Cross-section extraction using template fitting in T2K cross-section measurements

Margherita Buizza Avanzini
for the T2K collaboration

October 3rd 2023

Outline of the talk

1. Measuring a neutrino cross section
2. Binned likelihood fitters as unfolding methods allowing complex analyses
3. What can we learn from the post fit informations: lots of diagnostics
4. From the post fit results to the cross section extractions

List of T2K xsec publications that used binned LLH fitter

Phys. Rev. D 93, 112012 (2016),
CC0pi on CH

Phys. Rev. D 98, 012004 (2018),
CCINCL on CH

Phys. Rev. D 98, 032003 (2018),
STV, CC0piNp on CH

Phys. Rev. D 101, 112004 (2020),
joint O/C joint CC0pi

Phys. Rev. D 101, 112001 (2020),
joint $\nu/\bar{\nu}$ joint CC0pi on CH

Phys. Rev. D 102, 012007 (2020),
first $\bar{\nu}$ CC0pi on water

Phys. Rev. D 103, 112009 (2021),
first TKI CC1pi1p on CH

https://arxiv.org/abs/2303.14228,
first joint on/off-axis CC0pi on CH

Several joint analyses!

What is a cross section?

$$\frac{d\sigma}{dx_i dy_k} = \frac{N_{ik}^{signal}}{\epsilon_{ik}\Phi N_{nucleons}^{FV}} \times \frac{1}{\Delta x_i \Delta y_k}$$

What is a cross section?

After background subtraction and unfolding of detector effects

$$\frac{d\sigma}{\boxed{dx_i dy_k}} = \frac{\boxed{N_{ik}^{signal}}}{\boxed{\epsilon_{ik}}\Phi N_{nucleons}^{FV}} \times \frac{1}{\boxed{\Delta x_i \Delta y_k}}$$

true variables

efficiency correction

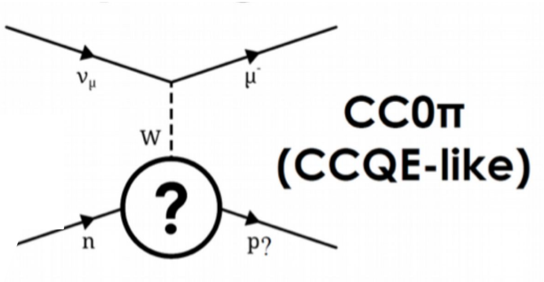
double (or more? or less?) differential

- Signal, to be defined considering the detector capabilities ⇒ **final state topology**
- Selected signal samples contain also some background ⇒ need of **background samples**
- Observables, to be chosen considering the detector capabilities (ND280 is tracking detector!) ⇒ **usually lepton and/or hadron kinematics**
- Limit interaction model dependence of the efficiency correction ⇒ perform **2D (or more) differential measurements**, phase space restriction,... *See more in S. Jenkins talk on thursday*
- Cross section usually extracted as a function of the true variables ⇒ **unfolding of detector effects**

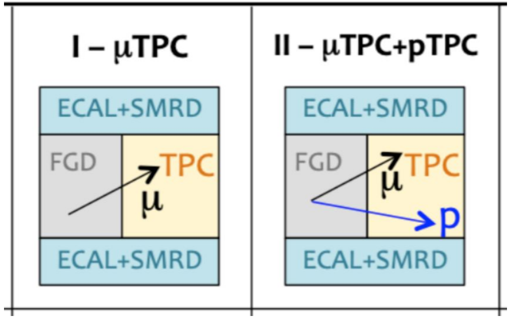
What do we need to measure a cross section?

(taking as an example the [Oxygen and Carbon CC0pi measurement from T2K])

A signal definition, for instance CC0pi, that lives in the truth space

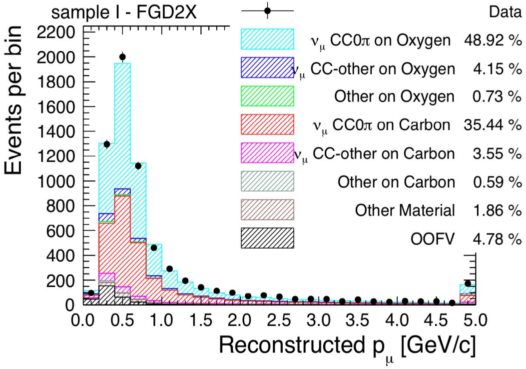


Signal selections, that apply on the reconstructed events with an adequate choice of **observables**, for instance lepton kinematics

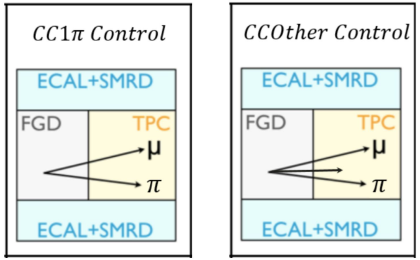


A Montecarlo prediction, that is the fundamental tool to:

1. have a first estimate of the background contamination and sample purity
2. move from the reco to the truth space (detector unfolding matrix and efficiency correction)
3. find the needed MC adjustments when compared to data



Background selections



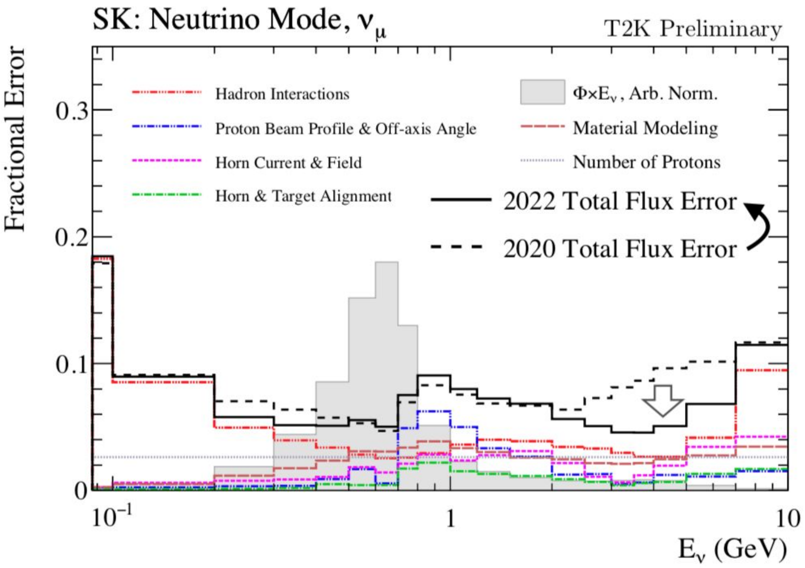
data driven constraint of the background remaining in the signal samples

What are the main systematic sources in a cross-section measurement?

Since your starting point is a Montecarlo prediction, you should take into account three main systematics sources that will affect it:

The neutrino flux prediction

In T2K, thanks to external experiment, we are able to quote the uncertainty on our flux predictions in bins of true neutrino energy



What are the main systematic sources in a cross-section measurement?

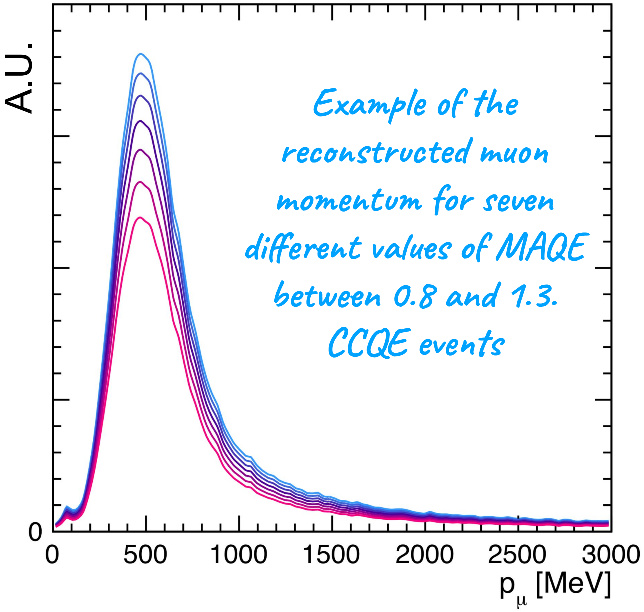
Since your starting point is a Montecarlo prediction, you should take into account three main systematics sources that will affect it:

The neutrino flux prediction

In T2K, thanks to external experiment, we are able to quote the uncertainty on our flux predictions in bins of true neutrino energy

The interaction model uncertainties

In T2K, we have a tool that allows to estimate the reweight to be applied to each event when we vary the value of specific parameters affecting the neutrino interaction predictions



What are the main systematic sources in a cross-section measurement?

Since your starting point is a Montecarlo prediction, you should take into account three main systematics sources that will affect it:

The neutrino flux prediction

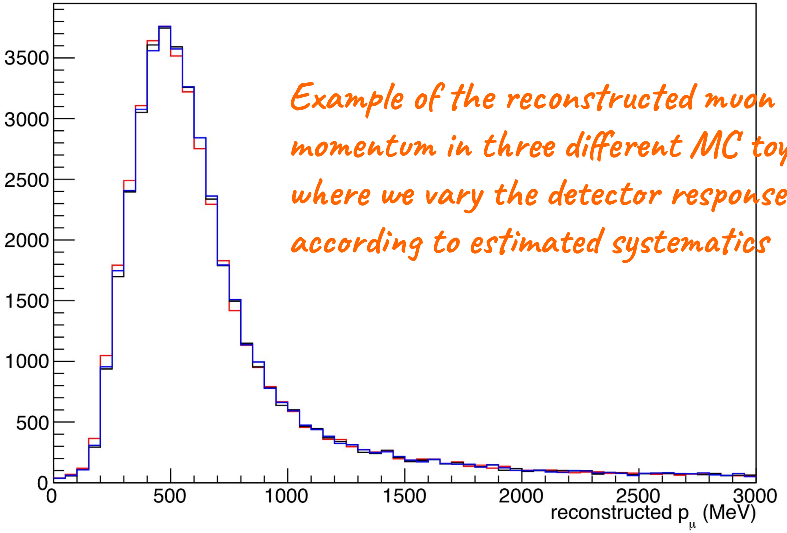
In T2K, thanks to external experiment, we are able to quote the uncertainty on our flux predictions in bins of true neutrino energy

The interaction model uncertainties

In T2K, we have a tool that allows to estimate the reweight to be applied to each event when we vary the value of specific parameters affecting the neutrino interaction predictions

The detector effects

In T2K, a series of detector systematics are estimated by comparing reconstruction results between the MC and data. Uncertainties can be propagated to the MC predictions



Example of the reconstructed muon momentum in three different MC toys where we vary the detector response according to estimated systematics

from [O and C CC0pi](#) analysis

What are the main systematic sources in a cross-section measurement?

