The Analyzer:
There and Back Again

Daniel Cherdack
University of Houston

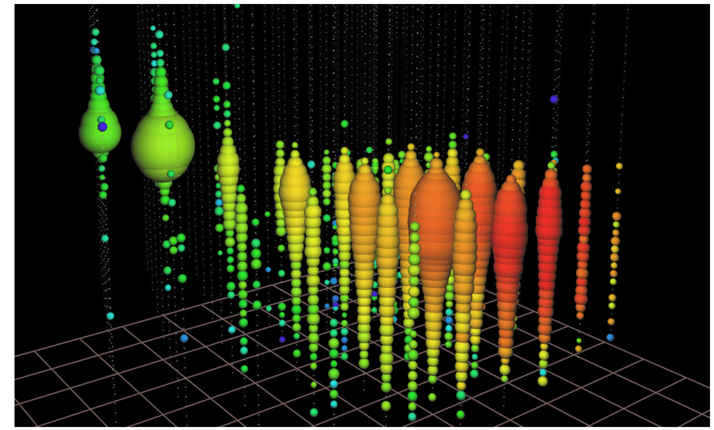
NuInt Satellite School
April 11th – 13th, 2024
Sao Paulo, Brazil

Choosing an Analysis

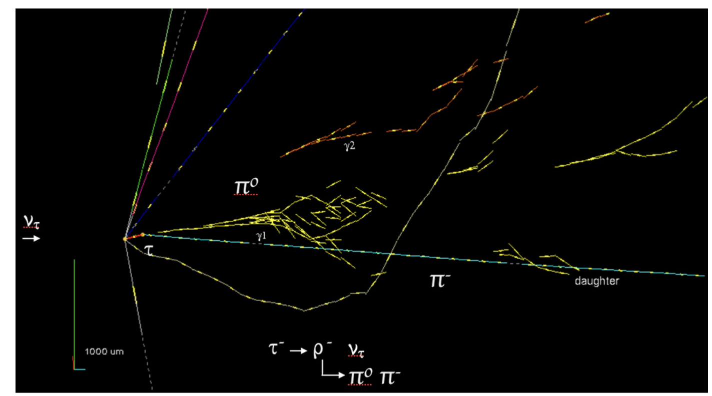
- As an analyzer you need to accomplish a few goals with your analysis
  - Get you PhD
    - Happy adviser
  - Learn something
    - Happy future employer
  - Contribute to you collaboration
    - Happy conveners
  - Contribute to science
    - Happy world
- Considerations
  - Is it interesting?
  - Is it possible?
  - What can you really measure?
  - What can you learn by doing this analysis?
    - Physics
    - Technique
    - Statistics
    - Modeling
    - Detector performance

Getting Started

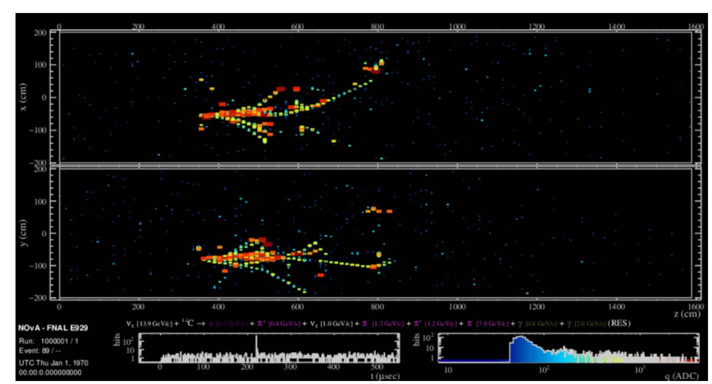
- Define some physics process of interest
  - Can it be measured with your detector?
  - Is your measurement competitive with previous results?
  - Is it valuable to the broader community?
- Count the number of occurrences in your detector over some exposure
- Subtract off anomalous counts (type I errors)
- Correct for missed counts (type II errors)
- Convert from counts to “cross section”



IceCUBE



OPERA



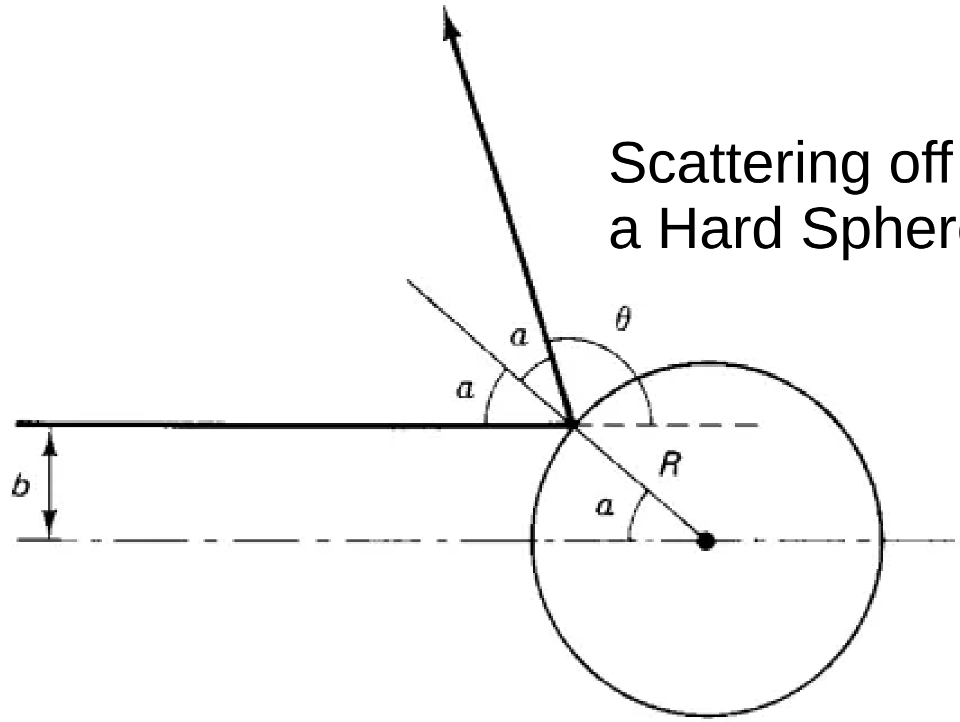
NOvA

ν_τ interactions

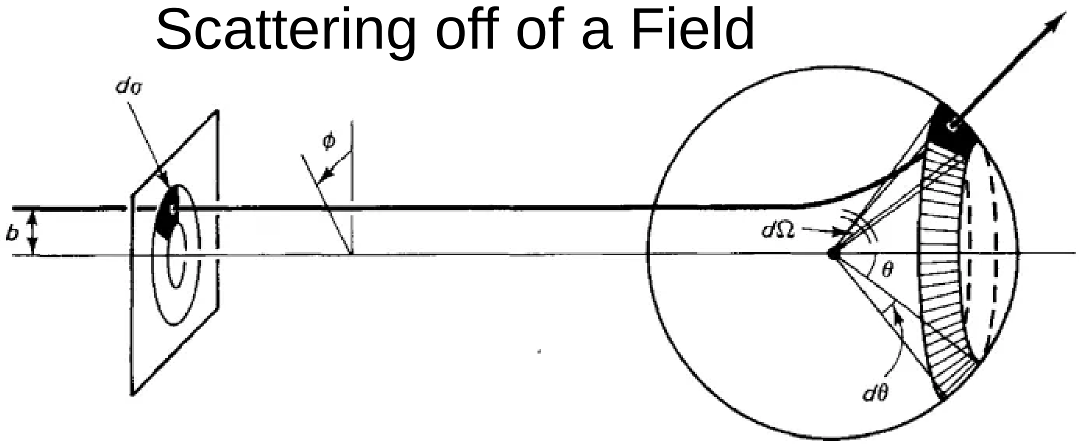
What is a Cross Section?

- Measure of probability of interaction occurring
- Given in units of area
  - Hard sphere scattering target
  - Analogy to cross sectional area
- Measurement of
  - Field (elastic)
  - Internal structure of the target (inelastic)

Scattering off of a Hard Sphere

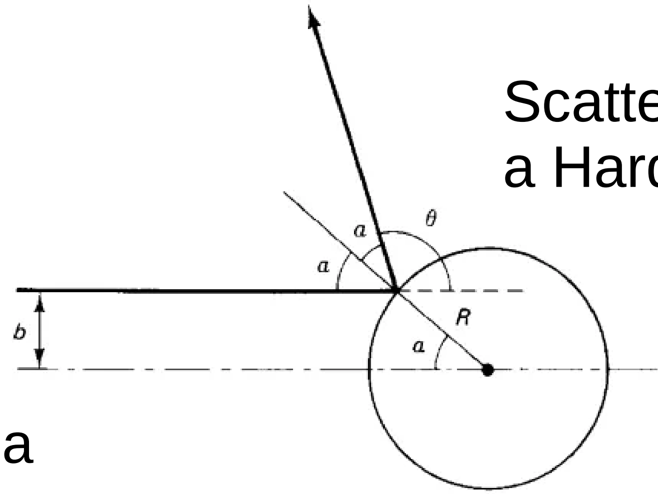


Scattering off of a Field

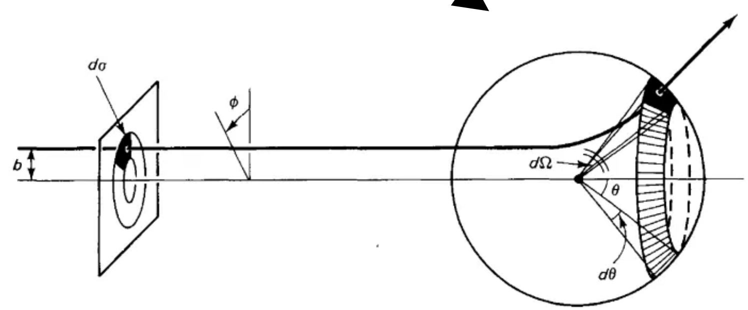


What is a Cross Section?

