter set (Section 6.4) and was adopted for MC-statistics reasons. None of these choices is final for the analysis.

Background-side systematics (atmospheric flux models, cosmic-ray hadronic interaction models, and similar) are irrelevant to this analysis end to end: backgrounds are modeled entirely from data through RA randomization (Section 9.9) of the observed event list at every stage (empirical null calibration for the hypothesis test, FC simulation grids for parameter estimation), with no MC entering the background under any construction here.

Signal acceptance is by construction model-dependent, and the $n_s \leftrightarrow \Phi$ conversion is the only place in the analysis where the per-declination signal acceptance $A_{\text{eff}}(\gamma, \delta)$ enters: the FC region's coverage statement lives entirely in (n_s, γ) space and is indifferent to acceptance. The present construction can operate in n_s alone precisely because the acceptance model stays fixed at the baseline for all trials, which keeps $n_s \leftrightarrow \Phi$ a fixed mapping. Under proper per-trial draws that symmetry breaks: the acceptance becomes a property of each trial's own systematic realization, and the conversion is no longer a single function. This is the same complication that drives the per-trial implementation plan above, and the acceptance treatment must be resolved together with it.

Once per-trial marginalization is implemented, the coverage statement will be in the prior-marginalized sense: averaged over the five-parameter prior support, the CI covers θ at the nominal level. The present per-event implementation does not yet deliver that statement. Its intervals behave like baseline intervals, as discussed above, and are therefore narrower than a full per-trial construction would give. The direction is certain even though the magnitude is not: the per-parameter SnowStorm slices (Section 6.4) show that the effective-area and angular-resolution responses to the systematic parameters are real and nonzero, so restoring the coherent per-trial draws adds genuine trial-to-trial variance on top of the baseline. By the law of total variance, that added between-trial spread widens the per-cell sampling distribution, and wider sampling distributions require wider intervals for nominal coverage; the per-event construction averages the systematics within each trial, suppresses that spread, and is correspondingly narrower. Only the size of the effect is unquantified. If the true detector configuration lies outside that support, or differs in a parameter that is not perturbed (anisotropy, depth profile), the framework provides no systematic correction in those directions.

[280] The specific prior shape (uniform, uncorrelated, ensemble) and the partial-perturbation choice summarized here are taken from internal collaboration documentation and analysis code; they are not stated in the published Galactic-plane analysis, which describes only the per-trial sampling of a realization of systematic parameters.

[281] IceCube Collaboration 2023, "Observation of high-energy neutrinos from the Galactic plane", supplement, "Systematic Uncertainties and their Impact".

12.8 Equivalence to standard Feldman–Cousins under sufficiency

The empirical-LR construction described above replaces the classical model-LR ranking with one built from the empirical sampling density of the MLE summary statistic. A natural question, complementary to the discussion of why Wilks-based intervals are inadequate (Section 12.1) above, is: when do the FC confidence regions from these two ranking choices coincide, and when do they diverge? The answer is given by a tight equivalence theorem that connects the question to the classical statistics of sufficient statistics:

Proposition 12.1. Let $L(x;\theta)$ be a parametric likelihood with MLE $\hat{\theta}(x)$, let $p(\hat{\theta};\theta)$ be the true sampling density of the MLE under truth θ, and let $\tilde{\theta}(\hat{\theta}) = \arg\max_\theta p(\hat{\theta};\theta)$ be the calibrated MLE. Define the two FC ranking functions:

- Model-LR: $R_L(x,\theta) = L(x;\theta)/L(x;\hat{\theta}(x))$
- Empirical-LR: $R_p(\hat{\theta},\theta) = p(\hat{\theta};\theta)/p(\hat{\theta};\tilde{\theta}(\hat{\theta}))$

The FC confidence regions inverted from R_L and R_p using empirical-quantile thresholds coincide at every observation x_{obs} and every confidence level $1-\alpha$ if and only if $\hat{\theta}$ is a sufficient statistic for θ.[282]

[282] There seems to be no exact statement of this equivalence in the literature, so we provide our own proof below, combining several standard results.

Proof. Throughout, the proof assumes the following regularity conditions. The MLE $\hat{\theta}(x)$ and the calibrated MLE $\tilde{\theta}(\hat{\theta})$ exist and are unique (well-defined argmax). The sampling density $p(\hat{\theta};\theta)$ of the random variable $\hat{\theta}(X)$ under $X \sim L(\cdot;\theta)$ exists and is positive on the support of $\hat{\theta}$, so the ranking ratios are well-defined. The data-distribution support is atomless under $L(\cdot;\theta)$, so rankings induce well-defined orderings of the upper level sets and quantiles are continuous in α. These are the standard textbook-level regularity conditions for theorems of this type and hold in the iid continuous-density setting.

($\Leftarrow$) Suppose $\hat{\theta}$ is sufficient. By the Fisher–Neyman factorization theorem[283] (historical statement: Halmos and Savage[284]), there exist functions g and h such that

$$L(x;\theta) = g(\hat{\theta}(x);\theta)\, h(x). \tag{12.13}$$

[283] Casella and Berger 2002, Theorem 6.2.6, p. 276.

[284] Halmos and Savage 1949, "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics".

Marginalizing h over the level set $\{x' : \hat{\theta}(x') = \hat{\theta}\}$ gives $p(\hat{\theta};\theta) = g(\hat{\theta};\theta)\, c(\hat{\theta})$ with $c(\hat{\theta})$ independent of θ, so g and p are proportional as functions of θ at fixed $\hat{\theta}$. Consequently

$$\tilde{\theta}(\hat{\theta}) = \arg\max_\theta\ p(\hat{\theta};\theta) = \arg\max_\theta\ g(\hat{\theta};\theta). \tag{12.14}$$

Combined with $\hat{\theta}(x) = \arg\max_\theta L(x;\theta) = \arg\max_\theta g(\hat{\theta}(x);\theta)$ (since h is θ-independent), this gives $\hat{\theta}(x) = \tilde{\theta}(\hat{\theta}(x))$: the model MLE and the calibrated MLE coincide. Substituting:

$$R_L(x,\theta) = \frac{g(\hat{\theta}(x);\theta)}{g(\hat{\theta}(x);\hat{\theta}(x))} = \frac{p(\hat{\theta}(x);\theta)}{p(\hat{\theta}(x);\tilde{\theta}(\hat{\theta}(x)))} = R_p(\hat{\theta}(x),\theta). \tag{12.15}$$

The two rankings coincide pointwise. The empirical thresholds $R_c(\theta)$ (the $(1-\alpha)$-quantiles over pseudo-experiments at truth θ) therefore agree, the acceptance regions agree, and the FC contours coincide at every x_{obs} and every α.

($\Rightarrow$) Suppose the FC contours coincide at every x_{obs} and every α. The acceptance regions $A_L(\theta;\alpha)$ and $A_p(\theta;\alpha)$ must then agree for every θ and every α. Equivalently, the rankings $R_L(\cdot,\theta)$ and $R_p(\hat\theta(\cdot),\theta)$ induce identical orderings of pseudo-experiments at every truth θ, since otherwise the top-$(1-\alpha)$ subsets would differ for some α. So $R_L(x,\theta)$ is a strictly monotone function of $R_p(\hat\theta(x),\theta)$ at fixed θ. Since R_p depends on x only through $\hat\theta(x)$, so does R_L:

$$\frac{L(x;\theta)}{L(x;\hat\theta(x))} = F(\hat\theta(x),\theta) \quad\Longleftrightarrow\quad L(x;\theta) = F(\hat\theta(x),\theta)\cdot L(x;\hat\theta(x)). \tag{12.16}$$

Setting $g(\hat\theta,\theta) = F(\hat\theta,\theta)$ and $h(x) = L(x;\hat\theta(x))$ gives $L(x;\theta) = g(\hat\theta(x),\theta)\,h(x)$, which is the Fisher–Neyman form. By Casella and Berger,[285] $\hat\theta$ is sufficient for θ. □

[285] Casella and Berger 2002, Theorem 6.2.6, p. 276.

Corollary 12.1 (equivalence under correct specification)**.** Under regularity conditions and correct model specification, the MLE is asymptotically efficient[286] and, more sharply, in the local asymptotic normality (LAN) sense, asymptotically sufficient.[287] The empirical-LR FC therefore reduces asymptotically to the model-LR FC, recovering as a special case the simulation-based inference LR-invariance result of Cranmer et al.[288].

[286] Casella and Berger 2002, Theorem 10.1.12, p. 472, Lehmann and Casella 1998, Theorem 5.1, p. 463.

[287] Le Cam 1986, *Asymptotic Methods in Statistical Decision Theory*, Ch. 10, Sec. 2, Theorem 1, p. 177, Vaart 1998, Theorem 7.10, p. 98.

[288] Cranmer, Pavez, and Louppe 2015, Theorem 1.

Corollary 12.2 (divergence under misspecification)**.** Under model misspecification, the MLE converges to a *pseudo-true* parameter $\theta^* \neq \vartheta$[289] and is generally not sufficient. The model-LR ranking is no longer asymptotically χ^2-distributed,[290] and the empirical-LR and model-LR FC constructions diverge: confidence-region centroids differ by the bias $\theta^* - \vartheta$, and shapes generally differ as well: in the asymptotic Gaussian-MLE regime, the model-LR contour is governed by the model Hessian while the empirical-LR contour reflects the actual sampling-distribution covariance of the MLE (the gap between the sandwich and Hessian estimators).[291] Both retain correct frequentist coverage by Neyman duality,[292] but only the empirical-LR construction centers the contour on the bias-corrected estimate $\tilde\theta(\hat\theta_{\text{obs}})$.

[289] White 1982, Theorem 2.2.

[290] Vuong 1989.

[291] White 1982, Theorems 3.2–3.3.

[292] Casella and Berger 2002, Theorem 9.2.2, p. 421, Dalmasso et al. 2024, Theorem 1, Feldman and Cousins 1998, Sec. II.

Implication for our regime

This places our construction in a clean setting. Under correctly-specified likelihoods, the empirical-LR FC and the classical model-LR FC give identical confidence regions. The difference is a relabeling of the centroid from $\hat\theta$ to $\tilde\theta = \hat\theta$. Under the misspecified PSF model, where the signal-recovery diagnostics (Section 9.10) directly demonstrate $\hat\theta \neq \vartheta$ in the median, the two constructions differ in centroid (by the empirical bias) and in shape (by the gap between sandwich and Hessian variance estimators). Empirical-LR FC delivers what we want in that regime: a confidence region with correct frequentist coverage centered on the bias-corrected

calibrated MLE. The model-LR FC, while still coverage-correct, reports a contour centered on the biased MLE, which propagates the bias into any downstream use of the parameter estimate (the primary operational motivation, Section 12.3, for going empirical in the first place).

12.9 One construction, two scopes

The two inferential machineries on this page (the empirical null calibration that produces the post-trial significance claim, and the Feldman–Cousins construction that produces the parameter estimates and confidence regions) are not independent constructions. They are the *same* Neyman-duality machinery applied at different scopes:

- **Significance:** Feldman–Cousins at the boundary atom $\theta = (n_s = 0, \cdot)$. Pseudo-experiments at this single truth point are the RA-resampled background trials. The ranking is the test statistic from the model-likelihood ratio $T(x) = 2\ln[\mathcal{L}(x;\hat{\theta})/\mathcal{L}(x;0)]$. The per-pixel empirical α-quantile of the BG-trial T-distribution gives the rejection threshold. The per-pixel p-value is the empirical fraction of BG trials with T above T_{obs}.
- **Parameter estimation:** Feldman–Cousins over the full (n_s, γ) truth grid. Pseudo-experiments at each truth cell are the per-cell calibration trials. The ranking is the empirical likelihood-ratio $R(x;\theta) = \hat{p}(x;\theta)/\hat{p}(x;\tilde{\theta}(x))$. The per-cell empirical α-quantile of the R-distribution defines the FC region. Profile-FC intervals on γ and on $\Phi(E_{\text{pivot}})$ extract per-axis confidence statements from the joint construction.

Both are Neyman-duality constructions. Both have $\geq 1 - \alpha$ coverage by the same theorem (Section 12.6). They differ in *scope* (one truth point vs. full grid) and in *ranking choice* (model-LR vs. empirical-LR). NOvA makes the structural identity explicit (Sec. 4.4): "the procedure can naturally address these binary tests (or discrete choices in general) since when applied to a single point the procedure becomes a classic likelihood ratio test with Monte Carlo used to determine the p-value."[293] The standard IceCube point source significance test is exactly that: Feldman–Cousins at a single truth point ($n_s = 0$), with the MC pseudo-experiments being the BG trials.

[293] NOvA Collaboration, "Monte Carlo method for constructing confidence intervals with unconstrained and constrained nuisance parameters in the NOvA experiment", p. 18.

The trial-count asymmetry

Each scope's pseudo-experiment budget reflects the inference it serves. Single-point tests can pour the entire trial budget into one truth cell. Grid-spanning constructions must distribute budget across knots. The NOvA paper states this explicitly:[294]

> "Since this procedure is only done at one point of the parameter space for each hypothesis test, we can afford to generate more FC pseudoexperiments (tens of thousands) and reach more accurate measurements

[294] NOvA Collaboration 2025, Sec. 4.4.

of the p-values and significances than for 1D and 2D confidence intervals."

— NOvA Collaboration, 2025, p. 18

And from the CI side,[295] immediately preceding:

[295] NOvA Collaboration 2025, Sec. 4.3.

"Between 1000 and 5000 FC pseudoexperiments are generated at each θ_i [...] given the very large number of FC pseudoexperiments that are required in the 3-sigma (and above) regions in order to accurately measure the corresponding small p-values, we choose to only perform the profile construction in regions where $\sqrt{\lambda_{\text{Wilks}}} < 20$ for 1-dimensional constraints and $\sqrt{\lambda_{\text{Wilks}}} < 12$ for 2-dimensional constraints."

— NOvA Collaboration, 2025, p. 17

Our analysis is the same structural fact at different absolute scales: $\sim 10^7$ pseudo-experiments per declination ring at the $n_s = 0$ atom (which gives the significance precision needed for a 5σ-grade deep-tail claim) versus $\sim 10^4$ per FC grid cell at the parameter-estimation scope (where the relevant CL levels are 1σ and 2σ, well-resolved by the per-cell empirical α-quantile at that trial budget). The asymmetry is not a quirk of how the analysis was historically structured—it is the natural consequence of the inference each scope is trying to support.

Ranking choice: model-LR for significance, empirical-LR for parameter estimation

The two scopes use different ranking choices in this analysis. The asymmetry is historical rather than methodologically required.

For significance, the analysis uses the model-LR ranking, the standard test statistic $T = 2 \ln L(\hat{\theta})/L(0)$. This is the standing IceCube point source convention, predating the empirical-LR FC framework developed for the parameter-estimation construction here. At the $n_s = 0$ truth atom, the FC pseudo-experiments are RA-resampled real data with no signal injection (structurally identical to the BG trials that drive the per-pixel α-quantile of T), so empirical-LR ranking would in principle apply there equally well, likely with better power (the empirical-LR ranking converges to the most powerful Neyman–Pearson likelihood ratio in the limit of infinite calibration statistics, while model-LR carries whatever PSF/PDF mismatches the model likelihood inherits). This symmetry went unrecognized until the development of the empirical-LR FC framework here for parameter estimation. Applying it to the per-pixel significance test as well is a candidate future improvement, not part of the current analysis.