Since your starting point is a Montecarlo prediction, you should take into account three main systematics sources that will affect it:

The neutrino flux prediction

In T2K, thanks to external experiment, we are able to quote the uncertainty on our flux predictions in bins of true neutrino energy

The interaction model uncertainties

In T2K, we have a tool that allows to estimate the reweight to be applied to each event when we vary the value of specific parameters affecting the neutrino interaction predictions

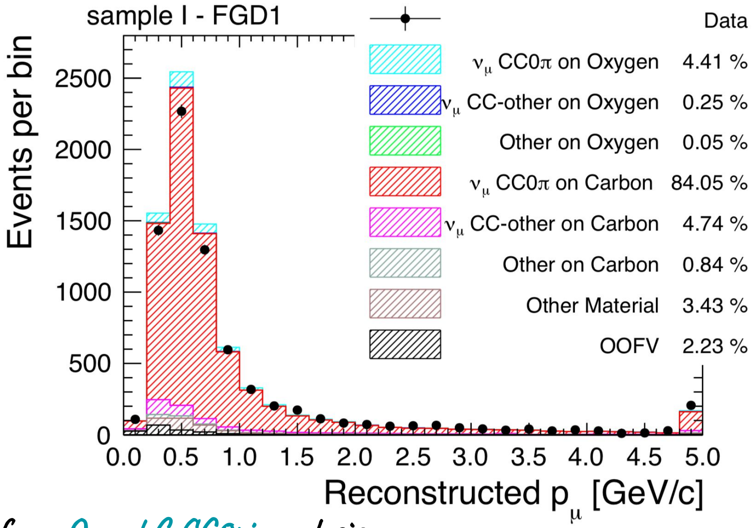
The detector effects

In T2K, a series of detector systematics are estimated by comparing reconstruction results between the MC and data. Uncertainties can be propagated to the MC predictions

$$\frac{d\sigma}{dx_i dy_k} = \frac{N^{signal}_{ik}}{\epsilon_{ik}\Phi N^{FV}_{nucleons}} \times \frac{1}{\Delta x_i \Delta y_k}$$

The effect of these uncertainties will propagate on several elements of the cross-section calculation

What are the info contained in the reco bins?



from O and C CC0pi analysis

We usually have several reconstructed signal samples as well as several reconstructed background samples

We usually bin reconstructed events in well reconstructable observables (like $\cos\theta_\mu$ and/or p_μ), that are also the variable we could use to extract the cross section

In a reconstructed CC0π ($\cos\theta_\mu$, p_μ) bin (j) we have N_j reco events:

$$N_j^{\mathrm{reco}} = \sum_i^{\mathrm{true\ bins}} \Big[\underbrace{c_i w_i^{\mathrm{signal}} N_i^{\mathrm{signal}}}_{\boxed{N_i^{\mathrm{signal}}}} + w_i^{\mathrm{bkg}} N_i^{\mathrm{bkg}} \Big] U_{ij}^{-1}$$

N_i^{signal} = what we want to extract

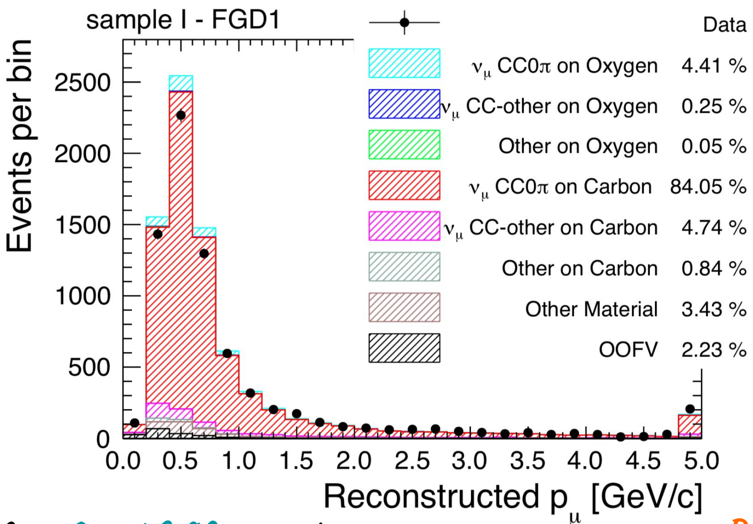
Num. of reco events in the reco bin j and sample s

Reweight due to the systematics effect

Num. of background events in the true bin i according to the MC

Smearing matrix to move from the truth to the reco bins

What are the info contained in the reco bins?



from O and C CC0pi analysis

We usually have several reconstructed signal samples as well as several reconstructed background samples

We usually bin reconstructed events in well reconstructable observables (like $\cos\theta_\mu$ and/or p_μ), that are also the variable we could use to extract the cross section

In a reconstructed CC0π ($\cos\theta_\mu$, p_μ) bin (j) we have N_j reco events:

Reweight due to the systematics effect

Num. of signal events in the true bin i according to the MC

Num. of background events in the true bin i according to the MC

$$N_j^{\text{reco}} = \sum_i^{\text{true bins}} \left[\underbrace{c_i w_i^{\text{signal}} N_i^{\text{signal}}} + w_i^{\text{bkg}} N_i^{\text{bkg}} \right] U_{ij}^{-1}$$

Num. of reco events in the reco bin j and sample s

$\boxed{N_i^{\text{signal}}}$ = what we want to extract

Reweight due to the systematics effect

Smearing matrix to move from the truth to the reco bins

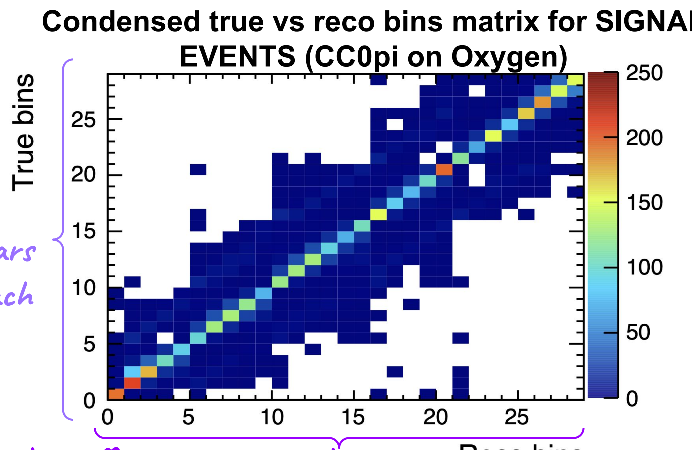
Zoom on the template parameters

Template parameters ARE the **parameters of interest** in our xsec analyses

They are **FREE** parameters that rescale the MC signal events (eventually corrected by some systematics) and thus they have the dominant effect (wrt the systematics parameters)

There is **one** template parameter **per truth signal bin** (in which you want to extract your cross section)

They thus **apply on the MC truth space and on MC truth bins of signal events** but they try **to adjust the data/MC agreement in the reco space** (the one that we really measure)



$$N_j^{\text{reco}} = \sum_i^{\text{true bins}} \left[c_i w_i^{\text{signal}} N_i^{\text{signal}} + w_i^{\text{bkg}} N_i^{\text{bkg}} \right] U_{ij}^{-1}$$

Num. of reco events in the reco bin j and sample s

Reweight due to the systematics effect

Num. of signal events in the true bin i according to the MC

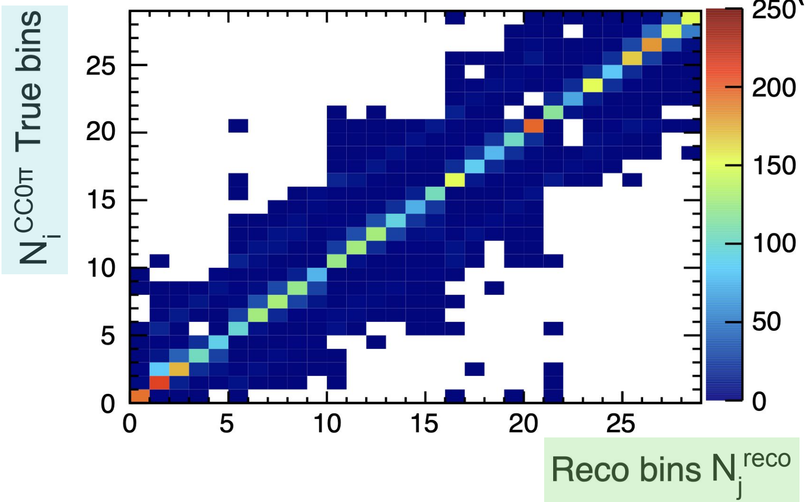
Num. of background events in the true bin i according to the MC

Data/MC correction, aka template parameters

Reweight due to the systematics effect

Smearing matrix to move from the truth to the reco bins

Zoom on the template parameters



from O and C CC0pi analysis

Reweight due to the systematics effect

Num. of signal events in the true bin i according to the MC

Num. of background events in the true bin i according to the MC

$$N_j^{\text{reco}} = \sum_i^{\text{true bins}} \left[c_i w_i^{\text{signal}} N_i^{\text{signal}} + w_i^{\text{bkg}} N_i^{\text{bkg}} \right] U_{ij}^{-1}$$

Num. of reco events in the reco bin j and sample s

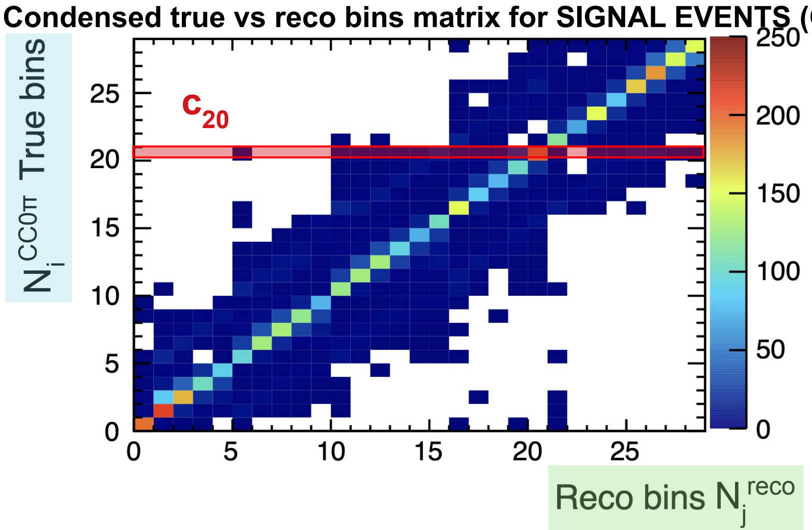
Data/MC correction, aka template parameters

Reweight due to the systematics effect

Smearing matrix to move from the truth to the reco bins

Zoom on the template parameters

Moving parameter c_{20} ⇔ moving the signal content of truth bin 20 ⇔ moving the signal content of ALL the reco bins corresponding to true bin 20 ⇔ agreement with data is checked in the reco space



from O and C CC0pi analysis

$$N_j^{\text{reco}} = \sum_i^{\text{true bins}} \left[c_i w_i^{\text{signal}} N_i^{\text{signal}} + w_i^{\text{bkg}} N_i^{\text{bkg}} \right] U_{ij}^{-1}$$