- Measure of probability of interaction occurring
- Given in units of area
  - Hard sphere scattering target
  - Analogy to cross sectional area
- Measurement of target properties



Scattering off of a Hard Sphere



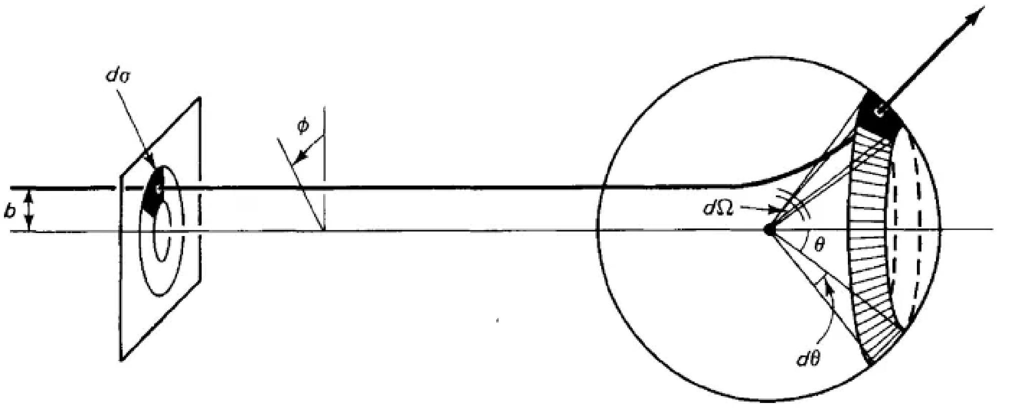
Scattering off of a Field

How do we turn counts in a detector into a cross section?

$$<\sigma_i> = \frac{N_i^{sel} - N_i^{bkg}}{T\,\Phi <\epsilon_i>}$$

What is a Cross Section?

- Measure of probability of interaction occurring
- Given in units of area
  - Hard sphere scattering target
  - Analogy to cross sectional area
- Measurement of target properties

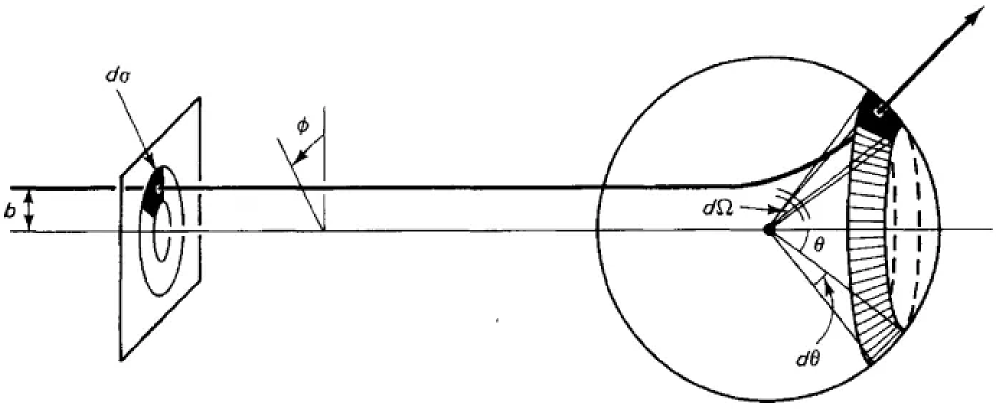


How do we turn counts in a detector into a cross section?

$$N_{sel}^{sig} = N_{sel} - N_{sel}^{bkg} = \int \sigma(E)\,\phi(E)\,T\,\epsilon(E)\,dE$$ ← Total number of events

What is a Cross Section?

- Measure of probability of interaction occurring
- Given in units of area
  - Hard sphere scattering target
  - Analogy to cross sectional area
- Measurement of target properties



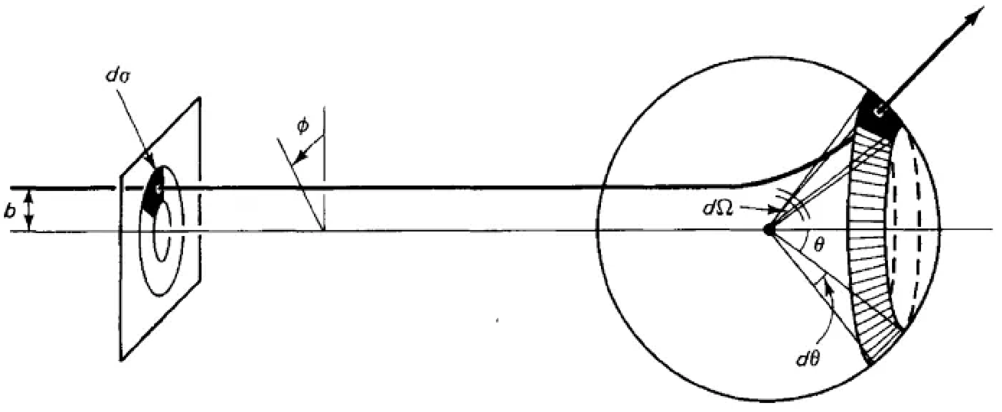
How do we turn counts in a detector into a cross section?

$$N_{sel}^{sig} = N_{sel} - N_{sel}^{bkg} = \int \sigma(E)\phi(E)T\,\epsilon(E)\,dE$$ ← Total number of events

$$N_i^{sel} - N_i^{bkg} = \int \sigma_i(E)\phi(E)T\,\epsilon_i(E)\,dE$$ ← Events in some final state kinematic region

What is a Cross Section?

- Measure of probability of interaction occurring
- Given in units of area
  - Hard sphere scattering target
  - Analogy to cross sectional area
- Measurement of target properties



How do we turn counts in a detector into a cross section?

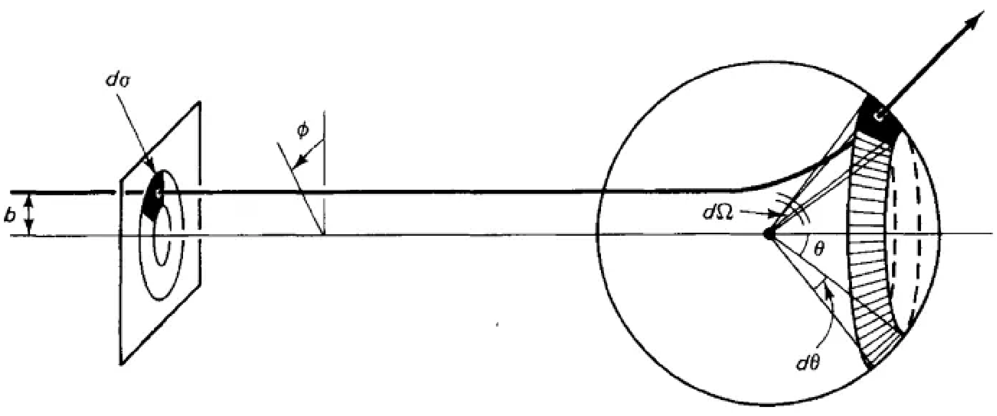
$$N_{sel}^{sig} = N_{sel} - N_{sel}^{bkg} = \int \sigma(E)\,\phi(E)\,T\,\epsilon(E)\,dE$$ ← Total number of events

$$N_i^{sel} - N_i^{bkg} = \int \sigma_i(E)\,\phi(E)\,T\,\epsilon_i(E)\,dE$$ ← Events in some final state kinematic region

$$N_i^{sel} - N_i^{bkg} = T <\sigma_i> <\epsilon_i> \int \phi(E)\,dE$$ ← Use energy averaged xsec and eff

What is a Cross Section?