For parameter estimation, the analysis uses the empirical-LR ranking, which absorbs the MLE bias under misspecification (Section 12.8) and centers the FC contour on the bias-corrected calibrated MLE rather than on the biased model MLE, the operational motivation laid out at the top of the parameter-estimation discussion (Section 12.3).

In the limit of an unbiased, well-specified likelihood, the two rankings collapse to the same FC region because they are equivalent under sufficiency (Section 12.8). The formal limit recovers a single inferential framework exactly.

Part III

Finding Astrophysical Neutrino Sources

13 Searching for Neutrino Sources

With the statistical machinery of Part II in place, the question that remains is a physical one: which objects in the universe should produce the neutrinos we are searching for, and why. This chapter steps back from the methods to the targets. It first sets out the physics that makes an astrophysical object a plausible neutrino source, then introduces the classes of object that carry that physics, and finally defines the list of candidates this work actually tests (Section 13.2).

13.1 Neutrino production and source classes

A high-energy neutrino is produced essentially only in hadronic interactions. When protons or heavier nuclei are accelerated to high energy and collide with surrounding gas (a proton–proton interaction) or radiation (a proton–photon interaction), the collisions produce pions of all three charges.[296] The neutral pions decay to two gamma rays, $\pi^0 \to \gamma\gamma$, while the charged pions decay through the same chain that yields the atmospheric neutrino flux (Section 2.2): $\pi^+ \to \mu^+\nu_\mu$ followed by $\mu^+ \to e^+\nu_e\bar{\nu}_\mu$, and the charge conjugates.[297] A hadronic accelerator therefore radiates gamma rays and neutrinos together. The electromagnetic processes that dominate most astrophysical emission produce no neutrinos, so a detected high-energy neutrino is unambiguous evidence that its source accelerates hadrons—the property that makes the neutrino a clean probe of cosmic-ray origins (Chapter 1).

This ties the search for neutrino sources to the century-old question of where cosmic rays come from.[298] Cosmic rays are charged, so galactic and intergalactic magnetic fields scramble their arrival directions and they no longer point back to their accelerators. The hadronic collisions that produce them, however, also produce neutrinos, which arrive undeflected. Wherever cosmic rays are accelerated and meet enough target gas or radiation, neutrinos are expected to follow.[299] The candidate sources for a neutrino search are therefore the objects believed to accelerate hadrons to the highest energies, whatever their type. This link between the origin of cosmic rays and neutrino emission, and with it the idea of doing astronomy with neutrinos, was recognized long before any detector could act on it: large Cherenkov detectors for cosmic neutrinos were proposed around 1960, both underground[300] and, by Markov, deep underwater.[301]

[296] Gaisser, Halzen, and Stanev 1995, "Particle astrophysics with high energy neutrinos".

[297] Halzen and Hooper 2002, "High-energy neutrino astronomy: the cosmic ray connection".

[298] Baade and Zwicky 1934, "Cosmic Rays from Super-Novae", Fermi 1949, "On the Origin of the Cosmic Radiation".

[299] Halzen and Hooper 2002.

[300] Greisen 1960, "Cosmic Ray Showers".

[301] Spiering 2012, "Towards High-Energy Neutrino Astronomy. A Historical Review".

Particle acceleration

For a source to produce neutrinos it must first accelerate the parent hadrons, and across these candidate objects the same mechanism does the work: *diffusive shock acceleration*[302], a first-order variant of the mechanism Fermi originally proposed.[303] A charged particle repeatedly crosses a shock front, gaining a small fractional energy at each crossing as it scatters off magnetic turbulence on either side, and the competition between this gradual energization and the probability of escaping downstream yields a power-law spectrum of accelerated particles: "The consequent energy spectrum is a power law with an index close to that observed for galactic cosmic rays".[304] This power-law form motivates the single spectral-index hypothesis used to describe candidate-source emission, and it underlies the power-law signal model we use throughout the search (Section 9.1).[305] The shocks occur in settings as different as the relativistic jets of active galaxies, the supernova remnants of star-forming galaxies, and the accretion flows feeding obscured nuclei, which is why such different objects share a place on the same candidate list, and why the classes that follow are organized by the engine that drives them.

The objects that follow are the source classes from which we construct our test catalog. Each is introduced as a physical type, with a few individual members named as examples. The rule that selects specific objects of each class into the catalog is defined in the next section (Section 13.2). The gamma-ray observations that anchor this discussion, and the source list itself, come from the Fermi Gamma-ray Space Telescope: its Large Area Telescope is a pair-conversion detector sensitive from about 20 MeV to over 300 GeV[306], and its 16-year Source List (FL16Y), an interim list of 7,220 sources ahead of the forthcoming 5FGL catalog, is the list from which our catalog is built.[307]

Active galactic nuclei

Active galactic nuclei (AGN) are the most numerous class among the candidates: galaxies whose central supermassive black hole accretes matter so vigorously that the nucleus outshines the surrounding starlight. A fraction of AGN launch collimated relativistic jets of magnetized plasma, and these jets are efficient particle accelerators. In the *unified model* of AGN, the wide variety of observed AGN types is attributed largely to one underlying object, an accreting black hole with its accretion disk and jet, viewed from different angles and at different accretion rates, rather than to intrinsically different objects.[308]

> "The appearance of active galactic nuclei (AGN) depends so strongly on orientation that our current classification schemes are dominated by random pointing directions instead of more interesting physical properties."
>
> — Urry and Padovani, 1995, p. 803

When the jet points close to the observer's line of sight, the emission is strongly Doppler-boosted and the object appears as a *blazar*, the brightest and most variable

[302] Axford, Leer, and Skadron 1977, "The acceleration of cosmic rays by shock waves", Blandford and Ostriker 1978, "Particle acceleration by astrophysical shocks", Ginzburg and Syrovatskii 1964, *The Origin of Cosmic Rays*.

[303] Fermi 1949.

[304] A. R. Bell, "The acceleration of cosmic rays in shock fronts - I", *Monthly Notices of the Royal Astronomical Society* 182, no. 2 (1978): p. 147.

[305] Drury 1983, "An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasmas".

[306] Atwood et al. 2009, "The Large Area Telescope on the Fermi Gamma-ray Space Telescope Mission".

[307] Fermi-LAT Collaboration 2026, "Fermi-LAT 16-year Source List".

[308] Urry and Padovani 1995, "Unified Schemes for Radio-Loud Active Galactic Nuclei".

class of AGN.[309] The boosting is relativistic beaming: the jet plasma streams toward us at close to the speed of light, so its radiation is concentrated into a narrow forward cone and shifted to higher energies, making an aligned jet appear far brighter and more rapidly variable than the same jet seen from the side. Blazars divide into two subclasses by their optical spectra, a split that traces a physical difference in how the black hole accretes. The presence or absence of strong emission lines reflects the accretion regime: "the presence of strong emitting lines is related to a transition in the accretion regime, becoming radiatively inefficient below a disc luminosity of the order of 1 per cent of the Eddington one".[310] The Eddington luminosity is the natural brightness limit of an accreting black hole, the level at which the outward radiation pressure balances the inward pull of gravity. Below roughly one per cent of it the flow radiates inefficiently and the emission lines are weak, while above it a bright, radiatively efficient disk drives strong ones. *BL Lacertae objects* (BL Lacs; FL16Y class BLL) show weak or absent emission lines, the signature of that radiatively inefficient flow, over a jet-dominated continuum. The class is named for its prototype, BL Lacertae, which lies at right ascension $22^{h}03^{m}$, declination $+42.3°$[311] and is itself one of the catalog's tested sources (Section 13.2). Several TeV-emitting BL Lacs also enter the catalog, among them Markarian 421, PKS 1424+240, and the southern PKS 2155-304.[312] *Flat-spectrum radio quasars* (FSRQs; FL16Y class FSRQ) show the strong, broad emission lines of the radiatively efficient regime and are typically more luminous. 3C 279, the first quasar detected in gamma rays by the EGRET telescope[313], is a bright representative. A substantial number of catalog blazars cannot be confidently placed in either subclass from the available data and are recorded simply as blazars of uncertain type (FL16Y class BCU).

Whether blazar jets emit neutrinos at a detectable level is an open question. Their gamma-ray emission can be explained by purely *leptonic* processes, in which relativistic electrons up-scatter low-energy photons to gamma-ray energies by *inverse Compton scattering*[314], with no neutrino production. A *hadronic* contribution, in which accelerated protons produce gamma rays and neutrinos together, is possible but not established for any individual blazar. Which subclass would dominate any neutrino output is unsettled. The single suggestive case is TXS 0506+056, a blazar from whose direction IceCube recorded a high-energy neutrino in temporal and spatial coincidence with a gamma-ray flare, and, separately, an excess of lower-energy neutrinos in archival data.[315] The flare coincidence itself was reported as "statistically significant at the level of 3 standard deviations".[316] Both numbers are best read as a-posteriori significances: the flare coincidence rests on correlation models selected only after the alert, by the reporting analysis's own account, and the archival excess, though corrected for the look-elsewhere effect of the time-window scan, applies to a source singled out by that same alert.[317]

When the same jetted AGN is viewed at a large angle to the jet, it appears instead as a *radio galaxy* (FL16Y class RDG): the misaligned, un-boosted counterpart of a blazar in the unified picture.[318] The catalog's representative of this class is M 87, the giant elliptical galaxy at the center of the Virgo cluster and one of the nearest radio galaxies.

A further class, the *narrow-line Seyfert 1 galaxies* (FL16Y class NLSy1), are AGN

[309] Urry and Padovani 1995.

[310] G. Ghisellini et al., "The transition between BL Lac objects and flat spectrum radio quasars", *Monthly Notices of the Royal Astronomical Society* 414, no. 3 (2011): p. 2674.

[311] Fermi-LAT Collaboration 2026.

[312] Fermi-LAT Collaboration 2026.

[313] Hartman et al. 1992, "Detection of high-energy gamma radiation from quasar 3C 279 by the EGRET telescope on the Compton Gamma Ray Observatory".

[314] Blumenthal and Gould 1970, "Bremsstrahlung, Synchrotron Radiation, and Compton Scattering of High-Energy Electrons Traversing Dilute Gases".

[315] IceCube Collaboration 2018, "Neutrino emission from the direction of the blazar TXS 0506+056 prior to the IceCube-170922A alert".

[316] IceCube Collaboration et al., "Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A", *Science* 361, no. 6398 (2018): p. 1.

[317] IceCube Collaboration et al. 2018, IceCube Collaboration 2018.

[318] Urry and Padovani 1995.

with comparatively low black-hole masses and high accretion rates; a minority of them launch relativistic jets and so emit gamma rays, which is why a few appear among the candidates drawn from gamma-ray catalogs.[319] PMN J0948+0022, the first narrow-line Seyfert 1 detected at gamma-ray energies, is the catalog's brightest example. Whether this class produces detectable neutrinos is unclear, and a recent modeling study of a jetted narrow-line Seyfert 1 favors a leptonic origin.[320]

[319] Fermi-LAT Collaboration 2009, "Radio-loud narrow-line Seyfert 1 as a new class of gamma-ray active galactic nuclei".

[320] Wang et al. 2023, "On the Hadronic Origin of High Energy Emission of γ-ray Loud Narrow-Line Seyfert 1 PKS 1502+036".

Obscured AGN and the role of the corona

The jetted AGN above are selected because they are gamma-ray bright. The strongest evidence to date for a steady extragalactic neutrino source[321] points instead to an object that is gamma-ray faint at the relevant energies: NGC 1068, a nearby *Seyfert galaxy* whose central engine is hidden behind dense, dusty gas (an *obscured*, or type-2, AGN).[322] In such systems the neutrinos are thought to originate not in a jet but in the hot *corona* of plasma immediately around the black hole, where protons accelerated near the accretion flow collide with the dense ambient photon and gas fields. This same density makes the region opaque to the GeV gamma rays that would accompany the neutrinos: the gamma rays are absorbed and reprocessed to lower energies before they escape, so a corona can be a strong neutrino source while remaining inconspicuous in the gamma-ray catalog used to build our candidate list. Such coronae are, in the language of these models, "hidden sources preventing the escape of GeV–TeV gamma rays".[323]

[321] IceCube Collaboration 2022a, "Evidence for neutrino emission from the nearby active galaxy NGC 1068".

[322] Antonucci and Miller 1985, "Spectropolarimetry and the nature of NGC 1068".

[323] Kohta Murase, Shigeo S. Kimura, and Peter Mészáros, "Hidden Cores of Active Galactic Nuclei as the Origin of Medium-Energy Neutrinos: Critical Tests with the MeV Gamma-Ray Connection", *Physical Review Letters* 125, no. 1 (2020): p. 011101-1.

This points to a selection bias worth stating plainly. A catalog ranked by gamma-ray brightness favors the sources whose high-energy emission escapes freely, which are predominantly the jetted, beamed AGN, and it systematically under-weights the obscured coronae that may be among the most propitious neutrino emitters. The lesson of NGC 1068 is that a focus on the brightest, jet-dominated gamma-ray sources had left the obscured, corona-dominated targets in Seyfert galaxies under-appreciated as neutrino emitters, even though the dense material that hides them from the gamma-ray catalogs is assumed to be the very target in which the neutrinos are produced. A dedicated search optimized for the obscured-Seyfert population is therefore a natural next step. Because these obscured Seyferts are bright in hard X-rays while faint in gamma rays, the natural selection basis is a hard X-ray catalog, such as the Swift-BAT AGN Spectroscopic Survey (BASS)[324], rather than a gamma-ray catalog. This is the choice made by the Northern Tracks X-ray AGN search.[325] That direction, together with a targeted Galactic-plane search, is taken up in the outlook (Chapter 15).

[324] Koss et al. 2022, "BASS XXII: The BASS DR2 AGN Catalog and Data".

[325] IceCube Collaboration 2026a, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube".

Star-forming galaxies

A second, physically distinct extragalactic class is the *starburst galaxies* (FL16Y class SBG): galaxies undergoing an intense episode of star formation[326], with the dense gas and high supernova rate that follow. Unlike AGN, these galaxies need no jet. Their many supernova remnants accelerate cosmic-ray protons throughout the galaxy. Each supernova drives a blast wave into the surrounding gas, and the

[326] Fermi-LAT Collaboration 2012a, "GeV Observations of Star-forming Galaxies with Fermi LAT".

expanding shock front is a site of diffusive shock acceleration (Section 13.1), the same mechanism that operates at the jets and accretion flows above. A galaxy forming stars rapidly hosts many such remnants at once, which is why supernova remnants are regarded as the primary accelerators of galactic cosmic rays.[327] The dense interstellar gas acts as a target thick enough that a large fraction of those protons interact before escaping, making the galaxy an efficient hadronic emitter, a so-called cosmic-ray calorimeter that can "convert efficiently cosmic-rays into pions, which in turn decay into high-energy neutrinos and photons".[328] Because their hadronic nature is comparatively secure, the catalog includes its star-forming galaxies on a less stringent basis than the jetted AGN. The nearby southern starbursts NGC 253 and NGC 4945 are among the catalog's members, together with the Large and Small Magellanic Clouds, the Milky Way's two satellite galaxies, whose proximity makes even their more modest star formation a worthwhile target. Because the construction includes every star-forming galaxy unconditionally, NGC 1068 was selected for its star formation rather than for the obscured nucleus the corona scenario invokes (Section 13.1). Its gamma-ray output is otherwise modest, so its inclusion was in that sense a coincidence. Only afterward did the neutrino measurement reveal a flux exceeding the potential TeV gamma-ray emission by at least an order of magnitude[329], far more than star formation alone could power, and leaving the obscured-AGN corona as the most plausible engine.