Num. of reco events in the reco bin j and sample s

Reweight due to the systematics effect

Num. of signal events in the true bin i according to the MC

Num. of background events in the true bin i according to the MC

Data/MC correction, aka template parameters

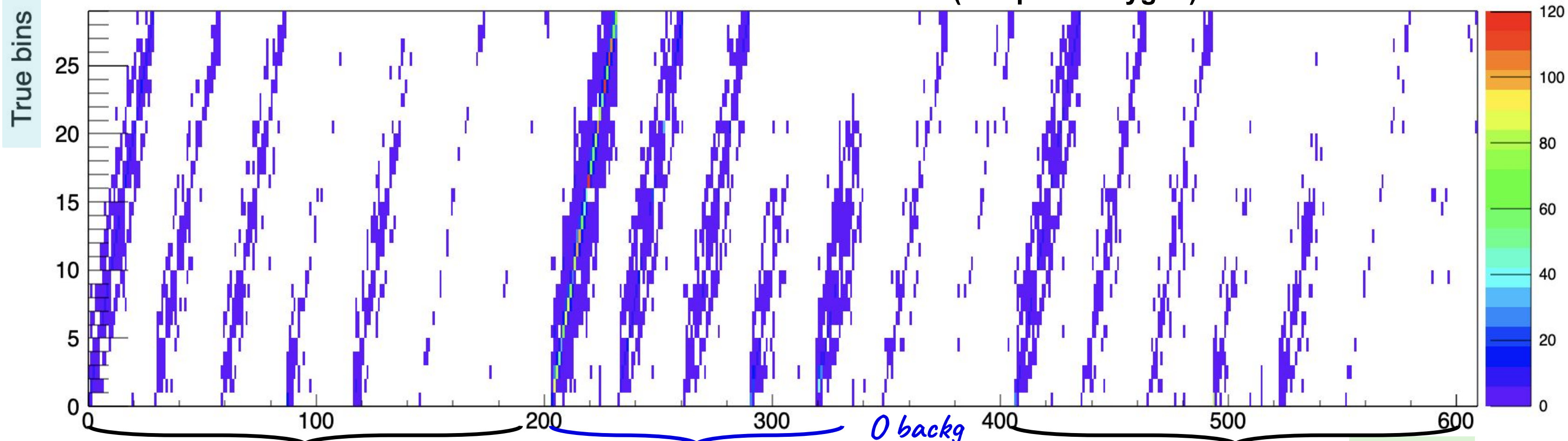
Reweight due to the systematics effect

Smearing matrix to move from the truth to the reco bins

Zoom on the template parameters

from O and C CC0pi analysis

Full true vs reco bins matrix for SIGNAL EVENTS (CC0pi on Oxygen)



C background samples

O signal samples

O backg samples

C background samples

Reweight due to the systematics effect

Num. of signal events in the true bin i according to the MC

Num. of background events in the true bin i according to the MC

$$N_j^{\mathrm{reco}} = \sum_i^{\text{true bins}} \left[c_i w_i^{\mathrm{signal}} N_i^{\mathrm{signal}} + w_i^{\mathrm{bkg}} N_i^{\mathrm{bkg}} \right] U_{ij}^{-1}$$

Num. of reco events in the reco bin j and sample s

Data/MC correction, aka template parameters

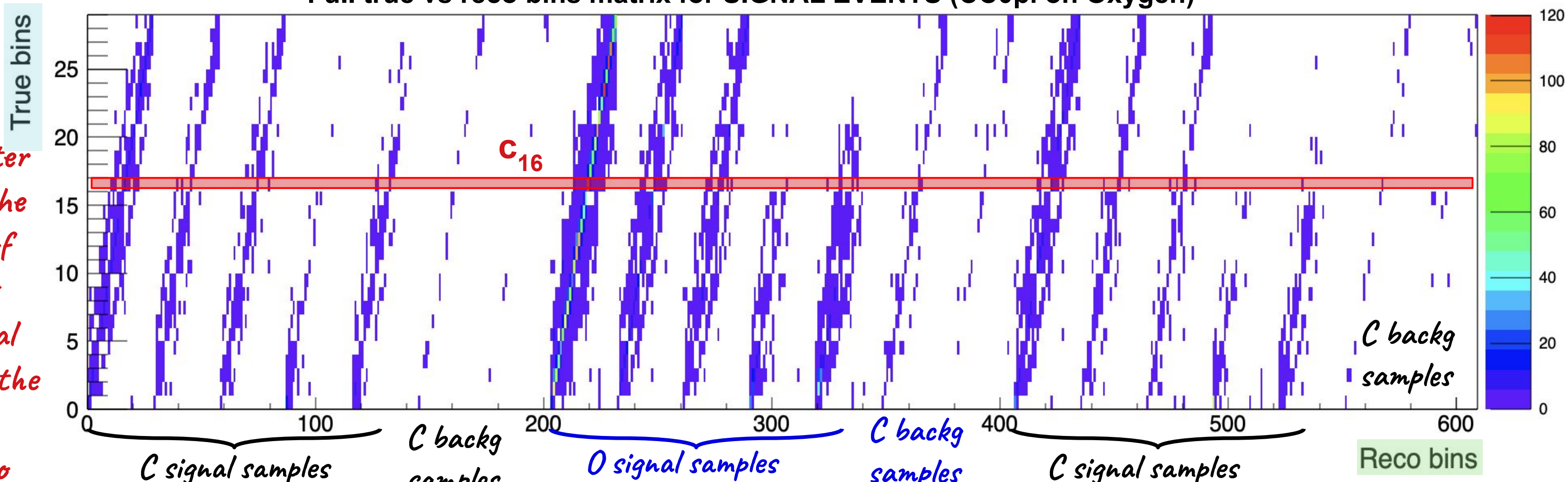
Reweight due to the systematics effect

Smearing matrix to move from the truth to the reco bins

Zoom on the template parameters

from O and C CC0pi analysis

Full true vs reco bins matrix for SIGNAL EVENTS (CC0pi on Oxygen)



C signal samples | C backg samples | O signal samples | C backg samples | C signal samples

Moving parameter c_{16} ⇔ moving the signal content of truth bin 16 ⇔ moving the signal content of ALL the reco bins corresponding to true bin 16 ⇔ agreement with data is checked i the reco space

Reweight due to the systematics effect | *Num. of signal events in the true bin i according to the MC* | *Num. of background events in the true bin i according to the MC*

$$N_j^{\mathrm{reco}} = \sum_i^{\text{true bins}} \left[c_i w_i^{\mathrm{signal}} N_i^{\mathrm{signal}} + w_i^{\mathrm{bkg}} N_i^{\mathrm{bkg}} \right] U_{ij}^{-1}$$

Num. of reco events in the reco bin j and sample s

Data/MC correction, aka template parameters

Reweight due to the systematics effect

Smearing matrix to move from the truth to the reco bins

Zoom on the template parameters

Smearing matrix for SIGNAL EVENTS (CC0pi on Oxygen)

True bins

Concretely :

- template parameters vary signal events in the truth space but this also affects the reco distributions where we can compare with real data (~ the unfolding matrix is recalculated at each iteration)
- template parameters are totally free ⇒ the unfolding process is totally unregularised (~ D'Agostini iterative unfolding with infinite num. of iterations)
- Regularisation can be introduced to smoothen the fit results ⇒ optional and checked to give results equivalent to the unregularised ones
- we work with a number of reconstructed bins much larger than the number of true bins ⇒ this limits the problem of "degenerate" solutions of the unfolding

See also Lukas' talk from yesterday

Moving parameter c_{16} ⇔ moving the signal content of truth bin 16 ⇔ moving the signal content of ALL the reco bins corresponding to true bin 16 ⇔ agreement with data is checked in the reco space

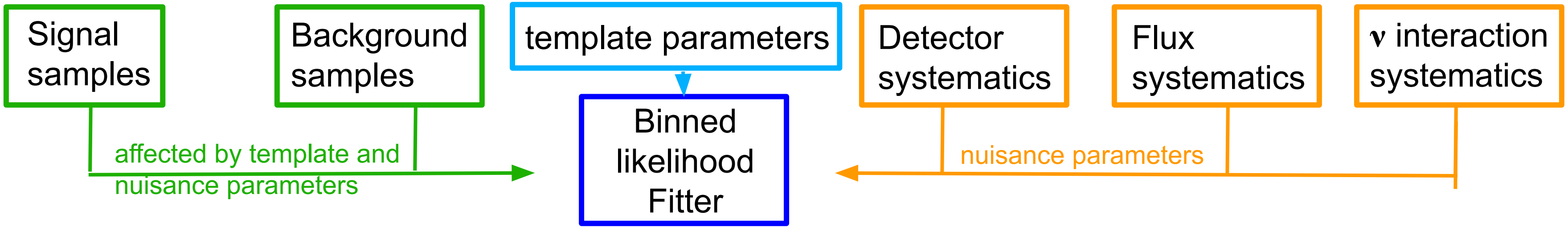
Num. of reco events in the reco bin j and sample s

i

Data/MC correction, aka template parameters

Smearing matrix to move from the truth to the reco bins

Cross section fitter



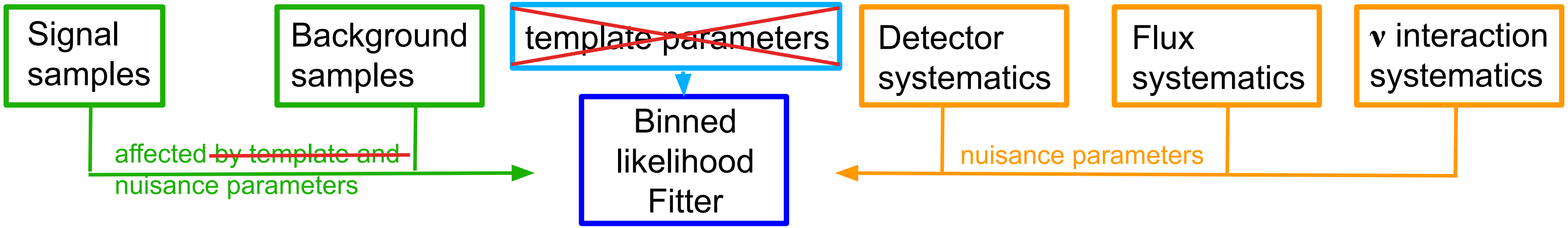
Adjust the **unconstrained template pars in the truth space** to match the MC and data distributions in the reco space (especially in the signal samples)

Adjust the constrained nuisance parameters to match the MC and data distributions ⇒ propagation of systematics uncertainties

Affect simultaneously the truth and the reco distributions ⇒ **unfolding** of the detector effects