- Measure of probability of interaction occurring
- Given in units of area
  - Hard sphere scattering target
  - Analogy to cross sectional area
- Measurement of target properties



How do we turn counts in a detector into a cross section?

$$N_{sel}^{sig} = N_{sel} - N_{sel}^{bkg} = \int \sigma(E)\,\phi(E)\,T\,\epsilon(E)\,dE$$ ← Total number of events

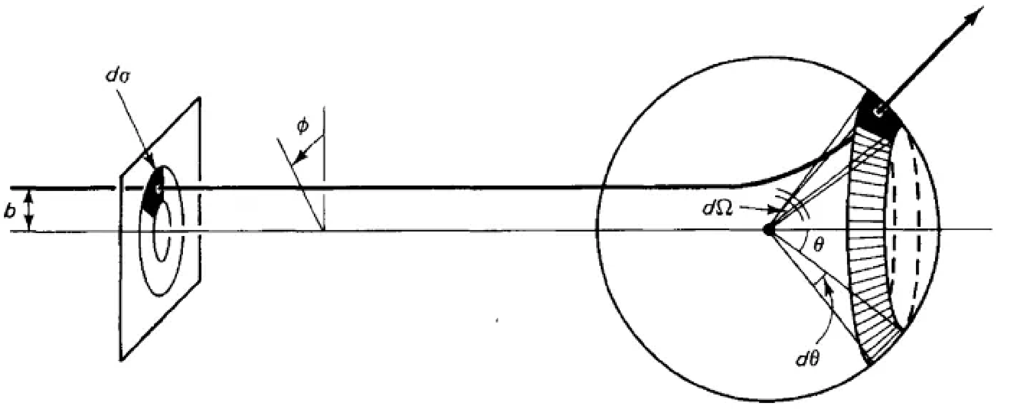
$$N_i^{sel} - N_i^{bkg} = \int \sigma_i(E)\,\phi(E)\,T\,\epsilon_i(E)\,dE$$ ← Events in some final state kinematic region

$$N_i^{sel} - N_i^{bkg} = T <\sigma_i> <\epsilon_i> \int \phi(E)\,dE$$ ← Use energy averaged xsec and eff

$$N_i^{sel} - N_i^{bkg} = T <\sigma_i> <\epsilon_i> \Phi$$ ← Φ is the integrated flux

What is a Cross Section?

- Measure of probability of interaction occurring
- Given in units of area
  - Hard sphere scattering target
  - Analogy to cross sectional area
- Measurement of target properties



How do we turn counts in a detector into a cross section?

$$N_{sel}^{sig} = N_{sel} - N_{sel}^{bkg} = \int \sigma(E)\,\phi(E)\,T\,\epsilon(E)\,dE$$ ← Total number of events

$$N_i^{sel} - N_i^{bkg} = \int \sigma_i(E)\,\phi(E)\,T\,\epsilon_i(E)\,dE$$ ← Events in some final state kinematic region

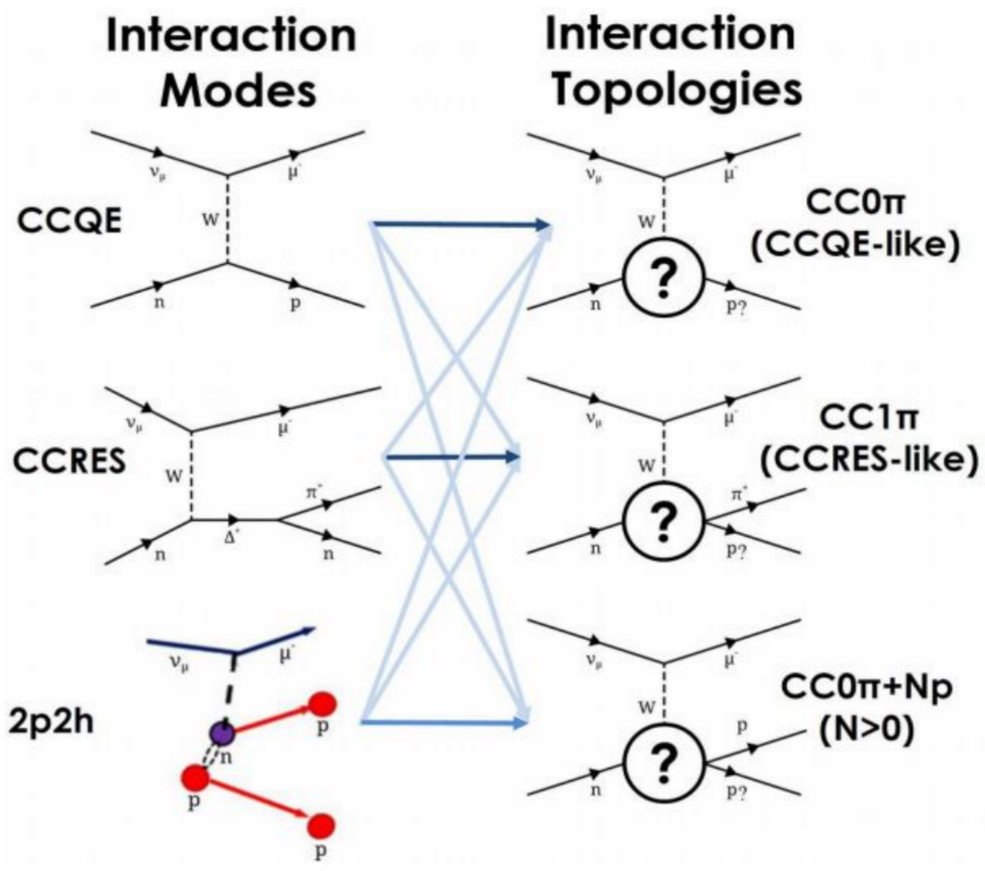
$$N_i^{sel} - N_i^{bkg} = T <\sigma_i> <\epsilon_i> \int \phi(E)\,dE$$ ← Use energy averaged xsec and eff

$$N_i^{sel} - N_i^{bkg} = T <\sigma_i> <\epsilon_i> \Phi$$ ← Φ is the integrated flux

$$<\sigma_i> = \frac{N_i^{sel} - N_i^{bkg}}{T\,\Phi <\epsilon_i>}$$ ← Solve for the energy averaged xsec

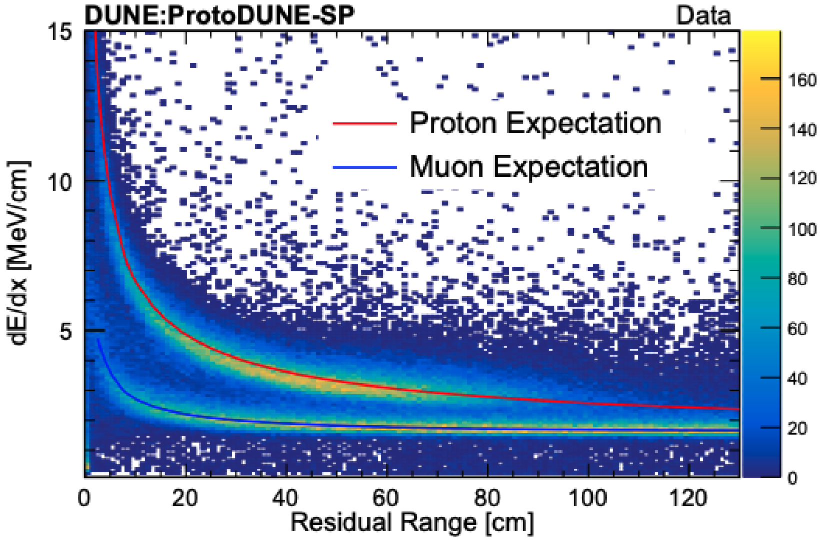
Selecting Your Signal

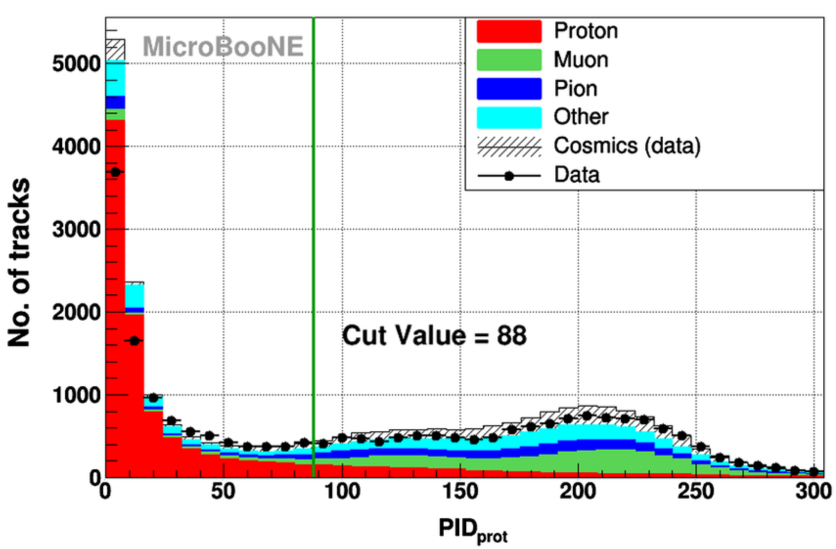
- What you **want to** measure vs what you **can** measure
  - Difference between the two must be corrected by model assumptions
  - Define your signal carefully <-- always based on truth information
  - Be honest and explicit about what you measure and the required corrections



