Generating Mock Data

DUNE Near Detector meeting
October 2 2019

Cristóvão Vilela

Mock data sets

- We have produced two Mock data sets for the TDR using multivariate reweighting.
- NuWro-reweight
  - Our GENIE MC is reweighted to match NuWro in a multidimensional true kinematic space.
  - Motivated by the LBNC request to run sensitivity studies on data from a different generator
  - We can’t put an alternative sample through the simulation+reconstruction chain in a reasonable amount of time, so use reweighting.
- Missing proton energy
  - Induce a change in Etrue->Erec that is difficult to identify with an on-axis LAr near detector.
  - Motivated by DUNE-PRISM studies: this type of mis-modelling gives biased oscillation parameters in a FD fit and this can be mitigated by a DUNE-PRISM data-driven fit.
- Different pre-processing, but reweighting procedure is the same.

BDT reweighting in a nutshell

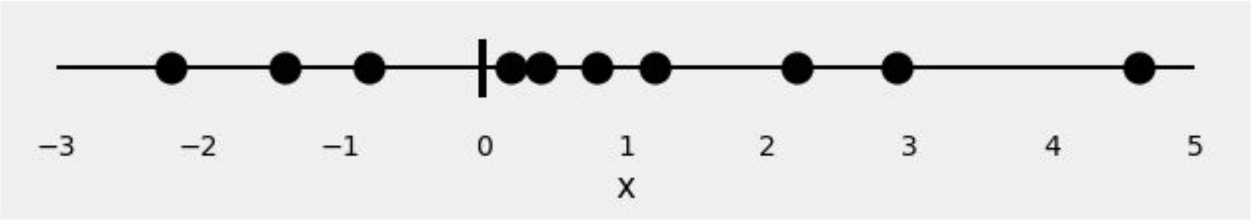
- For each mock data sample, we need two provide the BDT with two data sets: **origin** and **target**, or **nominal** and **mock**.
- The task of the BDT is to classify events as being drawn from the **origin** *vs* **target** distribution when given a set of variables (**features**) describing the event.
  - Think signal *vs* background in more common uses of BDTs in HEP.
- Given a training pair of **origin** and **target** distributions, where the events have a **label** in addition to **features**, we train the BDT by minimizing the **log loss,** aka **binary cross-entropy**: $-\mathcal{L} = y\log(p) + (1-y)\log(1-p)$
- Assign labels y = 0 for **target** and y = 1 for **origin** and the output of the BDT is:

$$BDT_{out} \approx p_{origin} \approx \frac{N_{origin}}{N_{origin}+N_{target}}$$

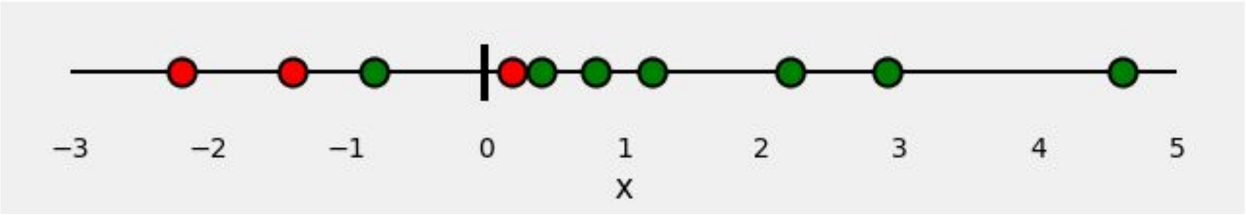
- And the reweighting function is given by: $w = \frac{N_{target}}{N_{origin}} \approx \frac{1}{BDT_{out}} - 1$

BDT reweighting in diagrams

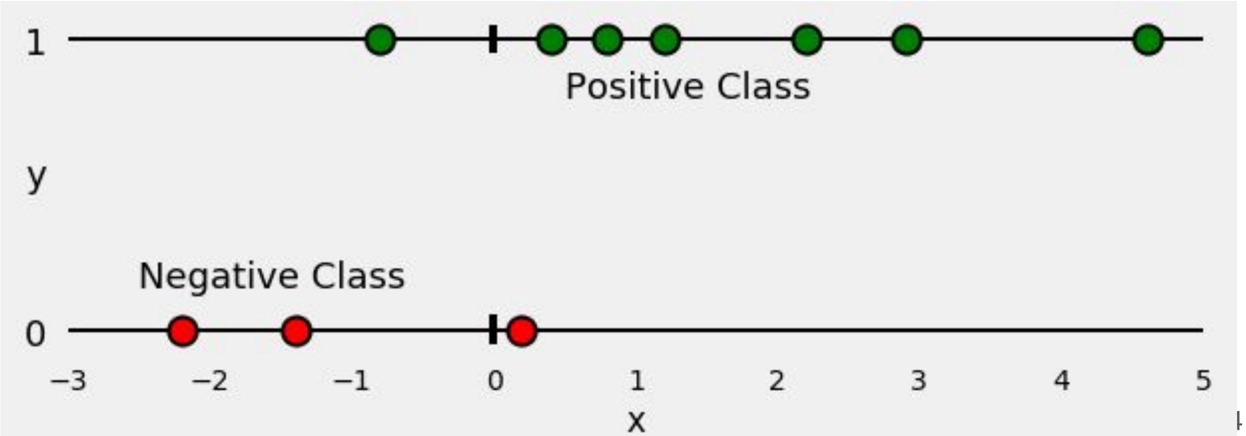
Unlabeled data.
Mix of **target** and **origin**.



Unlabeled data.
Mix of **target** and **origin**.



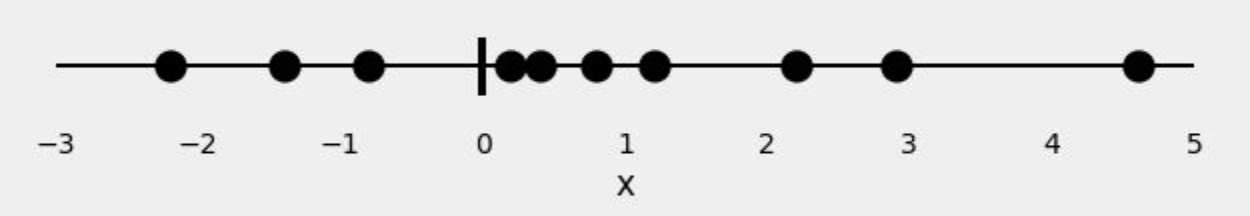
Labelled data.



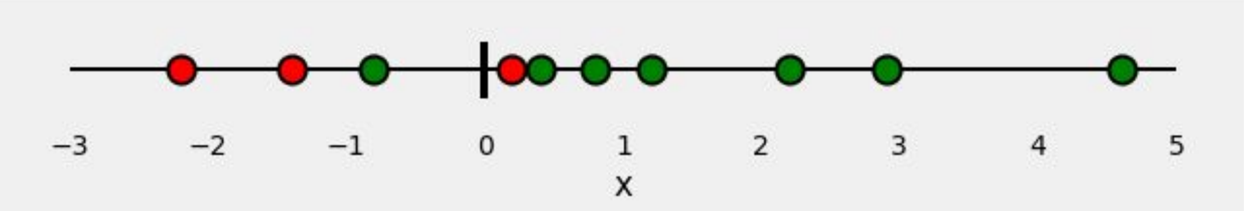
https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a

BDT reweighting in diagrams

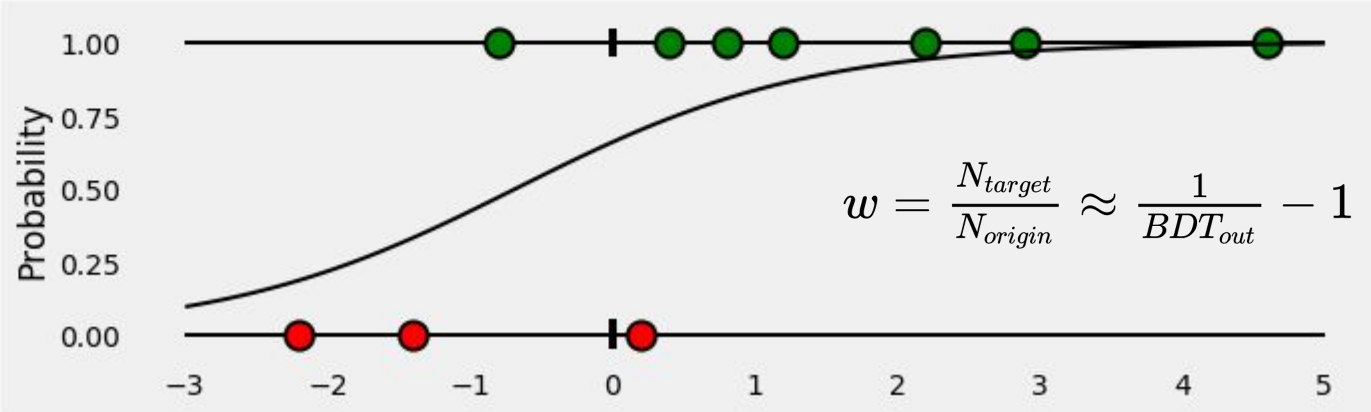
Unlabeled data.
Mix of **target** and **origin**.



Unlabeled data.
Mix of **target** and **origin**.



BDT output.



$$w = \frac{N_{target}}{N_{origin}} \approx \frac{1}{BDT_{out}} - 1$$

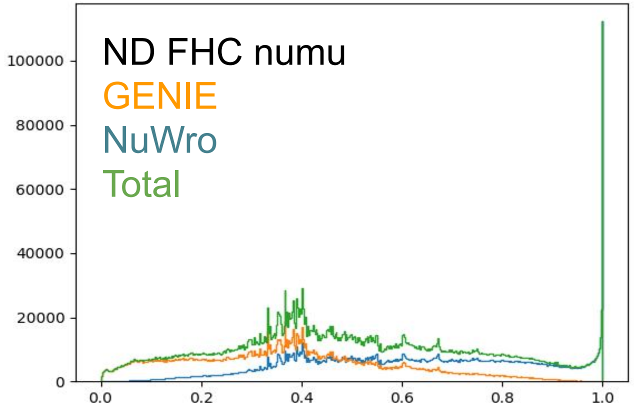
https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a

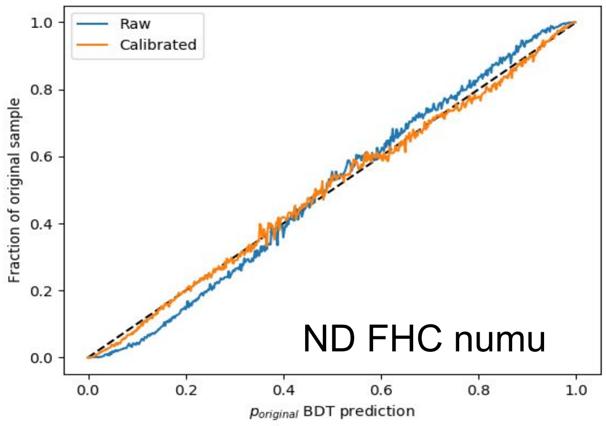
NuWro samples

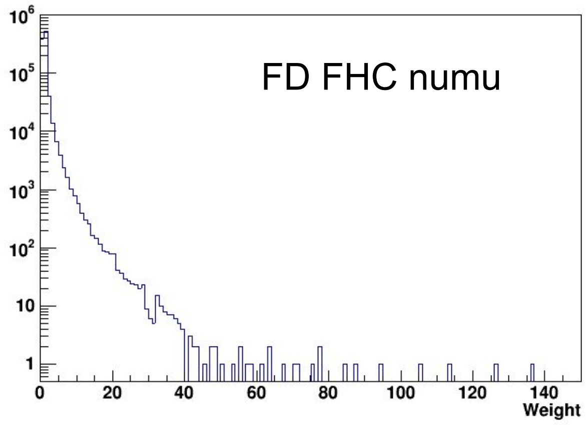
- NuWro events generated by Luke Pickering with the DUNE fluxes:
  - FD:
    - FHC: numu, nue
    - RHC: numubar, nuebar, numu, nue
  - ND:
    - FHC: numu
    - RHC: numubar, numu
- A set of 18 true variables is chosen as the space to reweight in:
  - Ev, lepton energy, angle between lepton and neutrino, Q2, W, x and y
  - Number of and total energy carried by:
    - Protons, neutrons, pi+, pi-, pi0 objects
  - Number of “em” objects
  - Ignore variables that do not have well-defined correspondence between generators:
    - E.g.: interaction mode, multiplicity of “other” and “nucleus” objects.
- BDTs are trained to classify events as “GENIE” or “NuWro” using these 18 variables as inputs.
  - One BDT per flux: 9 BDTs in total
- The linear BDT output is applied to GENIE events as a weight to get NuWro-like distributions.

BDT output

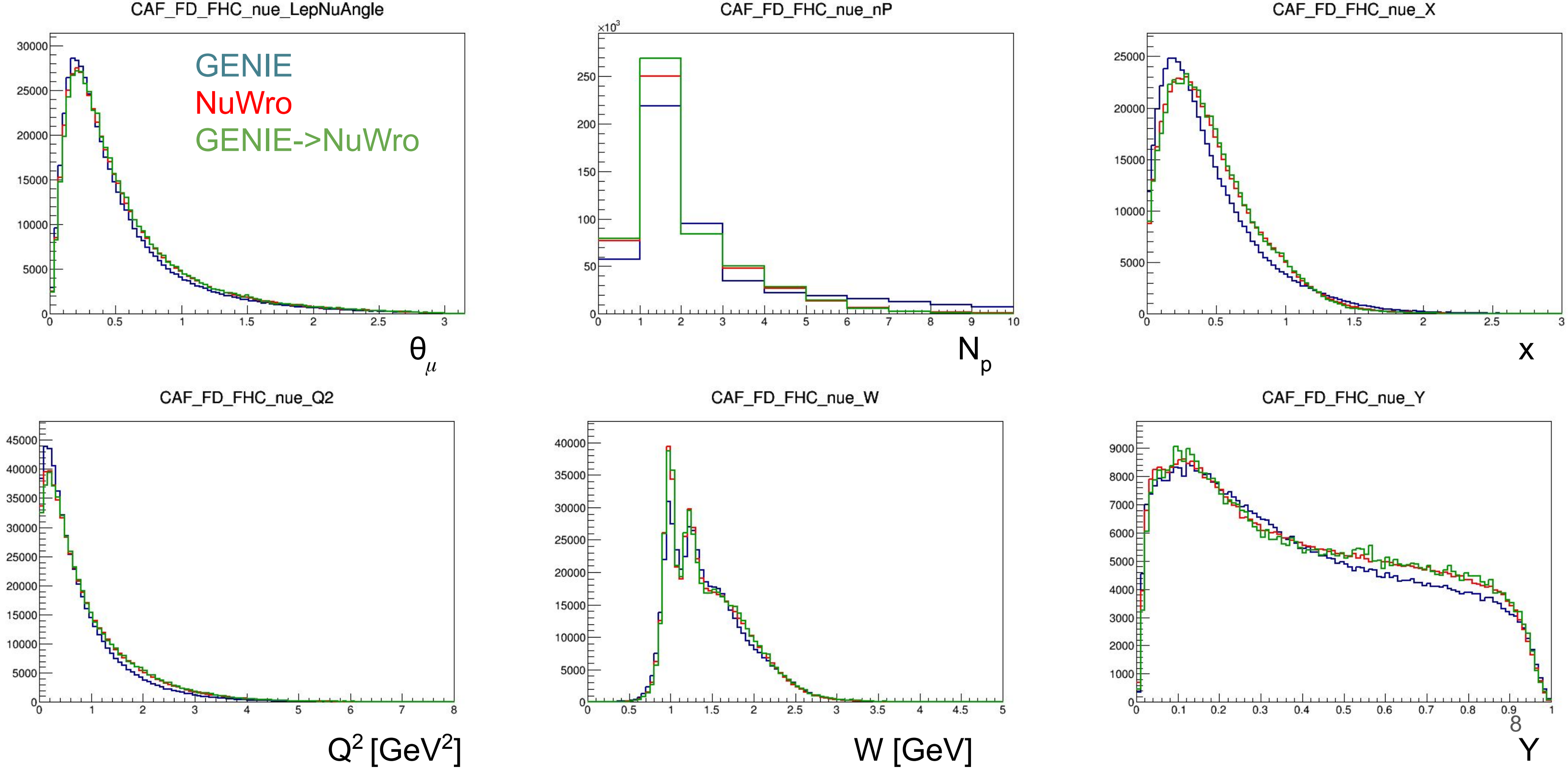
- As the BDT output will be used as a weight, it's important that it's linear.
  - Not a problem in typical classification tasks.
- While the output is designed to be linear, occasionally sigmoid-like features are present in the reliability plot.
  - Use Platt scaling to correct this - fit logistic function parameters that give linear output.