Simultaneously on signal and background samples ⇒ **background subtraction**

T2K cross section fitter vs T2K model tuning



~~Adjust the **unconstrained template pars in the truth space** to match the MC and data distributions in the reco space (especially in the signal samples)~~

Adjust the constrained nuisance parameters to match the MC and data distributions ⇒ propagation of systematics uncertainties

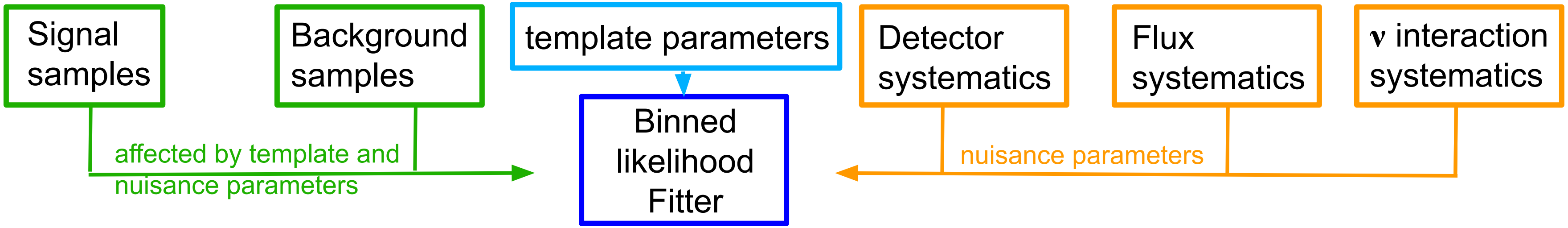
Affect ~~simultaneously the truth and~~ the reco distributions ~~⇒ **unfolding** of the detector effects~~

Simultaneously on signal and background samples ~~⇒ **background subtraction**~~

Cross section likelihood fitter is a complexification of the likelihood fitter approach that we use to tune the flux and neutrino interaction models for the T2K oscillation analysis

The MAIN difference consists in the presence of template parameters (~ free parameterisation of the SIGNAL cross section!) and of the unfolding process

Cross-section fitter



Adjust the **unconstrained template pars in the truth space** to match the MC and data distributions in the reco space (especially in the signal samples)

Adjust the constrained nuisance parameters to match the MC and data distributions ⇒ propagation of systematics uncertainties

Affect simultaneously the truth and the reco distributions ⇒ **unfolding** of the detector effects

Simultaneously on signal and background samples ⇒ **background subtraction**

By minimizing these two quantities with Minuit (Migrad/HESSE)

dominated by the effect of the template FREE parameters
(for signal samples)

$$-2\ln(\mathrm{L}^{\mathrm{stat}}) = \sum_{s}^{\text{sub-samples}} \sum_{j}^{\text{reco bins}} 2\left(\mathrm{N}_j^s - \mathrm{N}_j^{s,\,\mathrm{obs}} + \mathrm{N}_j^{s,\,\mathrm{obs}} \ln\frac{\mathrm{N}_j^{s,\,\mathrm{obs}}}{\mathrm{N}_j^s}\right)$$

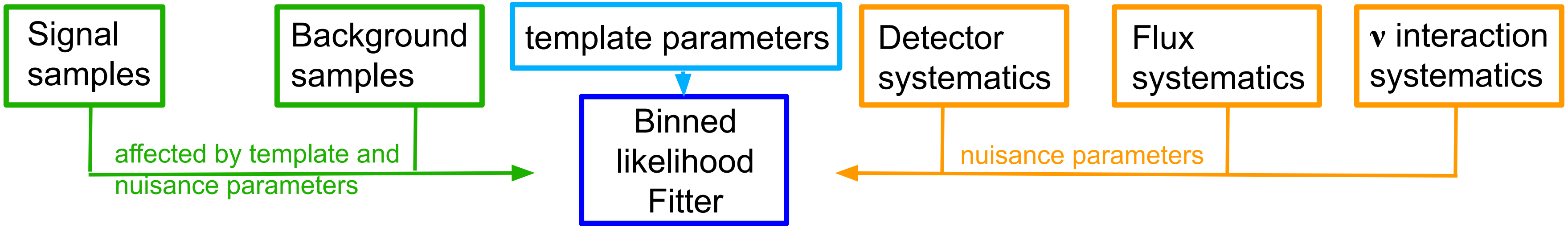
encodes all our prior knowledge of the systematics parameters

$$-2\ln(\mathrm{L}^{\mathrm{syst}}) = \sum_{p} (\vec{p} - \vec{p}_{\mathrm{prior}})\,(V_{\mathrm{cov}}^{\mathrm{syst}})^{-1}\,(\vec{p} - \vec{p}_{\mathrm{prior}})$$

NOTE: showing here the Poisson likelihood... but could also be in its Barlow-Beeston version

NOTE 2: it is a binned likelihood fitter, but concretely parameters apply on an event-by-event basis

Cross-section fitter

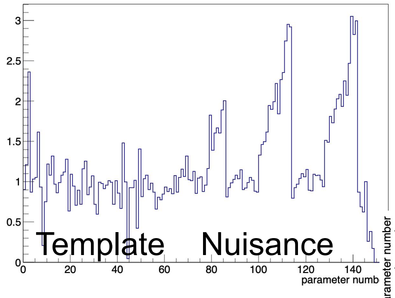


Adjust the **unconstrained template pars in the truth space** to match the MC and data distributions in the reco space (especially in the signal samples)

Adjust the constrained nuisance parameters to match the MC and data distributions ⇒ propagation of systematics uncertainties

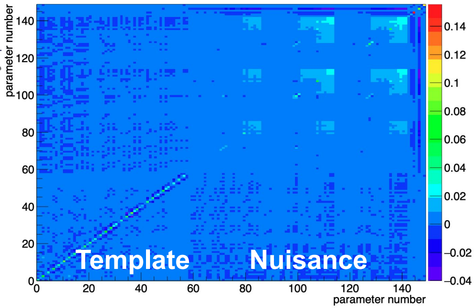
Affect simultaneously the truth and the reco distributions ⇒ **unfolding** of the detector effects

Simultaneously on signal and background samples ⇒ **background subtraction**



fitter outputs

Best fit parameter values and covariance matrix



from O and C CC0pi analysis

Cross-section extraction

Signal samples

Background samples

template parameters

Detector systematics

Flux systematics

ν interaction systematics

affected by template and nuisance parameters

Binned likelihood Fitter

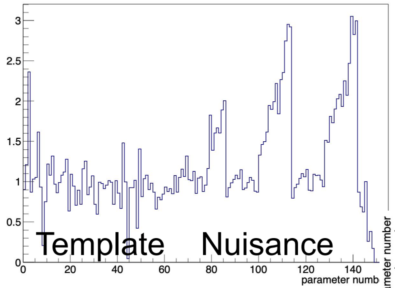
nuisance parameters

Adjust the **unconstrained template pars in the truth space** to match the MC and data distributions in the reco space (especially in the signal samples)

Adjust the constrained nuisance parameters to match the MC and data distributions ⇒ propagation of systematics uncertainties

Affect simultaneously the truth and the reco distributions ⇒ **unfolding** of the detector effects

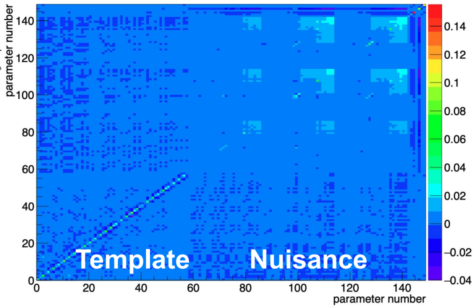
Simultaneously on signal and background samples ⇒ **background subtraction**



Best fit parameter values and covariance matrix

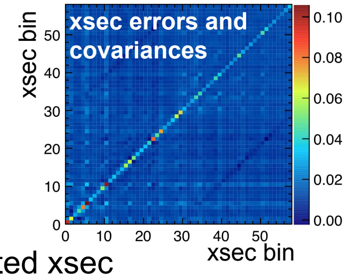
from O and C CC0pi analysis

fitter outputs

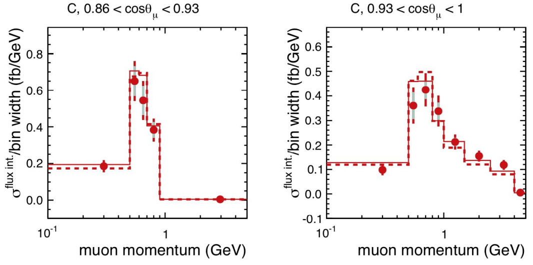


XSEC CALCULATION and ERROR PROPAGATION

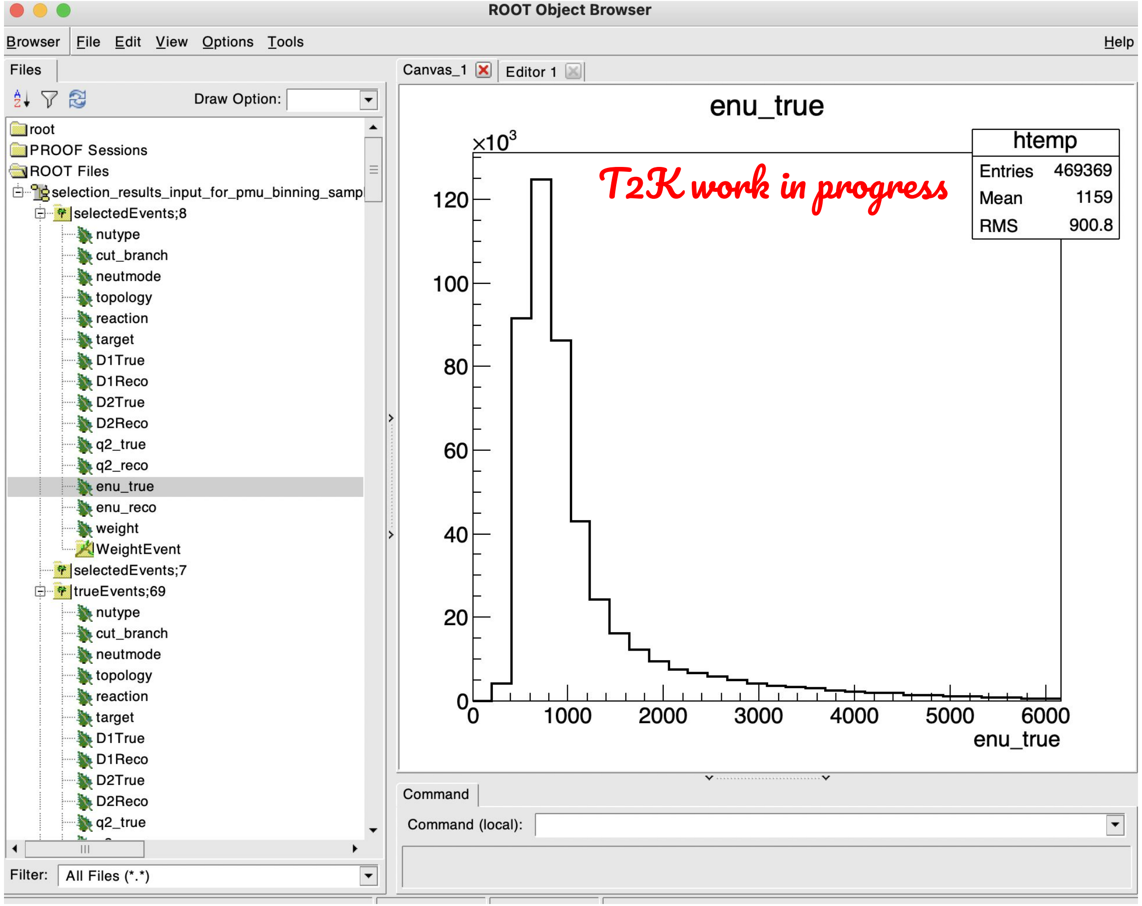
Done via MC toys, either on the input MC or **from the post fit covariance matrix.** Parameter uncertainties affect N_{signal}, efficiency, N_{tar} and the flux



Extracted xsec



Fitter inputs: samples and template parameters



A typical example of input root file for the cross section likelihood fitter.