Selecting Your Signal

- What you **want to** measure vs what you **can** measure
  - Difference between the two must be corrected by model assumptions
  - Define your signal carefully <-- always based on truth information
  - Be honest and explicit about what you measure and the required corrections
- Find reconstructed (measured) quantities that isolate your signal
  - Cuts: use each quantity individually
  - MVA: Identify regions of high signal density in N-dimensional quantity space
  - Reserve some quantities/regions for analysis





Selecting Your Signal

- What you **want to** measure vs what you **can** measure
  - Difference between the two must be corrected by model assumptions
  - Define your signal carefully   <-- always based on truth information
  - Be honest and explicit about what you measure and the required corrections
- Find reconstructed (measured) quantities that isolate your signal
  - Cuts: use each quantity individually
  - MVA: Identify regions of high signal density in N-dimensional quantity space
  - Reserve some quantities/regions for analysis
- Selection optimization
  - Figure of merit (FOM): quantity that scales with sensitivity
  - Good FOM depends on relative importance of systematic uncertainties
  - Propagation of error in terms of counts (N^{sig}), efficiency (ε), and purity (ρ), gives:

$$\sigma\left(N^{sig}\right) = \sqrt{\left(\frac{\partial N^{sig}}{\partial N_{sel}}\sigma\left(N_{sel}\right)\right)^2 + \left(\frac{\partial N^{sig}}{\partial N^{sig}}\sigma\left(N_{sig}\right)\right)^2 + \left(\frac{\partial N^{sig}}{\partial N_{bkg}}\sigma\left(N_{bkg}\right)\right)^2}$$

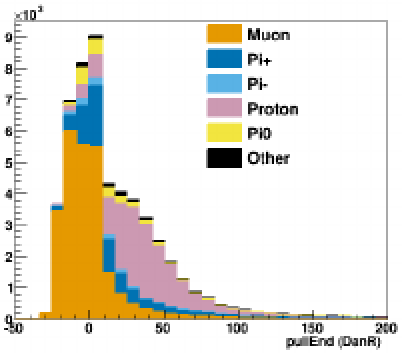
Selecting Your Signal

- What you **want to** measure vs what you **can** measure
  - Difference between the two must be corrected by model assumptions
  - Define your signal carefully <-- always based on truth information
  - Be honest and explicit about what you measure and the required corrections
- Find reconstructed (measured) quantities that isolate your signal
  - Cuts: use each quantity individually
  - MVA: Identify regions of high signal density in N-dimensional quantity space
  - Reserve some quantities/regions for analysis
- Selection optimization
  - Figure of merit (FOM): quantity that scales with sensitivity
  - Good FOM depends on relative importance of systematic uncertainties
  - Propagation of error in terms of counts (N^{sig}), efficiency (ε), and purity (ρ), gives:

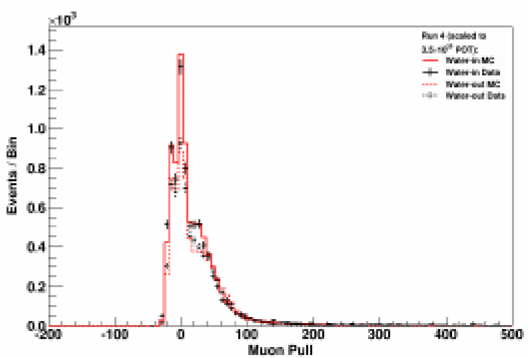
$$\left(\frac{\delta N^{sig}}{N^{sig}}\right)^2 = \left(\frac{1}{\epsilon \rho N^{sig}}\right) + \left(\left(\frac{1}{\rho} - 1\right) \sigma_{Bkg}\right)^2 + \sigma_{\epsilon}^2$$

$$\epsilon = \frac{N_{selected}^{signal}}{N^{signal}} \qquad \rho = \frac{N_{selected}^{signal}}{N_{selected}}$$