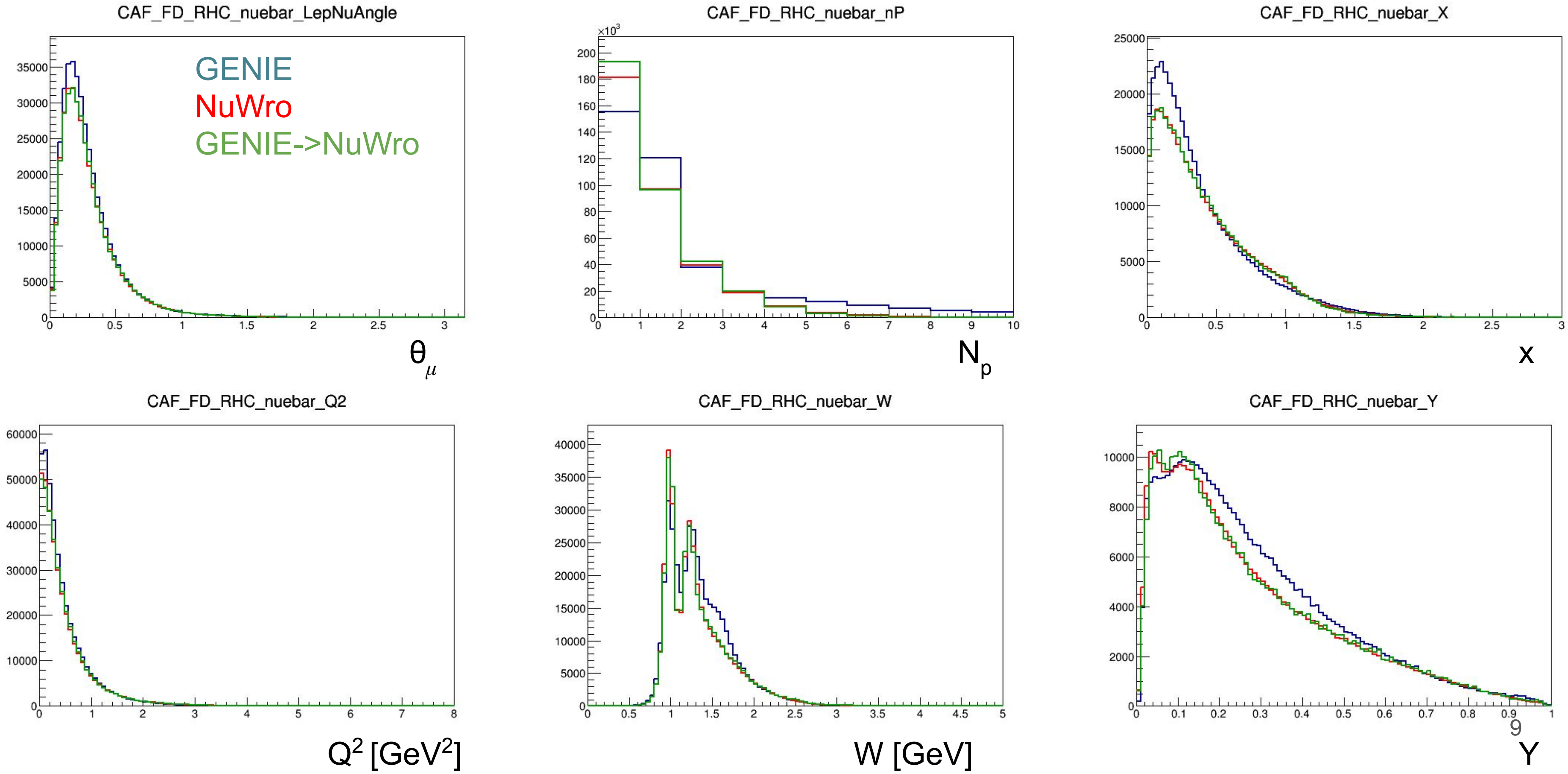


FD FHC nue



CAF_FD_FHC_nue_LepNuAngle
GENIE
NuWro
GENIE->NuWro
θ_μ
CAF_FD_FHC_nue_nP
×10³
N_p
CAF_FD_FHC_nue_X
x
CAF_FD_FHC_nue_Q2
Q^2 [GeV^2]
CAF_FD_FHC_nue_W
W [GeV]
CAF_FD_FHC_nue_Y
Y

FD RHC nuebar

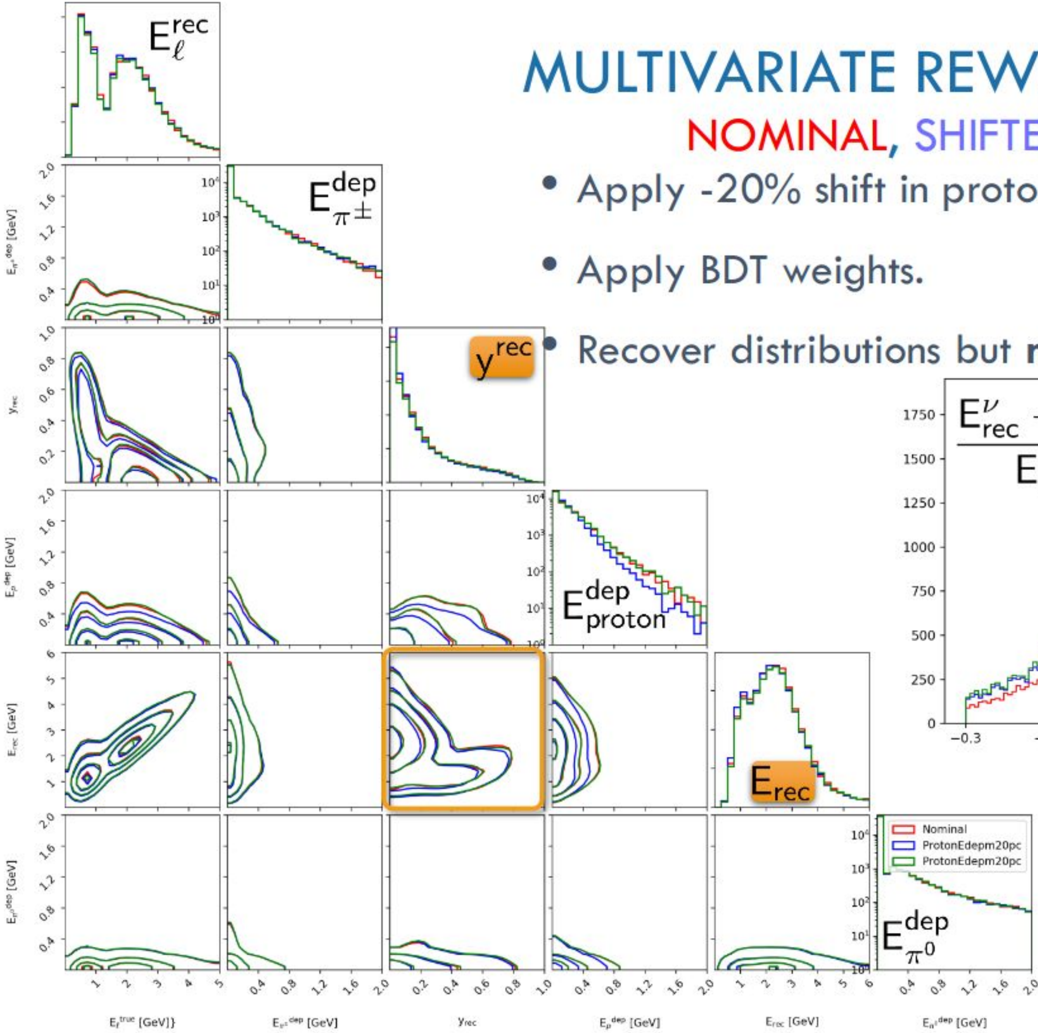


Missing proton energy fake data

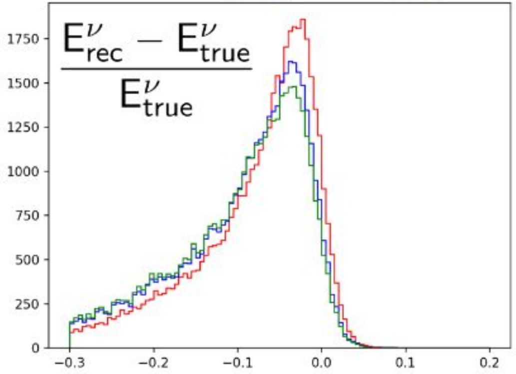
- The goal of this fake data set is to provide an example of mis-modelling that would be **difficult to measure** in an **on-axis** LAr detector and give b**iased oscillation parameter** estimation.
- Recipe:
    - Remove **20% of the proton energy** and add it to (largely invisible) neutrons.
        - In practice, we scale down the energy deposits in the LAr due to protons by 20%.
    - Reweight the shifted sample so that the on-axis ND **reconstructed** distributions agree with the nominal sample using a BDT.
    - Use additional BDT to capture the weights in **true** kinematic variables and propagate model to the far detector.
        - Interaction mode, neutrino energy, proton kinetic energy, elasticity.

MULTIVARIATE REWEIGHTING

NOMINAL, SHIFTED, FAKE DATA



- Apply -20% shift in proton deposited energy.
- Apply BDT weights.
- Recover distributions but **not** $E^{\nu}_{true} \rightarrow E^{\nu}_{rec}$.



More details in, e.g.:
https://indico.fnal.gov/event/16764/session/14/contribution/51/material/slides/0.pdf

Existing tools for DUNE: reweighting tools

- We have two sets of tools that use the XGBoost framework to train reweighting BDTs and a couple of examples of CAFAna implementations, for use in oscillation analysis.
- Reweighting our nominal MC to an alternative Generator using truth-level **features**: https://github.com/cvilelasbu/GeneratorReweight/
  - Two python scripts:
    - One pre-processes the data (CAF files + alternative model in CAF-style TTree) and stores everything in a large HDF5. Also deals with relative normalization of flux.
    - Training script reads HDF5 and runs XGBoost.
- Using a hacked version of our MC as the alternative model (e.g., 20% missing proton energy): https://github.com/cvilelasbu/MagicRW
  - Works like the above, but has a lot more built-in functionality to propagate changes in the model correctly. E.g., changing proton energy variable affects Erec.
  - A couple of examples implemented, including variables of interest for MPD like transverse variables -- but please check it makes sense before using!

Existing tools for DUNE: CAFAna implementation

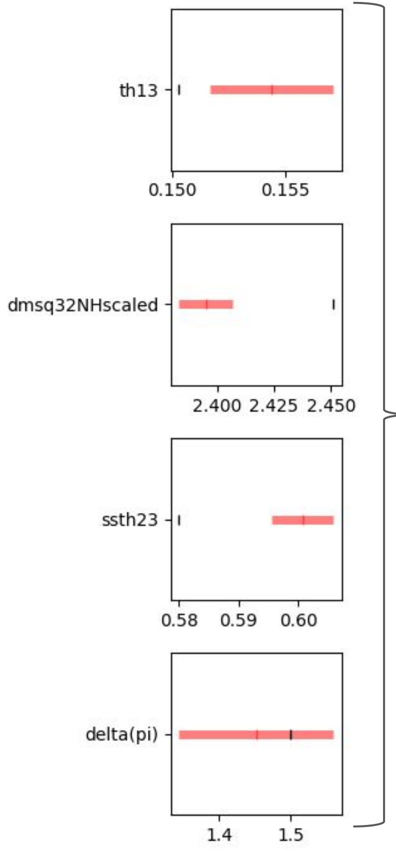
- Convert the XGBoost output into C code using treelite
    - https://github.com/dmlc/treelite
- Wrap treelite output in a C++ class:
    - https://github.com/cvilelasbu/ClassifyTreeLite/
- Implement reweighting as a systematic in CAFAna (L. Pickering):
    - Example: https://github.com/DUNE/lblpwgtools/blob/strong_and_stable/code/CAFAna/CAFAna/Systs/NuWroReweightFakeData.h

HELP!

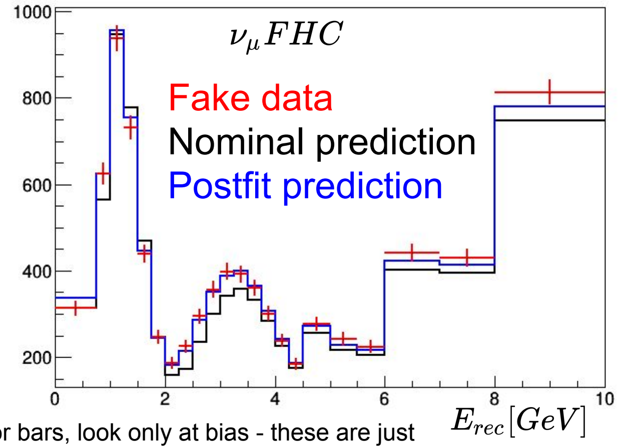
- Get in touch with:
    - CV
    - J. Wolcott (sorry!)
    - L. Pickering

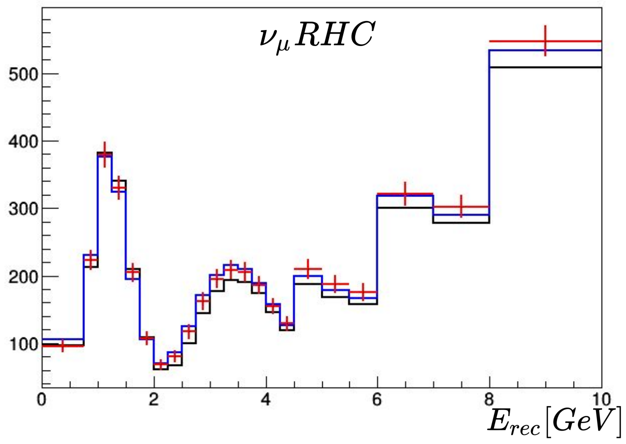
Backup

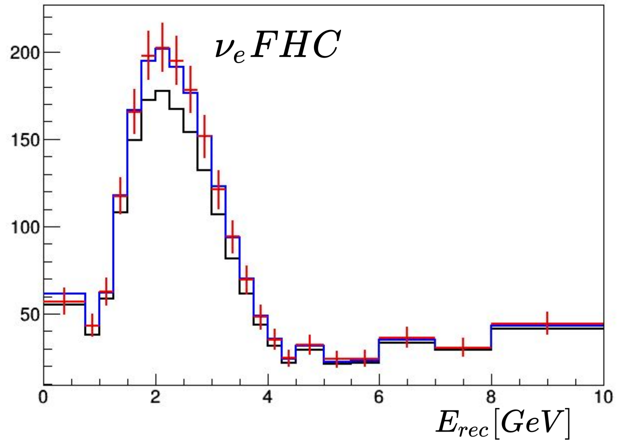
Fake data fit with latest analysis tools

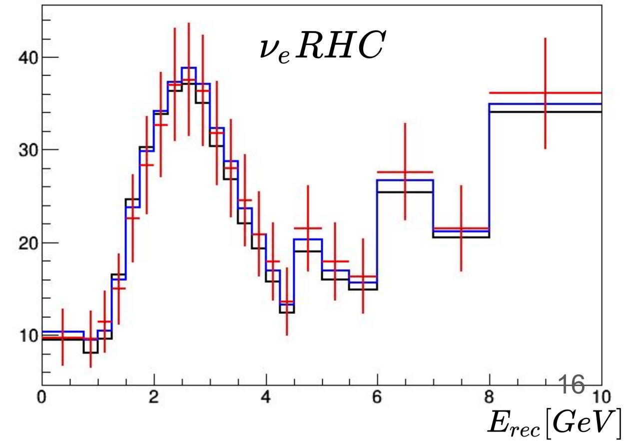


Ignore error bars, look only at bias - these are just the gaussian-like uncertainties at the best-fit point