Per each event you need common information like: the neutrino type, the sample number, the true topology and reaction, the target type, the reco and truth kinematics variables, you are interested in,...

Here p_{μ} and $\cos\theta_{\mu}$ are indicated as D1 and D2

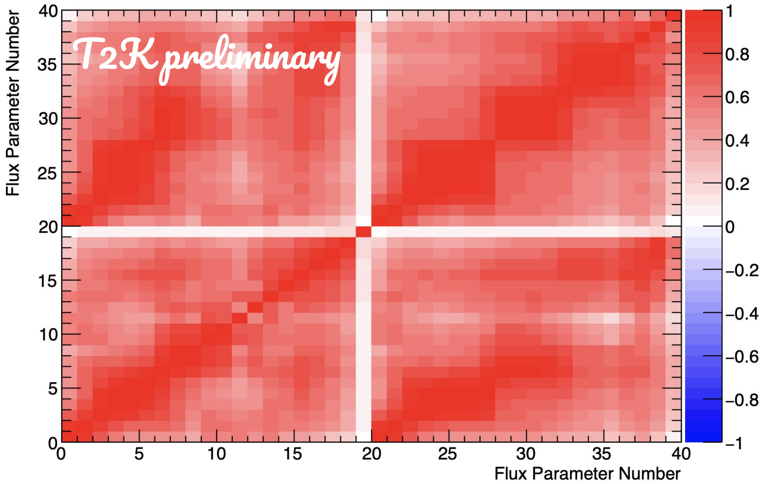
Template parameters are introduced as weight parameters by simply defining the binning in truth variables that you want to use for the cross section extraction

How we describe the FLUX and DETECTOR systematics

In both cases we have a series of **weight (normalisation) parameters that can <u>gaussianly</u> move around their nominal value (1) according to a prior covariance matrix.**

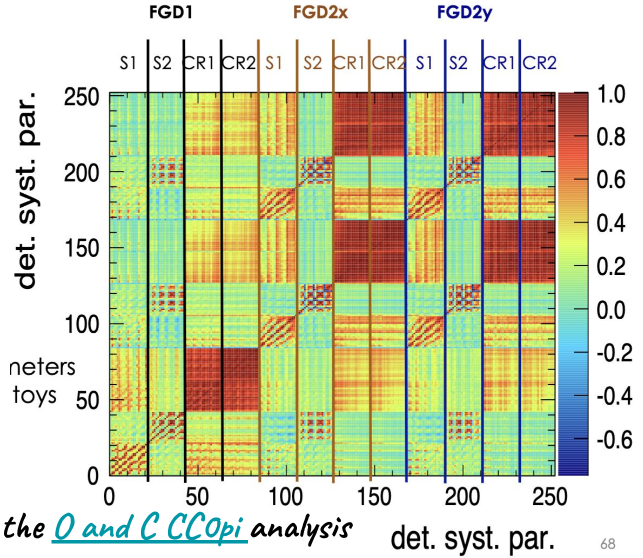
- **Flux parameters** are defined as one per **bin of true neutrino energy** (10 to 40 pars, depending on the analysis)
- **Detector parameters** are defined as **one per reconstructed bin** (~50 to ~1000 pars, depending on the analysis)

Example of flux correlation matrix



From the first on/off-axis CC0pi joint analysis

Example of detector correlation matrix

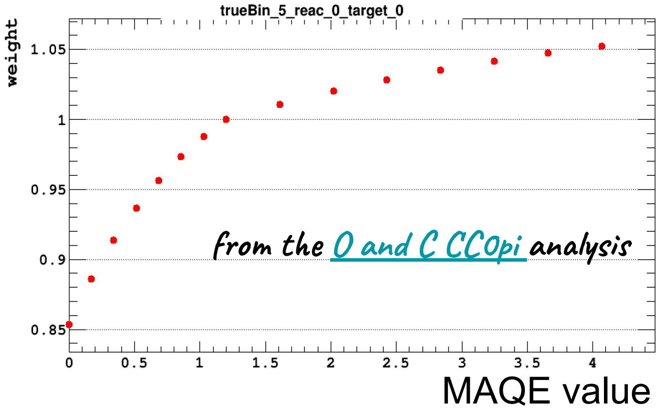


from the [O and C CC0pi](#) analysis

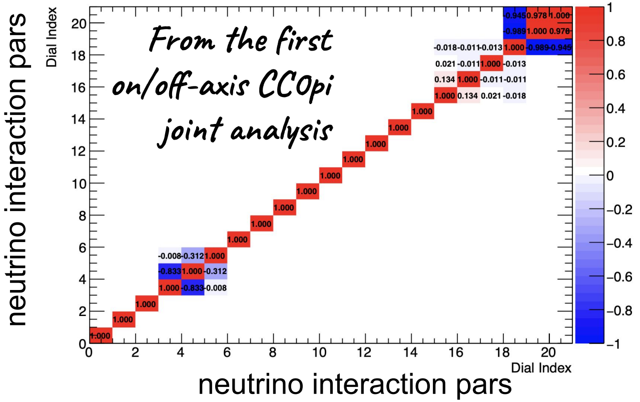
How we describe the ν INTERACTION systematics

Per each interaction parameter we construct so-called splines or response functions. They are usually, binned, i.e. they are constructed per reco or truth kinematics bin, per true neutrino interaction, per sample, per target nuclei. They encode the weight to be applied at each bin per each possible value of the parameter

MAQE

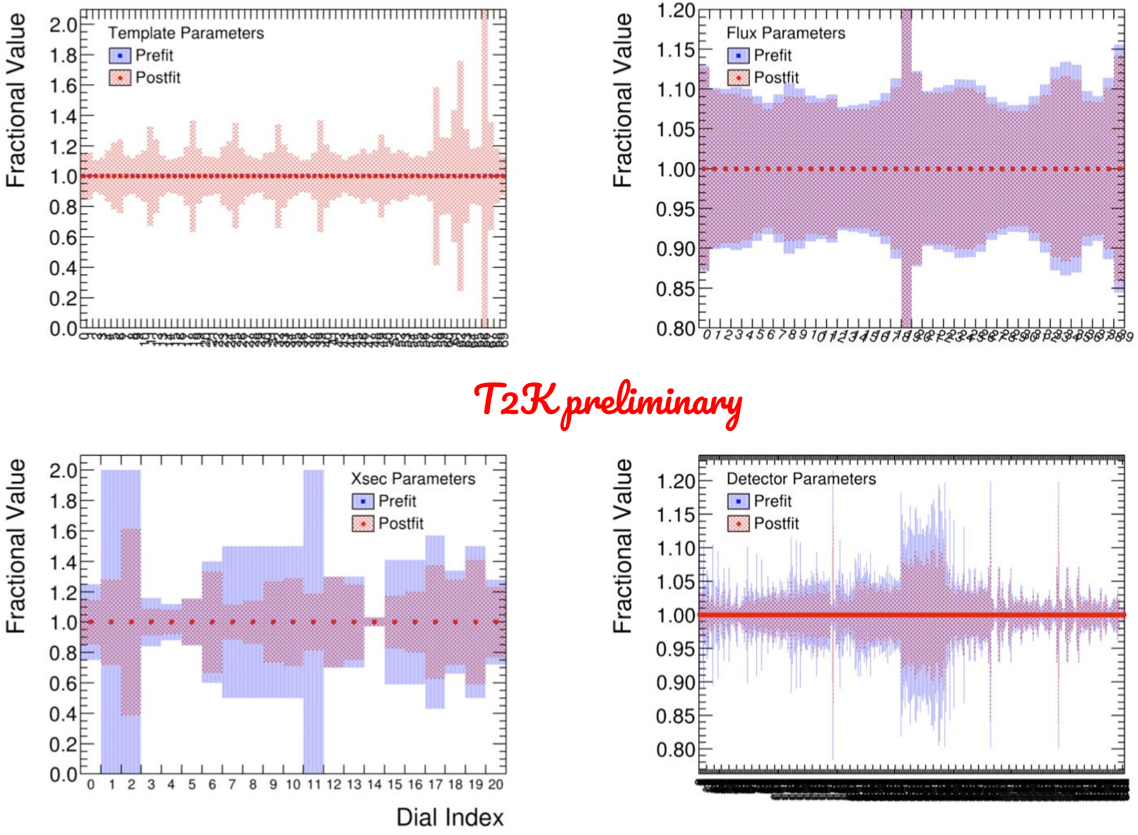


Signal sample: mu in TPC
True Bin 5 ($0.6 < \cos\theta_\mu < 0.75$, $0<p_\mu<0.35$)
True interaction: CCQE
Target: Oxygen



Once more, input covariance matrices encode the prior error of each parameter and the possible correlation between them