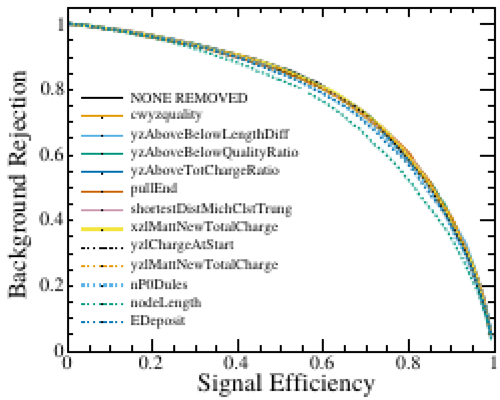
Selecting Your Signal

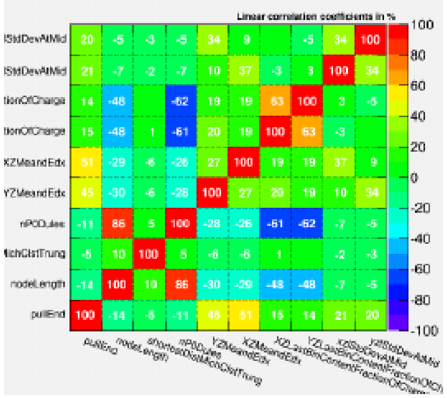


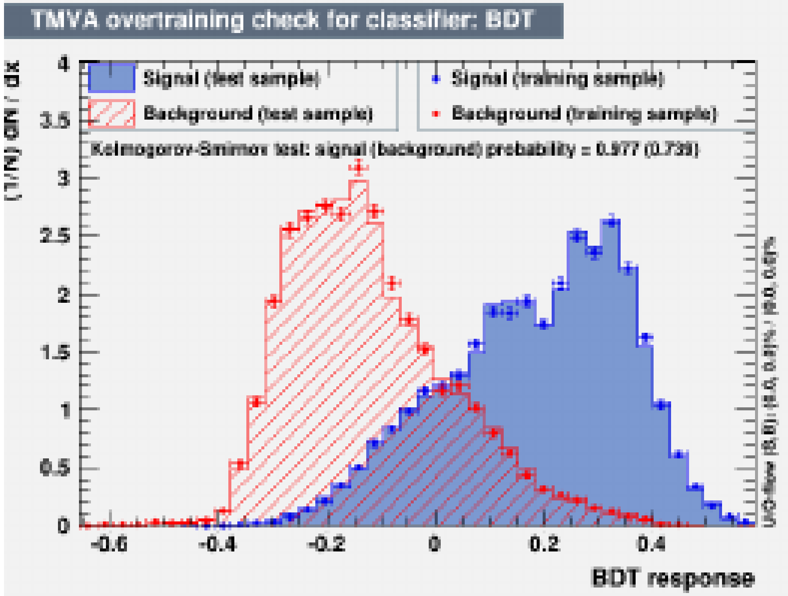
(a) Particle Breakdown

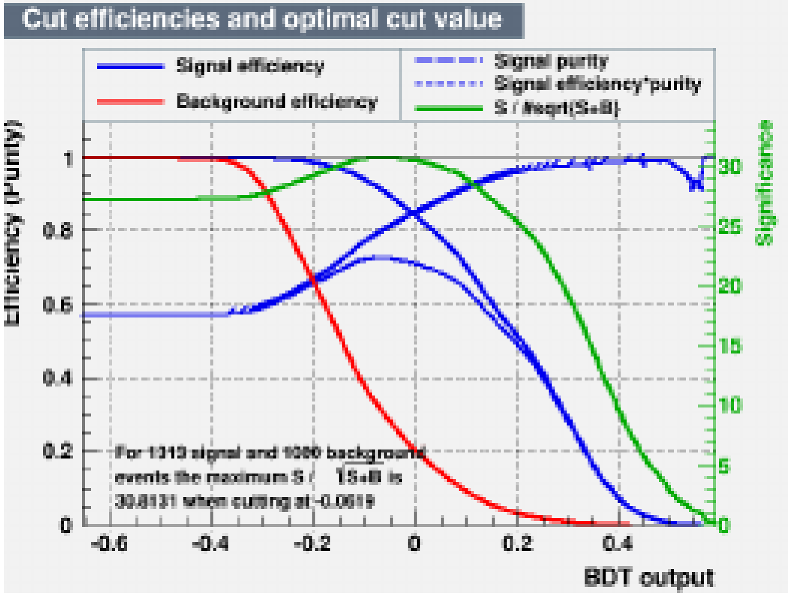


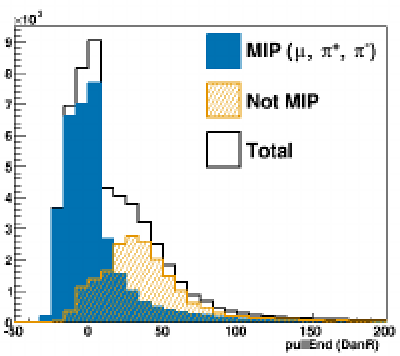
(b) Data MC Comparison



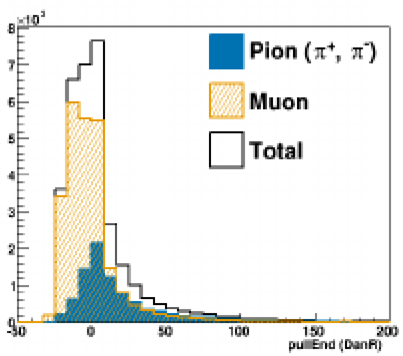








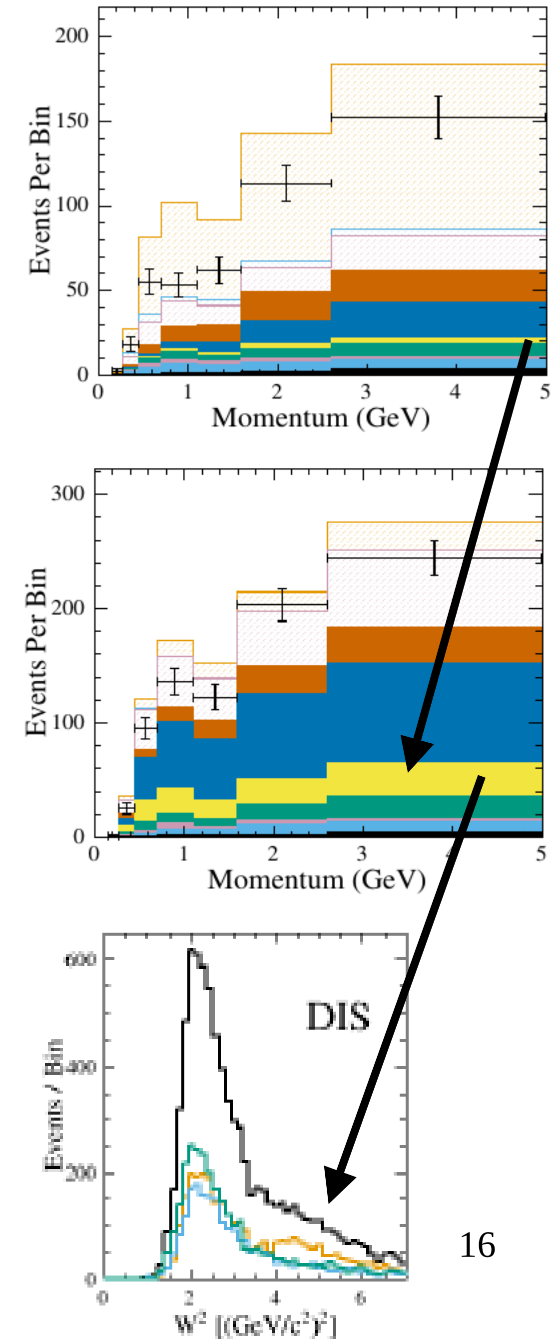
(c) MIP Signal Background Comparison



(d) MuPi Signal Background Comparison

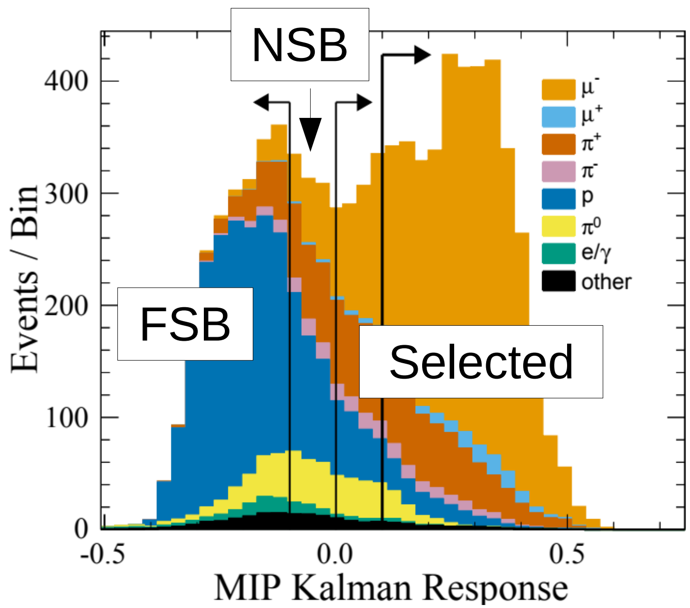
Constraining Your Backgrounds

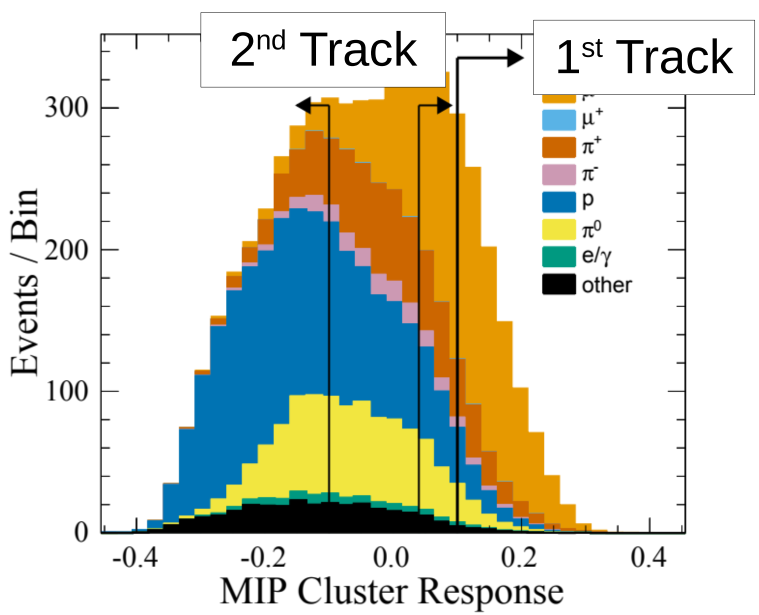
- Signal --> Defined by true quantities
- Selection --> Defined by reconstructed quantities
- These will not match up so you will have Backgrounds
- Need to minimize: $(1/\rho - 1)\ \sigma_{Bkg}$
  - Improve purity --> redo selection
  - Reduce σ_{Bkg} --> constrain/measure background
- Sideband samples:
  - Background becomes your “Signal”
  - Backgrounds must “match” i.e. come from same events
  - Constrained background model (template) with Sideband
  - Propagate constraints to the Selected sample
  - Best technique: combined fit of Selected and Sideband samples
    - Two steps in one (constrain and propagate)
    - Correlated errors automatically included