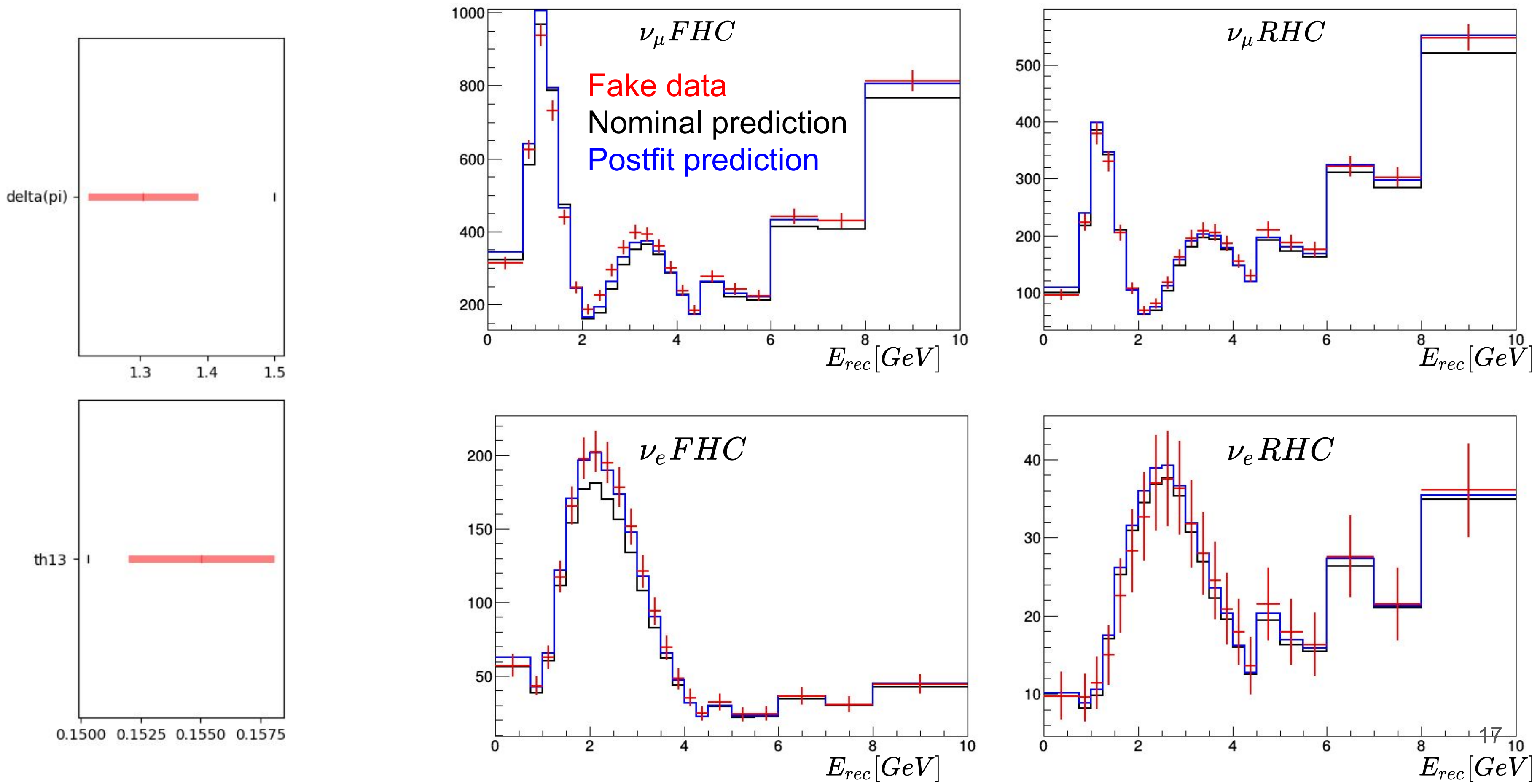




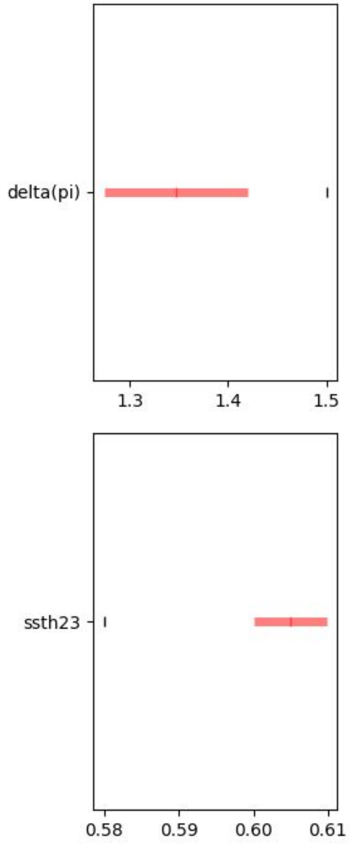


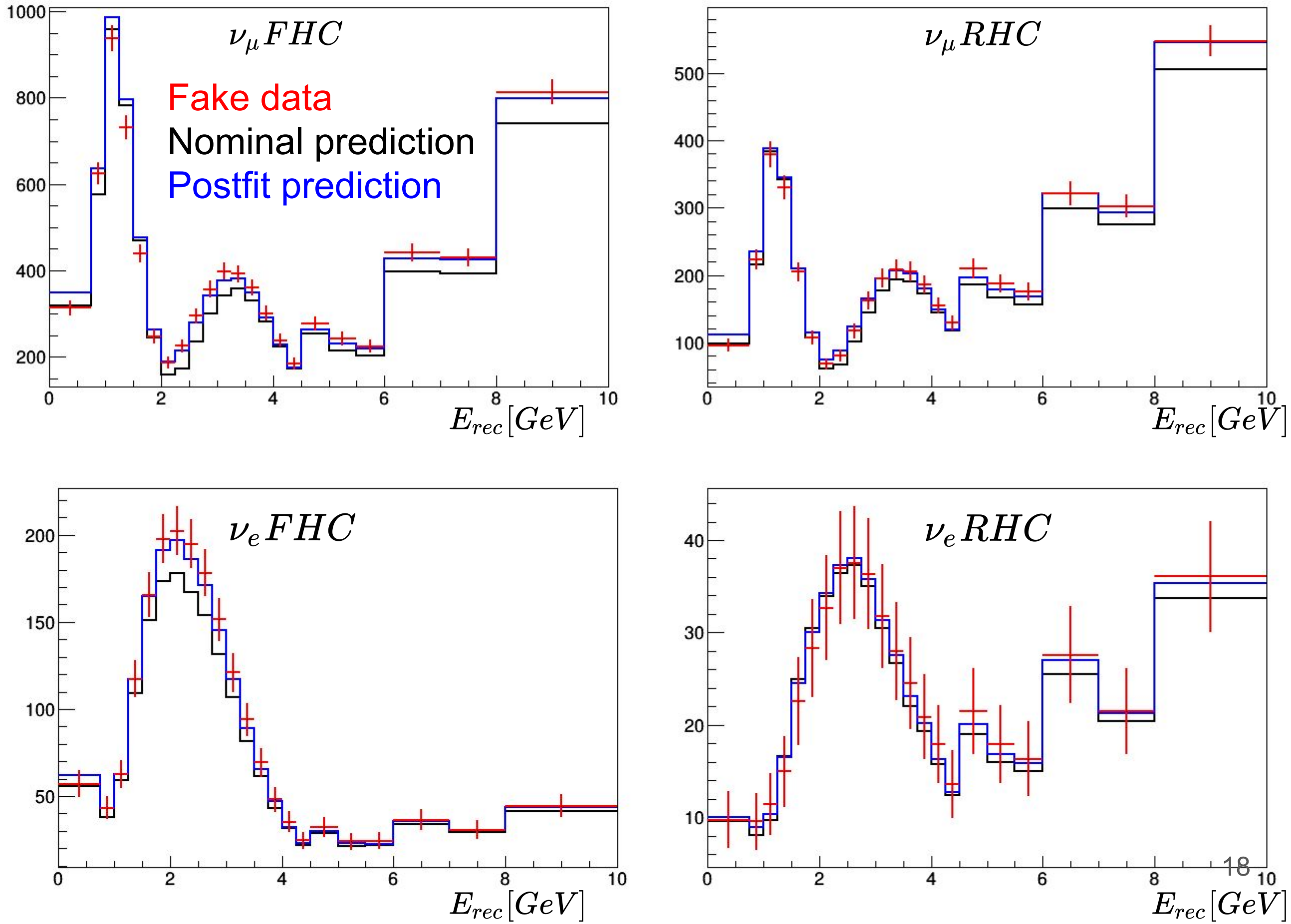


All oscillation parameters fixed other than delta and th13

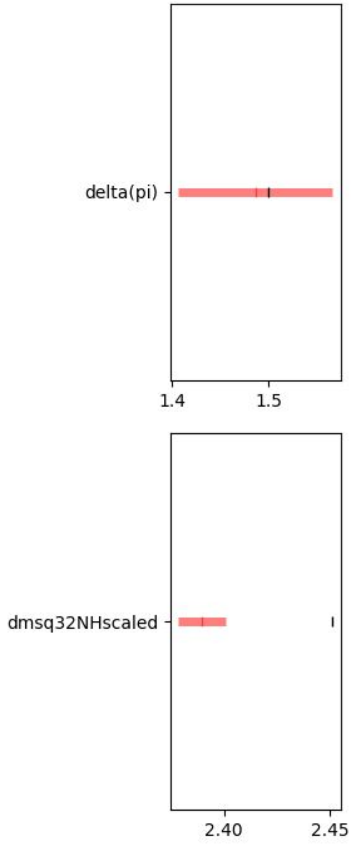


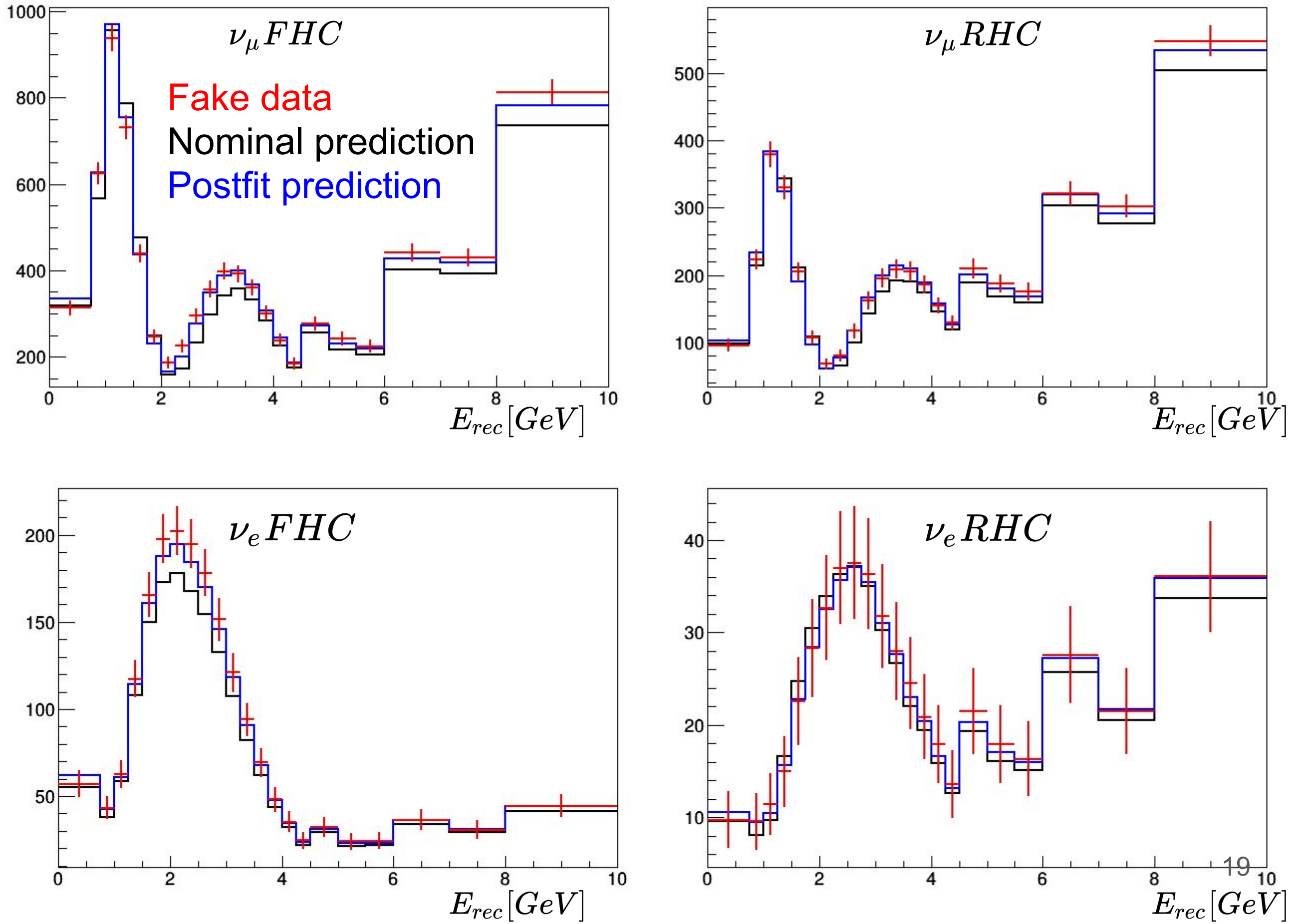
All oscillation parameters fixed other than delta and th23





All oscillation parameters fixed other than delta and dmsq32





Why is the deltaCP bias small - is this just a fluke?

- Toy example:
- For a global energy scale transformation: $E \to E' = aE$
- From disappearance we get a biased mass-squared splitting: $\Delta m^2_{32} \to \Delta m'^2_{32} = a\Delta m^2_{32}$
    - Such that numu survival probability stays invariant.
        - i.e., energy scale shift is absorbed by oscillation parameters.

$$P_{\mu\to x} \approx 1 - \left(\cos^4\theta_{13}\cdot\sin^2 2\theta_{23} + \sin^2\theta_{23}\cdot\sin^2 2\theta_{13}\right)\sin^2\left(\frac{\Delta m^2 L}{4E_\nu}\right)$$

$$\Delta m^2 \approx \Delta m^2_{32} \approx \Delta m^2_{31}$$

Why is the deltaCP bias small - is this just a fluke?

- Ignoring the solar term, can write the deltaCP dependence as:
- $K sin(\Delta_{21}) sin(\Delta_{31})[cos(\delta_{CP}) cos(\Delta_{32}) - sin(\delta_{CP}) sin(\Delta_{32})]$

  $\Rightarrow K \frac{cos(\delta_{CP}+\Delta_{32})}{2}[1 - cos(\Delta_{32} + 2\Delta_{21})]$

  with $K = 8c_{13}^2 s_{12} s_{13} s_{23} c_{12} c_{23}$

- Now apply energy scale transformation and use transformed Δ_{32}:
  - $\Delta_{32} \rightarrow \Delta'_{32} = \Delta m'^2_{32} \frac{L}{4E'} = a\Delta m^2_{32} \frac{L}{aE} = \Delta_{32}$
- Appearance probability is invariant under:
  - $E \rightarrow E' = aE$ and $\Delta m^2_{32} \rightarrow \Delta m'^2_{32} = a\Delta m^2_{32}$
- To first order, deltaCP measurements are robust wrt energy scale in a joint LBL fit.
  - Disappearance parameter measurements are not.

$$P(\nu_\mu \rightarrow \nu_e) = 4c_{13}^2 s_{13}^2 s_{23}^2 \sin^2\Delta_{31}$$ Leading term — θ_{13}

$$+8c_{13}^2 s_{12}s_{13}s_{23}(c_{12}c_{23}\cos\delta - s_{12}s_{13}s_{23})\cos\Delta_{32}\sin\Delta_{31}\sin\Delta_{21}$$ CPC

$$-8c_{13}^2 c_{12}c_{23}s_{12}s_{13}s_{23}\sin\delta\sin\Delta_{32}\sin\Delta_{31}\sin\Delta_{21}$$ CPV

$$+4s_{12}^2 c_{13}^2(c_{12}^2 c_{23}^2 + s_{12}^2 s_{23}^2 s_{13}^2 - 2c_{12}c_{23}s_{12}s_{23}s_{13}\cos\delta)\sin^2\Delta_{21}$$ Solar

$c_{ij} = \cos\theta_{ij}, s_{ij} = \sin\theta_{ij}$

$\Delta_{ij} = \Delta m^2_{ij} \frac{L}{4E_\nu}$

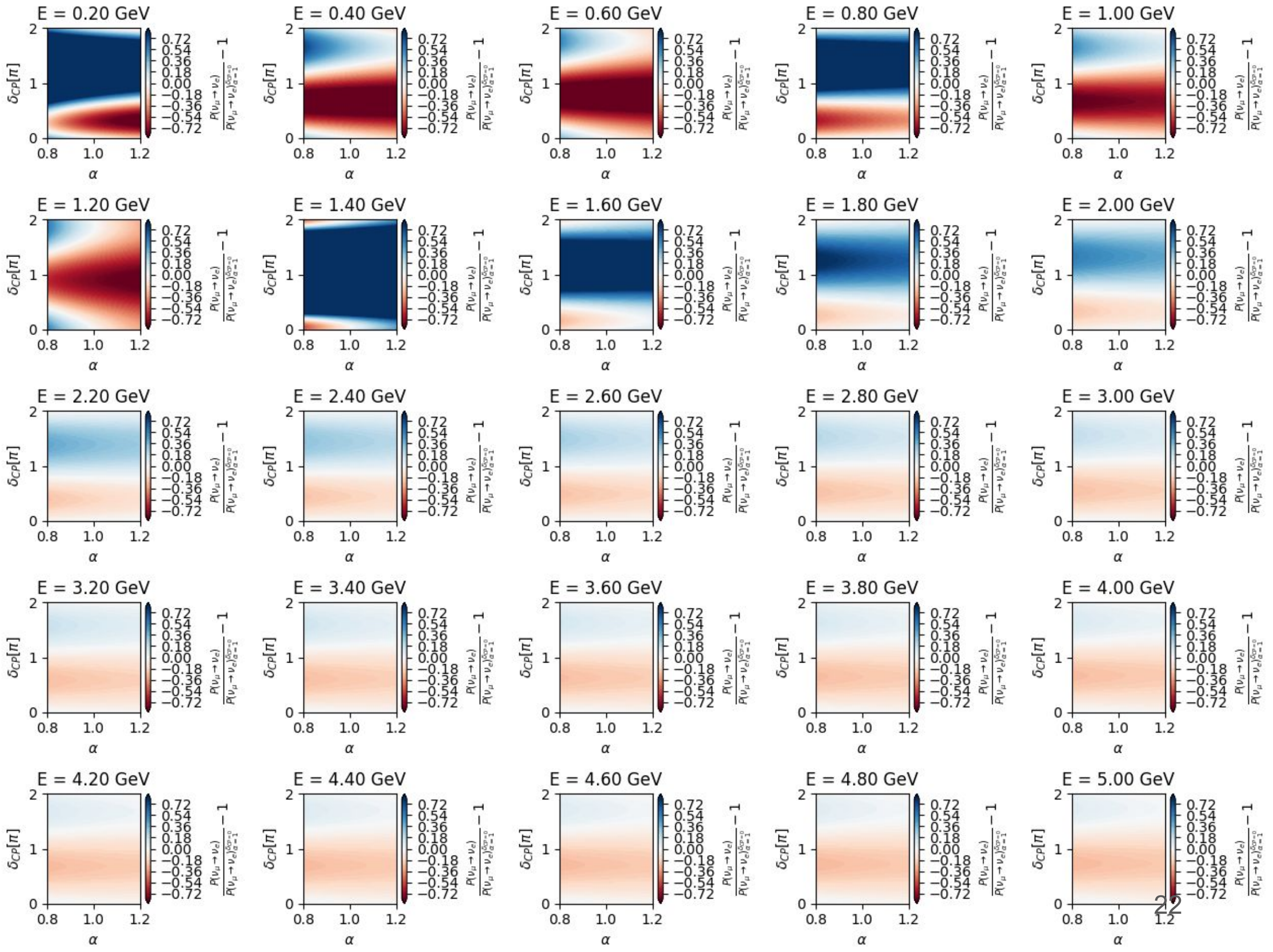
replace δ by $-\delta$ for $P(\overline{\nu_\mu} \rightarrow \overline{\nu_e})$