[327] Ginzburg and Syrovatskii 1964.

[328] Abraham Loeb and Eli Waxman, "The cumulative background of high energy neutrinos from starburst galaxies", *Journal of Cosmology and Astroparticle Physics* 2006, no. 05 (2006): p. 1.

[329] IceCube Collaboration 2022a.

13.2 The gamma source list construction

The statistical machinery that tests this catalog is built in Chapter 11, there the catalog is nothing but a list of (δ, α) positions. Now it is time to define the list. We build it around a deliberately optimistic assumption: that the GeV gamma-ray emission of a candidate source is predominantly hadronic, so that its measured gamma-ray flux tracks its neutrino flux. Under this assumption the brightest gamma-ray sources are the most promising neutrino candidates, and the candidates can be ranked by their measured gamma-ray flux. The assumption is known to be optimistic, as discussed in the preceding section. No attempt is made to remove sources whose gamma rays are likely leptonic, and the construction does not single out the obscured Seyfert galaxies, bright in X-rays, that the corona scenario favors (Section 13.1).

The reason for accepting this assumption, rather than building a more physically motivated list, is methodological. The selection is a fixed *algorithm*, defined before any source significances are known. An experiment like IceCube retests essentially the same data many times as its samples and methods improve, and a source list chosen with knowledge of earlier results—subconsciously or not—would risk self-triggering: rediscovering the analyzer's own prior fluctuations and mistaking them for signal. Fixing the selection rule in advance removes that experimenter bias.

This construction is the established precedent for IceCube point-source searches. It was introduced in the ten-year, time-integrated search for point sources[330] and carried forward to each new event sample since: the Northern Tracks NGC 1068

[330] IceCube Collaboration 2020, "Time-integrated Neutrino Source Searches with 10 years of IceCube Data".

analysis[331], the DNNCascade Galactic-plane analysis, which also tested a catalog of GeV gamma-ray emitters built by the same construction[332], the ESTES analysis[333], the Northern Tracks X-ray AGN search[334], and now the Lightning Tracks sample of this work. For the first analysis of a new sample we follow it, and defer the more physically targeted constructions, such as a search optimized for the obscured Seyferts or for the Galactic plane, to future work (Chapter 15).

[331] IceCube Collaboration 2022a.

[332] IceCube Collaboration 2023, "Observation of high-energy neutrinos from the Galactic plane".

[333] IceCube Collaboration 2026d, "Time-Integrated Southern-Sky Neutrino Source Searches with 10 Years of IceCube Starting-Track Events at Energies Down to 1 TeV".

[334] IceCube Collaboration 2026a.

Source list construction

The source list is built by a fixed, reproducible procedure originated by Nahee Park for the earlier IceCube source selections based on gamma-ray catalogs and applied here to the sample of this work. Fixing the procedure in advance, rather than choosing sources by hand, is what removes experimenter bias from the search. The catalog holds a fixed total of 110 sources, a size set by the trial-correction target so that a source significant at 5σ before the trial correction remains roughly 4σ after it. The galactic candidates are selected first and the remaining places are filled from extragalactic Fermi sources, so the 2 galactic and 108 extragalactic are an outcome of the selection, not a preset split.

The two galactic sources are selected first, drawn from the gamma-ray sources measured by LHAASO, a wide-field observatory sensitive to the highest-energy gamma rays. Its first catalog reports that "43 sources are detected with ultra-high energy (E > 100 TeV) emission".[335] Each candidate's measured gamma-ray spectrum is converted, under the same optimistic hadronic assumption, into a predicted neutrino flux and compared to the differential 3σ discovery potential at the source declination. The source is kept if the prediction exceeds the sensitivity. Both of LHAASO's detector arrays (a lower-energy water-Cherenkov array and a higher-energy square-kilometer array) are used as measured, rather than extrapolating the high-energy fit down into the low-energy range, where it overpredicts the flux. Most of these candidates are pulsar wind nebulae whose gamma-ray emission is likely leptonic. They are kept all the same: the construction is optimistic by design and makes no attempt to filter out the likely-leptonic sources. A measurement still earns its place, since even a non-detection sets a limit on any hadronic component. Two survive: the Crab nebula and MGRO J1908+06. Others, such as Vela X, are excluded for their large angular extent and the resulting source confusion, and the Galactic Center region is left to a dedicated future search (Chapter 15) rather than forced onto this list.

[335] LHAASO Collaboration, "The First LHAASO Catalog of Gamma-Ray Sources", *The Astrophysical Journal Supplement Series* 271, no. 1 (2024): p. 2.

The remaining budget (110 minus the two galactic sources) is filled algorithmically from the 16-year Fermi-LAT source list[336] within $|\delta| < 80°$, across the source classes introduced in Section 13.1: the blazars, the minor radio-loud active-galaxy subtypes pooled together, and the star-forming and normal galaxies. Each source is weighted by its Fermi photon flux above 1 GeV times the declination-dependent sensitivity factor $\min(\mathrm{DP})/\mathrm{DP}(\delta)$, where $\mathrm{DP}(\delta)$ is the LT + DNNC 3σ discovery potential at $\gamma = 2.37$ (the Northern Tracks diffuse spectral index), interpolated linearly in $\sin\delta$. The budget is then filled by class. Starburst and normal galaxies are included unconditionally (hadronic emitters), which is why NGC 1068 is in the list

[336] Fermi-LAT Collaboration 2026.

despite its relatively low gamma-ray flux. BL Lacs and FSRQs each contribute their top fraction by weight. The minor other-AGN subtypes are pooled and contribute their top fraction (too few individually for a per-class fraction). Uncertain blazars (BCU) are kept down to the weight of the faintest selected blazar (too numerous to fraction). The shared fraction is solved so the total lands on the budget. The selection is whole-sky, so the north–south balance follows from the sensitivity curve, not from a forced hemisphere split.

One caveat attaches to the extragalactic selection. Blazars are strongly variable, so the gamma-ray flux recorded in any catalog reflects the particular epoch of its observations rather than a steady output. Matching that epoch to the neutrino data set would require a dedicated reanalysis of the gamma-ray observations and is not attempted here. The catalog flux is used as the best available time-averaged proxy for the source's emission.

The resulting catalog contains 110 sources (96 northern, 14 southern): 61 BLL, 28 FSRQ, 8 SBG, 4 GAL, 3 BCU, 2 RDG, 2 NLSy1, and the 2 galactic sources. All are tested under the same point source hypothesis with a free power-law spectral index as the all-sky scan (see the per-pixel hypothesis testing, Section 10.2). Figure 13.1 shows the sky positions of all 110 sources. The complete per-source listing, with coordinates and selection weights, is collected in Table A.2, and the per-source results for the sources that return a positive fit are reported in the results table of Chapter 14.

The catalog's closest pair, B2 1215+30 and PG 1218+304 (both at $\delta \approx 30.1°$), lie only 0.76° apart. Should both cross the evidence threshold, attributing the signal to one rather than the other would be ill-posed: two sources within each other's point-spread function cannot be cleanly separated, a resolution limit shared by every method rather than a defect of this one. At this separation the pair is borderline for the Lightning Tracks sample: a hard-spectrum source, whose higher-energy events sharpen the point-spread function, would likely be resolvable, whereas a soft-spectrum one likely would not.

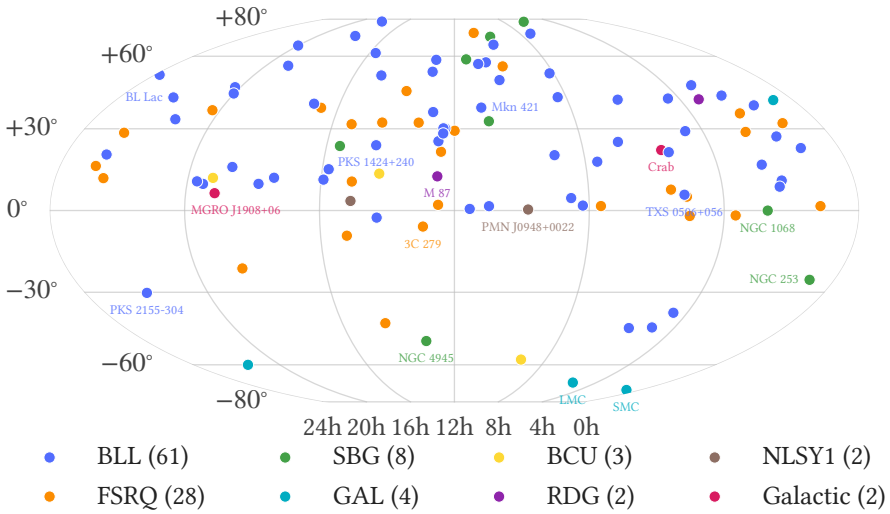

Figure 13.1: Sky positions of the 110 catalog sources in equatorial coordinates, colored by source category. Notable sources are annotated.

14

Projected Results

With the statistical machinery now in place and the sources we want to look for defined, what remains is to fix the exact procedure for examining the data, and then, finally, to look and see what nature has in store. This is what the whole effort has been building toward. In this version of the dissertation the real-data results are not yet included: the unblinding awaits the collaboration review described below.

14.1 The unblinding protocol

The unblinding procedure is fixed before the data are revealed. We perform two searches: an all-sky scan of the combined LT + DNNC sample, as developed in Chapter 10, and a catalog search applying the same point-source method (Chapter 11) at the positions of the source list defined in Section 13.2. The catalog search applies a single global trial correction across all 110 sources, without a hemisphere split, while the all-sky scan splits at the muon horizon ($\delta = -5°$). Both convert the scan maximum to a post-trial p-value through the empirical trial correction (Section 10.3), Gumbel-extrapolated in the deep tail beyond the reach of the 10^5 background scans, since the 5σ discovery threshold lies well past their empirical floor. Every source is reported with the same deliverables (a bias-corrected point estimate and the Feldman–Cousins interval on $\Phi(E_{\mathrm{pivot}})$ and γ), regardless of whether it rejects the background-only hypothesis, with rejection flagged separately. Where more than one source crosses threshold, post-trial p-values follow the step-down procedure (Section 11.2).

Fixing the procedure in advance is what makes the result trustworthy, and one feedback loop makes the requirement concrete: the background top-up. If any observed test statistic (for a sky-scan pixel or for a catalog source) has fewer than 10 sampled background trials above it, the per-ring calibration at that declination is topped up with further trials until at least ten lie above the observed value.

One notable exception to what must be fixed in advance is the set of Feldman–Cousins confidence regions. Because they do not affect the unblinded significances (only the parameter estimation) and give valid coverage under whichever signal model determines their thresholds (baseline or systematics-perturbed alike), they can be produced post hoc for whatever choice we report: all of them are valid, and there is no look-elsewhere effect in choosing among them. What cannot be altered are the measurement-side PDFs, which fix the unblinded p-values and must stay frozen. The signal side, by contrast, is free. We could equally report several

flux estimates under different systematic assumptions. For now we report only the baseline. As set out in the Feldman–Cousins systematics treatment (Section 12.7), the plan is to also produce wider coverage regions that fold in per-trial systematic perturbations.

14.2 Collaboration review and unblinding

A blind analysis earns its credibility only if the decision to open the box is made for the right reasons, and seen to be made for the right reasons. The pre-specified protocol of the previous section fixes what will be reported before the data are examined. Before that protocol is executed, the analysis itself is checked through a structured internal review, carried out in stages by progressively wider and more independent groups. The process is worth describing in its own right, because it functions as a form of peer review conducted entirely before the work is ever submitted to a journal.

The analysis begins inside a working group, a standing subset of the collaboration organized around one class of measurements, in this case the search for astrophysical neutrino sources. The analyzer presents the analysis to the working group during its initial design phase, while the design can still be shaped by the group's scrutiny.

As the analysis matures, it is handed to dedicated reviewers. The working group first appoints an internal referee, a member tasked with examining the analysis in depth. Once that review is well advanced, a second reviewer is added from outside the working group, usually from an unrelated part of the collaboration. This outside reviewer is what makes the review genuinely independent—scrutiny by someone with no stake in the result, and no investment in the choices that produced it, is far more likely to surface a mistaken assumption or an unexamined corner than review from within.

Independent review is only meaningful if the analysis can actually be checked, so everything needed to reproduce it is recorded: the procedure itself, the exact versions of the software, and the datasets used. Designated technical reviewers sign off on the data handling and the code. The standard is reproducibility in the literal sense: an independent person, working from the documentation alone, could repeat the analysis and recover the same numbers.

When the reviewers and the working group are satisfied, the analyzer formally requests permission to unblind. The request is presented to the full collaboration at a regular review meeting, and presenting it opens a fixed comment window of two weeks,[337] during which any member of the collaboration may raise an objection or ask for an additional check. There are currently 420 collaborators on IceCube's author list,[338] all of whom will be considered co-authors of this work. The request is granted only after every comment has been addressed to the collaboration's satisfaction. Only at that point is the box opened. Unblinding is therefore not an act performed by the analyzer in private but a decision ratified by the whole

[337] At the time of writing, the first of those two weeks has passed for the analyses presented here, and permission to unblind is expected to be granted on July 2, 2026, five days before this dissertation is defended.

[338] As of June 23, 2026; IceCube author list, `authorlist.icecube.wisc.edu`.

collaboration, with the analysis having survived every objection anyone cared to raise.

Opening the box does not end the review. The realized result is put through the cross-checks fixed in advance, documented, and presented again, first to the working group and then to the collaboration, so that the final numbers receive the same scrutiny as the method that produced them. Publication is a further stage. Before a manuscript is drafted, the collaboration must approve an outline of the intended paper: the journal it will target and why, the handful of conclusions a reader should take away, the figures, and the central physics message. That outline opens its own comment window, and only once it is approved is the analyzer cleared to begin writing, with a dedicated publication committee assisting from there.

The cumulative effect is that, by the time a result reaches an external journal, it has already passed through several independent layers of review: an internal referee, an independent reviewer from outside the working group, a reproducibility check of the code and data, and at least two collaboration-wide comment periods in which any member could object. It is reasonable to regard this as a peer review carried out before submission.

14.3 Projected results from mock unblindings

What follows are projections from *mock unblindings*, in which a known signal is injected into a background realization and the full analysis is run on the resulting synthetic data.

The mock is built source by source. For each catalog source, the signal strength measured by the published Northern Tracks point-source analysis[339] is converted to the expectation for the combined sample through the ratio of acceptances, and that number of signal events is injected, without Poisson fluctuation, on top of an RA-randomized realization of the data. The full catalog analysis is then run on the injected sample. To average over the arbitrary choice of background realization and of which simulated events are drawn, the procedure is repeated over 10,000 independent realizations, each with its own randomization and signal draw, and the per-source results below are reported as the median across realizations, with the [16, 84] percentile spread as an uncertainty. The all-sky search is mocked in the same spirit, with a diffuse Galactic-plane template and NGC 1068 injected into each background realization. That injection is described with the all-sky result (Section 14.3).