Example of fitter outputs: fit parameter in an Asimov fit



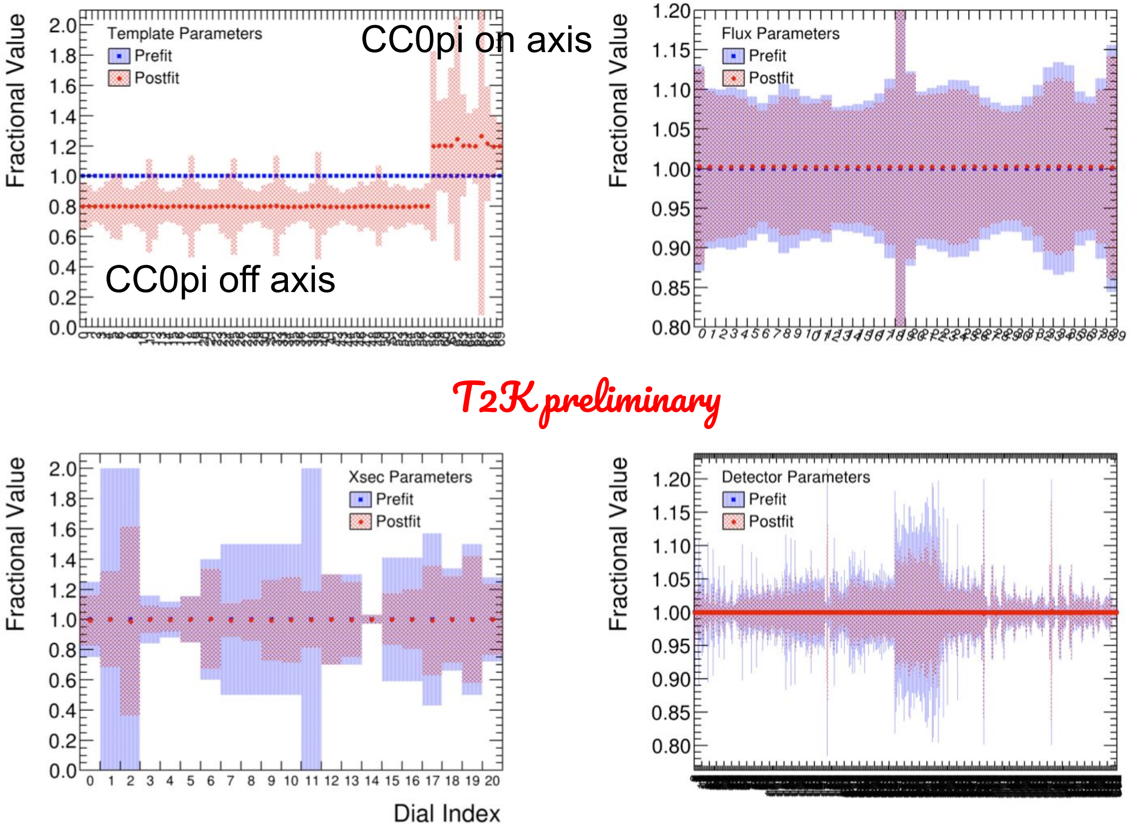
Asimov fit (where MC prior = fake data) **are a good way to test the overall fit machinery and the overall analysis sensitivity.**

You expect:

- post fit parameter values to be identical to the prefit values
- post fit parameter errors to be ≤ the prefit errors

From the first on/off-axis CC0pi joint analysis

Example of fitter outputs: fit parameters in a Fake data fit



From the first on/off-axis CC0pi joint analysis

Performing fake data studies is the way to answer:

- Is the fitter able to reproduce the fake data truth?
- Is my systematic model introducing enough freedom to the fitter to correctly adjust parameter values w/o biasing the signal results?

In a likelihood fitter you can check that the parameters in the fit move towards the expected values

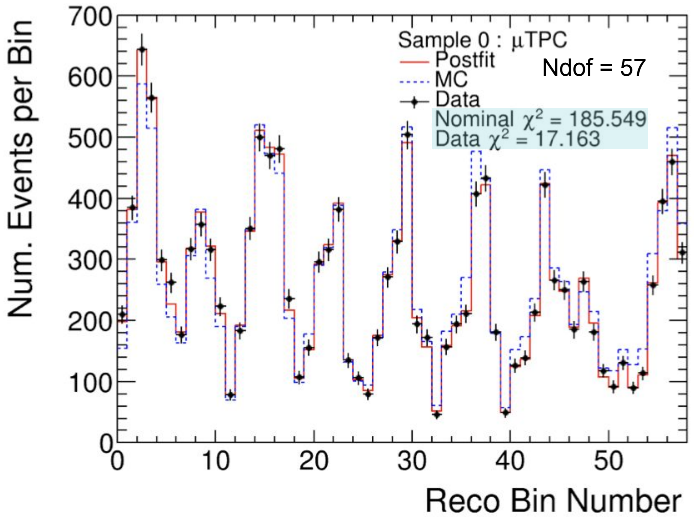
For instance, in this example the fake data is obtained by modifying the signal cross section w.r.t. the nominal values:

- *CC0pi off-axis x 0.8*
- *CC0pi on-axis x 1.2*

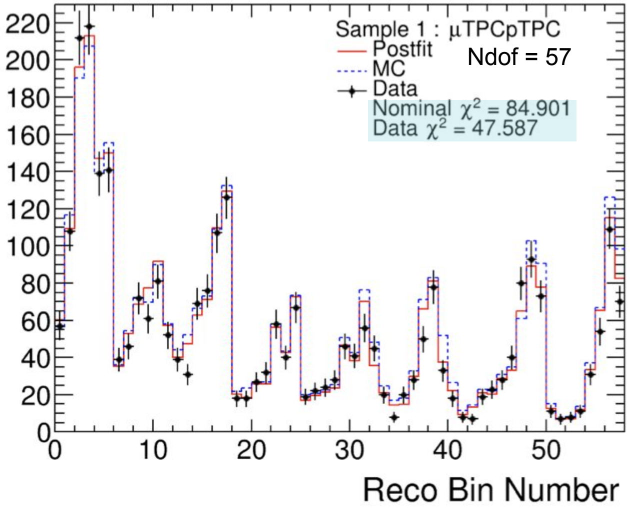
⇒ this is exactly what the post fit template pars values reproduce

Example of fitter outputs: number of events is reco space and postfit -2lnL (χ^2)

Just a subset of reco samples



T2K preliminary



From the first on/off-axis CC0pi joint analysis, data fit

A series of diagnostics of the results:

- Checking the **agreement in the reco space** (num of events per bin)
- quoting the value of the **total -2lnL (χ^2)**
- quoting the value of the **χ^2 per sample** ⇒ this is a way to verify if there is a particularly problematic sample

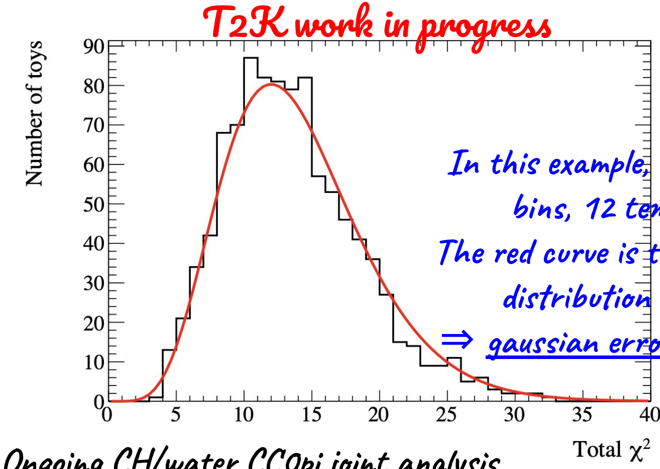
Additional fit diagnostic: p-value

So-called **"coverage studies"** are realized by performing a series of fits on a series of fake data (FD) distributions. FD are obtained by simultaneously:

- statistically varying the content of each reconstructed bin (Poisson, bin-to-bin independent)
- throwing all the systematics parameters based on their prior values and errors

You then construct the **χ^2 curve using the postfit value of the -2lnL** of each fit. This gives you:

- the **effective number of degrees of freedom** (~N_{bin} - $N_{Template\ pars}$)
- the possibility to **quote the p-value** once you perform the fit on real data (is the analysis model compatible with the data fit?)

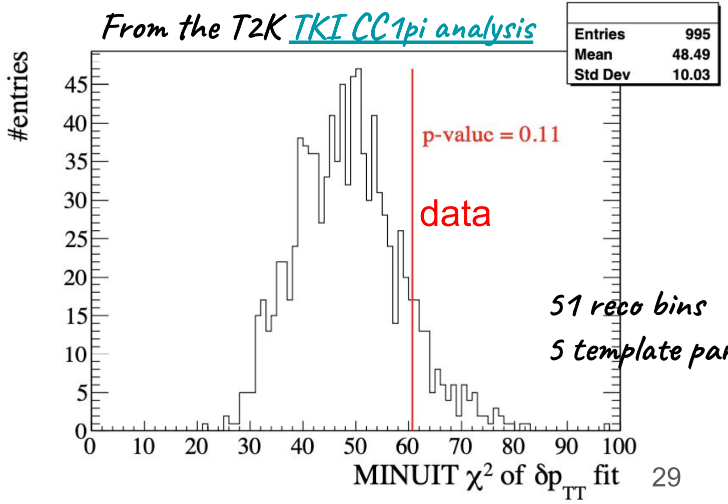


In this example, we have 26 reco bins, 12 template pars
The red curve is the theoretical χ^2 distribution for 14 DOF
⇒ gaussian error assumption ok!