CP violating term introduced by interference among three-flavor mixing

Delta CP energy scale robustness - neutrinos

$$E \rightarrow E' = aE$$

$$\Delta m^2_{32} \rightarrow \Delta m'^2_{32} = a\Delta m^2_{32}$$



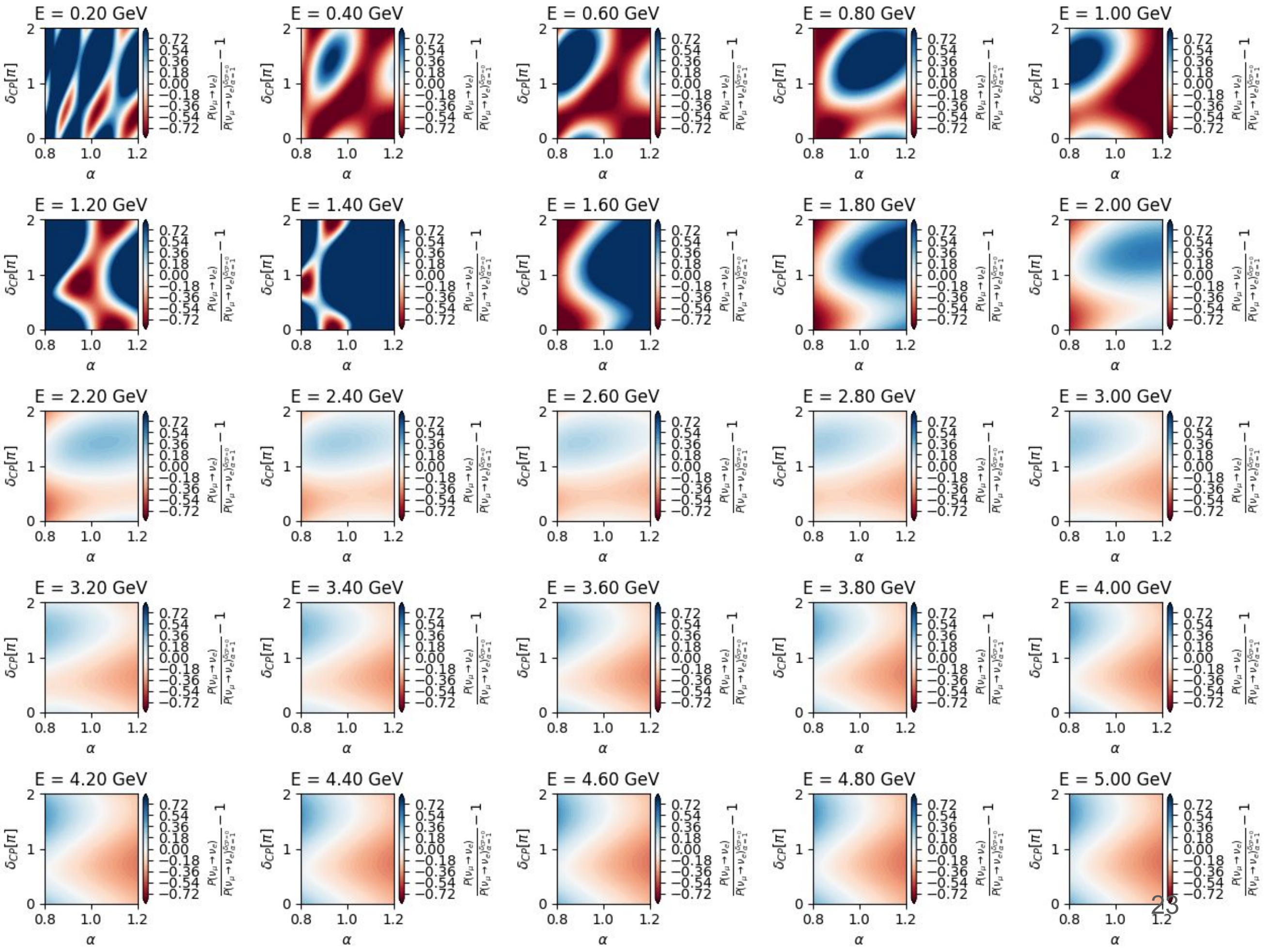
Probabilities from Prob3++ with:

$\sin^2\theta_{12}$ = 0.310
$\sin^2\theta_{13}$ = 0.02241
$\sin^2\theta_{23}$ = 0.580
Δm_{21} = 7.39e-5 eV^2
Δm_{Atm} = 2.525e-3 eV^2

Delta CP energy scale robustness - neutrinos

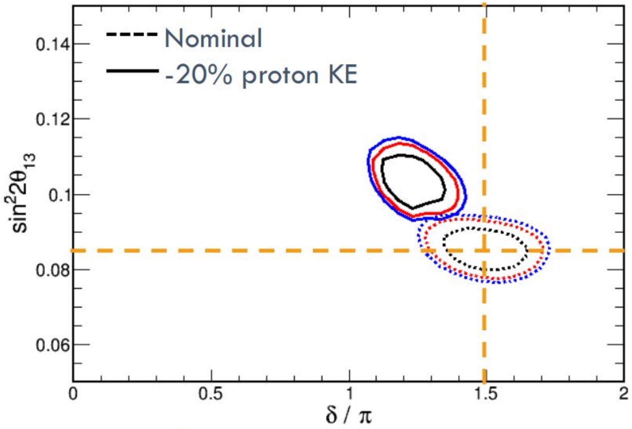
$E \rightarrow E' = aE$

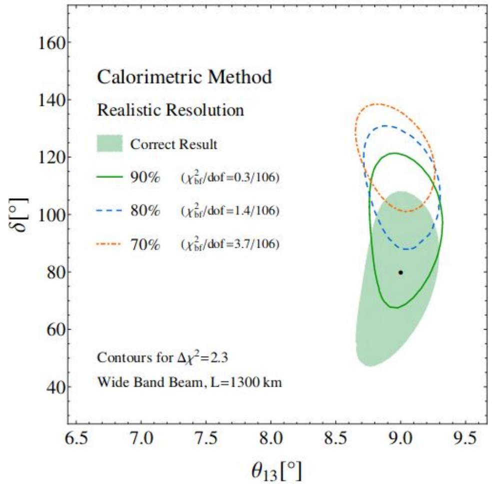
True atmospheric mass splitting known.



Probabilities from Prob3++ with:
$\sin^2\theta_{12} = 0.310$
$\sin^2\theta_{13} = 0.02241$
$\sin^2\theta_{23} = 0.580$
$\Delta m_{21} = 7.39\text{e-}5\ \text{eV}^2$
$\Delta m_{Atm} = 2.525\text{e-}3\ \text{eV}^2$

So what about this?



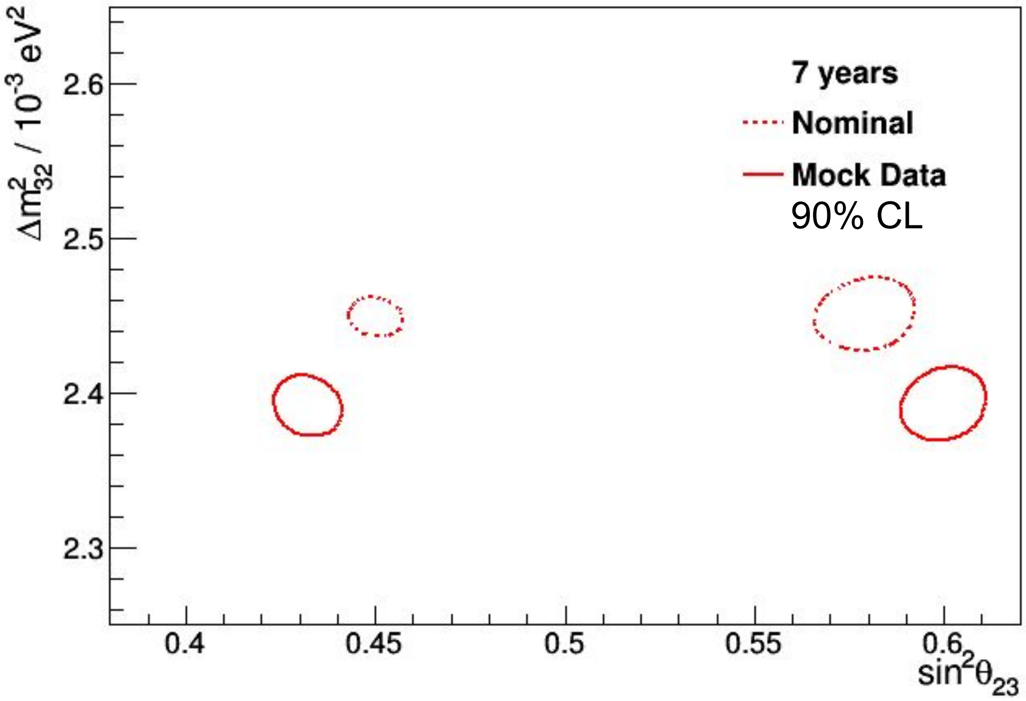


- In previous deltaCP bias plots we had fixed disappearance parameters at the nominal.
- Our intuition was that biased disappearance parameters would, if anything, **contribute** to deltaCP bias.
- Looks like this is a common assumption…

"Since the atmospheric parameters are fixed to their current best-fit values, and we are only interested in the δCP sensitivity,there is no need to include νμ and ν̄μ disappearance channels in our analysis."

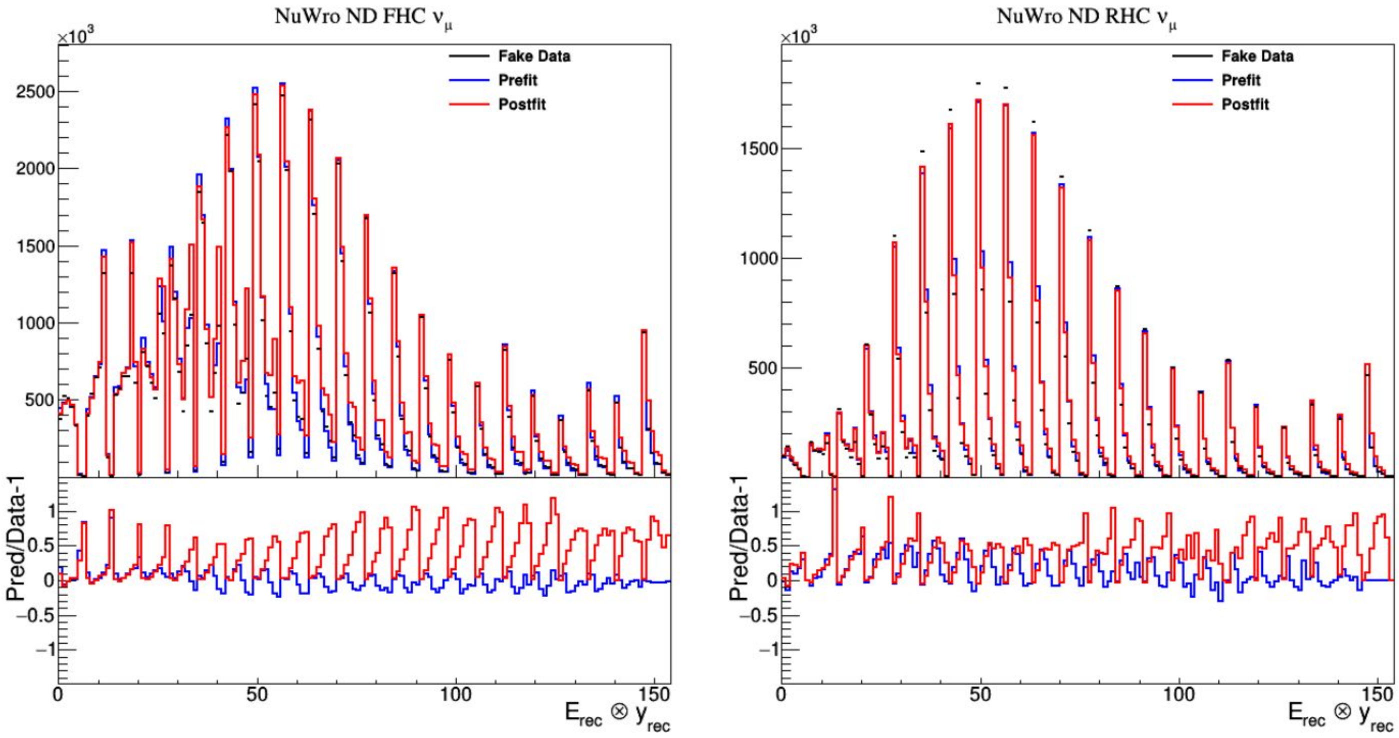
Phys. Rev. D 92, 091301 (2015)

Disappearance parameter bias with 20% missing proton energy



Near detector fits

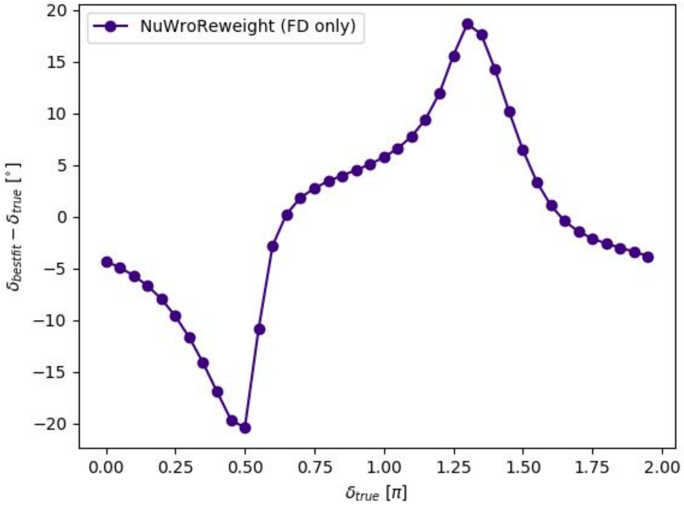
If nature was NuWro we would know something was up: χ^2 ~ 11000