[339] IceCube Collaboration 2026a, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube".

The projection injects each source at the best-fit n_s its precedent analysis returned and treats that strength as genuine signal. This is an optimistic assumption. The precedent per-source results establish individually significant evidence only for NGC 1068[340]; no other catalog source reaches that level, so the remaining best-fit n_s are excesses consistent with background. The projected significances, both the per-source results (Table 14.1) and the population significances of the binomial

[340] IceCube Collaboration 2026a.

tests, are therefore optimistic, crediting the catalog with signal that the precedent data do not establish.

Remark 14.1. This preview is not strictly blind. Because the injected signal strengths are taken from a prior Northern Tracks measurement, and roughly 88% of Northern Tracks events also enter the present LT sample (Figure 9.1), the mock is correlated with what the real unblinding will show. The analysis is final, and the source catalog was fixed, by a pre-registered selection algorithm, long before any significant excess was observed. What such a preview must never do is influence source *selection*: choosing or reweighting catalog targets in light of their previewed significance would be post-hoc selection and would invalidate the trial correction. That is not done here: the catalog is fixed independently of any preview.

The mock results inherit the simulation's model of the detector, its effective area and its angular and energy response. They are projections of the analysis's power at the assumed source fluxes, not coverage-bearing forecasts of the true fluxes, and should be read in that light throughout this section.

The all-sky mock results

The all-sky scan was run as a mock unblinding over many background realizations. One example is shown here. Each realization combines an independent right-ascension randomization of the data, for the background, with an independent Poisson draw of the injected signal from the assumed fluxes: a diffuse Galactic-plane template at the flux measured by the Galactic-plane observation[341] and NGC 1068 at the best-fit flux of the Northern Tracks analysis ($\Phi_0 = 4.7 \times 10^{-11}$ TeV^{-1} cm^{-2} s^{-1} at 1 TeV, $\gamma = 3.4$)[342]. The Galactic-plane emission is injected from each of three spatial templates in turn (the π^0 template and the KRA$_\gamma$ template[343] at cutoffs of 5 and 50 PeV), and the example shown uses the KRA$_\gamma$-50 template, injected at 0.37 of its model flux, the best-fitting normalization of the Galactic-plane analysis[344]. The combined sample is scanned at $N_{\text{side}} = 512$. A pre-trial p-value is computed at every pixel, and the hottest pixel in each hemisphere (split at $\delta = -5°$) is converted to a post-trial significance through the empirical trial correction, with the two-hemisphere correction of the unblinding protocol (Section 14.1).

Figure 14.1 shows the resulting mock sky map for that realization: the injected NGC 1068 is recovered as the hottest northern spot, and the injected Galactic-plane template is visible along the plane. There is very substantial variation between realizations that is not captured by this single example.

[341] IceCube Collaboration 2023, "Observation of high-energy neutrinos from the Galactic plane", Table 1.

[342] IceCube Collaboration 2026a, Sec. 4.2.

[343] Gaggero et al. 2015, "The gamma-ray and neutrino sky: A consistent picture of Fermi-LAT, Milagro, and IceCube results".

[344] IceCube Collaboration 2023, Table 1.

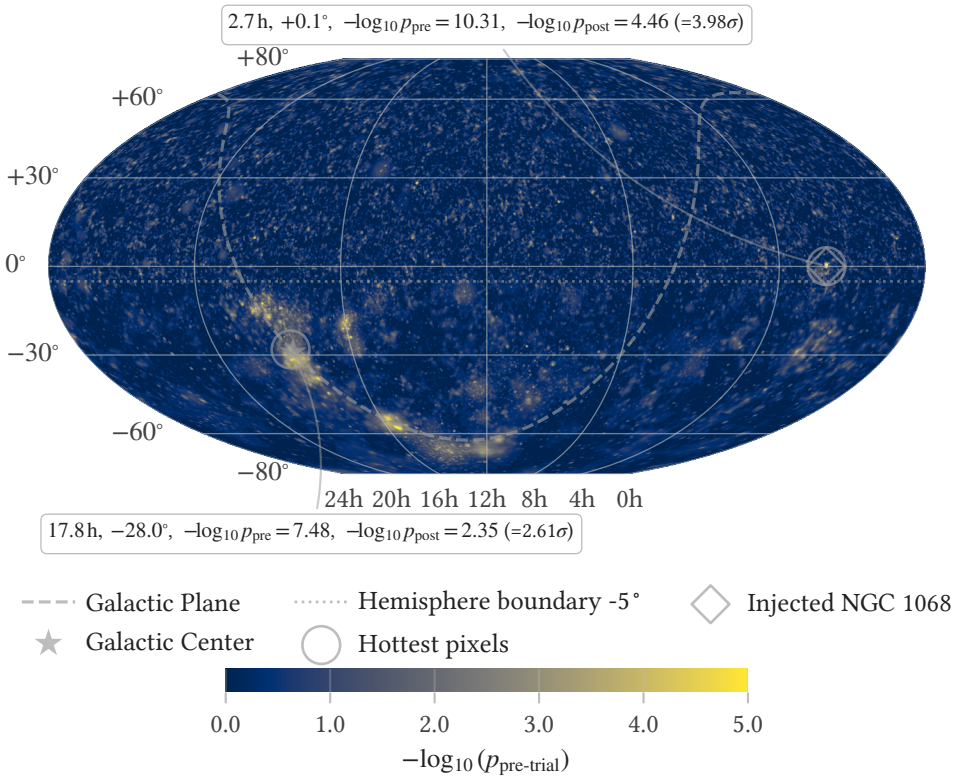

Figure 14.1: Mock all-sky pre-trial p-value sky map (projected, not real): the combined-sample scan on one background realization with a diffuse Galactic-plane (KRA_γ-50) template and NGC 1068 injected (Section 14.3). The hottest spots are annotated with their pre- and post-trial significances.

[345] IceCube Collaboration 2026a, Sec. 4.2.

The mock catalog results

The projected catalog result for the benchmark source, NGC 1068, is illustrated by its noise-free Feldman–Cousins diagnostics at the best-fit flux of the Northern Tracks analysis ($\gamma = 3.4$, $\Phi_0 = 4.7 \times 10^{-11}$ TeV^{-1} cm^{-2} s^{-1} at 1 TeV), converted to the combined-sample expectation through the per-declination signal acceptance[345]. Figure 14.2 shows the Feldman–Cousins confidence region. Figure 14.3 shows the pivot-energy diagnostic, where the projection width on $\Phi(E)$ is narrowest at the pivot energy $E_{\text{pivot}} \approx 1.4$ TeV. Both are projections computed from simulation before unblinding. The realized region and the observed parameter estimates follow at unblinding.

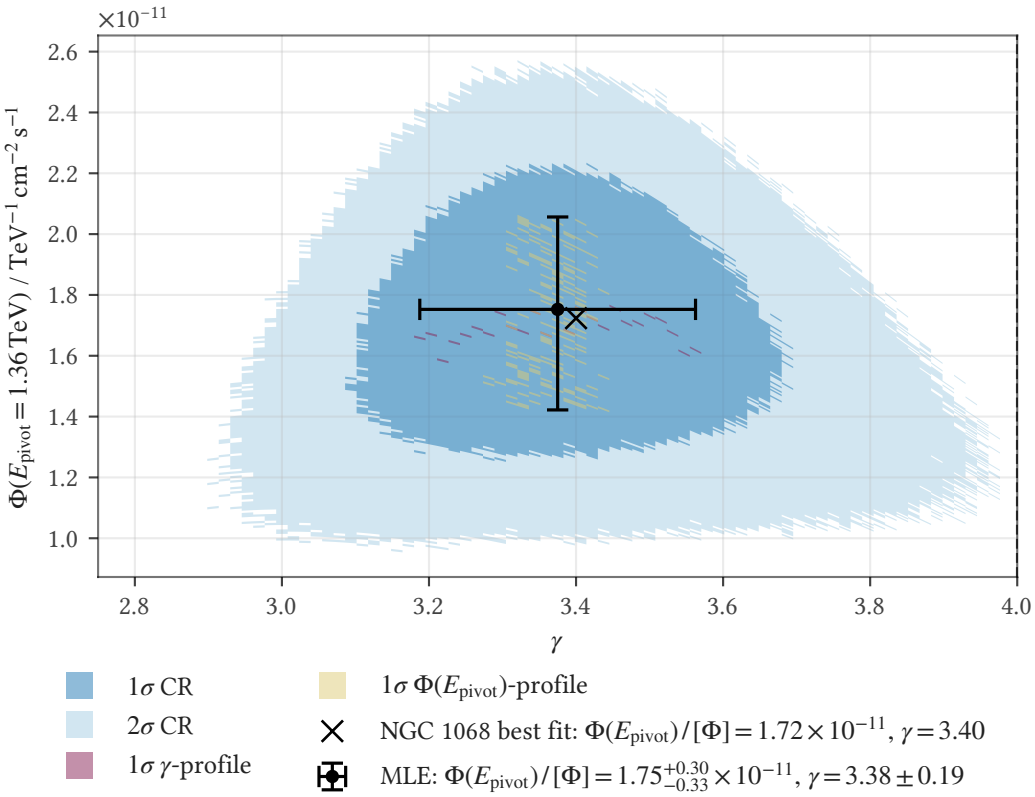

Figure 14.2: Noise-free Feldman–Cousins 1σ and 2σ confidence regions for NGC 1068 at the best-fit flux of the Northern Tracks analysis ($\sin\delta = 0$). Horizontal axis: γ; vertical axis: Φ at the pivot energy. The injected truth is marked with an x and the calibrated MLE $(\tilde{n}_s, \tilde{\gamma})$ with a filled circle, whose error bars are the 1D profile-FC intervals (Berger-Boos threshold) on γ and $\Phi(E_{\text{pivot}})$.

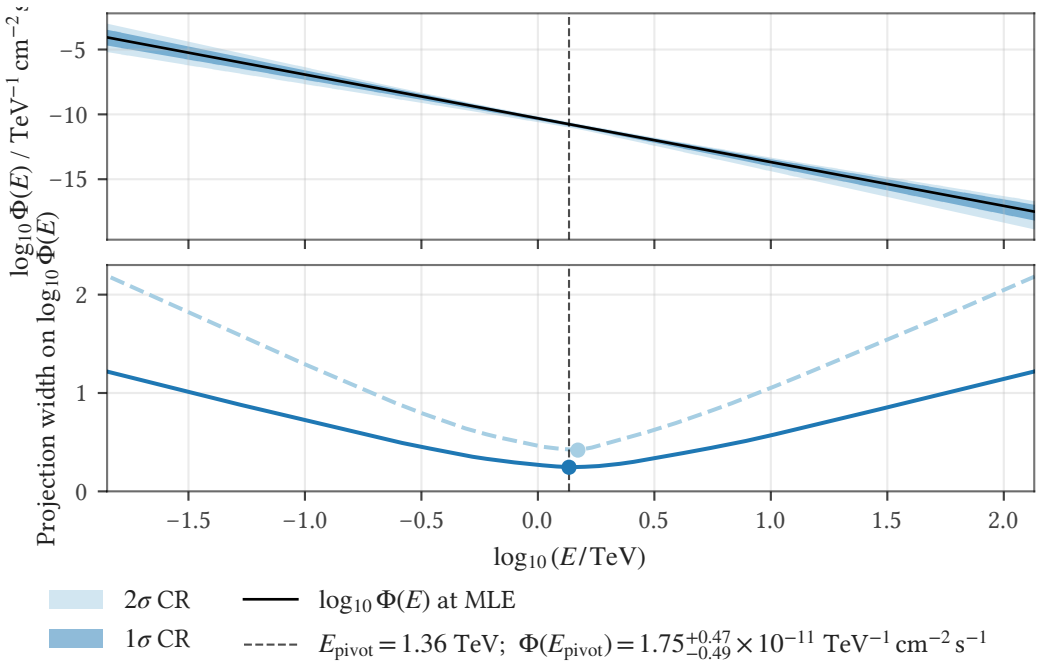

Figure 14.3: Noise-free pivot-energy diagnostic for NGC 1068 at the best-fit flux of the Northern Tracks analysis ($\sin\delta = 0$). Top panel: the $\log_{10}\Phi(E)$ envelope (shaded bands = [min, max] over cells inside the 1σ and 2σ 2D-FC regions at each E), with the central trace at the calibrated MLE. Bottom panel: the projection width Width$[\log_{10}\Phi(E)]$ of the 1σ and 2σ regions at each test energy, minimized at $\log_{10} E_{\mathrm{pivot}}$.

The catalog analysis is run on the mock-injected sample described in Section 14.3, evaluating all catalog positions at their exact coordinates. Table 14.1 lists the sources with a positive median catalog post-trial significance, alongside their median pre-trial local significance. The remaining catalog sources are consistent with background after the catalog trial correction. The full source list is tabulated in the appendix (Section 13.2). We do not report recovered signal counts or spectral indices here: without the per-source simulation grids the recovered estimates are not bias-corrected, so only the significances are quoted.

NGC 1068 is the most significant catalog source, with a median catalog post-trial significance of 5.63σ [4.38, 6.86] (median pre-trial local significance 6.39σ). PKS 1424+240 follows at 4.04σ [2.58, 5.36], and TXS 0506+056 at 1.28σ [−2.54, 3.28], whose interval dips below zero in the realizations where it underfluctuates the background. These three are the only catalog sources with a positive median post-trial significance.

Table 14.1: Per-source results from the catalog mock unblinding (Section 14.3), over an ensemble of 10,000 mock realizations. For each source we quote the median pre-trial local significance and the median catalog post-trial significance.

Source	Local σ	Post-trial σ
NGC 1068	6.39	5.63 [4.38, 6.86]
PKS 1424+240	5.02	4.04 [2.58, 5.36]
TXS 0506+056	3.05	1.28 [−2.54, 3.28]

A population (binomial) test complements the single-source step-down (Section 11.3). On the projected catalog (an ensemble of 10,000 fixed-count realizations, each statistic quoted as the ensemble median with its [16, 84] percentile interval), the full-catalog binomial reaches a local significance of 7.13σ [6.24, 8.09]. Trial-correcting for the rank scan, the global significance is reported under two analytic conventions: a conservative Sidak correction[346] over the 110 tested ranks gives 6.45σ [5.46, 7.50], and the deep-tail effective trial factor ($T \approx 30$) gives 6.64σ [5.68, 7.67]. The available 10^5 background scans are too few to calibrate the global empirically at this significance, where the empirical p-value floors, so the direct calibration yields only a one-sided Clopper–Pearson lower bound[347] of $\geq 4.01\sigma$ (floored in 9,870 of 10,000 realizations). The analytic conventions place the true global near 6.4–6.6σ.

[346] Šidák 1967, "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions".

[347] Clopper and Pearson 1934, "The use of confidence or fiducial limits illustrated in the case of the binomial".