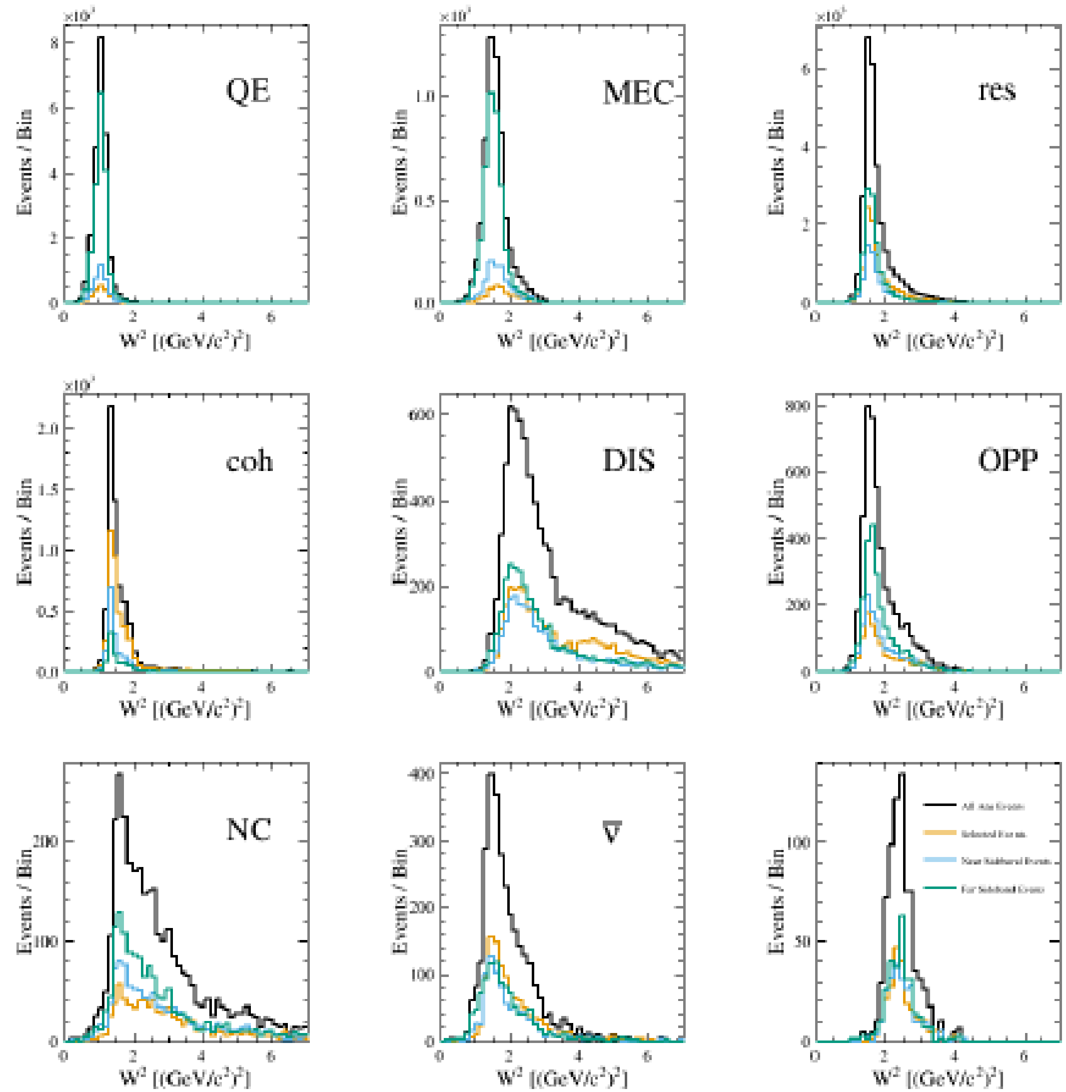


Constraining Your Backgrounds

- All
- Selected
- NSB
- FSB







Understand Your Efficiencies

- Propagated error on the efficiency cannot be reduced by measurements!
- Plot you eff in all important 1D and 2D spaces
  - Cut/MVA variables
  - Potential analysis variables
  - True particle kinematics (ex. p_x - cosθ_x)
  - True event kinematics (Q2, W)
- Efficiency binning should be narrower than analysis bins
- Not so narrow that statistical errors on efficiency dominate

- If ε is not flat in *x* measure *x* <-- Change across bin smaller that stat flucts
- If you can't measure *x*, cut the phase space (region ε is not flat in *x*)
- If you can't cut the phase space, correctly propagate the error, σ_{ε}

Understand Your Efficiencies

Calculate the cross section with:

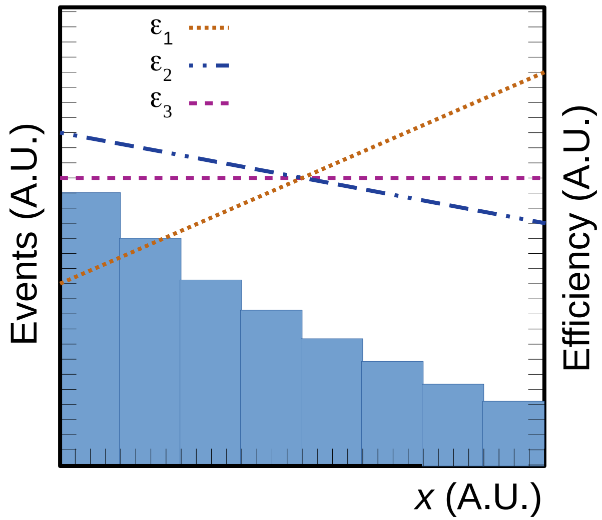
$$\sigma = \sum_i \frac{N_i - B_i}{\Phi T \epsilon_i}$$

However,

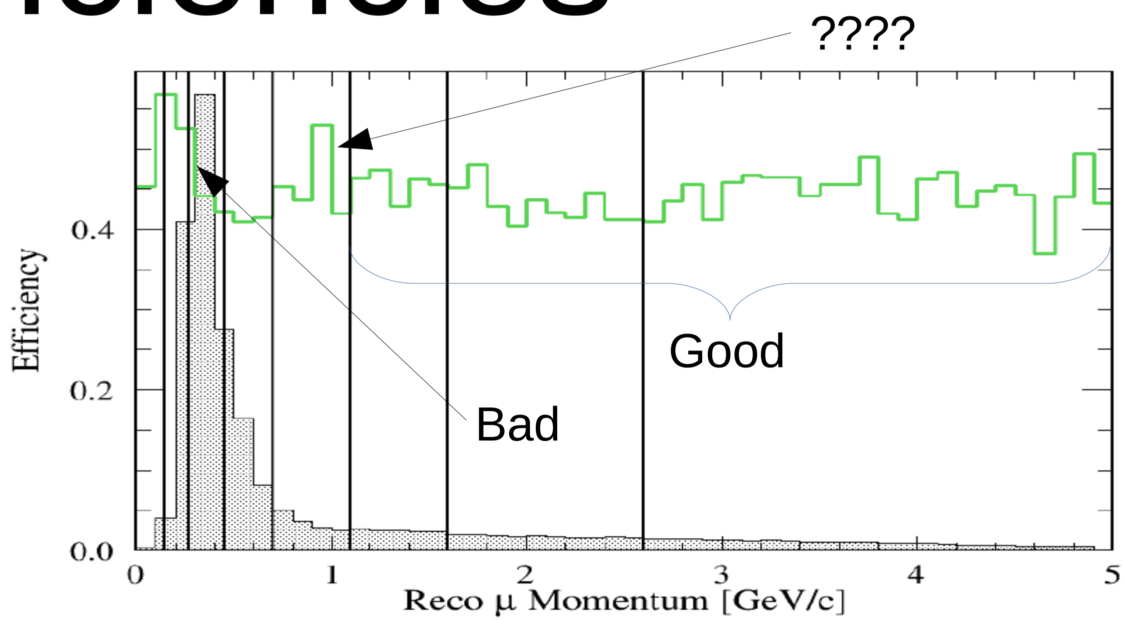
$$\sum_i a_i \times b_i \neq \sum_i a_i \times \sum_i b_i$$

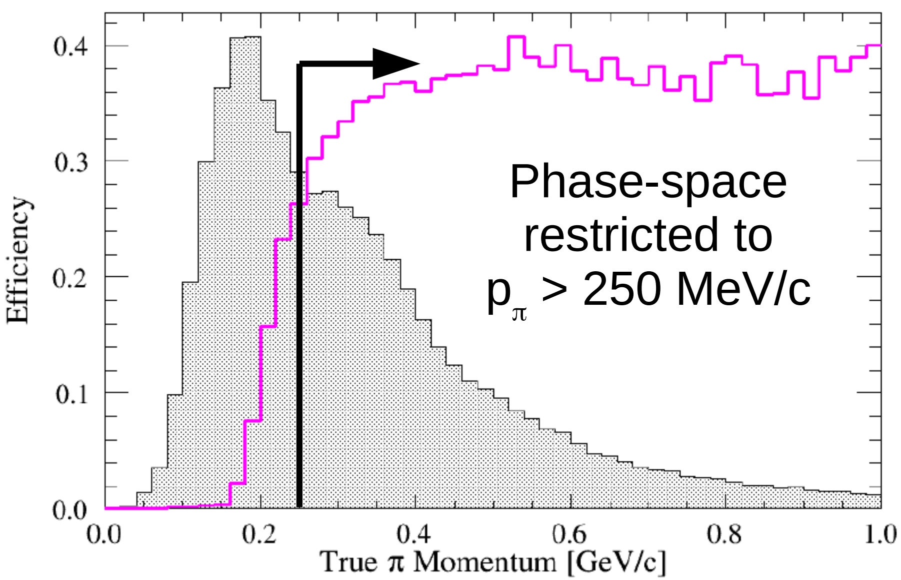
With $a_i = N_i - B_i$, and $b_i = \frac{1}{\epsilon_i}$

Unless a or b does not change with i



All 3 efficiencies have the same average, but will produce very different corrections bin by bin