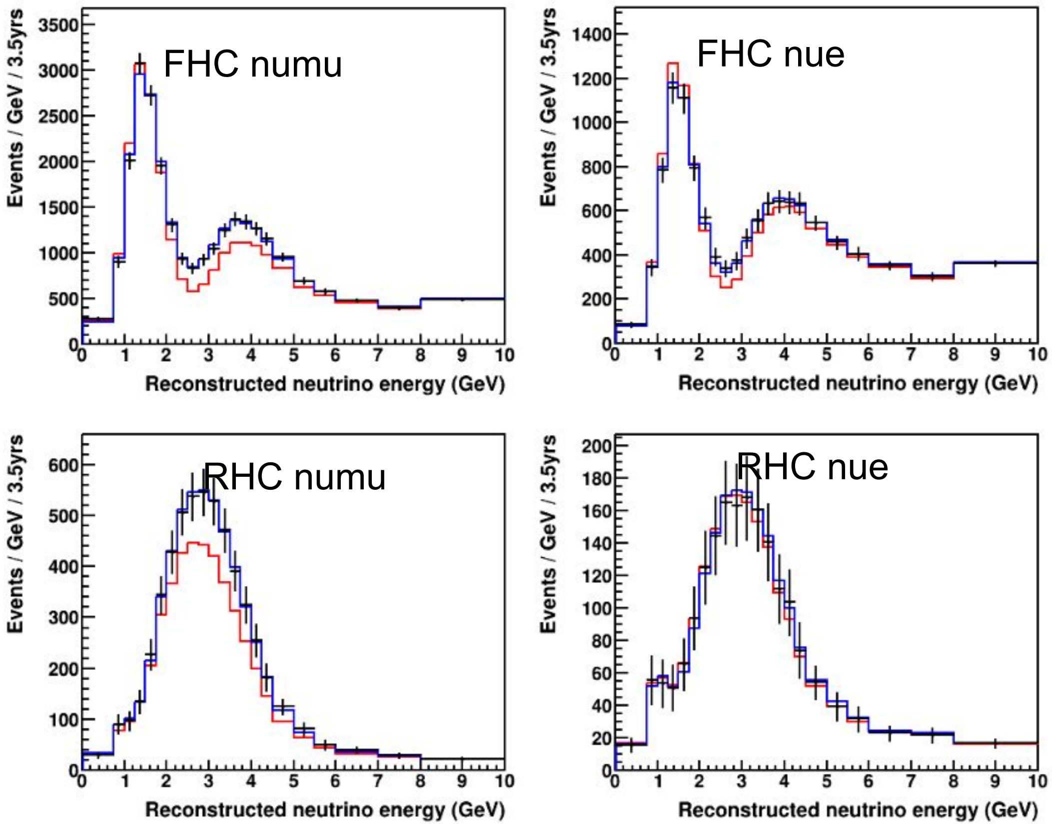


FD-only fit

Without a near detector we wouldn't…

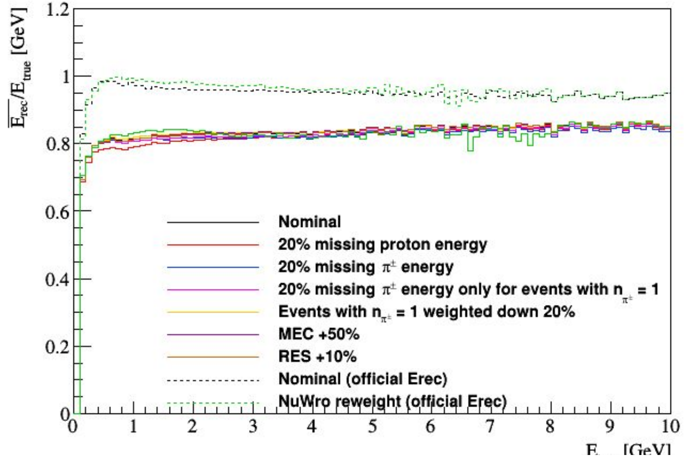
$\chi^2 \sim 10$

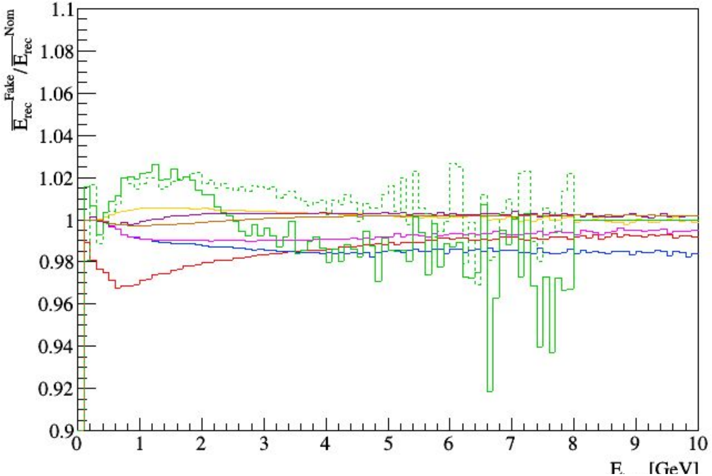




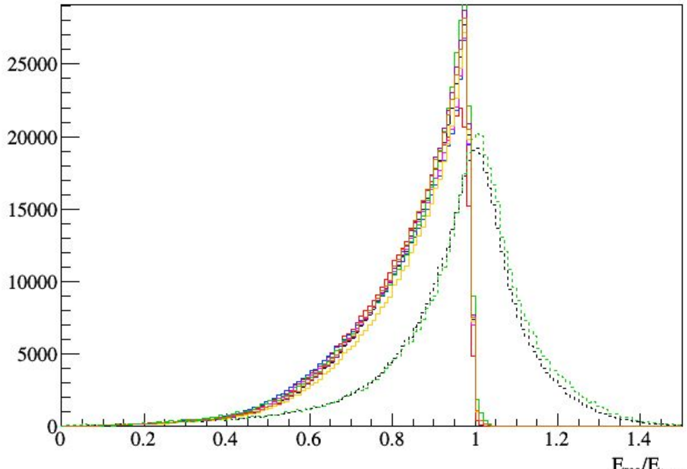
C. Marshall, yesterday

Missing proton energy alternatives - nue FHC

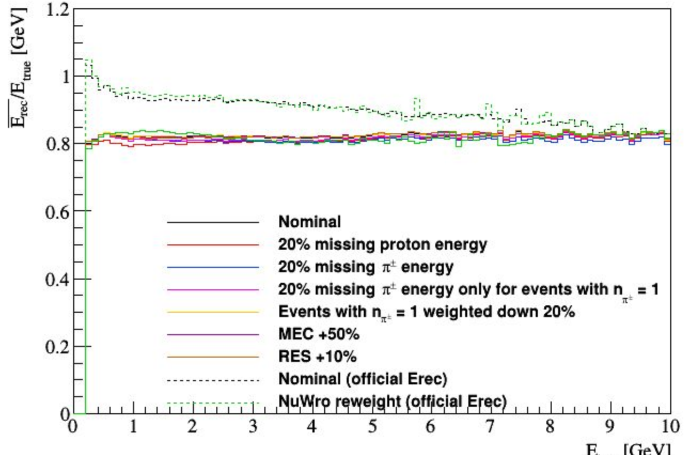


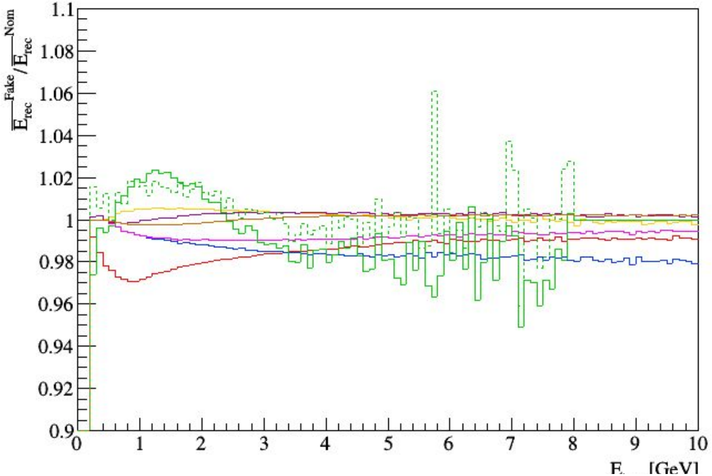


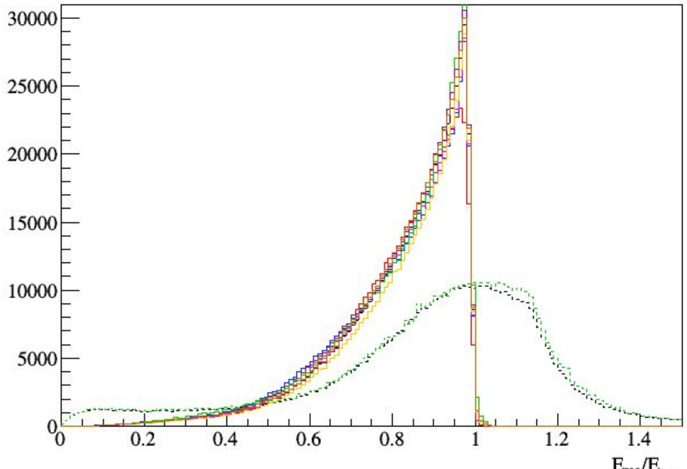
- I think we might need something that changes more violently around 1 - 4 GeV.
  - So that it doesn't look like an energy scale in the region where oscillation effects are larger.
    - And maybe that way the effect on oscillation parameters doesn't cancel out so much.
- Missing proton energy and NuWro seem to be the most violent of these…
  - More ideas?



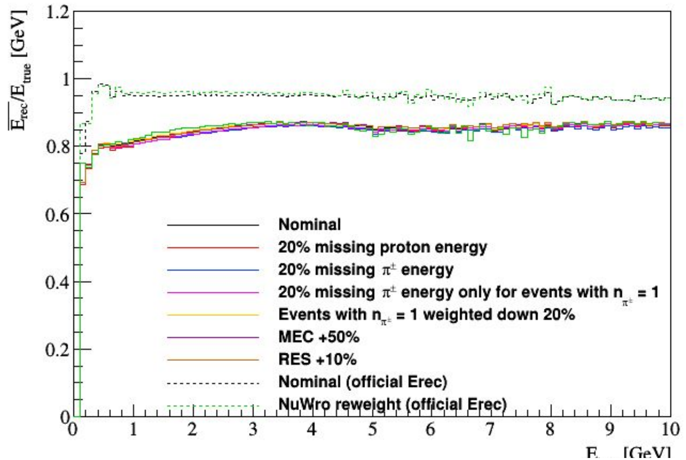
Missing proton energy alternatives - numu FHC

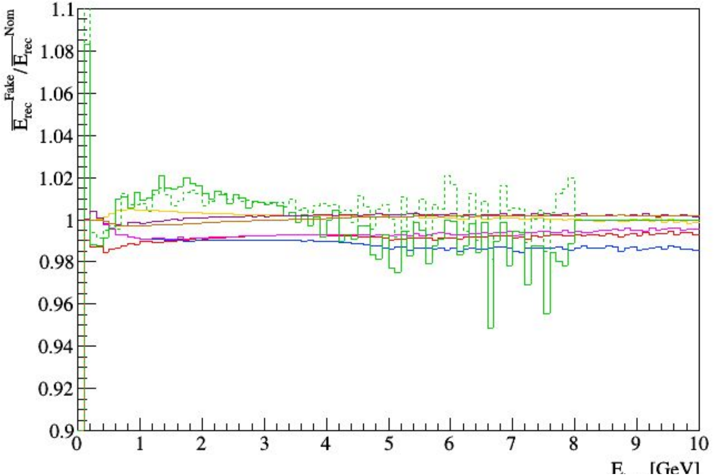


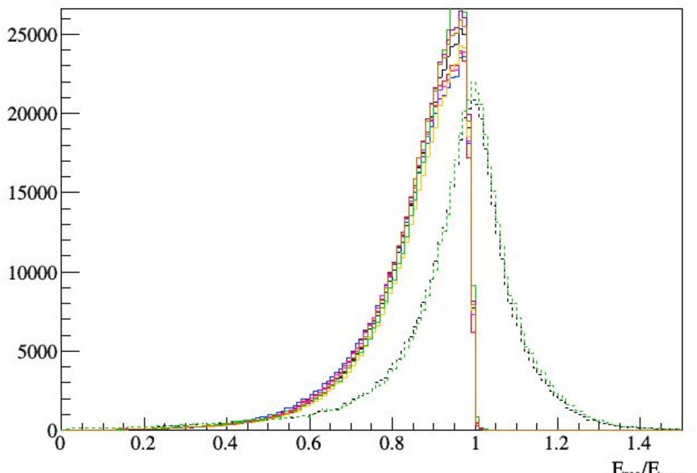




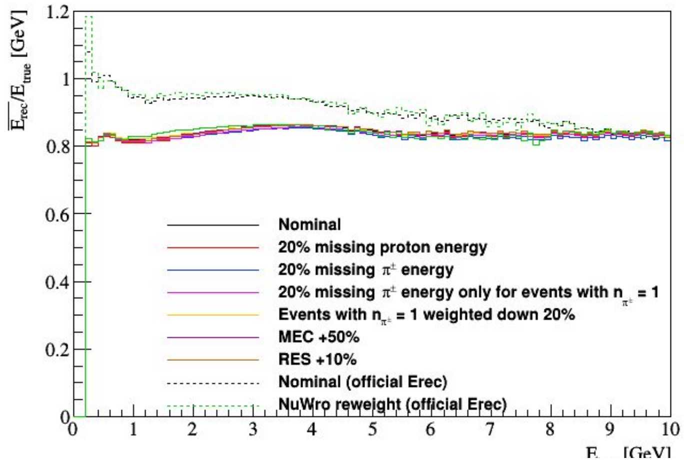
Missing proton energy alternatives - nue RHC

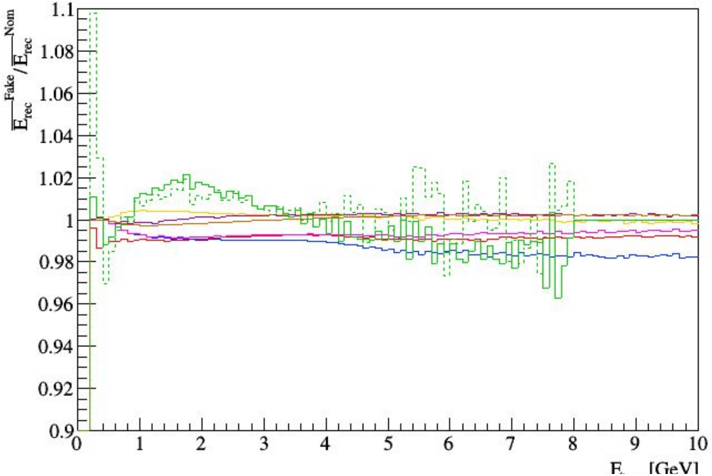


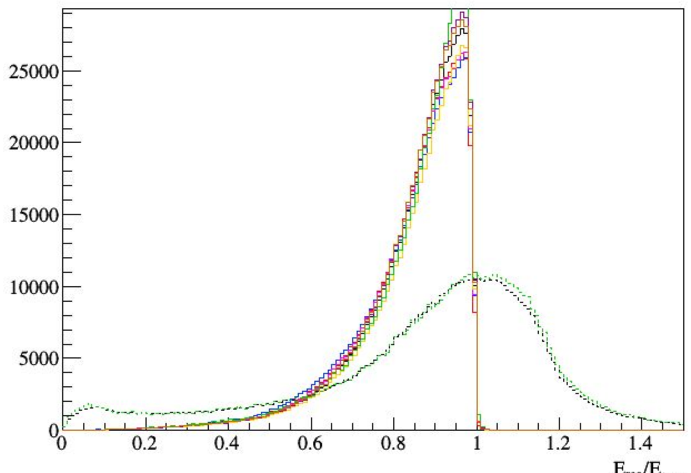




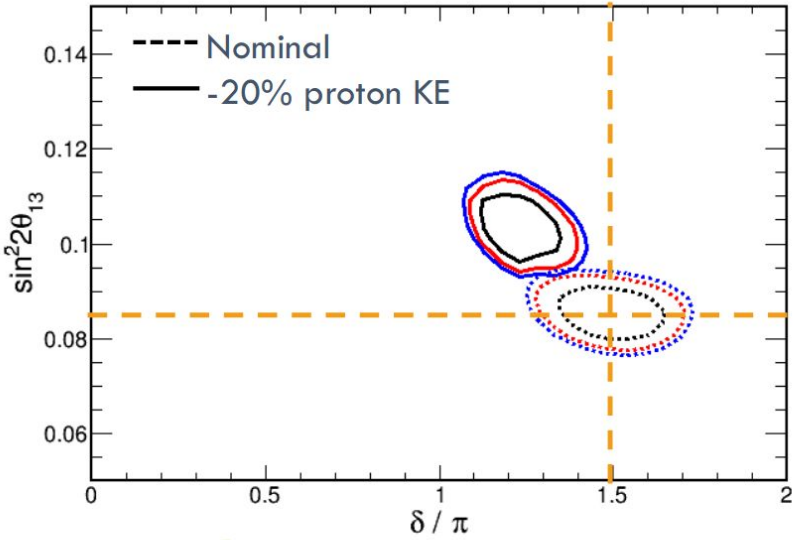
Missing proton energy alternatives - numu RHC

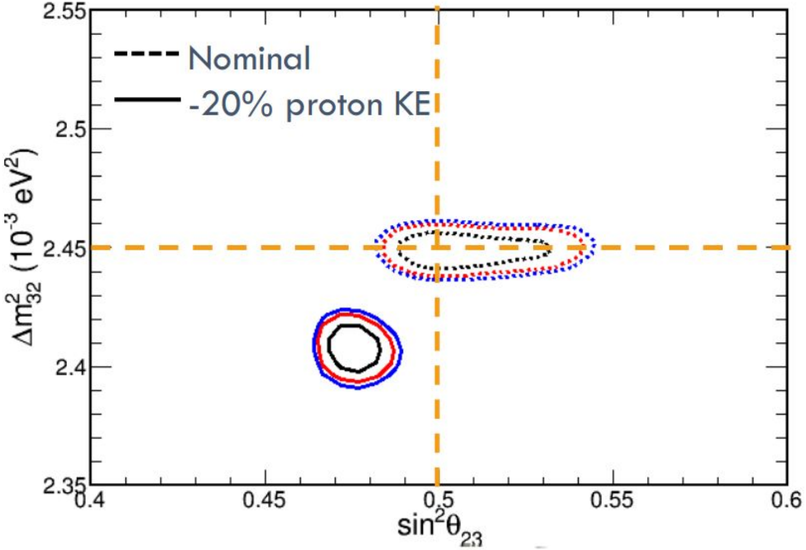






This is what we have presented before

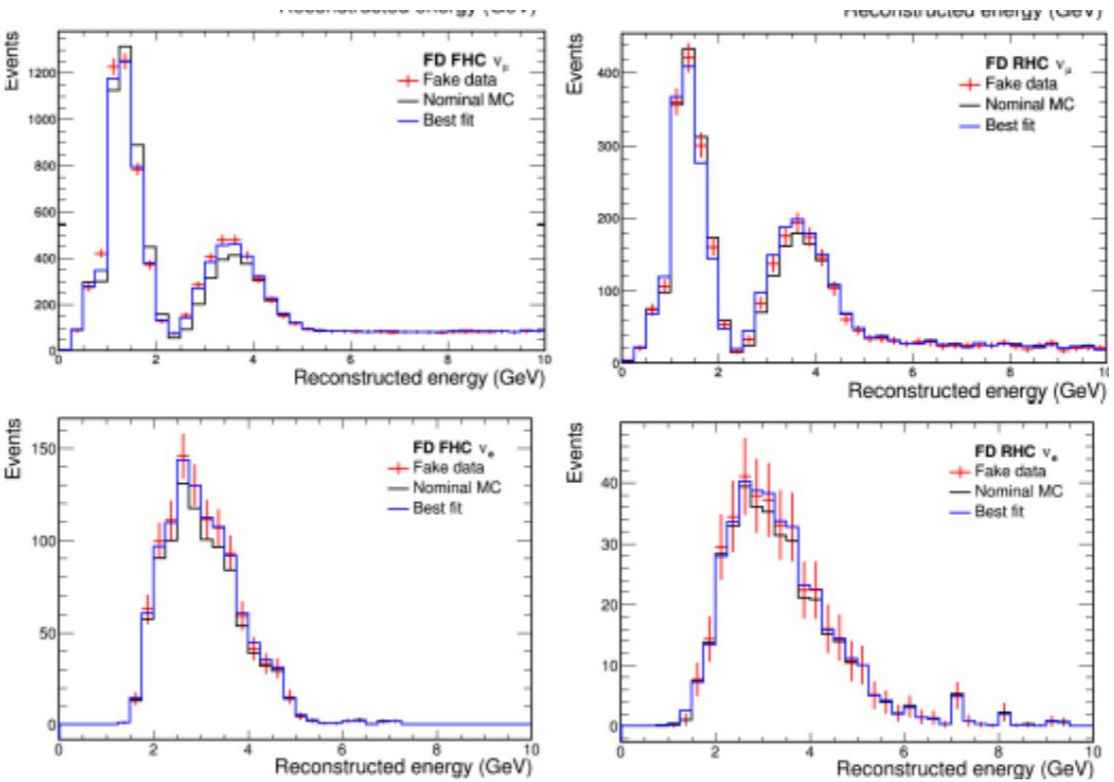




- Mass-squared bias: ~0.04e-3 eV*eV
- sinsq(theta_23) bias: ~0.025
- deltaCP bias: ~ 0.3 pi

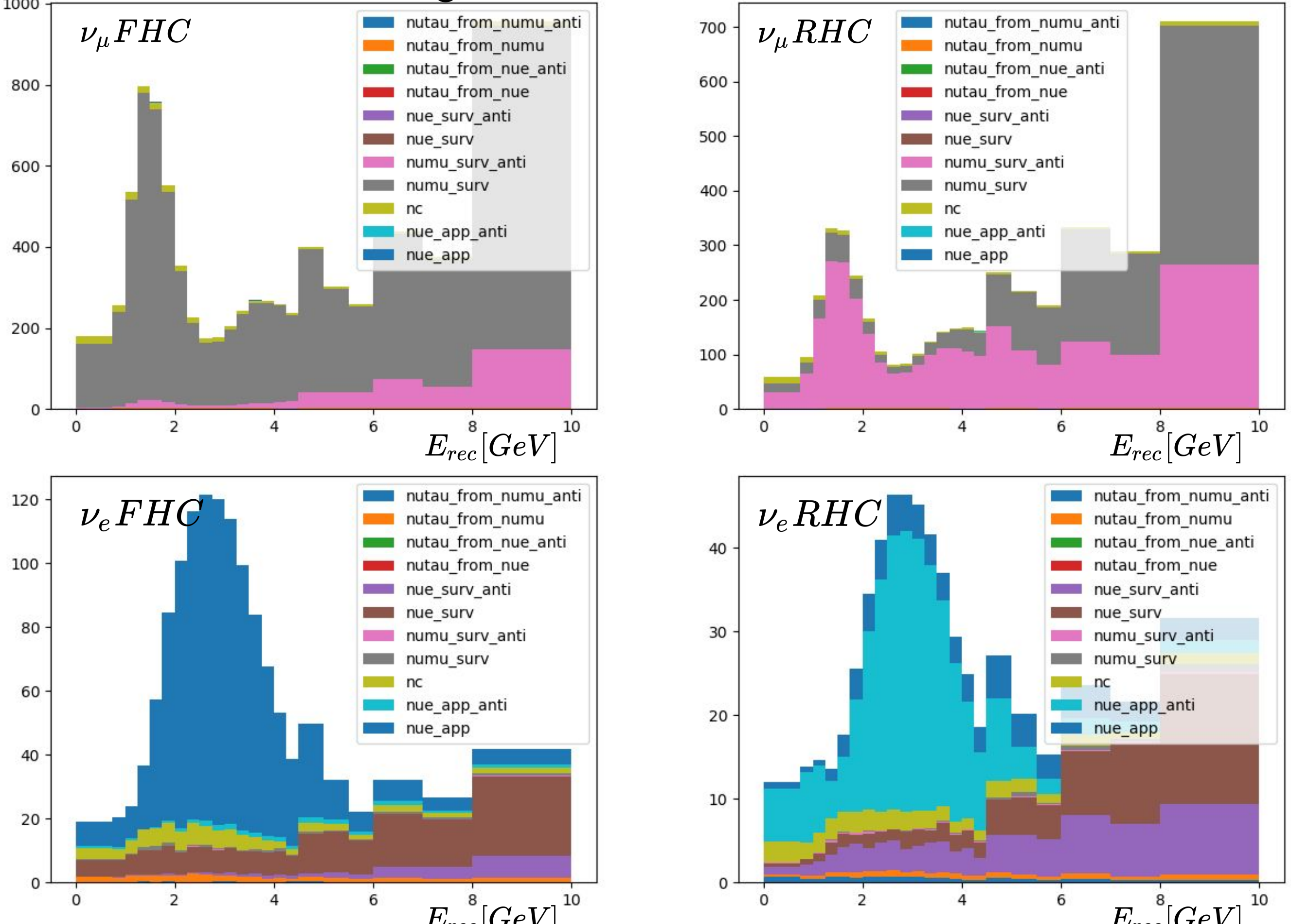
These are the spectra we have showed before