Because the full-catalog test includes the individually significant sources, conflating an individual detection with a population result, the *residual* binomial, restricted to the sources the step-down does not individually reject (Section 11.3), isolates the sub-threshold population. Its local significance is 5.20σ [4.29, 6.12], its global 4.24σ under Sidak ($T = 108$) and 4.58σ under the effective trial factor ($T \approx 23$), with an empirical Clopper–Pearson lower bound of $\geq 3.90\sigma$ (floored in 4,891 of 10,000 realizations). We quote the residual projection only for comparison; it is not included in the unblinding plan (Section 14.1).

The most significant sub-population is small: the rank scan selects $k^* = 1$ in 17.8% of realizations, $k^* = 2$ in 30.8%, $k^* = 3$ in 11.1%, $k^* = 4$ in 6.0%, and $k^* \geq 5$ in 34.4%. NGC 1068 is the rank-1 source in 79.3% of realizations. The sub-populations that most often set the minimum are dominated by NGC 1068 and PKS 1424+240 (Table 14.2).

Table 14.2: The ten sub-populations that most often set the full-catalog binomial minimum over the projected mock ensemble (10,000 realizations). For each, k^* is the rank at the minimum, the share is the fraction of realizations in which that exact set sets it, and the local significance is the ensemble median with its [16, 84] percentile interval. All of these projected significances inherit the optimism of the mock injection, which treats each source's precedent best-fit n_s as genuine signal (Section 14.3).

sub-population setting the minimum	k^*	share	local σ
NGC 1068, PKS 1424+240	2	28.9%	7.40 [6.44, 8.38]
NGC 1068	1	17.2%	6.96 [6.14, 7.79]
NGC 1068, PKS 1424+240, TXS 0506+056	3	5.2%	7.57 [6.64, 8.63]
NGC 1068, TXS 0506+056	2	1.3%	7.47 [6.49, 8.44]
NGC 1068, OJ 014, PKS 1424+240	3	0.7%	6.80 [6.20, 7.76]
NGC 1068, PKS 1424+240, PKS 1502+106	3	0.6%	7.45 [6.36, 8.57]
MG2 J201534+3710, NGC 1068, PKS 1424+240	3	0.6%	7.26 [6.28, 8.18]
3C 454.3, NGC 1068, PKS 1424+240	3	0.6%	7.09 [6.65, 7.57]
PKS 1424+240	1	0.6%	6.49 [6.00, 7.36]
B3 0609+413, NGC 1068, PKS 1424+240	3	0.5%	7.27 [6.20, 7.88]

15

Conclusions and Outlook

This dissertation began with a single goal: to improve IceCube's sensitivity to astrophysical neutrino point sources. The route we took is the one argued in Chapter 1. With the detector and the physics questions fixed, sensitivity is advanced by technical work on the event selection and on the statistical machinery that turns events into measurements. We have followed that route to its end. The result is Lightning Tracks, a new track sample we built from a single processing pipeline and optimized directly for point-source sensitivity. It splits into disjoint starting and through-going components (Chapter 4), each selected with modern machine-learning methods. On its own it is the most sensitive point-source sample assembled to date across almost all declinations; combined with the DNN Cascade sample it reaches the best sensitivity available anywhere in the sky, improving on the best previous samples by up to a factor of four in the southern sky, where it is the first competitive track selection, and by up to 30 percent in the north, improvements that come largely from an energy threshold we pushed down to roughly 100 GeV (Chapter 9). Around the sample we built the analysis framework that turns its events into measurements, from the calibration of the per-event angular uncertainties through to the Feldman–Cousins construction for parameter estimation (Chapter 7 through Chapter 12). With it we carried out two searches: an all-sky scan and a search over a catalog of candidate sources selected by their gamma-ray emission (Chapter 13).

None of this came for free. The selection processing and sensitivity optimization, together with the tens of billions of background trials and roughly 300 thousand sky scans behind these searches, consumed more than 54 million CPU-hours on Michigan State University's High-Performance Computing Center, a substantial economic and environmental cost that should be acknowledged. The figure is still rising, as the Feldman–Cousins simulation grids for per-source parameter estimation are being built.

We cannot yet say what those searches found. The analysis is finished and its procedure was fixed before the data were examined, but the box is still closed. The real-data results remain sealed pending the collaboration review described in Chapter 14, and drawing definitive conclusions without them would be premature. The conclusions that follow are written to be revisited: this chapter and Chapter 14 will both be updated substantially once the data are unblinded.

15.1 Lightning Tracks as a general-purpose sample

The searches of Chapter 13 are the first use of Lightning Tracks, and likely not the most consequential one. The sample is the most sensitive instrument for neutrino point-source searches available today, and it was built to be used: it was released to the whole IceCube collaboration. Its adoption has been quick. At the time of writing, essentially every point-source analysis under development in IceCube already runs on Lightning Tracks, for the simple reason that the most sensitive sample is the rational choice for almost any point-source application.

We expect this to hold well past the present moment. Once the inertia of established pipelines is overcome, most of IceCube's point-source work in the coming years will be built on this sample. That, far more than the handful of analyses any single analyzer can carry out by themselves, is the lasting contribution of this work, and what scientific collaboration is all about.

We can hope for more still. The principles behind the selection—filtering taken to the extreme, sensitivity over purity as the objective, empirical optimization of the cuts, and a single pipeline resolved into topology-matched components—are set out in Chapter 1 and developed through Part I, documented there both as a record of what we did here and as a template others can build on. The neutrino telescopes now being designed will each need event selections of their own; should the methods presented here help shape those of experiments such as P-ONE[348] and, if it is built, IceCube-Gen2,[349] the contribution of this work may outlast the sample itself.

[348] Agostini et al. 2020, "The Pacific Ocean Neutrino Experiment".

[349] IceCube-Gen2 Collaboration 2021, "IceCube-Gen2: the window to the extreme Universe".

15.2 Outlook

No analysis is ever truly finished. There is always more that could be done. What follows are some of the possibilities for this work, several of them already planned.

Further sample improvements

The most immediate extension is to the sample itself. Lightning Tracks admits only track-like events, but the same pipeline, with its topology classifier widened to accept cascades, would select every event topology at once—Lightning Events rather than Lightning Tracks. A cascade component is not expected to add much over the DNN Cascade sample, yet a single selection spanning all flavors and topologies would be worth having in its own right: it unifies the processing and makes future additions of livetime straightforward. Nor are the cascade gains a foregone conclusion. The improvement over Northern Tracks was itself unexpected, and the DNN Cascade sample never went through the declination-dependent, physics-weighted sensitivity optimization of Chapter 5, so there may be more to extract there as well.

A second target is the energy reconstruction. The selection currently takes its energy estimate from MuEX, which a more advanced reconstruction would

replace. As Chapter 9 shows by decomposing the signal likelihood into its spatial and energy parts, the energy term contributes little to point-source detection power (Section 9.12), so a better estimate is not expected to improve it. It could still help parameter estimation for hard sources, where the spectrum carries more of the weight.

Methodological improvements

The statistical treatment can be refined independently of the sample. The per-event energy density could be estimated by kernel-density methods,[350] and the angular-uncertainty calibration could be made to depend on the assumed source spectrum, both sharpening the per-event likelihood.

The time-window selection points beyond a simple re-tuning. Re-optimizing the cuts for the shorter windows that transient searches require is trivial, since the selection machinery already exists; the more interesting step is to apply the window cut dynamically within the analysis itself, perhaps even as a free parameter in the likelihood, rather than fixing it before the sample enters the analysis.

A further direction returns to the factorized signal density of Chapter 9, which treats the spatial and energy responses as independent. A unified posterior over direction and energy, an idea due to Chiara Bellenghi (personal communication) and already raised in Section 8.1, would infer the two jointly. As Chapter 9 notes, the larger gain would be for the spatial model, while the joint treatment would also supply the uncertainty the energy estimate now lacks.

The obstacle is computational. The signal density enters the likelihood for every event, and the test statistic must be sampled from a large number of background realizations under the null, so evaluating a normalizing-flow model per event is almost certainly infeasible in practice. A workable version would need a parameterization cheap enough to sit inside that loop. It would have to be expressed in observable energy, because the mapping from true neutrino energy to observed lepton energy depends on the assumed spectrum (Section 8.3), the same dependence that drives the pull correction (Section 8.4). It may be possible to factor that dependence out as a parameterization of its own within the likelihood. The direction is worth pursuing, but the first step is a toy Monte Carlo study of how much there is to gain, since carrying the full method into practice would take considerable effort.

A more radical possibility still takes the empirical Feldman–Cousins construction of Chapter 12 to its extreme. That construction already replaces the model likelihood-ratio ordering with one built from the empirical sampling density of the maximum-likelihood estimate, computed from simulation, though the estimate it ranks still comes from the actual likelihood. The likelihood-free frequentist inference program[351] goes further, dispensing with the parametric likelihood altogether, replacing it with an advanced probabilistic model trained on simulation while retaining frequentist coverage by construction. This is a far more drastic step than anything above, and would demand correspondingly extreme scrutiny.

[350] A kernel-density implementation was developed while investigating the energy-PDF pathologies in the context of signal subtraction (Section 9.7) but was not adopted for this analysis, set aside primarily to avoid the delay that taking a new method through collaboration review would incur.

[351] Dalmasso et al. 2024, "Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning for Reliable Simulator-Based Inference", Cranmer, Brehmer, and Louppe 2020, "The frontier of simulation-based inference".

Joint-constraint analysis of Galactic-plane point sources and diffuse emission

[352] IceCube Collaboration 2023, "Observation of high-energy neutrinos from the Galactic plane".

The DNN Cascade observation of Galactic-plane neutrino emission rejected the isotropic-background null against diffuse-template alternatives,[352] but those alternatives carried no individual-source component, so the morphology of the emission (a truly diffuse field versus a population of discrete sources) is left unresolved. One search we are planning addresses that question directly: a joint-constraint analysis that fits the Galactic plane's diffuse emission and its point-source population together.

Resolving the morphology calls for a composite hypothesis in which both signal components enter a single likelihood alongside the background. Generalizing the per-source mixture of Chapter 9 to two signal terms,

$$\begin{aligned}\mathcal{L}(x;\theta_{\text{diff}},\theta_{\text{src}}) = \prod_{i=1}^{N}\Bigg[&\frac{n_{\text{diff}}}{N}\,\mathcal{S}_i^{\text{diff}}(\theta_{\text{diff}}) \\ &+ \frac{n_{\text{src}}}{N}\,\mathcal{S}_i^{\text{src}}(\theta_{\text{src}}) \\ &+ \frac{N-n_{\text{diff}}-n_{\text{src}}}{N}\,\mathcal{B}_i\Bigg],\end{aligned} \tag{15.1}$$

where $\mathcal{S}^{\text{diff}}$ and $\mathcal{S}^{\text{src}}$ are the diffuse and source densities and n_{diff}, n_{src} their event counts, tied to the data through the shared total N.

The source term cannot afford a free parameter for every Galactic source: a Feldman–Cousins construction requires a simulation grid over the joint parameter space, and that grid stays tractable only for a handful of parameters. One idea is to collapse the population into a few templates. The diffuse component might retain a gamma-ray template, as in the original analysis, while a single template built from the H.E.S.S. Galactic-plane survey[353] could stand in for the source population, its neutrino spectrum fixed by a hadronic pion-decay model so that one template carries both the spatial morphology and the energy dependence of the sources. One or a few such templates, tested jointly with the diffuse component, could keep the construction within reach.

[353] H.E.S.S. Collaboration 2018, "The H.E.S.S. Galactic Plane Survey".

A Feldman–Cousins construction on the joint space $\Theta_{\text{diff}} \times \Theta_{\text{src}}$ then constrains both contributions within a single coverage-preserving procedure, and tests the competing morphologies against one another as likelihood-ratio orderings R evaluated at the relevant null. The test against isotropic background sets both counts to zero, $R(x; n_{\text{diff}} = 0, n_{\text{src}} = 0)$. The source-contribution test fixes the source count to zero with the diffuse left free, $R(x; n_{\text{src}} = 0)$, and the diffuse-contribution test fixes the diffuse count to zero with the sources left free, $R(x; n_{\text{diff}} = 0)$. The source-contribution test profiles the diffuse normalization as a nuisance,

$$R(x; n_{\text{src}} = 0) = \frac{\sup_{\theta_{\text{diff}}} \mathcal{L}(x; \theta_{\text{diff}}, n_{\text{src}} = 0)}{\sup_{\theta_{\text{diff}},\,\theta'_{\text{src}}} \mathcal{L}(x; \theta_{\text{diff}}, \theta'_{\text{src}})}, \tag{15.2}$$

the numerator the best fit with the sources absent and the diffuse free, the denominator the global best fit over both components. Each ranking R is the reciprocal of

the likelihood ratio that defines our significance test statistic, so that $T = -2 \ln R$ recovers the familiar form (Chapter 12) and $\ln R$ is the quantity evaluated in practice. This is the same per-source test Fermi-LAT applies for the 4FGL catalog,[354] and the same profiling NOvA applies to its Feldman–Cousins intervals,[355] now with the diffuse component fitted from the same data rather than fixed. That distinction is the methodological point. Fixing the diffuse flux at a single externally measured best-fit value and injecting it into the background trials makes the null model-dependent and, we argue, breaks strict-frequentist coverage. Profiling it within the same likelihood that sets the per-source significance keeps the construction valid. Beyond the binary tests, the same construction delivers the full Feldman–Cousins confidence region in the $(\theta_{\mathrm{diff}}, \theta_{\mathrm{src}})$ plane, a joint constraint on the two contributions at confidence level $1 - \alpha$. It also answers conditional questions directly: the section of that region at a fixed diffuse flux gives the range of source contributions jointly consistent with the data, a statement of the form "were the diffuse flux at this level, the source contribution would lie within that interval."

[354] Fermi-LAT Collaboration 2020, "Fermi Large Area Telescope Fourth Source Catalog".

[355] NOvA Collaboration 2025, "Monte Carlo method for constructing confidence intervals with unconstrained and constrained nuisance parameters in the NOvA experiment", Sec. 2.2.

X-ray Seyfert full-sky search

The most compelling future analysis is saved for last: a full-sky search for neutrino emission from Seyfert galaxies selected by their X-ray emission—the obscured active galactic nuclei of which NGC 1068 is the prototype. It would extend to the full sky the approach of Chiara Bellenghi's recent Northern Tracks search for the same source class[356], now that Lightning Tracks makes the southern hemisphere competitive.

The source selection would carry over from the northern analysis, where candidate Seyferts are drawn from the BAT AGN Spectroscopic Survey (BASS)[357] and ranked by their hard X-ray flux, the band that penetrates the obscuring material around these sources and traces the accreting corona. That northern selection was defined precisely: "Hence, we select all sources with at least 20% of the X-ray flux (F_X) of NGC 1068. [...] Specifically, we choose to use the hard X-ray component of the spectrum in the 20 to 50 keV band".[358] We would apply the same prescription, extended into the southern sky that Lightning Tracks opens up.