Don't measure π momentum well

Find flat region and redefine signal

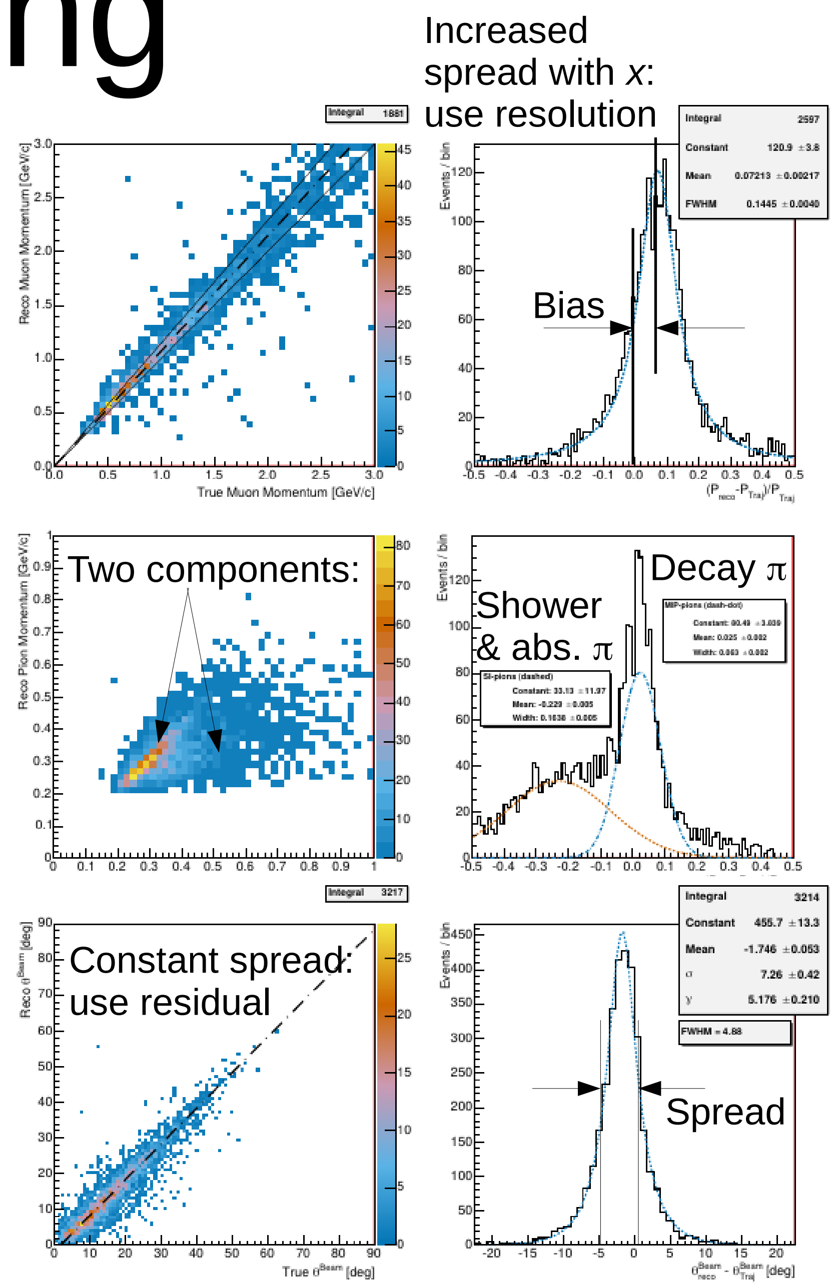
p_π > 250 MeV/c

Detector Systematics

- Need to understand all your observables
  - Everything you cut on (implicit and explicit)
  - All analysis variables
  - Data quality
- How does my inability to model *X* affect my analysis?
  - Smearing (true vs reco), resolution (reco-true)/true, and/or residual (reco - true)
  - How many events cross cut boundaries
  - Artificially induced and/or missed counts
- Organize and evaluate by **cause**, not by **effect** (if possible); example:
  - Fiducial Mass:
    - Number of targets
    - Uncertainty on material amounts in the fiducial volume
  - Fiducial volume:
    - Vertex migration --> Vertex resolution near fiducial volume boundary
    - Out of detector volume --> mis-reconstruction or neutral particle S.I.

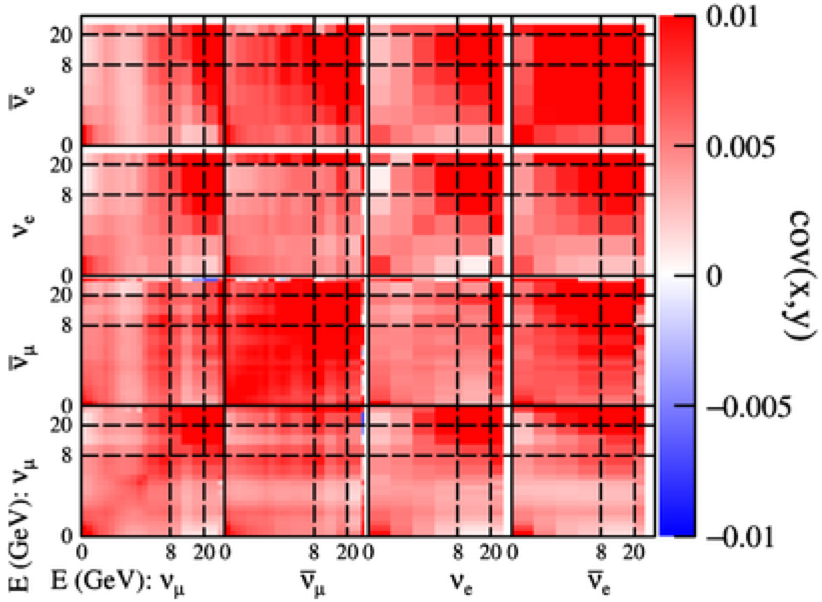
Analysis Variables and Binning

- What can be accurately measured / well calibrated?
  - Diagonal smearing?
  - Low bias?
  - Good resolution?
  - Modeled analytically (e.g. Gaussian or Cauchy)
- Statistics: Bin stat error ~ $\sqrt{N_i}$
- Efficiency: Flat across each bin?
- True space vs reco space (unfolding)



Flux and Cross Section Systematics

- What affects your analysis samples?
- Flux:
  - Covariance matrix in Eν-flavor-beam space.
  - Do you need to worry about wrong flavor or wrong sign contamination?
  - To PCA or not to PCA?

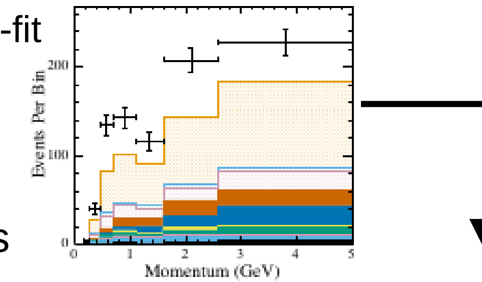


- Cross section:
  - Signal models: Generally you measure, so no error on N_{sel}^{sig} ... but ...
    - FSI induced topology changes (e.g. pion absorption)
    - Efficiency correction
  - Background models: analysis sample template fluctuations, by channel
  - Nuclear models: Nucleon momentum distribution, FSI
  - Secondary interactions: Do your your FS particles interact with your detector?

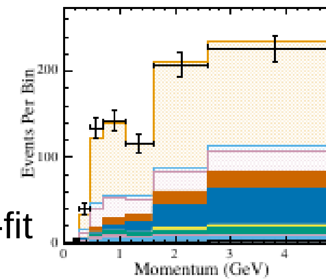
Cross Section Extraction

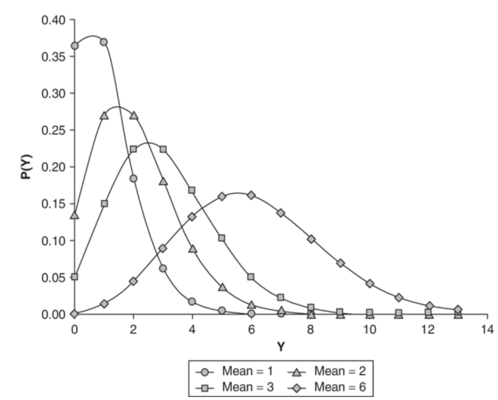
- Fitting --> parameter estimation techniques
  - Fit parameters:
    - Signal – unconstrained (i.e. no bias towards prior)
    - Backgrounds – model based templates with shape altering parameters
    - Need more bins than fit parameters
    - Only one unconstrained parameter per bin
  - Developing your test statistics (χ^2)
    - Gaussian vs Poisson statistics (Wilke's theorem) & MC statistics
    - Systematics and the penalty term:
      - Penalty terms has the form: $(\delta/\sigma)^2$, where δ is the parameter value change and σ is the gaussian width
      - Same χ^2 contribution as a combined fit with the external data used to estimate the prior, σ
      - Correlated priors: $\boldsymbol{\delta} C^{-1} \boldsymbol{\delta}^T$, where $\boldsymbol{\delta}$ is a vector of δ, and C represents an N-dimensional gaussian
- Unfolding and related techniques
  - Relates the measurement in reco space to the model in true space
  - Relies on the MC based smearing prediction (detector model)
  - Tends to have many degenerate solutions --> Ill posed problem
- Fitter outputs
  - Best-fit parameter values
  - Best-fit test statistic (χ^2) value
  - Parameter uncertainties and covariances