Ongoing CH/water CC0pi joint analysis

See more in N. Latham talk later

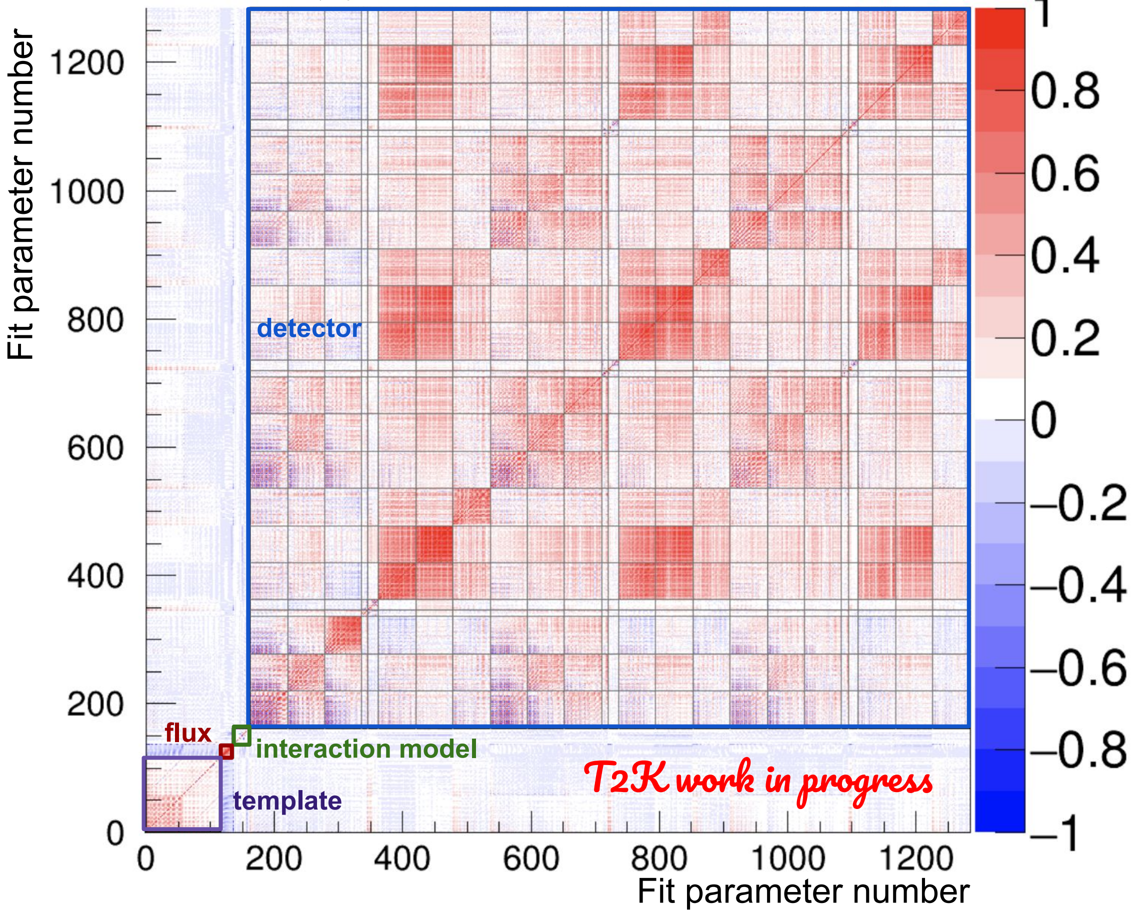
From the T2K TKI CC1pi analysis



51 reco bins
5 template pars

Post fit correlation and covariance matrix

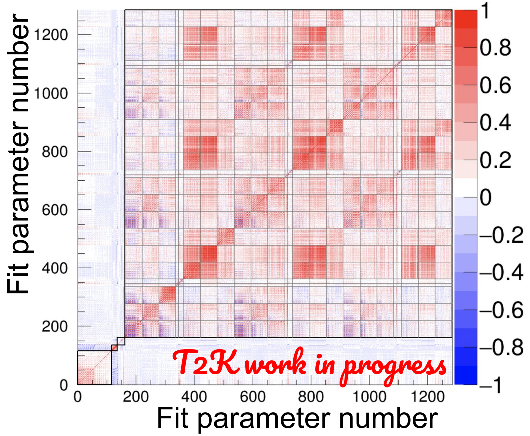
From an ongoing O/C CC0pi joint analysis



In addition to the post fit parameters values and errors, the fit also provide the post fit **covariance and correlation matrices** that encode the covariances and correlations among all the parameters.

This is a fundamental ingredient that we need **to move from the "postfit space" to the "cross-section space"**

The cross-section extraction and the error propagation: the concept

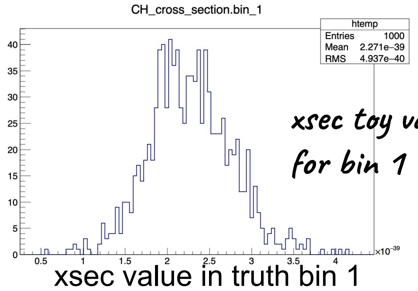


Post fit errors, covariances and values of each parameter are the fundamental ingredients to calculate the value of the cross section in each true bin and estimate the corresponding error.
This is done **throwing toys** starting from the post fit covariance matrix and parameter values: at each toy a new set of template+nuisance parameters is estimated by **gaussianly varying the best fit value of each parameter according to the post fit covariance matrix**

$$\frac{d\sigma}{dx_i dy_k} = \frac{N^{signal}_{ik}}{\epsilon_{ik}\Phi N^{FV}_{nucleons}} \times \frac{1}{\Delta x_i \Delta y_k}$$

Per each toy, **all the quantities entering in the cross-section calculation are recalculated and used to estimate the cross-section value in each bin**. Over a big number of toys, the cross section in each bin is estimated as the mean (or the value obtained at the best fit), while its corresponding error is the RMS of the distribution

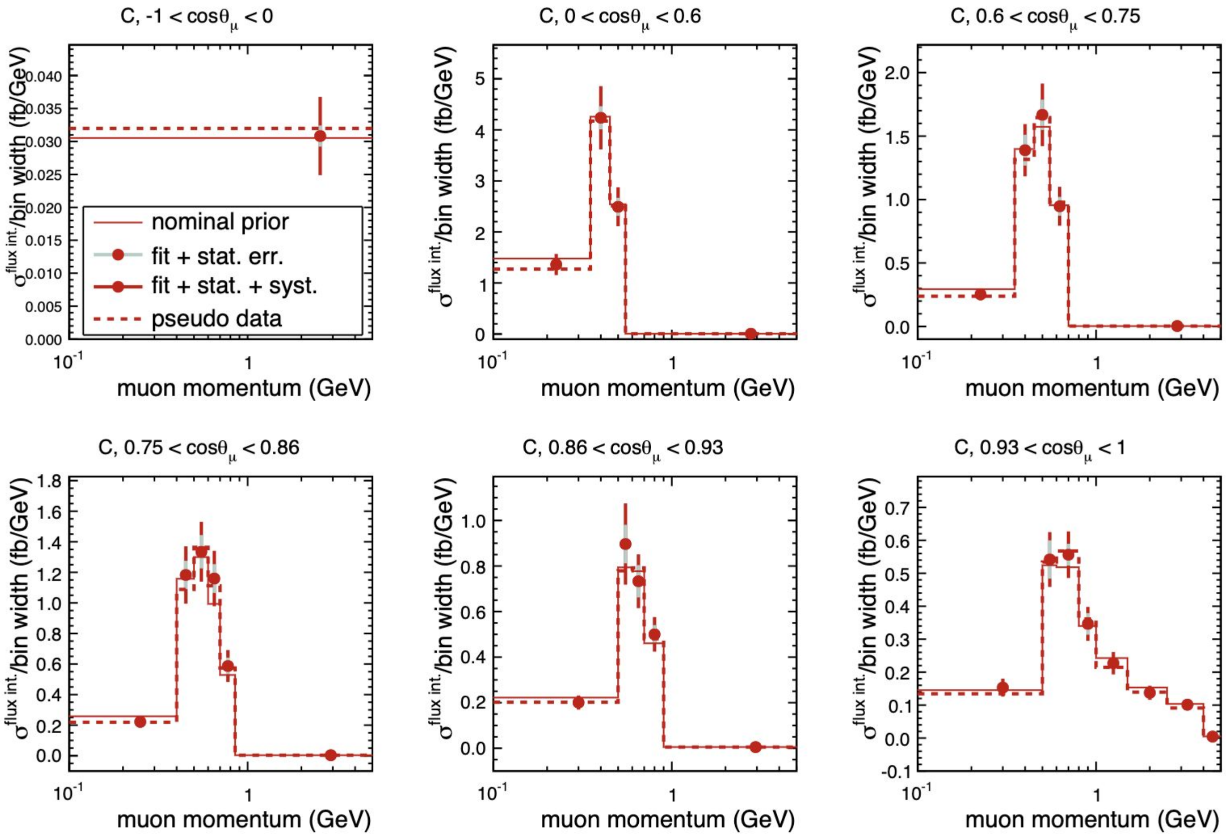
T2K work in progress



xsec toy values for bin 1

Ongoing CH/water CC0pi joint analyses

Extracting the cross section: fake data



Ndof = 58

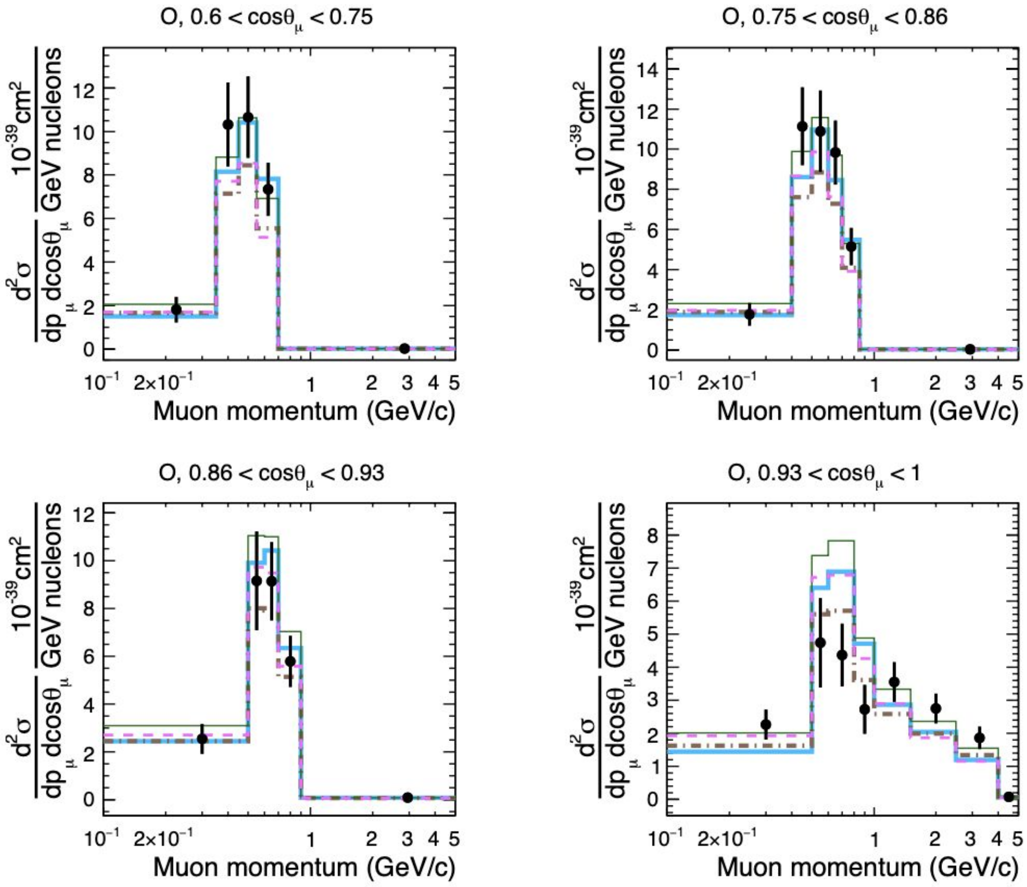
χ^2_{nominal} = 47

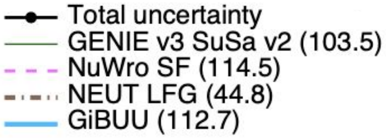
$\chi^2_{\text{fake data}}$ = 24

$$\chi^2_{\text{tot}} = \sum_i \sum_j \left(\frac{\mathrm{d}\sigma^{\text{model}}}{\mathrm{d}x_i} - \left\langle \frac{\mathrm{d}\sigma^{\text{meas.}}}{\mathrm{d}x_i} \right\rangle \right) \cdot (V^{-1})_{ij} \left(\frac{\mathrm{d}\sigma^{\text{model}}}{\mathrm{d}x_j} - \left\langle \frac{\mathrm{d}\sigma^{\text{meas.}}}{\mathrm{d}x_j} \right\rangle \right)$$