Background was not included on the nue samples (see next slide)

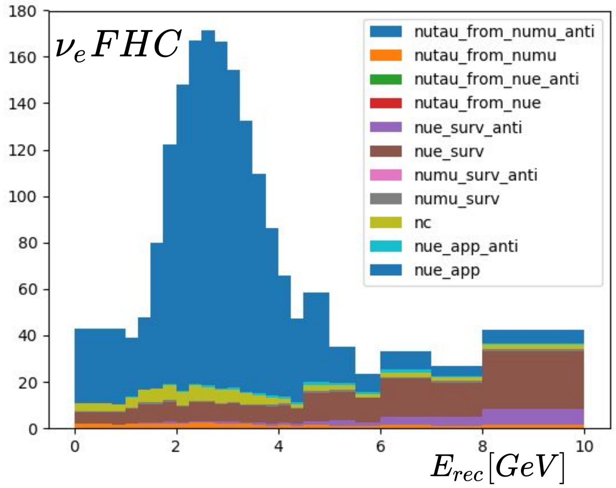


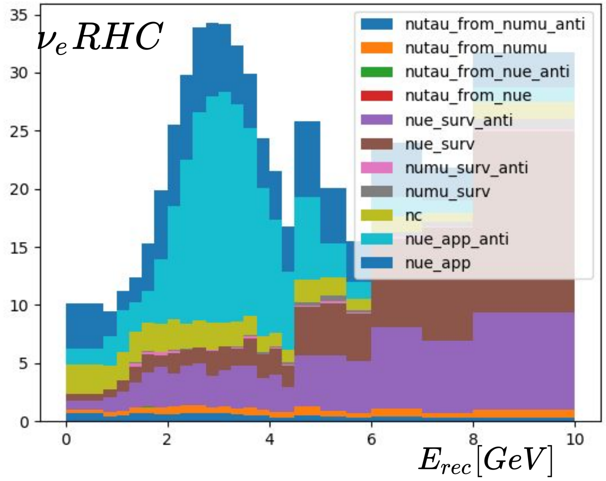
This is what the backgrounds look like

Also, new selection and binning

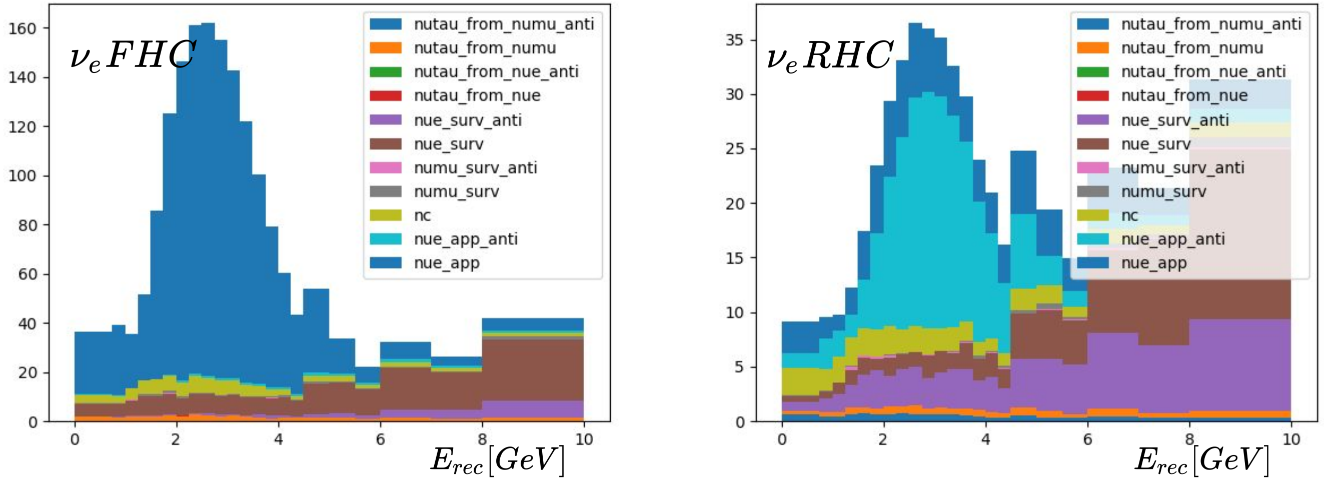


DeltaCP = 1.5 pi





DeltaCP = 1.2 pi



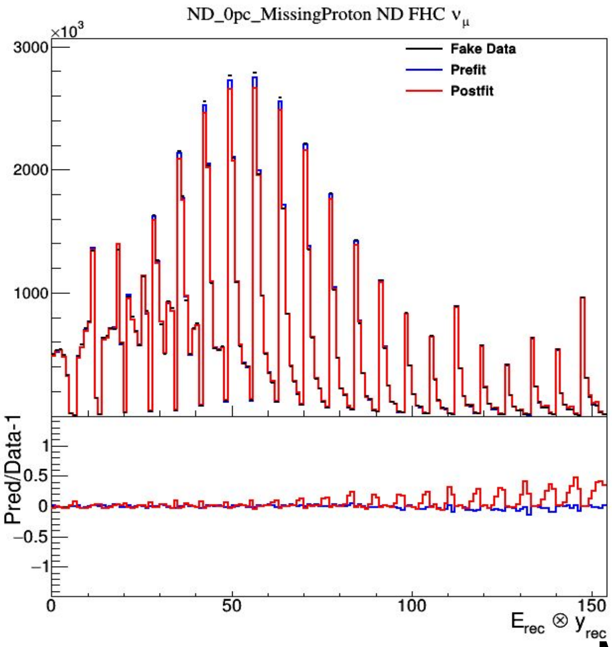
Fits to missing proton energy fake data

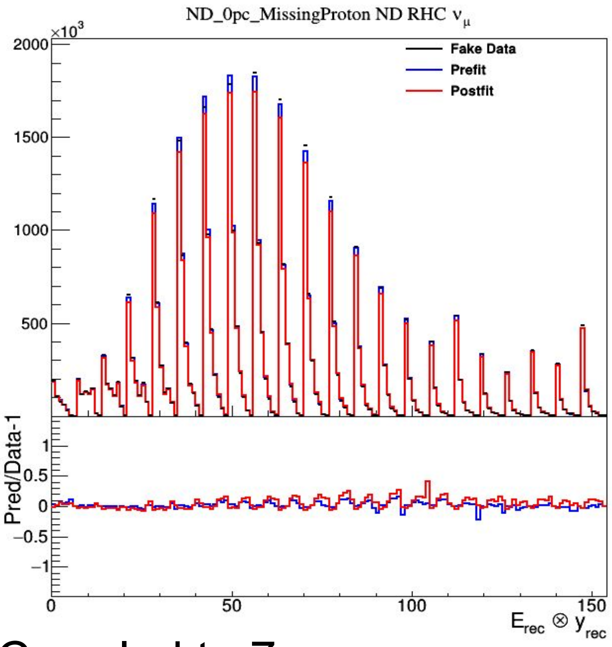
- Since last collaboration meeting, we integrated the missing proton fake data in the latest analysis tools and updated with latest inputs.

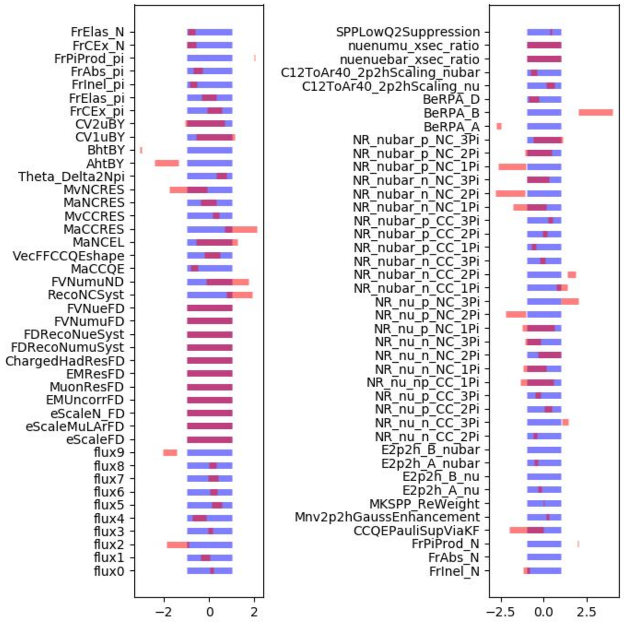
- Found that we don’t have enough near detector MC statistics to run full exposure ND+FD fits to fake data.
- Also found that while this fake data set introduces large biases in disappearance parameters, the effect on deltaCP is smaller than previously thought.

Near detector MC statistics

- We currently scale up our existing ND MC statistics (equivalent to ~4 months) to full exposure (x 20).
- This has a small effect on Asimov fits, but breaks fake data fit, likely due to event migration between bins.



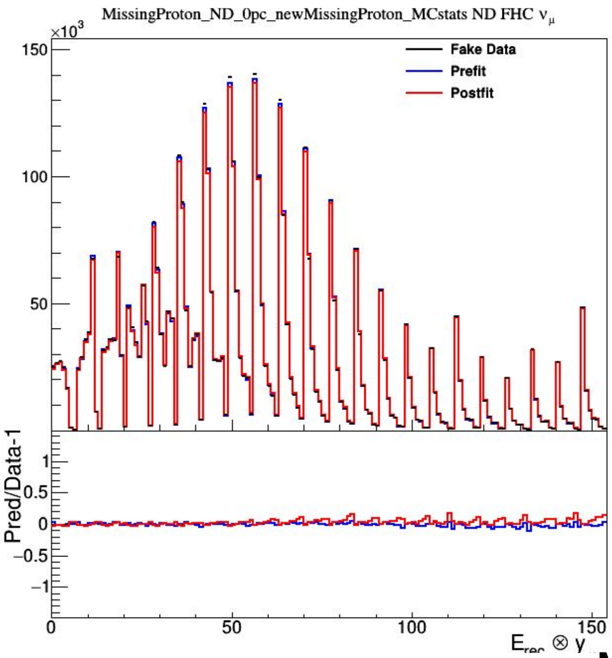


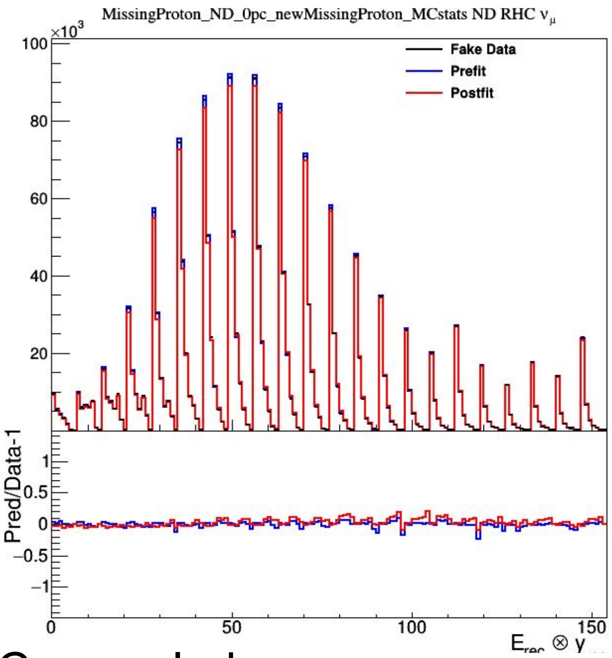


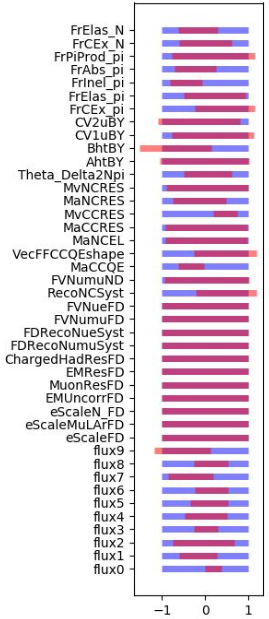
MC scaled to 7 years exposure

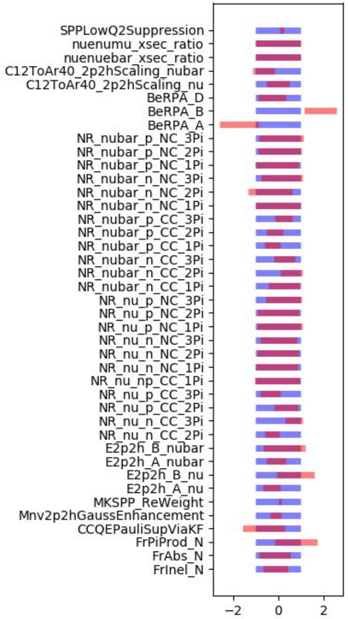
Near detector MC statistics

- With unscaled MC, get expected result from ND fake data fit.





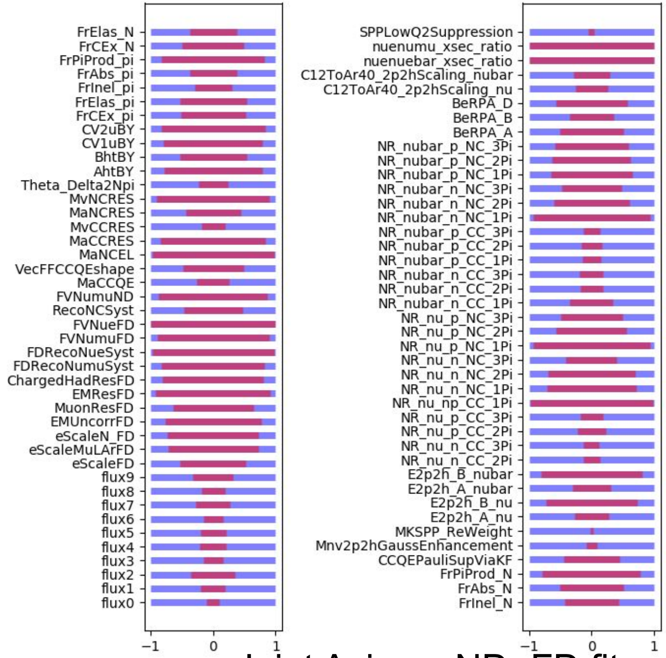




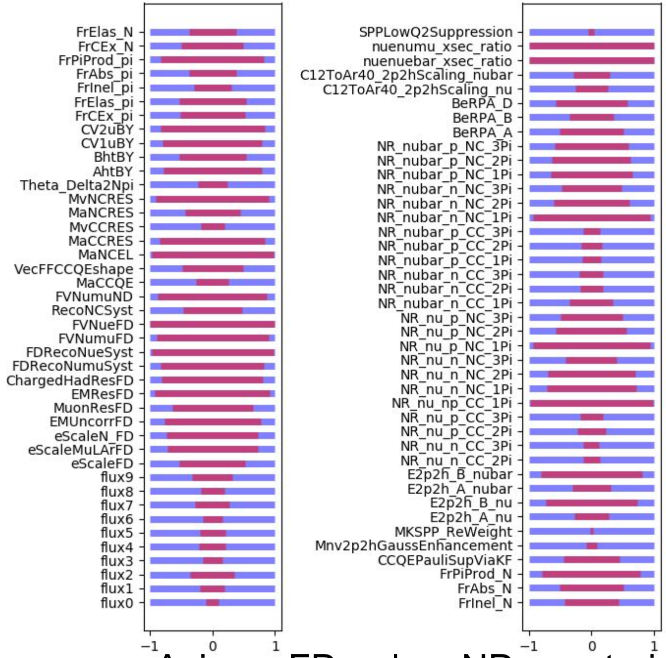
MC unscaled

Solution to limited ND MC statistics

- Generate more MC (Chris M., in progress)
- In the short term, run FD-only fake data fits with ND constraint on systematic parameters from 7 year exposure Asimov fit.