A spectral subtlety bears on how the search should be designed and deserves further study. The Northern Tracks measurement fits NGC 1068 to a soft power law[359], but such a fit may capture only the steep high-energy edge of an AGN-corona spectrum that hardens substantially toward lower energies. The northern population test was run with both signal hypotheses (a free-index power law and a fixed corona model), and the corona model, carrying no free spectral parameters, came out less significant than the power law. A possible reading is that even if the corona picture is the correct one, fixing the spectral shape exactly is too restrictive. A more flexible curved spectrum, such as a log-parabola with one or two free shape parameters—an idea due to Francis Halzen (personal communication)—might recover what the rigid model gives up. Whether it would is hard to say without explicit model studies, which should precede any commitment to a spectral hypothesis.

[356] IceCube Collaboration 2026a, "Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube".

[357] Koss et al. 2022, "BASS XXII: The BASS DR2 AGN Catalog and Data".

[358] Chiara Bellenghi, "The emergence of a new sky: First associations of IceCube high-energy neutrinos with Active Galactic Nuclei" (PhD thesis, Technische Universität München, 2024), p. 117.

[359] IceCube Collaboration 2026a.

Neither search is part of this dissertation. Its main contribution is the sample, and the analysis framework built around it already reaches well beyond the initial plan. These richer, more involved searches are the natural next step.

Supplementary Tables

A

Table A.1: Experimental SLT and TLT datasets used in the Lightning Tracks selection (seasons 2011–2022), with per-season run count, livetime, and event counts at the filter and final-selection levels. The filter and final-selection rates are stable across seasons (filter ∼20.6 mHz for SLT, ∼68.6 mHz for TLT; final ∼1.74 mHz for SLT, ∼5.56 mHz for TLT) and are omitted here.

Season	Runs	Livetime (d)	Sample	Filter events	Final events
2011	1101	332.71	SLT	591 665	49 792
			TLT	1 955 185	158 346
2012	1089	323.58	SLT	579 779	48 441
			TLT	1 967 744	156 786
2013	1189	341.42	SLT	607 606	51 420
			TLT	2 058 149	165 157
2014	1198	360.48	SLT	641 979	54 167
			TLT	2 156 126	173 239
2015	1176	361.87	SLT	641 851	54 273
			TLT	2 146 815	173 994
2016	1116	353.82	SLT	633 230	53 876
			TLT	2 119 401	170 564
2017	1305	406.45	SLT	721 487	60 431
			TLT	2 386 987	194 841
2018	1172	365.39	SLT	649 314	54 719
			TLT	2 165 173	175 753
2019	979	306.51	SLT	558 449	46 892
			TLT	1 885 796	149 282
2020	1111	359.99	SLT	641 663	53 745
			TLT	2 113 250	172 623
2021	1340	427.74	SLT	756 587	64 498
			TLT	2 455 923	203 693
2022	1503	478.79	SLT	851 663	72 127

continued on next page

Table A.1 (continued)

Season	Runs	Livetime (d)	Sample	Filter events	Final events
			TLT	2 788 839	229 416
All	14 279	4418.76	SLT	7 875 273	664 381
			TLT	26 199 388	2 123 694

Table A.2: The complete catalog of 110 candidate neutrino sources, grouped by source class and ordered within each class by selection weight. Positions are equatorial (J2000), in degrees. $F_{>1\,\mathrm{GeV}}$ is the source's Fermi-LAT photon flux above 1 GeV, in units of $10^{-8}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$; $w_\delta = \min(\mathrm{DP})/\mathrm{DP}(\delta)$ is the declination-dependent sensitivity factor, normalized to the most sensitive declination. The selection weight that sets the within-class ranking is their product, $F_{>1\,\mathrm{GeV}}\, w_\delta$ (Section 13.2). Source classes follow the gamma-ray catalogs. The two galactic sources are selected on differential sensitivity rather than this weight and are listed without a flux or weight.

Source	Class	RA (°)	Dec (°)	$F_{>1\,\mathrm{GeV}}$	w_δ
BL Lac	BL Lac	330.69	+42.28	4.41	0.81
Mkn 421	BL Lac	166.12	+38.21	3.33	0.81
PG 1553+113	BL Lac	238.93	+11.19	1.61	0.94
S5 0716+71	BL Lac	110.49	+71.34	2.02	0.54
TXS 0518+211	BL Lac	80.45	+21.21	1.16	0.91
3C 66A	BL Lac	35.67	+43.04	1.21	0.80
PKS 1424+240	BL Lac	216.76	+23.80	0.98	0.90
PKS 0235+164	BL Lac	39.67	+16.62	0.93	0.93
Mkn 501	BL Lac	253.48	+39.76	1.00	0.80
S2 0109+22	BL Lac	18.03	+22.75	0.86	0.91
B2 1215+30	BL Lac	184.48	+30.12	0.87	0.87
TXS 0506+056	BL Lac	77.36	+5.70	0.76	0.95
PKS 0735+17	BL Lac	114.54	+17.71	0.72	0.93
1H 1013+498	BL Lac	153.78	+49.43	0.77	0.76
MG1 J021114+1051	BL Lac	32.81	+10.86	0.60	0.95
1ES 1959+650	BL Lac	300.01	+65.15	0.94	0.60
TXS 0141+268	BL Lac	26.15	+27.09	0.57	0.89
OT 081	BL Lac	267.89	+9.65	0.52	0.90
S4 0814+42	BL Lac	124.57	+42.38	0.58	0.81
4C +01.28	BL Lac	164.61	+1.56	0.48	0.97
OJ 287	BL Lac	133.71	+20.11	0.47	0.93
ON 246	BL Lac	187.57	+25.30	0.47	0.89
7C 2010+4619	BL Lac	303.03	+46.49	0.53	0.78
RGB J2243+203	BL Lac	340.99	+20.36	0.44	0.92
S4 0954+65	BL Lac	149.70	+65.57	0.65	0.60

continued on next page

Table A.2 (continued)

Source	Class	RA (°)	Dec (°)	$F_{>1\,\mathrm{GeV}}$	w_δ
1ES 0647+250	BL Lac	102.70	+25.05	0.44	0.88
PKS 0426-380	BL Lac	67.17	-37.94	1.80	0.20
B3 0133+388	BL Lac	24.14	+39.10	0.45	0.79
B3 0609+413	BL Lac	93.23	+41.37	0.44	0.80
OJ 014	BL Lac	122.86	+1.78	0.36	0.98
PKS 2155-304	BL Lac	329.71	-30.23	1.89	0.19
PG 1246+586	BL Lac	192.09	+58.34	0.51	0.67
1H 1720+117	BL Lac	261.28	+11.88	0.36	0.93
S5 1803+784	BL Lac	270.18	+78.47	0.78	0.43
TXS 1902+556	BL Lac	285.80	+55.68	0.45	0.74
PG 1218+304	BL Lac	185.35	+30.17	0.37	0.87
PKS 0537-441	BL Lac	84.71	-44.09	1.59	0.20
GB6 J1542+6129	BL Lac	235.75	+61.50	0.47	0.64
1ES 0806+524	BL Lac	122.46	+52.31	0.38	0.76
GB6 J1037+5711	BL Lac	159.43	+57.20	0.41	0.71
MG2 J043337+2905	BL Lac	68.41	+29.10	0.33	0.87
NVSS J154824+145702	BL Lac	237.11	+14.96	0.31	0.92
TXS 1452+516	BL Lac	223.63	+51.41	0.36	0.78
PKS 0829+046	BL Lac	127.97	+4.49	0.28	0.98
S4 1749+70	BL Lac	267.15	+70.10	0.47	0.57
W Comae	BL Lac	185.38	+28.24	0.32	0.82
S4 1250+53	BL Lac	193.31	+53.02	0.35	0.74
MG4 J200112+4352	BL Lac	300.30	+43.88	0.32	0.80
NVSS J141826-023336	BL Lac	214.60	-2.56	0.30	0.85
RX J1931.1+0937	BL Lac	292.79	+9.63	0.28	0.90
4C +41.11	BL Lac	65.99	+41.84	0.30	0.80
PKS 0447-439	BL Lac	72.35	-43.84	1.21	0.20
87GB 194024.3+102612	BL Lac	295.70	+10.56	0.25	0.94
B2 2114+33	BL Lac	319.07	+33.66	0.29	0.81
4C +47.08	BL Lac	45.91	+47.27	0.30	0.77
NVSS J184425+154646	BL Lac	281.12	+15.78	0.25	0.91
Ton 116	BL Lac	190.81	+36.46	0.26	0.86
PKS B1130+008	BL Lac	173.19	+0.57	0.23	0.95
TXS 1055+567	BL Lac	164.66	+56.47	0.31	0.72
ZS 0214+083	BL Lac	34.32	+8.62	0.23	0.96
1ES 2344+514	BL Lac	356.77	+51.70	0.28	0.77
3C 454.3	FSRQ	343.49	+16.15	5.85	0.92
CTA 102	FSRQ	338.15	+11.73	3.42	0.93
3C 279	FSRQ	194.04	-5.79	3.61	0.69
Ton 599	FSRQ	179.89	+29.25	1.94	0.87
PKS 1502+106	FSRQ	226.10	+10.49	1.67	0.94
4C +01.02	FSRQ	17.16	+1.59	1.53	0.97
4C +21.35	FSRQ	186.23	+21.38	1.08	0.90
PKS 1424-41	FSRQ	216.98	-42.10	4.84	0.20

continued on next page

Table A.2 (continued)

Source	Class	RA (°)	Dec (°)	$F_{>1\,\mathrm{GeV}}$	w_δ
PKS 1510-089	FSRQ	228.21	-9.10	2.83	0.32
4C +28.07	FSRQ	39.47	+28.80	1.01	0.86
MG2 J201534+3710	FSRQ	303.92	+37.18	1.02	0.82
OP 313	FSRQ	197.63	+32.35	0.94	0.86
4C +38.41	FSRQ	248.82	+38.14	0.98	0.82
4C +55.17	FSRQ	149.42	+55.38	0.99	0.75
B2 1520+31	FSRQ	230.55	+31.74	0.71	0.87
PKS 0736+01	FSRQ	114.82	+1.61	0.60	0.97
B2 0218+357	FSRQ	35.28	+35.94	0.65	0.86
B3 1343+451	FSRQ	206.39	+44.89	0.68	0.80
OQ 334	FSRQ	215.64	+32.39	0.61	0.86
PKS 0336-01	FSRQ	54.88	-1.77	0.55	0.96
S5 1044+71	FSRQ	162.12	+71.73	0.95	0.53
PKS 1830-211	FSRQ	278.41	-21.06	2.78	0.18
4C +31.03	FSRQ	18.22	+32.14	0.56	0.88
S3 0458-02	FSRQ	75.30	-1.97	0.53	0.93
3C 273	FSRQ	187.27	+2.05	0.47	0.99
PKS 0502+049	FSRQ	76.34	+5.00	0.48	0.96
B2 2234+28A	FSRQ	339.10	+28.49	0.53	0.84
OG 050	FSRQ	83.17	+7.55	0.46	0.93
PKS 0903-57	BCU	136.22	-57.58	1.51	0.21
PKS B1413+135	BCU	214.00	+13.34	0.28	0.93
TXS 1913+115	BCU	288.80	+11.83	0.25	0.93
NGC 1275	Radio gal.	49.96	+41.51	3.17	0.80
M 87	Radio gal.	187.71	+12.39	0.16	0.94
PMN J0948+0022	NLSy1	147.22	+0.37	0.17	0.94
PKS 1502+036	NLSy1	226.29	+3.45	0.13	0.98
M 82	Starburst	148.97	+69.67	0.10	0.58
NGC 1068	Starburst	40.66	-0.02	0.05	0.93
NGC 4945	Starburst	196.36	-49.46	0.09	0.19
Arp 220	Starburst	233.72	+23.51	0.02	0.91
NGC 253	Starburst	11.89	-25.30	0.08	0.19
Arp 299	Starburst	172.16	+58.55	0.01	0.67
NGC 2146	Starburst	94.50	+78.33	0.01	0.44
NGC 3424	Starburst	162.92	+32.88	0.01	0.83
LMC	Galaxy	80.00	-68.75	1.30	0.17
SMC	Galaxy	14.50	-72.75	0.28	0.18
M 31	Galaxy	10.81	+41.20	0.03	0.80
NGC 7059	Galaxy	321.89	-60.01	0.01	0.20
Crab	PWN	83.63	+22.01	–	–
MGRO J1908+06	PWN	287.05	+6.26	–	–

Bibliography

Page and locator references in the citations throughout this document refer to the *published* version of each work. Where an arXiv (or other preprint) link is provided for convenient access, its pagination may differ from the published version cited here.

Agostini, Matteo, et al. "The Pacific Ocean Neutrino Experiment". *Nature Astronomy* 4, no. 10 (2020): 913–915. https://doi.org/10.1038/s41550-020-1182-4.

Ahn, Eun-Joo, et al. "Cosmic ray interaction event generator Sibyll 2.1". *Physical Review D* 80 (2009): 094003. https://doi.org/10.1103/PhysRevD.80.094003. arXiv: 0906.4113.

AMANDA Collaboration. "Muon track reconstruction and data selection techniques in AMANDA". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 524, no. 1-3 (2004): 169–194. https://doi.org/10.1016/j.nima.2004.01.065.

— . "Optical properties of deep glacial ice at the South Pole". *Journal of Geophysical Research: Atmospheres* 111, no. D13 (2006): D13203. https://doi.org/10.1029/2005JD006687.

Anderson, T. W., and D. A. Darling. "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes". *The Annals of Mathematical Statistics* 23, no. 2 (1952): 193–212. https://doi.org/10.1214/aoms/1177729437.

Andrews, D. W. K., and X. Cheng. "Estimation and Inference with Weak, Semi-Strong, and Strong Identification". *Econometrica* 80, no. 5 (2012): 2153–2211. https://doi.org/10.3982/ECTA9456.

Antonucci, R. R. J., and J. S. Miller. "Spectropolarimetry and the nature of NGC 1068". *The Astrophysical Journal* 297 (1985): 621. https://doi.org/10.1086/163559.

Argüelles, Carlos A., et al. "Unified atmospheric neutrino passing fractions for large-scale neutrino telescopes". *Journal of Cosmology and Astroparticle Physics* 2018, no. 07 (2018): 047. https://doi.org/10.1088/1475-7516/2018/07/047. arXiv: 1805.11003.

Atwood, W. B., et al. “The Large Area Telescope on the Fermi Gamma-ray Space Telescope Mission”. *The Astrophysical Journal* 697, no. 2 (2009): 1071–1102. https://doi.org/10.1088/0004-637X/697/2/1071. arXiv: 0902.1089.

Auld, G., and I. Papastathopoulos. “Extremal clustering in non-stationary random sequences”. *Extremes* 24, no. 4 (2021): 725–752. https://doi.org/10.1007/s10687-021-00418-2.

Axford, W. I., E. Leer, and G. Skadron. “The acceleration of cosmic rays by shock waves”. In *Proceedings of the 15th International Cosmic Ray Conference (Plovdiv)*, 11:132–137. 1977.

Baade, W., and F. Zwicky. “Cosmic Rays from Super-Novae”. *Proceedings of the National Academy of Sciences* 20, no. 5 (1934): 259–263. https://doi.org/10.1073/pnas.20.5.259.

Barr, G. D., et al. “Uncertainties in Atmospheric Neutrino Fluxes”. *Physical Review D* 74, no. 9 (2006): 094009. https://doi.org/10.1103/PhysRevD.74.094009. arXiv: astro-ph/0611266.