Pre-fit



Post-fit

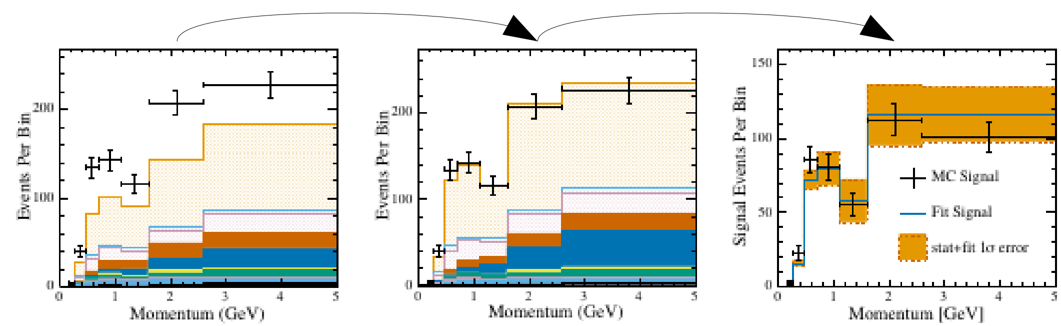




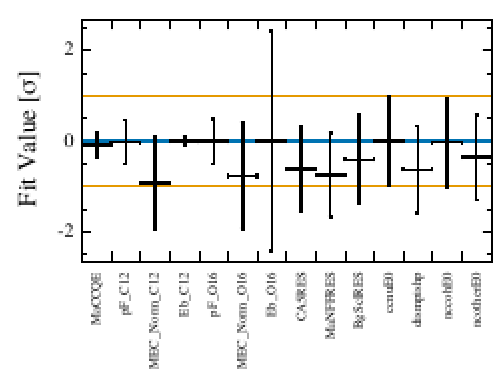
Mock Data Studies

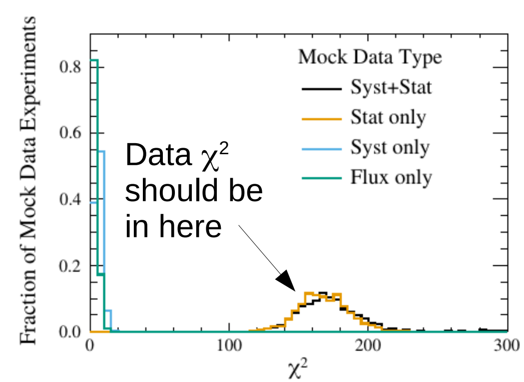
- Asimov studies (closure tests)
  - Do all the dials work as intended --> no crazy or unphysical weights
  - No local minima --> change starting parameter values
- Changing the model with fit parameters
  - Understand fitter response to parameters shifts
  - Make MD by adjusting dials by 1σ
  - Make MD by adjusting groups of dials by known amounts
- Alternate models
  - Nuclear models (RFG, LFG, SF, etc)
  - Interaction models (RS --> MK, RS --> BS)
  - Generators (NEUT --> GENIE, NuWro)
  - Crazy weights (test robustness, no physical meaning)
- Random throws
  - Statistical (random throw from a Poisson distribution)
  - Systematic (random parameter throws)
  - Statistical + systematic (all together now)

- What to look for:
  - Fits converge properly
  - Best-fit MC matches the data
  - Parameter values (best fit values, pulls, errors, and correlations)
  - No degenerate parameters (have same effect on analysis spectra)
  - What are you sensitive to? What aren't you sensitive to?
  - Distribution of χ^2 for random throws



Pre-fit   Post-fit   Signal Comp.



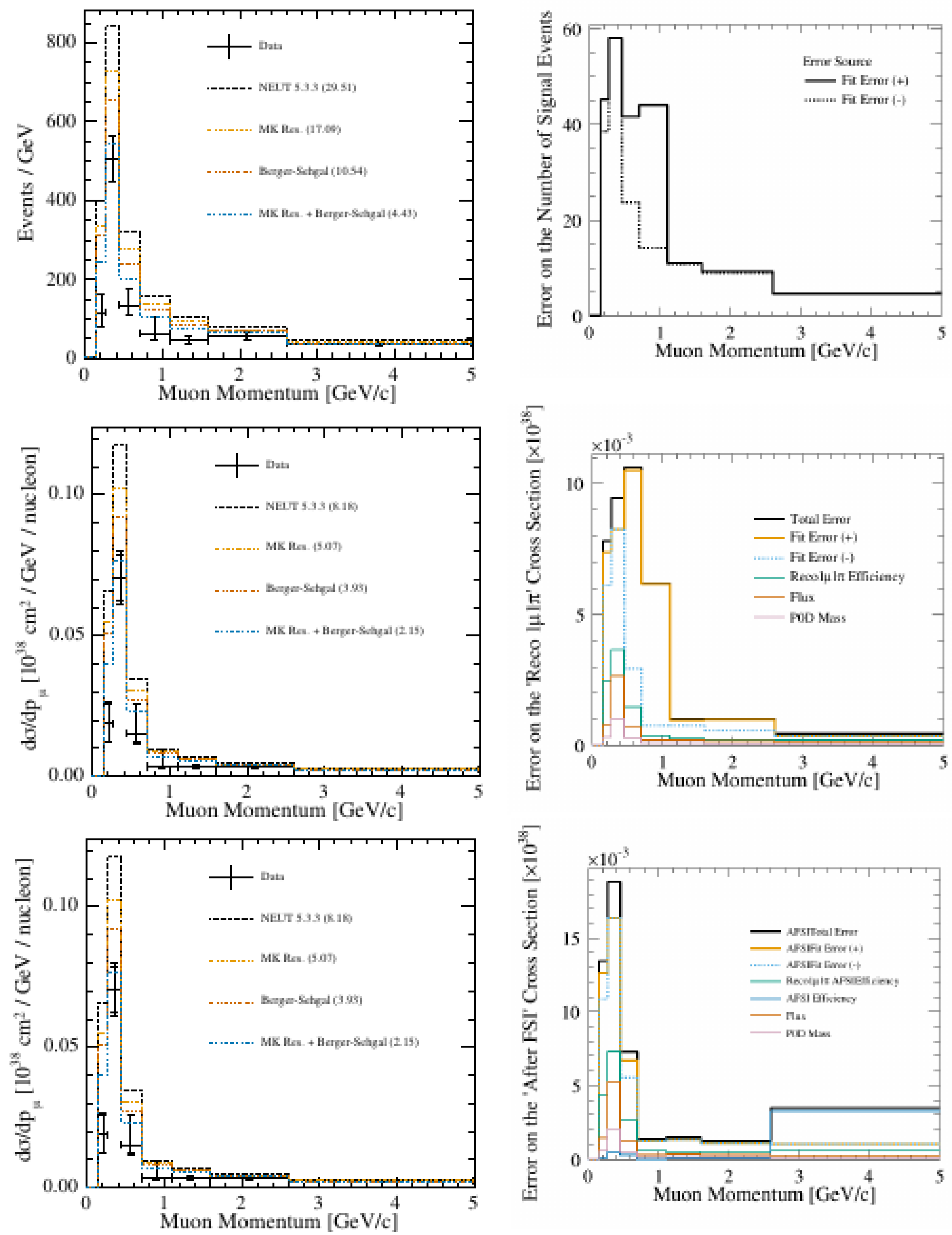


Calculate the Cross Section

- Back to where we started:

$$<\sigma_i> \ = \ \frac{N_i - B_i}{T\,\Phi <\epsilon_i>}$$

  - Fit gives you $N_i - B_i$, with errors
  - Throw full fit covariance matrix (w/ nuisance parameters) to get error on N_i's
  - Fiducial Mass in AMU gives T
  - Error on T from survey info, etc
  - Flux w/ error comes from beam group
  - Efficiency w/ error should be in hand from selection studies
  - Apply efficiency corrections before combining bins!
  - Bins with negative signal content?



Box Opening & Model Comparisons

- Mock data studies tell us:
  - Sensitivity to fit parameters
  - Range of potential data where the fitter works
  - Test statistic distribution / ndof
- Staged box opening procedure:
  - Open one sample at a time
  - Compare with mock data results
    - Data within range studied (pre-fit)?
    - Best-fit parameters in post-fit range?
    - Best-fit χ^2 in range?
  - If consistent, move on
  - If problematic, reevaluate
  - Easy to move forward, but hard to go back without introducing bias

- Model Comparisons
  - Depends on the type of result
  - Unfolded:
    - Easy to compare post publication
    - More model dependence
  - Forward folded:
    - Less corrections needed
    - Compare post publication --> ReMU
  - What is in your MC?
  - What can your MC be reweighted to?
  - Data releases and NUISANCE

Comments,
Questions,
Discussion?

Official Plots and Data Release

- Official plots
  - Plots from your TN
  - To be used in:
    - Conference talks
    - Publications
  - Include (but not limited to):
    - Main physics results
    - Error analysis
    - MC signal characterization
    - Selected samples and efficiencies
    - Detector response and characterization
  - Posted separately to t2k.org
  - Plots in ROOT and graphical format
  - Plot caption/description
  - Script exists to build html formatting

- Data Release
  - Formats
    - ROOT file
    - Text file
    - Tables in paper appendix
  - Code
    - Script that reads in and uses data products
    - NUISANCE implementation
  - Contents
    - Plots of data
    - Output covariance matrix
    - Input flux covariance matrix
  - Not a fully solved problem and continuing to develop