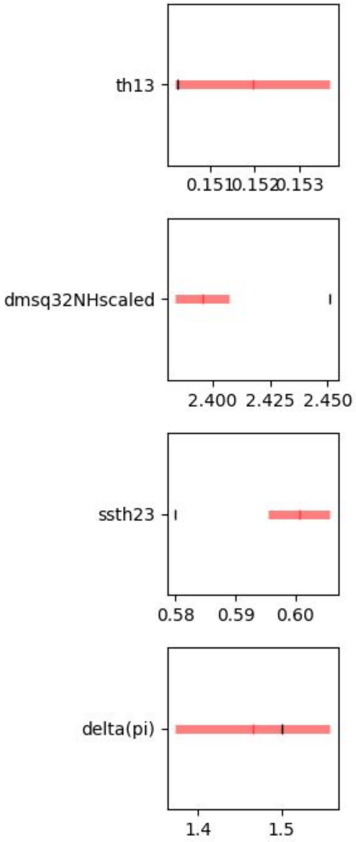


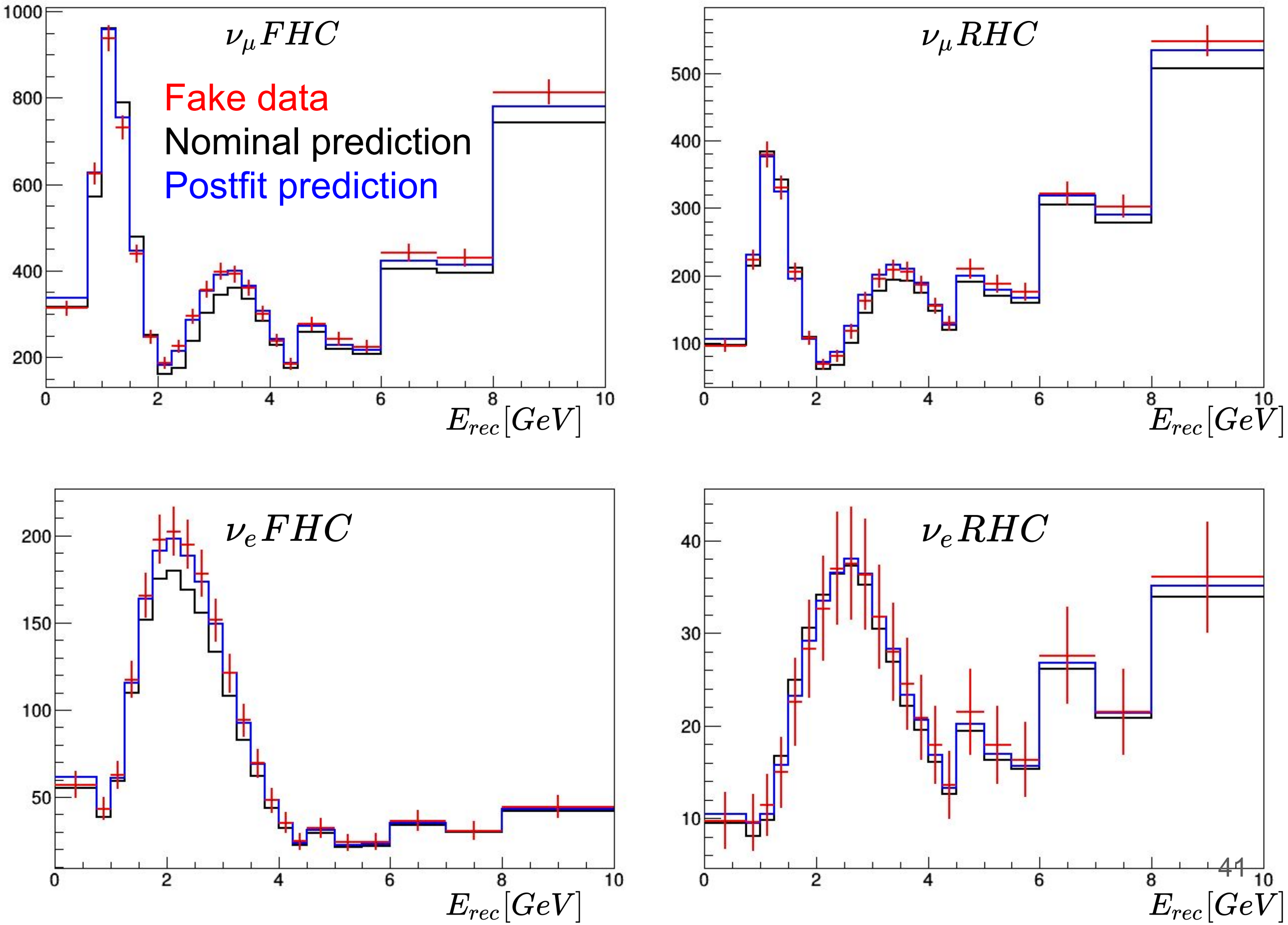
Joint Asimov ND+FD fit



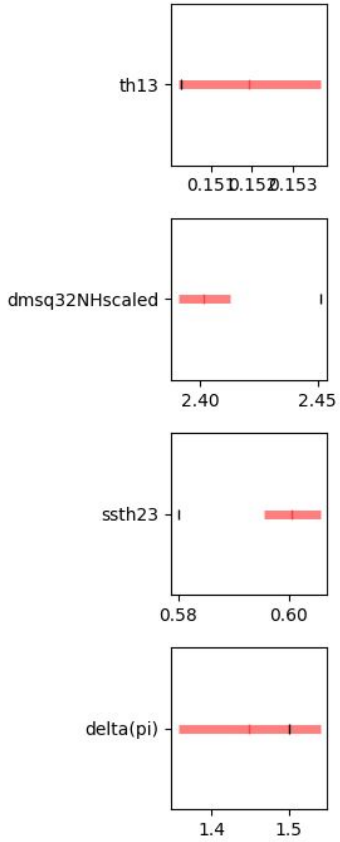
Asimov FD-only + ND constraint

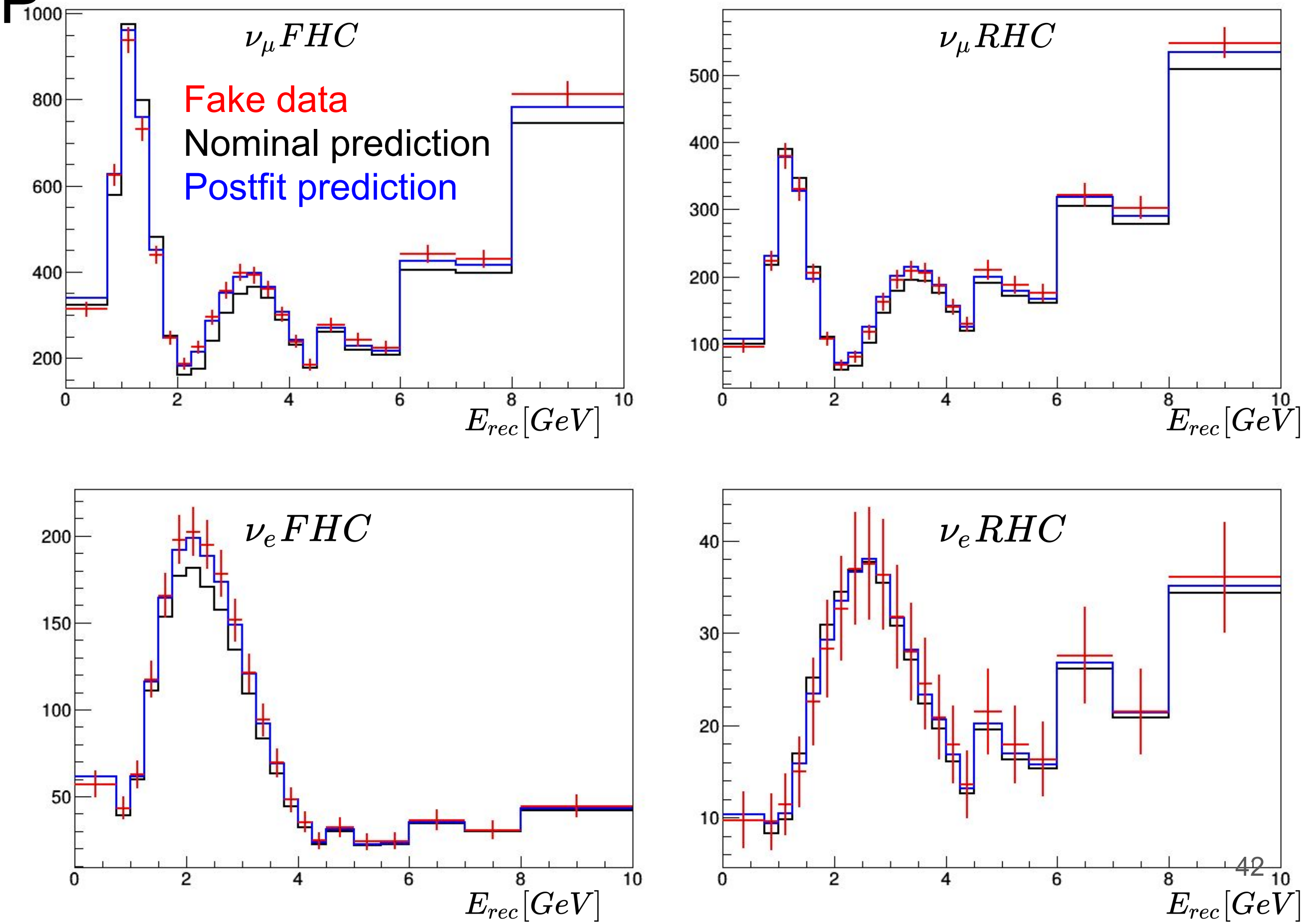
With th13 constrained to NuFit



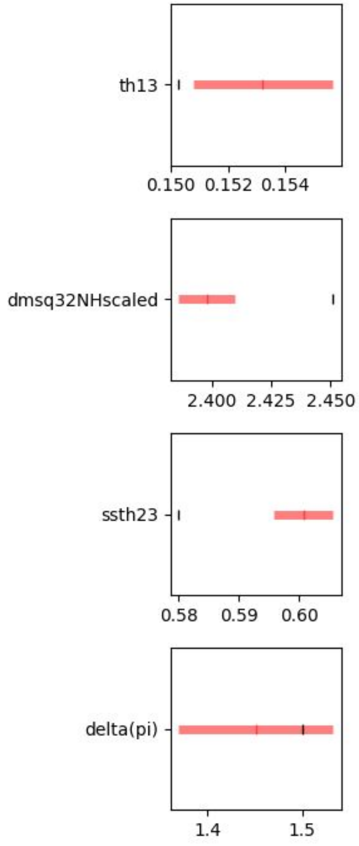


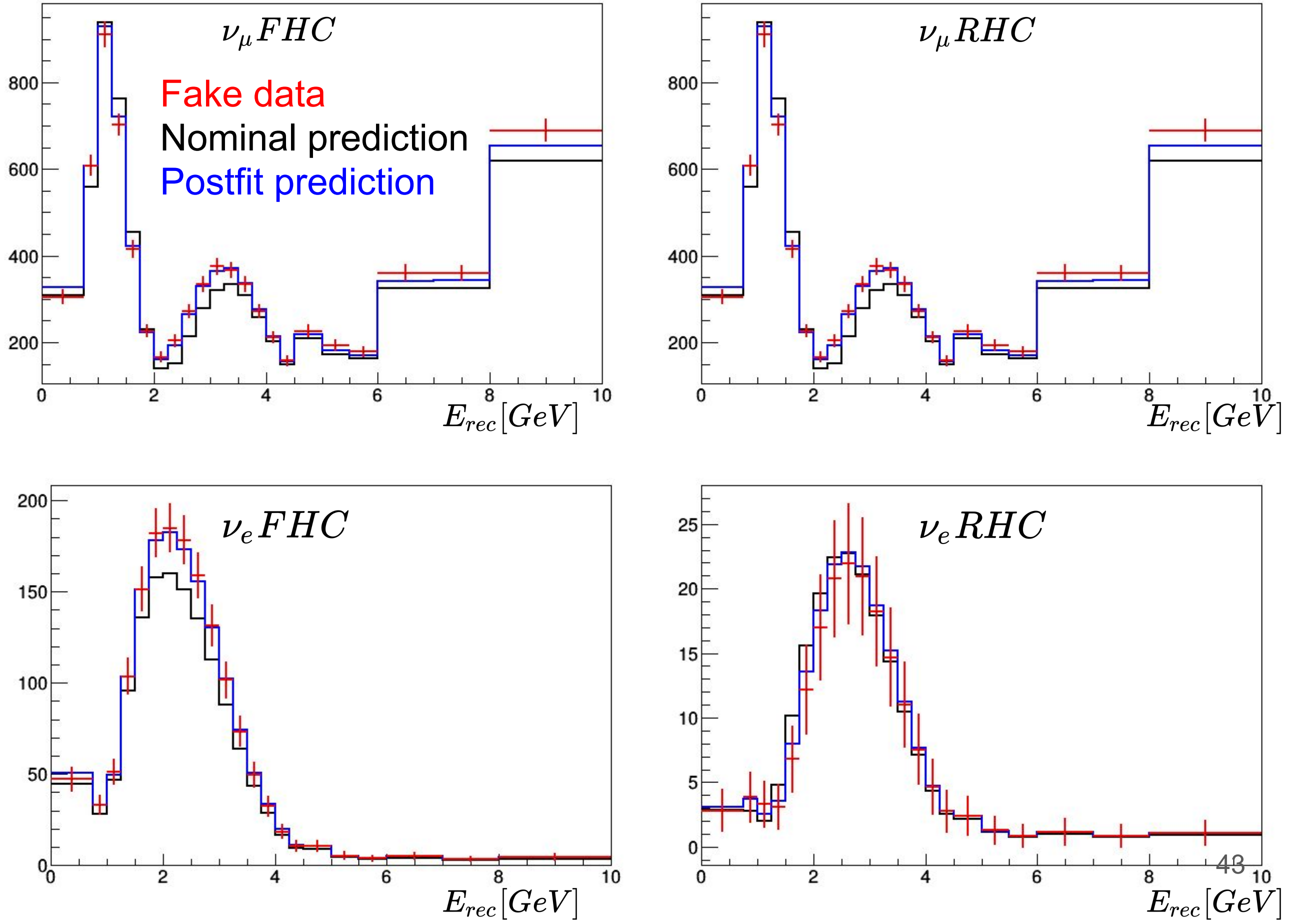
7 years exposure all oscillation parameters, NuFit constraint on all except deltaCP