Bell, A. R. “The acceleration of cosmic rays in shock fronts - I”. *Monthly Notices of the Royal Astronomical Society* 182, no. 2 (1978): 147–156. https://doi.org/10.1093/mnras/182.2.147.

Bellenghi, Chiara. “The emergence of a new sky: First associations of IceCube high-energy neutrinos with Active Galactic Nuclei”. PhD thesis, Technische Universität München, 2024.

Berger, R. L., and D. D. Boos. “P Values Maximized Over a Confidence Set for the Nuisance Parameter”. *Journal of the American Statistical Association* 89, no. 427 (1994): 1012–1016. https://doi.org/10.1080/01621459.1994.10476841.

Blandford, R. D., and J. P. Ostriker. “Particle acceleration by astrophysical shocks”. *The Astrophysical Journal* 221 (1978): L29. https://doi.org/10.1086/182658.

Blumenthal, George R., and Robert J. Gould. “Bremsstrahlung, Synchrotron Radiation, and Compton Scattering of High-Energy Electrons Traversing Dilute Gases”. *Reviews of Modern Physics* 42, no. 2 (1970): 237–270. https://doi.org/10.1103/RevModPhys.42.237.

Casella, G., and R. L. Berger. *Statistical Inference*. 2nd ed. Duxbury Press; Chapman / Hall/CRC, 2002. https://doi.org/10.1201/9781003456285.

Chen, S. X. “Beta kernel estimators for density functions”. *Computational Statistics and Data Analysis* 31, no. 2 (1999): 131–145. https://doi.org/10.1016/S0167-9473(99)00010-9.

— . “Probability density function estimation using gamma kernels”. *Annals of the Institute of Statistical Mathematics* 52 (2000): 471–480. https://doi.org/10.1023/A:1004165218295.

Chernoff, H. “On the Distribution of the Likelihood Ratio”. *The Annals of Mathematical Statistics* 25, no. 3 (1954): 573–578. https://doi.org/10.1214/aoms/1177728725.

Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter. “Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)”. In *International Conference on Learning Representations (ICLR)*. 2016. arXiv: 1511.07289.

Clopper, C. J., and E. S. Pearson. “The use of confidence or fiducial limits illustrated in the case of the binomial”. *Biometrika* 26, no. 4 (1934): 404–413. https://doi.org/10.1093/biomet/26.4.404.

Cousins, R. D., and V. L. Highland. “Incorporating systematic uncertainties into an upper limit”. *Nuclear Instruments and Methods in Physics Research A* 320, **numbers** 1–2 (1992): 331–335. https://doi.org/10.1016/0168-9002(92)90794-5.

Cowan, G., et al. “Asymptotic formulae for likelihood-based tests of new physics”. *European Physical Journal C* 71, no. 2 (2011): 1554. https://doi.org/10.1140/epjc/s10052-011-1554-0.

Cranmer, K., J. Brehmer, and G. Louppe. “The frontier of simulation-based inference”. *Proceedings of the National Academy of Sciences* 117, no. 48 (2020): 30055–30062. https://doi.org/10.1073/pnas.1912789117.

Cranmer, K., J. Pavez, and G. Louppe. “Approximating Likelihood Ratios with Calibrated Discriminative Classifiers”, 2015. arXiv: 1506.02169. https://arxiv.org/abs/1506.02169.

Dalmasso, N., R. Izbicki, and A. B. Lee. “Confidence Sets and Hypothesis Testing in a Likelihood-Free Inference Setting”. In *Proceedings of the 37th International Conference on Machine Learning*, 119:2323–2334. PMLR. 2020. arXiv: 2002.10399.

Dalmasso, N., et al. “Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning for Reliable Simulator-Based Inference”. *Electronic Journal of Statistics* 18, no. 2 (2024). https://doi.org/10.1214/24-EJS2307.

Davies, R. B. “Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative”. *Biometrika* 64, no. 2 (1977): 247–254. https://doi.org/10.2307/2335690.

— . “Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternatives”. *Biometrika* 74, no. 1 (1987): 33–43. https://doi.org/10.2307/2336019.

Drury, L. O’C. “An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasmas”. *Reports on Progress in Physics* 46, no. 8 (1983): 973–1027. https://doi.org/10.1088/0034-4885/46/8/002.

Duchon, Jean. “Splines Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces”. In *Constructive Theory of Functions of Several Variables*, 571:85–100. Lecture Notes in Mathematics. Springer, 1977. https://doi.org/10.1007/BFb0086566.

Eldan, Ronen, and Ohad Shamir. “The Power of Depth for Feedforward Neural Networks”. In *29th Annual Conference on Learning Theory (COLT)*, 49:907–940. Proceedings of Machine Learning Research. 2016. arXiv: 1512.03965.

Enberg, Rikard, Mary Hall Reno, and Ina Sarcevic. “Prompt neutrino fluxes from atmospheric charm”. *Physical Review D* 78, no. 4 (2008): 043005. https://doi.org/10.1103/PhysRevD.78.043005. arXiv: 0806.0418.

Fedynitch, Anatoli, et al. “Calculation of conventional and prompt lepton fluxes at very high energy”. *EPJ Web of Conferences* 99 (2015): 08001. https://doi.org/10.1051/epjconf/20159908001. arXiv: 1503.00544.

Feldman, G. J., and R. D. Cousins. “Unified approach to the classical statistical analysis of small signals”. *Physical Review D* 57, no. 7 (1998): 3873–3889. https://doi.org/10.1103/PhysRevD.57.3873.

Fermi, Enrico. “On the Origin of the Cosmic Radiation”. *Physical Review* 75, no. 8 (1949): 1169–1174. https://doi.org/10.1103/PhysRev.75.1169.

Fermi-LAT Collaboration. “Fermi Large Area Telescope Fourth Source Catalog”. *The Astrophysical Journal Supplement Series* 247, no. 1 (2020): 33. https://doi.org/10.3847/1538-4365/ab6bcb.

— . “Fermi-LAT 16-year Source List”, 2026. arXiv: 2602.22148.

— . “GeV Observations of Star-forming Galaxies with Fermi LAT”. *The Astrophysical Journal* 755, no. 2 (2012a): 164. https://doi.org/10.1088/0004-637X/755/2/164. arXiv: 1206.1346.

— . “Radio-loud narrow-line Seyfert 1 as a new class of gamma-ray active galactic nuclei”. *The Astrophysical Journal* 707, no. 2 (2009): L142–L147. https://doi.org/10.1088/0004-637X/707/2/L142.

— . “The Fermi Large Area Telescope on Orbit: Event Classification, Instrument Response Functions, and Calibration”. *The Astrophysical Journal Supplement Series* 203, no. 1 (2012b): 4. https://doi.org/10.1088/0067-0049/203/1/4. arXiv: 1206.1896.

Fisher, R. A. “Dispersion on a Sphere”. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 217, no. 1130 (1953): 295–305. https://doi.org/10.1098/rspa.1953.0064.

Fisher, R. A., and L. H. C. Tippett. “Limiting forms of the frequency distribution of the largest or smallest member of a sample”. *Mathematical Proceedings of the Cambridge Philosophical Society* 24, no. 2 (1928): 180–190. https://doi.org/10.1017/S0305004100015681.

Frank, I., and Ig. Tamm. "Coherent Visible Radiation of Fast Electrons Passing Through Matter". In *I. E. Tamm: Selected Papers*, 29–35. 1937. Reprint of Comptes Rendus (Dokl.) Acad. Sci. URSS 14, 109–114 (1937). Springer Berlin Heidelberg, 1991. ISBN: 9783642746284. https://doi.org/10.1007/978-3-642-74626-0_2.

Gaggero, Daniele, et al. "The gamma-ray and neutrino sky: A consistent picture of Fermi-LAT, Milagro, and IceCube results". *The Astrophysical Journal Letters* 815, no. 2 (2015): L25. https://doi.org/10.1088/2041-8205/815/2/L25. arXiv: 1504.00227.

Gaisser, T. K., F. Halzen, and T. Stanev. "Particle astrophysics with high energy neutrinos". *Physics Reports* 258, no. 3 (1995): 173–236. https://doi.org/10.1016/0370-1573(95)00003-Y. arXiv: hep-ph/9410384.

Gaisser, T. K., T. Stanev, and S. Tilav. "Cosmic ray energy spectrum from measurements of air showers". *Frontiers of Physics* 8, no. 6 (2013): 748–758. https://doi.org/10.1007/s11467-013-0319-7.

Gaisser, Thomas K. "Spectrum of cosmic-ray nucleons, kaon production, and the atmospheric muon charge ratio". *Astroparticle Physics* 35, no. 12 (2012): 801–806. https://doi.org/10.1016/j.astropartphys.2012.02.010. arXiv: 1111.6675.

Gandhi, Raj, et al. "Neutrino interactions at ultrahigh energies". *Physical Review D* 58 (1998): 093009. https://doi.org/10.1103/PhysRevD.58.093009. arXiv: hep-ph/9807264.

Gazizov, Askhat, and Marek Kowalski. "ANIS: High energy neutrino generator for neutrino telescopes". *Computer Physics Communications* 172 (2005): 203–213. https://doi.org/10.1016/j.cpc.2005.03.113. arXiv: astro-ph/0406439.

Ghisellini, G., et al. "The transition between BL Lac objects and flat spectrum radio quasars". *Monthly Notices of the Royal Astronomical Society* 414, no. 3 (2011): 2674–2689. https://doi.org/10.1111/j.1365-2966.2011.18578.x.

Ginzburg, V. L., and S. I. Syrovatskii. *The Origin of Cosmic Rays*. Pergamon Press, 1964.

Glashow, Sheldon L. "Resonant Scattering of Antineutrinos". *Physical Review* 118, no. 1 (1960): 316–317. https://doi.org/10.1103/PhysRev.118.316.

Gnedenko, B. "Sur la distribution limite du terme maximum d'une série aléatoire". *Annals of Mathematics (Second Series)* 44, no. 3 (1943): 423–453. https://doi.org/10.2307/1968974.

Gorski, K. M., et al. "HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere". *The Astrophysical Journal* 622, no. 2 (2005): 759–771. https://doi.org/10.1086/427976.

Greisen, K. "Cosmic Ray Showers". *Annual Review of Nuclear Science* 10, no. 1 (1960): 63–108. https://doi.org/10.1146/annurev.ns.10.120160.000431.

Griffiths, David. *Introduction to Elementary Particles*. 2nd ed. Wiley-VCH, 2008. ISBN: 9783527406012. https://doi.org/10.1002/9783527618460.

H.E.S.S. Collaboration. "Gamma-ray blazar spectra with H.E.S.S. II mono analysis: the case of PKS 2155-304 and PG 1553+113". *Astronomy & Astrophysics* 600 (2017): A89. https://doi.org/10.1051/0004-6361/201629427. arXiv: 1612.05111. https://arxiv.org/abs/1612.05111.

— . "The H.E.S.S. Galactic Plane Survey". *Astronomy & Astrophysics* 612 (2018): A1. https://doi.org/10.1051/0004-6361/201732098. arXiv: 1804.02432.

Halmos, P. R., and L. J. Savage. "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics". *The Annals of Mathematical Statistics* 20, no. 2 (1949): 225–241. https://doi.org/10.1214/aoms/1177730032.

Halzen, Francis, and Dan Hooper. "High-energy neutrino astronomy: the cosmic ray connection". *Reports on Progress in Physics* 65 (2002): 1025–1078. https://doi.org/10.1088/0034-4885/65/7/201. arXiv: astro-ph/0204527.

Hartman, R. C., et al. "Detection of high-energy gamma radiation from quasar 3C 279 by the EGRET telescope on the Compton Gamma Ray Observatory". *The Astrophysical Journal* 385 (1992): L1. https://doi.org/10.1086/186263.

He, Kaiming, et al. "Deep Residual Learning for Image Recognition". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. IEEE, 2016. https://doi.org/10.1109/CVPR.2016.90. arXiv: 1512.03385.

Heck, D., et al. *CORSIKA: A Monte Carlo Code to Simulate Extensive Air Showers*. Report FZKA 6019. Forschungszentrum Karlsruhe, 1998. https://publikationen.bibliothek.kit.edu/270043064.

Herrmann, K., M. Hofert, and J. G. Neslehová. "Limiting Behavior of Maxima under Dependence", 2024. arXiv: 2405.02833. https://arxiv.org/abs/2405.02833.

Holm, S. "A Simple Sequentially Rejective Multiple Test Procedure". *Scandinavian Journal of Statistics* 6, no. 2 (1979): 65–70. https://www.jstor.org/stable/4615733.

Hünnefeld, Mirco. "Observation of high-energy neutrinos from the Milky Way". PhD thesis, TU Dortmund University, 2023. https://doi.org/10.17877/DE290R-24043.

IceCube Collaboration. "A Convolutional Neural Network based Cascade Reconstruction for the IceCube Neutrino Observatory". *Journal of Instrumentation* 16 (2021a): P07041. https://doi.org/10.1088/1748-0221/16/07/P07041. arXiv: 2101.11589.

— . “A muon-track reconstruction exploiting stochastic losses for large-scale Cherenkov detectors”. *Journal of Instrumentation* 16, no. 08 (2021b): P08034. https://doi.org/10.1088/1748-0221/16/08/P08034. arXiv: 2103.16931.

— . “All-sky Neutrino Point-source Search with IceCube Combined Track and Cascade Data”. *The Astrophysical Journal*, 2025. https://doi.org/10.3847/1538-4357/ae113f. arXiv: 2507.07275.

— . “All-sky Search for Time-integrated Neutrino Emission from Astrophysical Sources with 7 yr of IceCube Data”. *The Astrophysical Journal* 835 (2017a): 151. https://doi.org/10.3847/1538-4357/835/2/151. arXiv: 1609.04981.

— . “Calibration and Characterization of the IceCube Photomultiplier Tube”. *Nuclear Instruments and Methods in Physics Research A* 618 (2010): 139–152. https://doi.org/10.1016/j.nima.2010.03.102. arXiv: 1002.2442.

— . “Characterization of the Astrophysical Diffuse Neutrino Flux using Starting Track Events in IceCube”. *Physical Review D* 110 (2024a): 022001. https://doi.org/10.1103/PhysRevD.110.022001. arXiv: 2402.18026.

— . “Cosmic Ray Composition and Energy Spectrum from 1–30 PeV Using the 40-String Configuration of IceTop and IceCube”. *Astroparticle Physics* 42 (2013a): 15–32. https://doi.org/10.1016/j.astropartphys.2012.11.003. arXiv: 1207.3455.

— . “Efficient propagation of systematic uncertainties from calibration to analysis with the SnowStorm method in IceCube”. *Journal of Cosmology and Astroparticle Physics* 2019, no. 10 (2019): 048. https://doi.org/10.1088/1475-7516/2019/10/048. arXiv: 1909.01530.

— . “Evidence for High-Energy Extraterrestrial Neutrinos at the IceCube Detector”. *Science* 342 (2013b): 1242856. https://doi.org/10.1126/science.1242856. arXiv: 1311.5238.

— . “Evidence for neutrino emission from the nearby active galaxy NGC 1068”. *Science* 378, no. 6619 (2022a): 538–543. https://doi.org/10.1126/science.abg3395. arXiv: 2211.09972.