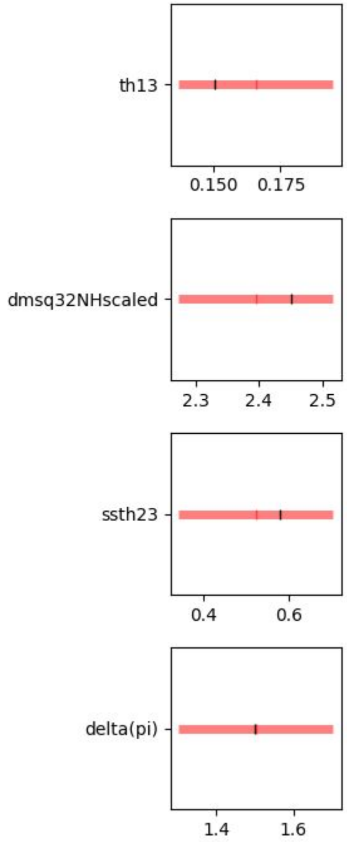


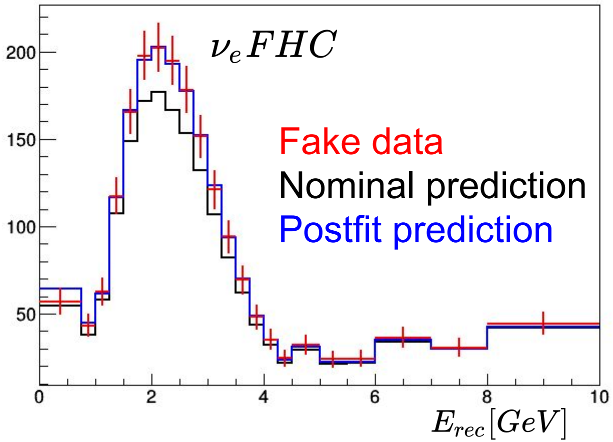
Without backgrounds

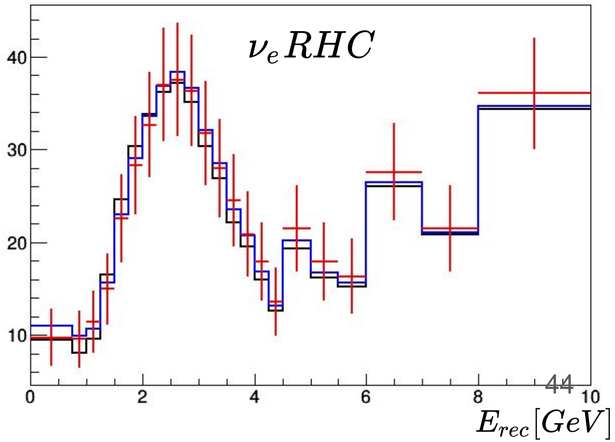




Appearance only, th13 unconstrained

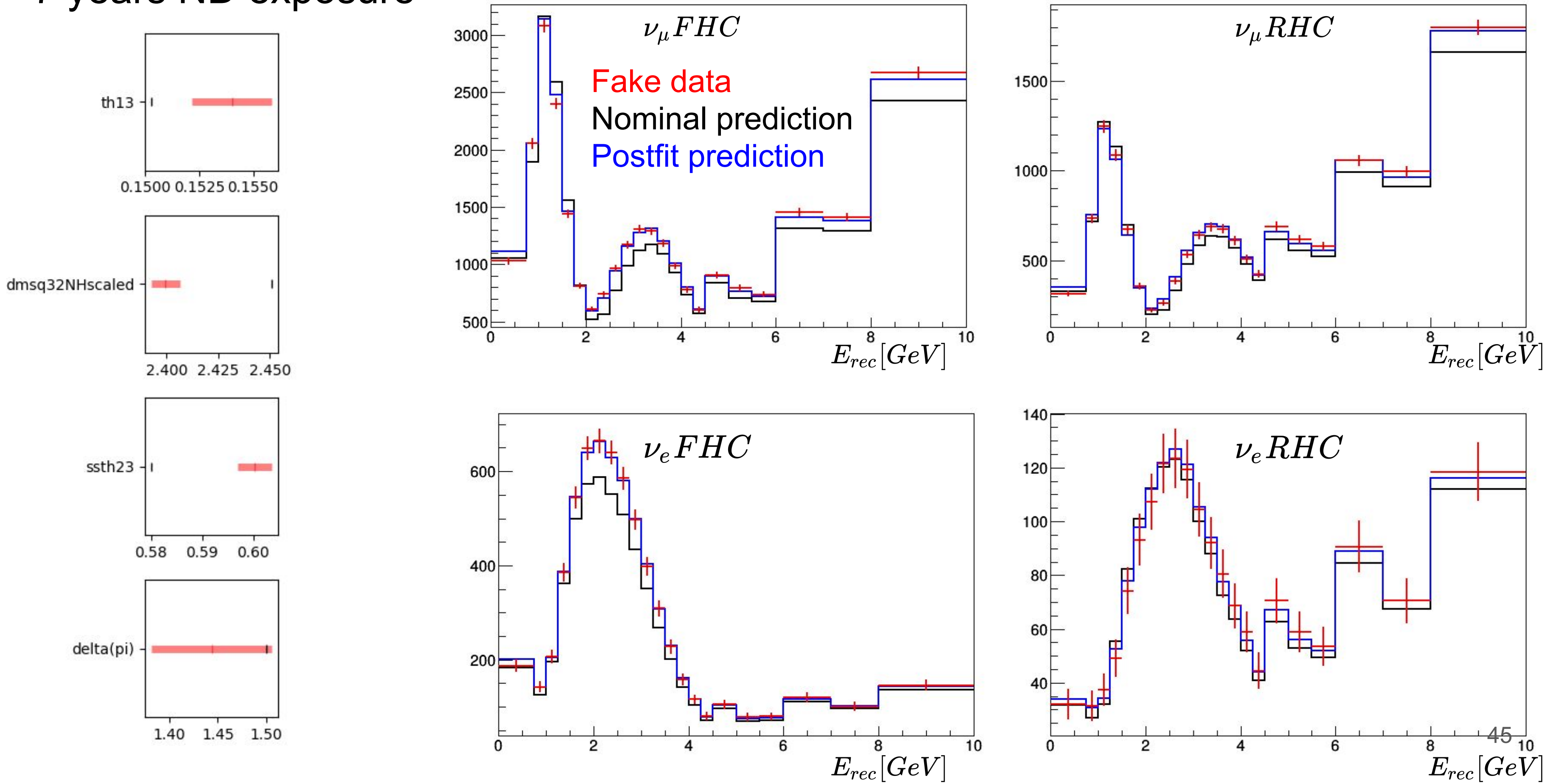




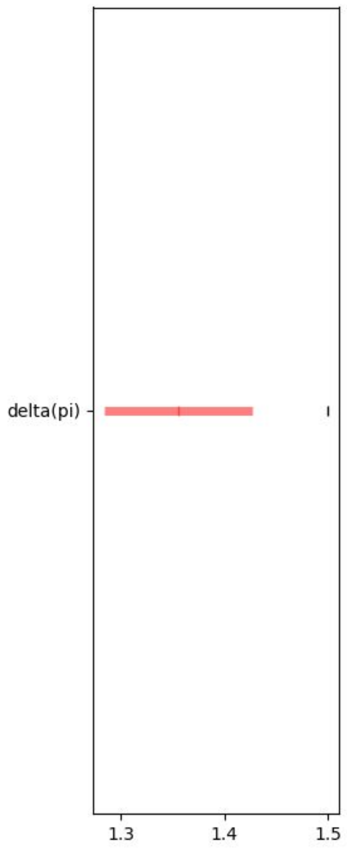


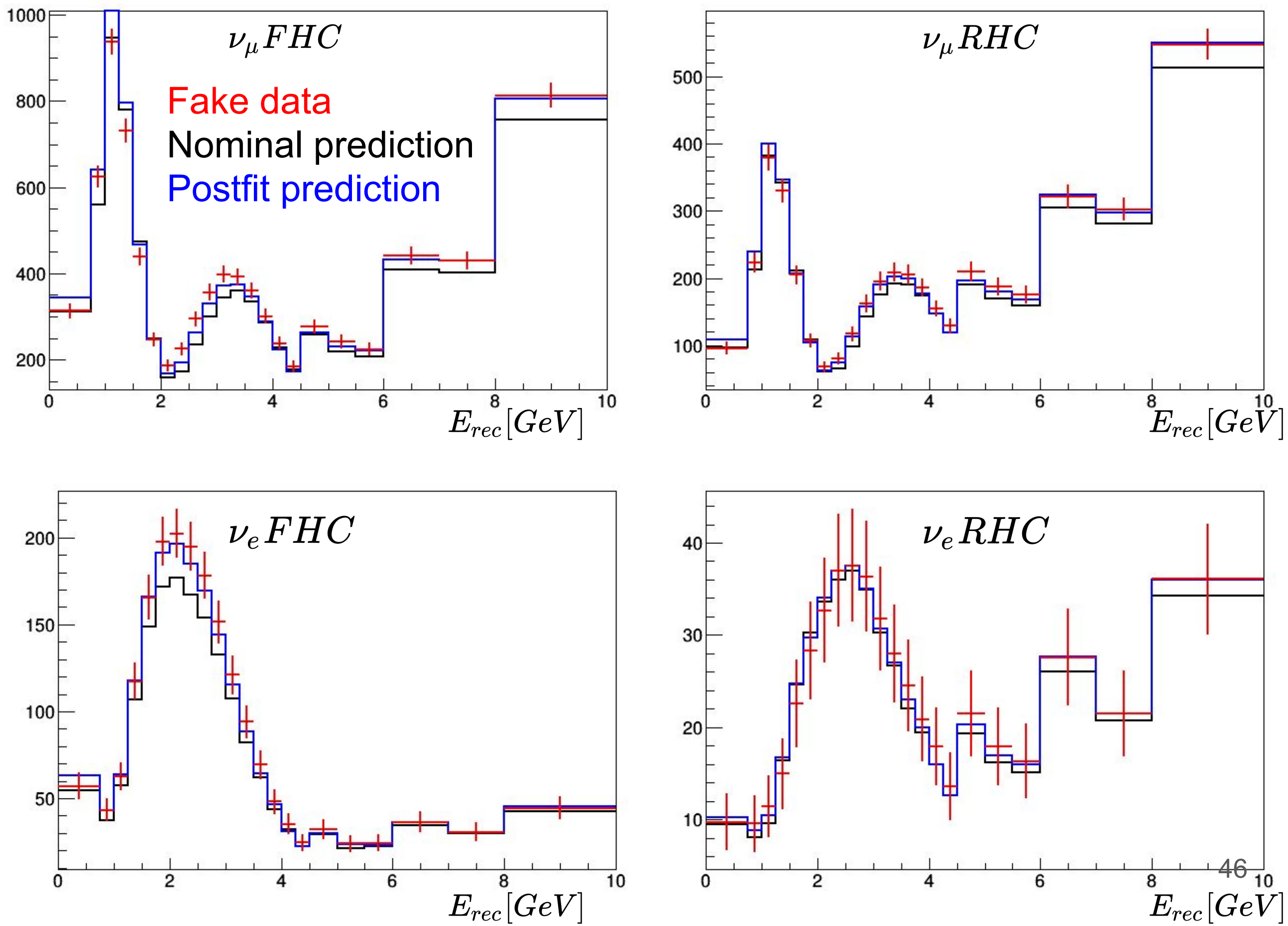
15 years exposure all oscillation parameters fitted

7 years ND exposure



All oscillation parameters fixed other than delta

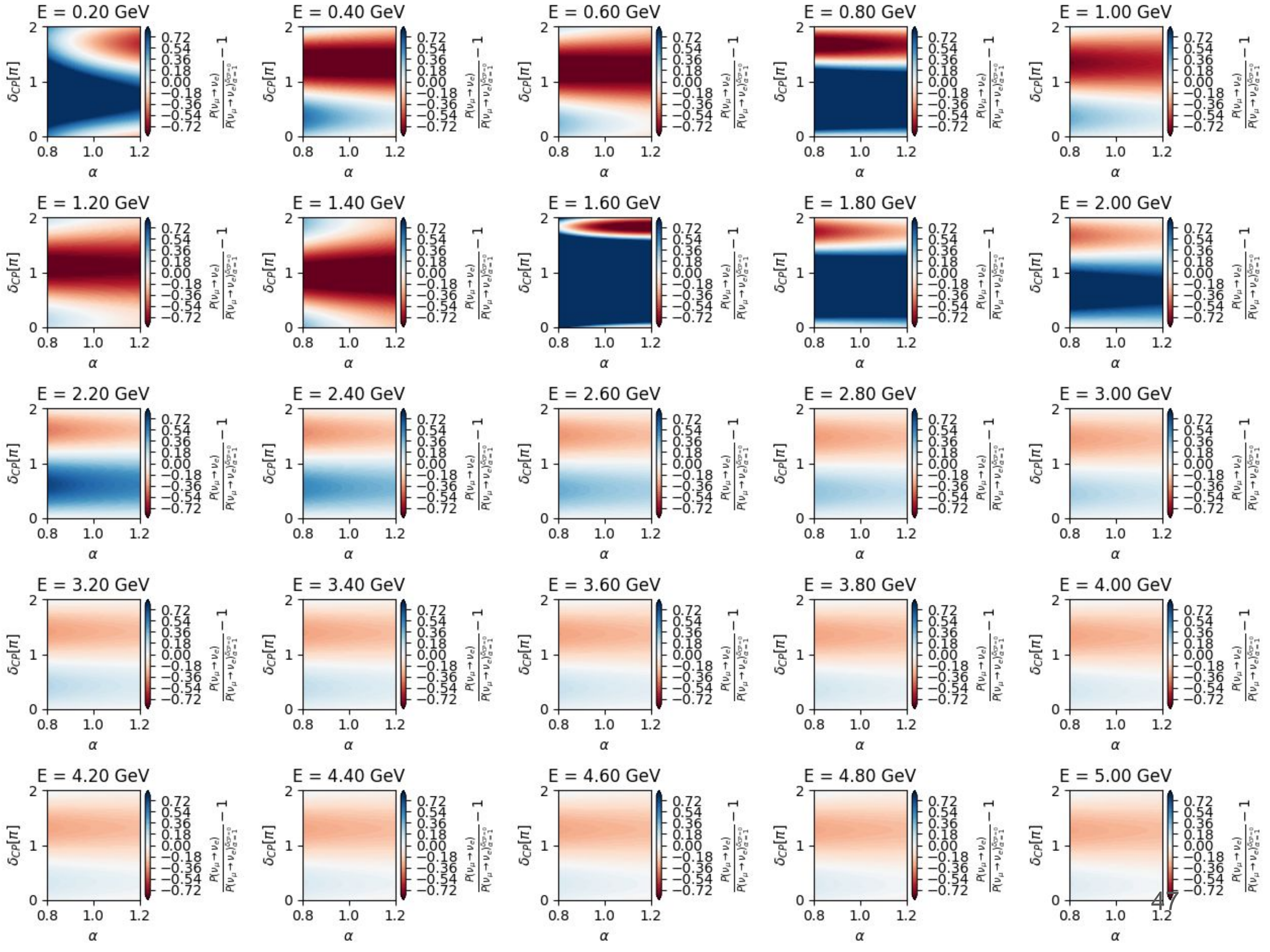




Delta CP energy scale robustness - antineutrinos

$$E \rightarrow E' = aE$$

$$\Delta m^2_{32} \rightarrow \Delta m'^2_{32} = a\Delta m^2_{32}$$



Probabilities from Prob3++ with:

$\sin^2\theta_{12}$ = 0.310

$\sin^2\theta_{13}$ = 0.02241

$\sin^2\theta_{23}$ = 0.580

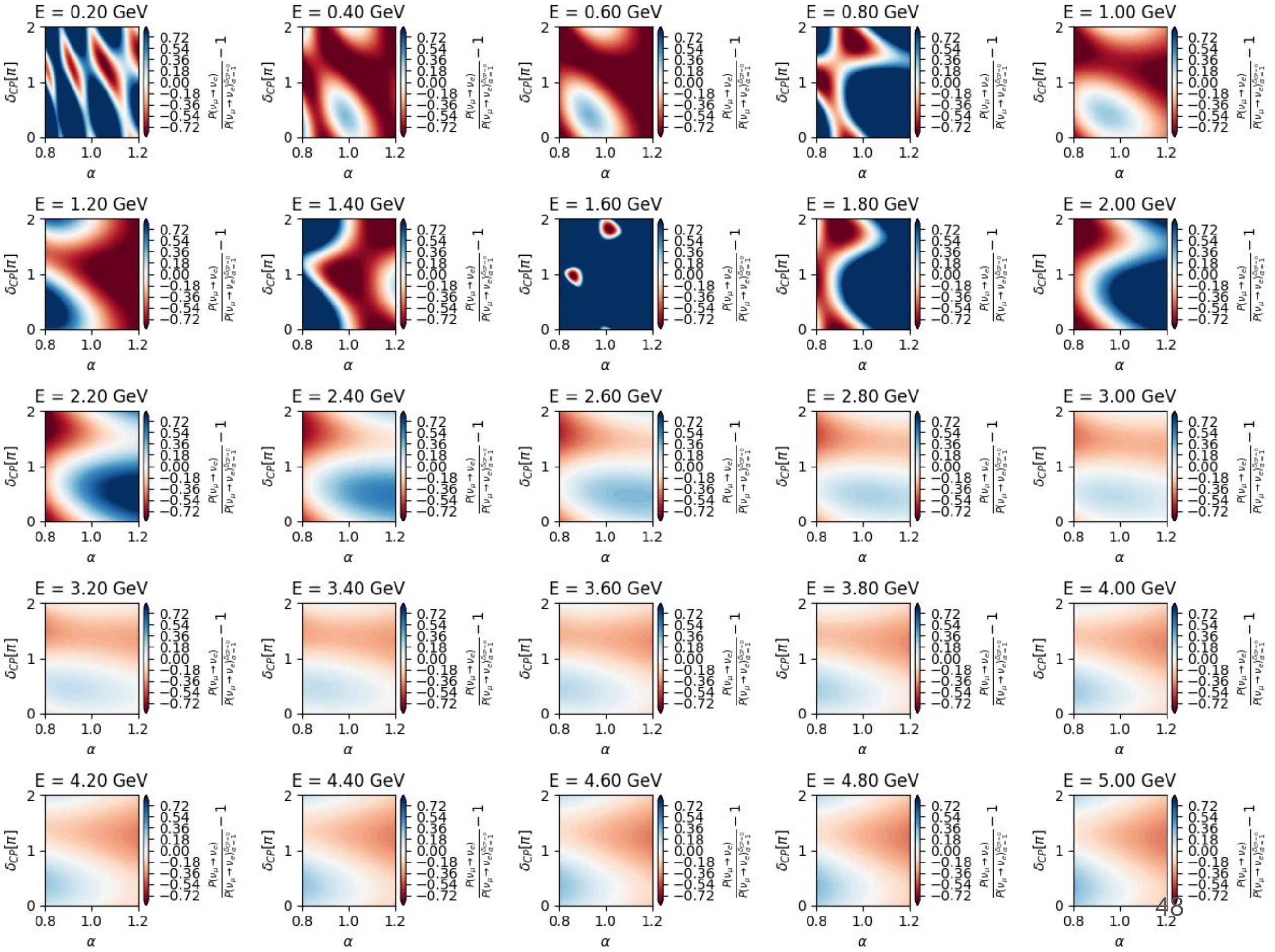
Δm_{21} = 7.39e-5 eV^2

Δm_{Atm} = 2.525e-3 eV^2

Delta CP energy scale robustness - antineutrinos

$$E \to E' = aE$$

True atmospheric mass splitting known.

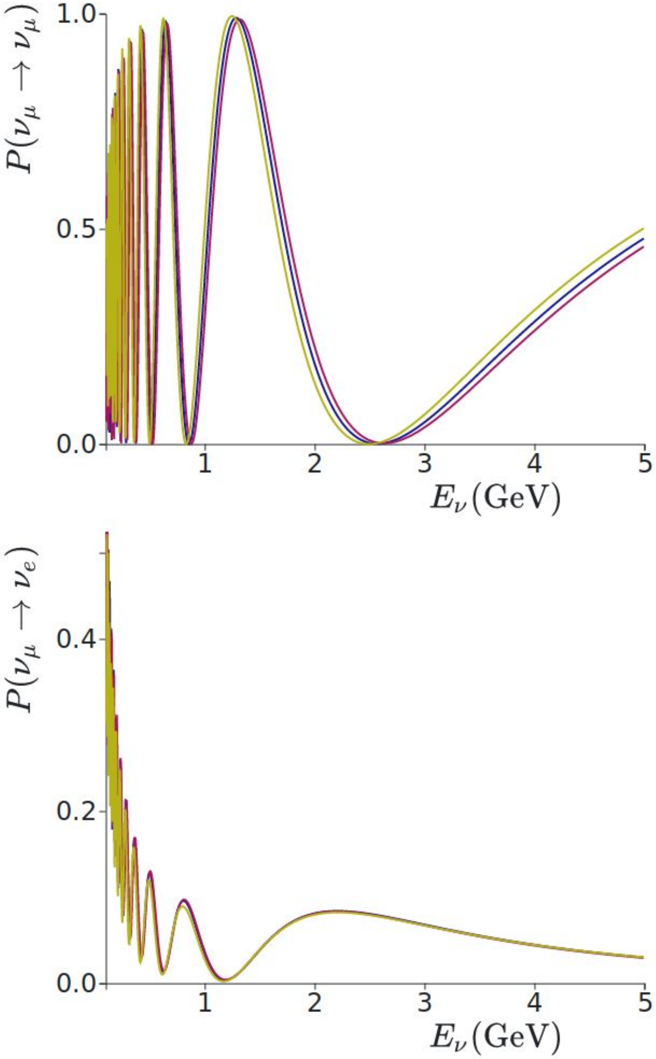


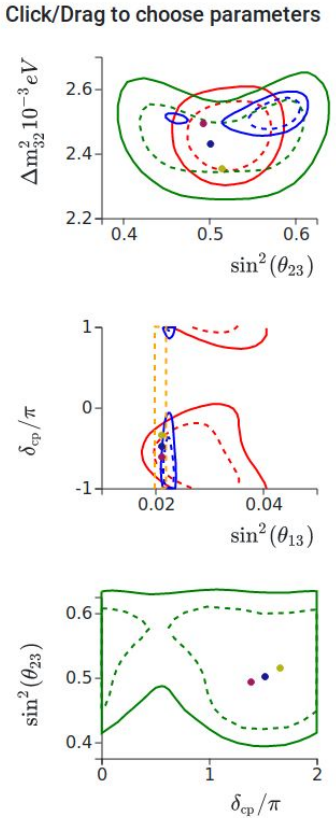
Probabilities from Prob3++ with:
$\sin^2\theta_{12}$ = 0.310
$\sin^2\theta_{13}$ = 0.02241
$\sin^2\theta_{23}$ = 0.580
Δm_{21} = 7.39e-5 eV^2
Δm_{Atm} = 2.525e-3 eV^2

Degeneracies

Neutrinos

- Disappearance parameters can be degenerate with deltaCP.





Made with L. Pickering's plotting tool.