— . “Evidence for Neutrino Emission from X-Ray-bright Active Galactic Nuclei with IceCube”. *The Astrophysical Journal Letters* 1000, no. 1 (2026a): L26. https://doi.org/10.3847/2041-8213/ae4aad. arXiv: 2510.13403.

— . “Improved Characterization of the Astrophysical Muon-Neutrino Flux with 9.5 Years of IceCube Data”. *The Astrophysical Journal* 928 (2022b): 50. https://doi.org/10.3847/1538-4357/ac4d29. arXiv: 2111.10299.

— . “In situ estimation of ice crystal properties at the South Pole using LED calibration data from the IceCube Neutrino Observatory”. *The Cryosphere* 18, no. 1 (2024b): 75–102. https://doi.org/10.5194/tc-18-75-2024.

IceCube Collaboration. “Measurement of South Pole ice transparency with the IceCube LED calibration system”. *Nuclear Instruments and Methods in Physics Research A* 711 (2013c): 73–89. https://doi.org/10.1016/j.nima.2013.01.054. arXiv: 1301.5361.

— . “Neural posterior estimation of the neutrino direction in IceCube using transformer-encoded normalizing flows on the sphere”. In preparation, 2026b.

— . “Neutrino emission from the direction of the blazar TXS 0506+056 prior to the IceCube-170922A alert”. *Science* 361, no. 6398 (2018): 147–151. https://doi.org/10.1126/science.aat2890.

— . “Observation and Characterization of a Cosmic Muon Neutrino Flux from the Northern Hemisphere Using Six Years of IceCube Data”. *The Astrophysical Journal* 833, no. 1 (2016): 3. https://doi.org/10.3847/0004-637X/833/1/3. arXiv: 1607.08006.

— . “Observation of High-Energy Astrophysical Neutrinos in Three Years of IceCube Data”. *Physical Review Letters* 113 (2014): 101101. https://doi.org/10.1103/PhysRevLett.113.101101. arXiv: 1405.5303.

— . “Observation of high-energy neutrinos from the Galactic plane”. *Science* 380, no. 6652 (2023): 1338–1343. https://doi.org/10.1126/science.adc9818. arXiv: 2307.04427.

— . “Physics potential of the IceCube Upgrade for atmospheric neutrino oscillations”. *Physical Review D* 113 (2026c): 072009. https://doi.org/10.1103/nnjw-jp1n. arXiv: 2509.13066.

— . “Testing the Pointing of IceCube Using the Moon Shadow in Cosmic-Ray-Induced Muons”. In *Proceedings of the 37th International Cosmic Ray Conference (ICRC 2021)*. PoS 1087. 2021c. arXiv: 2108.04093.

— . “The IceCube Neutrino Observatory: Instrumentation and Online Systems”. *Journal of Instrumentation* 12, no. 03 (2017b): P03012. https://doi.org/10.1088/1748-0221/12/03/P03012. arXiv: 1612.05093.

— . “The IceCube Realtime Alert System”. *Astroparticle Physics* 92 (2017c): 30–41. https://doi.org/10.1016/j.astropartphys.2017.05.002. arXiv: 1612.06028.

— . “Time-integrated Neutrino Source Searches with 10 years of IceCube Data”. *Physical Review Letters* 124, no. 5 (2020): 051103. https://doi.org/10.1103/PhysRevLett.124.051103. arXiv: 1910.08488.

— . “Time-Integrated Southern-Sky Neutrino Source Searches with 10 Years of IceCube Starting-Track Events at Energies Down to 1 TeV”. *The Astrophysical Journal* 998, no. 1 (2026d): 37. https://doi.org/10.3847/1538-4357/ae2c86. arXiv: 2501.16440.

IceCube Collaboration et al. “Multimessenger observations of a flaring blazar coincident with high-energy neutrino IceCube-170922A”. *Science* 361, no. 6398 (2018). https://doi.org/10.1126/science.aat1378.

IceCube-Gen2 Collaboration. "IceCube-Gen2: the window to the extreme Universe". *Journal of Physics G: Nuclear and Particle Physics* 48, no. 6 (2021): 060501. https://doi.org/10.1088/1361-6471/abbd48.

Ioffe, Sergey, and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 37:448–456. Proceedings of Machine Learning Research. PMLR, 2015. arXiv: 1502.03167.

Jones, M. C. "Simple boundary correction for kernel density estimation". *Statistics and Computing* 3 (1993): 135–146. https://doi.org/10.1007/BF00147776.

King, Ivan. "The Structure of Star Clusters. I. An Empirical Density Law". *The Astronomical Journal* 67 (1962): 471–485. https://doi.org/10.1086/108756.

Kingma, Diederik P., and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In *International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2015. 2015. arXiv: 1412.6980.

Koehne, J.-H., et al. "PROPOSAL: A tool for propagation of charged leptons". *Computer Physics Communications* 184, no. 9 (2013): 2070–2090. https://doi.org/10.1016/j.cpc.2013.04.001.

Koss, Michael J., et al. "BASS XXII: The BASS DR2 AGN Catalog and Data". *The Astrophysical Journal Supplement Series* 261, no. 1 (2022): 2. https://doi.org/10.3847/1538-4365/ac6c05. arXiv: 2207.12432.

Kullback, S., and R. A. Leibler. "On Information and Sufficiency". *The Annals of Mathematical Statistics* 22, no. 1 (1951): 79–86. https://doi.org/10.1214/aoms/1177729694.

Le Cam, Lucien. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer New York, 1986. ISBN: 9781461293699. https://doi.org/10.1007/978-1-4612-4946-7.

Leadbetter, M. R., G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer-Verlag, 1983. https://doi.org/10.1007/978-1-4612-5449-2.

Learned, John G., and Sandip Pakvasa. "Detecting ν_τ oscillations at PeV energies". *Astroparticle Physics* 3 (1995): 267–274. https://doi.org/10.1016/0927-6505(94)00043-3. arXiv: hep-ph/9405296.

Lehmann, E. L., and G. Casella. *Theory of Point Estimation*. 2nd ed. Springer Texts in Statistics. Springer-Verlag, 1998. https://doi.org/10.1007/b98854.

Leshno, Moshe, et al. "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". *Neural Networks* 6, no. 6 (1993): 861–867. https://doi.org/10.1016/S0893-6080(05)80131-5.

LHAASO Collaboration. “The First LHAASO Catalog of Gamma-Ray Sources”. *The Astrophysical Journal Supplement Series* 271, no. 1 (2024): 25. https://doi.org/10.3847/1538-4365/acfd29.

Loeb, Abraham, and Eli Waxman. “The cumulative background of high energy neutrinos from starburst galaxies”. *Journal of Cosmology and Astroparticle Physics* 2006, no. 05 (2006): 003–003. https://doi.org/10.1088/1475-7516/2006/05/003.

Loshchilov, Ilya, and Frank Hutter. “Decoupled Weight Decay Regularization”. In *International Conference on Learning Representations (ICLR)*. 2019. arXiv: 1711.05101.

Majumdar, S. N., A. Pal, and G. Schehr. “Extreme value statistics of correlated random variables: A pedagogical review”. *Physics Reports* 840 (2020): 1–32. https://doi.org/10.1016/j.physrep.2019.10.005.

Mardia, Kanti V., and Peter E. Jupp. *Directional Statistics*. John Wiley & Sons, 2000. https://doi.org/10.1002/9780470316979.

Masserano, L., et al. “Simulator-Based Inference with Waldo: Confidence Regions by Leveraging Prediction Algorithms and Posterior Estimators for Inverse Problems”. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 206:2960–2974. Proceedings of Machine Learning Research. PMLR, 2023. arXiv: 2205.15680. https://arxiv.org/abs/2205.15680.

Mattox, J. R., et al. “The Likelihood Analysis of EGRET Data”. *The Astrophysical Journal* 461 (1996): 396–407. https://doi.org/10.1086/177068.

McElfresh, Duncan, et al. “When Do Neural Nets Outperform Boosted Trees on Tabular Data?” In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*. Curran Associates, 2023. arXiv: 2305.02997.

Moffat, A. F. J. “A Theoretical Investigation of Focal Stellar Images in the Photographic Emulsion and Application to Photographic Photometry”. *Astronomy and Astrophysics* 3 (1969): 455–461.

Mölder, Felix, et al. “Sustainable data analysis with Snakemake”. *F1000Research* 10 (2021): 33. https://doi.org/10.12688/f1000research.29032.2.

Murase, Kohta, Shigeo S. Kimura, and Peter Mészáros. “Hidden Cores of Active Galactic Nuclei as the Origin of Medium-Energy Neutrinos: Critical Tests with the MeV Gamma-Ray Connection”. *Physical Review Letters* 125, no. 1 (2020a). https://doi.org/10.1103/PhysRevLett.125.011101.

— . “Hidden Cores of Active Galactic Nuclei as the Origin of Medium-Energy Neutrinos: Critical Tests with the MeV Gamma-Ray Connection”. *Physical Review Letters* 125, no. 1 (2020b). https://doi.org/10.1103/PhysRevLett.125.011101.

Neyman, J. “Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability”. *Philosophical Transactions of the Royal Society A* 236, no. 767 (1937): 333–380. https://doi.org/10.1098/rsta.1937.0005.

Neyman, J., and E. S. Pearson. “On the Problem of the Most Efficient Tests of Statistical Hypotheses”. *Philosophical Transactions of the Royal Society A* 231 (1933): 289–337. https://doi.org/10.1098/rsta.1933.0009.

NOvA Collaboration. “Monte Carlo method for constructing confidence intervals with unconstrained and constrained nuisance parameters in the NOvA experiment”. *Journal of Instrumentation* 20 (2025): T02001. https://doi.org/10.1088/1748-0221/20/02/T02001. arXiv: 2207.14353.

Particle Data Group. “Review of Particle Physics”. Particle Data Group (RPP); the offline copy is the complete published review (artifact swapped from the standalone Statistics chapter, 2026-06-12). *Physical Review D* 110 (2024): 030001. https://doi.org/10.1103/PhysRevD.110.030001. https://pdg.lbl.gov/2024/download/PhysRevD.110.030001.pdf.

Paszke, Adam, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 8024–8035. Curran Associates, 2019. arXiv: 1912.01703.

Protassov, R., et al. “Statistics, Handle with Care: Detecting Multiple Model Components with the Likelihood Ratio Test”. *The Astrophysical Journal* 571, no. 1 (2002): 545–559. https://doi.org/10.1086/339856.

Romano, Joseph P, and Michael Wolf. “Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing”. *Journal of the American Statistical Association* 100, no. 469 (2005): 94–108. https://doi.org/10.1198/016214504000000539.

Rongen, M., and D. Chirkin. “Advances in IceCube ice modelling and what to expect from the Upgrade”. On behalf of the IceCube Collaboration. VLVnT-2021 proceedings. *Journal of Instrumentation* 16, no. 09 (2021): C09014. https://doi.org/10.1088/1748-0221/16/09/C09014. arXiv: 2108.03291.

Rongen, Martin. “Calibration of the IceCube Neutrino Observatory”. PhD thesis, RWTH Aachen University, 2019. arXiv: 1911.02016.

Rosenblatt, M. “Remarks on a Multivariate Transformation”. *The Annals of Mathematical Statistics* 23, no. 3 (1952): 470–472. https://doi.org/10.1214/aoms/1177729394.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. *Nature* 323, no. 6088 (1986): 533–536. https://doi.org/10.1038/323533a0.

Schöneberg, Sebastian. “The spectrum of atmospheric neutrinos above GeV energies”. URN: urn:nbn:de:hbz:294-52689. PhD thesis, Ruhr-Universität Bochum, 2016. https://hss-opus.ub.ruhr-uni-bochum.de/opus4/frontdoor/index/index/docId/5268.

Schwanekamp, Hendrik, Ramona Hohl, Dmitry Chirkin, Tom Gibbs, Alexander Harnisch, Claudio Kopper, Peter Messmer, Vishal Mehta, Alexander Olivas, Benedikt Riedel, Martin Rongen, David Schultz, and Jakob van Santen. "Accelerating IceCube's Photon Propagation Code with CUDA". *Computing and Software for Big Science* 6, no. 1 (2022): 4. https://doi.org/10.1007/s41781-022-00080-8.

Sclafani, Stephen. "Observation of Neutrinos from the Milky Way Galaxy". PhD thesis, Drexel University, 2023. https://doi.org/10.17918/00001635.

Šidák, Z. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions". *Journal of the American Statistical Association* 62, no. 318 (1967): 626–633. https://doi.org/10.1080/01621459.1967.10482935.

Smirnov, N. "Table for Estimating the Goodness of Fit of Empirical Distributions". *The Annals of Mathematical Statistics* 19, no. 2 (1948): 279–281. https://doi.org/10.1214/aoms/1177730256.

Spiering, Christian. "Towards High-Energy Neutrino Astronomy. A Historical Review". *The European Physical Journal H* 37, no. 3 (2012): 515–565. https://doi.org/10.1140/epjh/e2012-30014-2. arXiv: 1207.4952.

Staiger, D., and J. H. Stock. "Instrumental Variables Regression with Weak Instruments". *Econometrica* 65, no. 3 (1997): 557–586. https://doi.org/10.2307/2171753.

Tegenfeldt, F., and J. Conrad. "On Bayesian Treatment of Systematic Uncertainties in Confidence Interval Calculation". *Nuclear Instruments and Methods in Physics Research A* 539 (2005): 407–413. https://doi.org/10.1016/j.nima.2004.09.037. arXiv: physics/0408039.

Telgarsky, Matus. "Benefits of depth in neural networks". In *29th Annual Conference on Learning Theory (COLT)*, 49:1517–1539. JMLR: Workshop and Conference Proceedings. 2016. arXiv: 1602.04485.

Urry, C. Megan, and Paolo Padovani. "Unified Schemes for Radio-Loud Active Galactic Nuclei". *Publications of the Astronomical Society of the Pacific* 107 (1995): 803. https://doi.org/10.1086/133630.

Vaart, A. W. van der. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. https://doi.org/10.1017/CBO9780511802256.

Vuong, Q. H. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses". *Econometrica* 57, no. 2 (1989): 307–333. https://doi.org/10.2307/1912557.

Wand, M. P., and M. C. Jones. *Kernel Smoothing*. Chapman / Hall/CRC, 1995. https://doi.org/10.1201/b14876.

Wang, Zhen-Jie, et al. “On the Hadronic Origin of High Energy Emission of γ-ray Loud Narrow-Line Seyfert 1 PKS 1502+036”. *The Astrophysical Journal* 942, no. 1 (2023): 51. https://doi.org/10.3847/1538-4357/aca1b9. arXiv: 2211.07070.

Wasserman, L. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York, 2006. https://doi.org/10.1007/0-387-30623-4.

Westfall, Peter H., and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, 1993. ISBN: 0-471-55761-7.

White, H. “Maximum Likelihood Estimation of Misspecified Models”. *Econometrica* 50, no. 1 (1982): 1–25. https://doi.org/10.2307/1912526.

Wilks, S. S. “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses”. *The Annals of Mathematical Statistics* 9, no. 1 (1938): 60–62. https://doi.org/10.1214/aoms/1177732360.