from the O and C CC0pi analysis, fake data fit

Extracting the cross section: real data





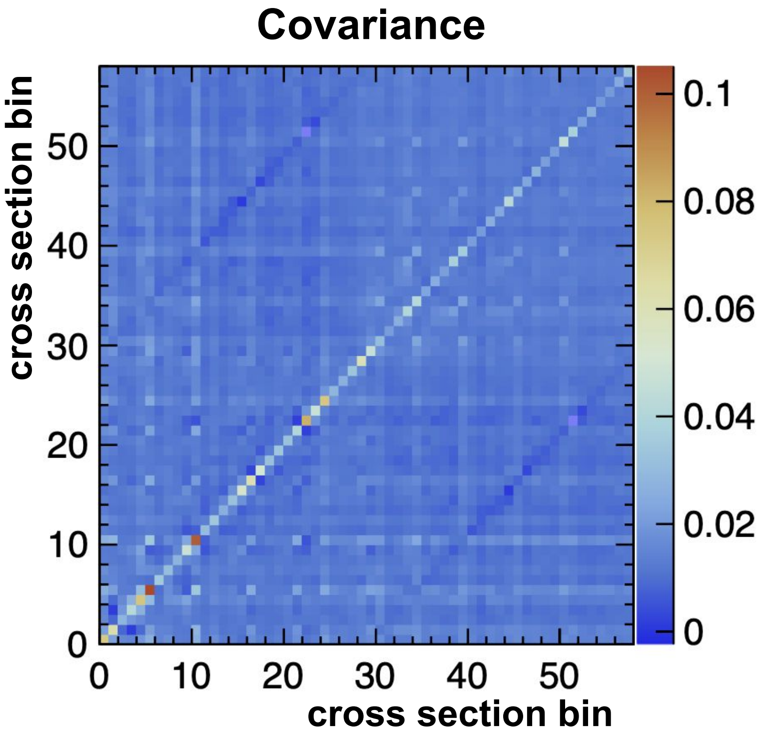
Ndof = 58

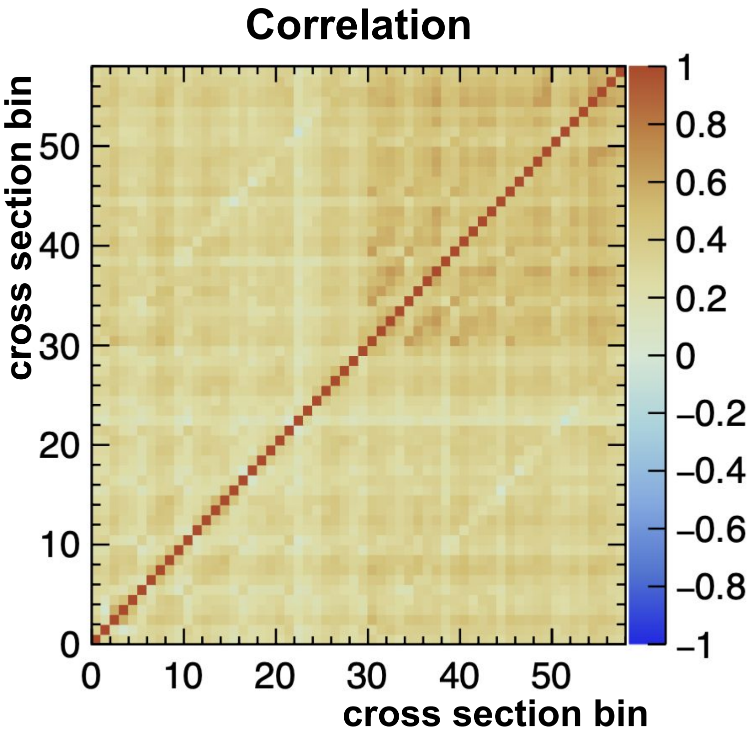
$$\chi^2_{\mathrm{tot}} = \sum_i \sum_j \left(\frac{\mathrm{d}\sigma^{\mathrm{model}}}{\mathrm{d}x_i} - \left\langle \frac{d\sigma^{\mathrm{meas.}}}{\mathrm{d}x_i} \right\rangle \right) \cdot (V^{-1})_{ij} \left(\frac{\mathrm{d}\sigma^{\mathrm{model}}}{\mathrm{d}x_j} - \left\langle \frac{d\sigma^{\mathrm{meas.}}}{\mathrm{d}x_j} \right\rangle \right)$$

from the [O and C CC0pi](#) analysis, data fit

XSEC covariance matrices

from the O and C CC0pi analysis, data fit





The other fundamental ingredient you need to provide in order to safely compare your xsec results with model predictions

Final remarks

Template fitting is used in T2K for cross section extraction since ~ 8 years, including a series of joint analyses, and will continue to be the reference unfolding method

This is an unfolding method allowing a series of useful **diagnostic** about the correctness of your results

In the years, we refined and **optimized a series of techniques:** coverage studies, error propagation, data release

Same for the fitter itself: we started with a software (manual) sharing between collaborators, moving then to the Super-xsLLHFitter and recently to GUNDAM, an open source tool that encodes all the features (+ many others!) of the T2K cross section fitter and of the T2K near detector fitter used for the oscillation analysis

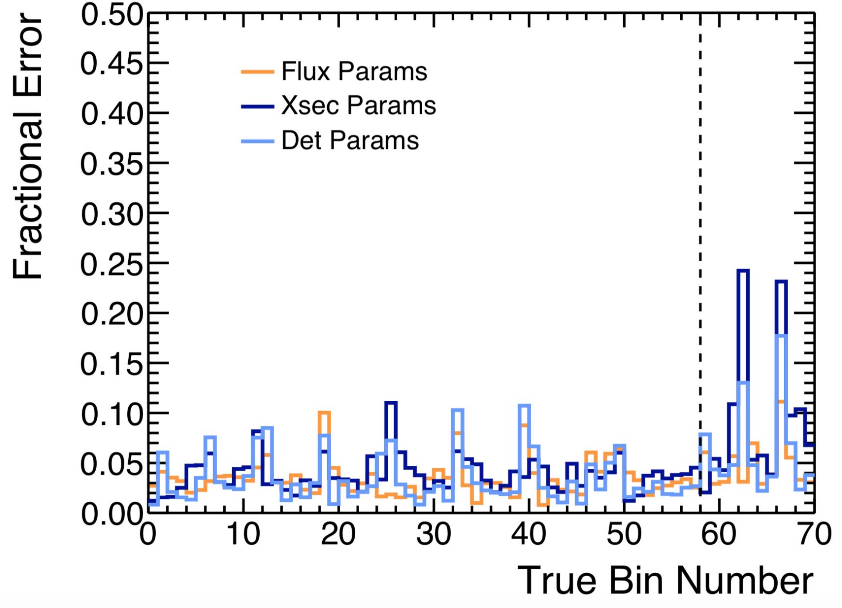
See L. Munteanu's talk tomorrow

See next talk (N. Latham) for additional details on T2K template fitters for xsec!

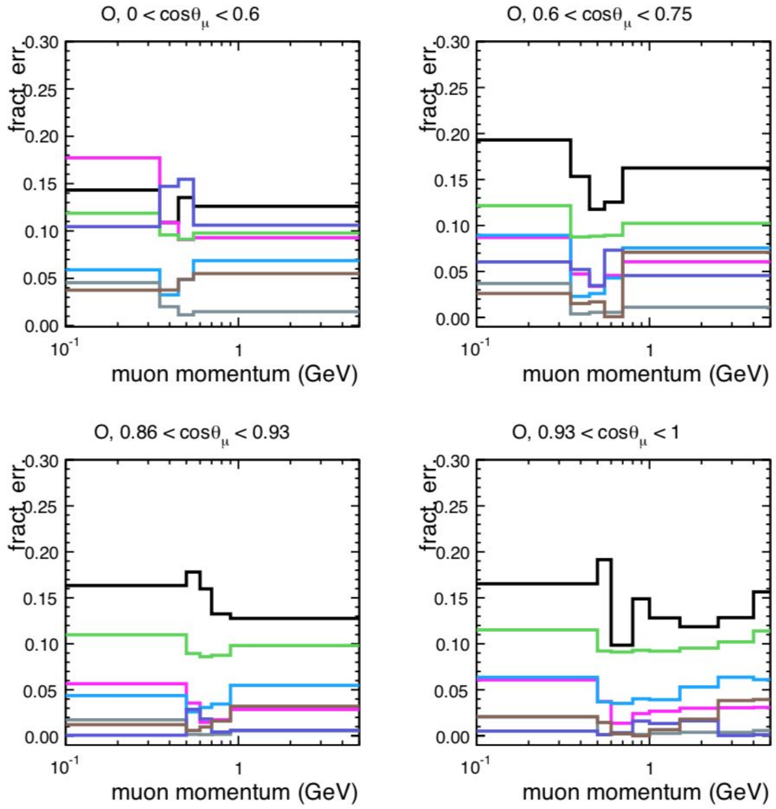

T2K collaboration
THANKS!
Tokai, July 2023

Systematics error contributions

from the joint on/off-axis analysis



from the joint O/C analysis



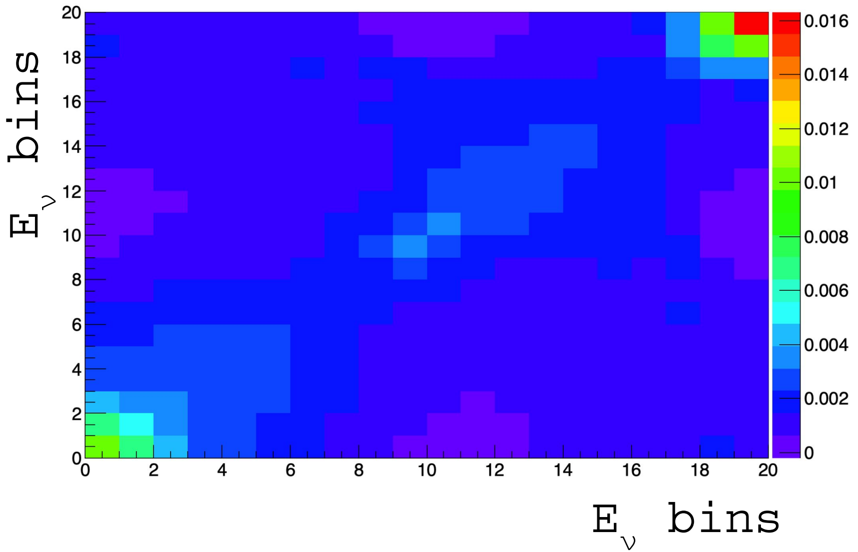
How we describe the FLUX systematics

Taking as an example used inputs for the ongoing water and CH CC0pi analysis of T2K

Uncertainty on the flux can affect:

- the shape of the spectrum, i.e. the number of events per bin (numerator in the xsec)
- The total integrated flux (denominator in the xsec)

Input flux covariance matrix **(ν_μ in ν_μ beam)**



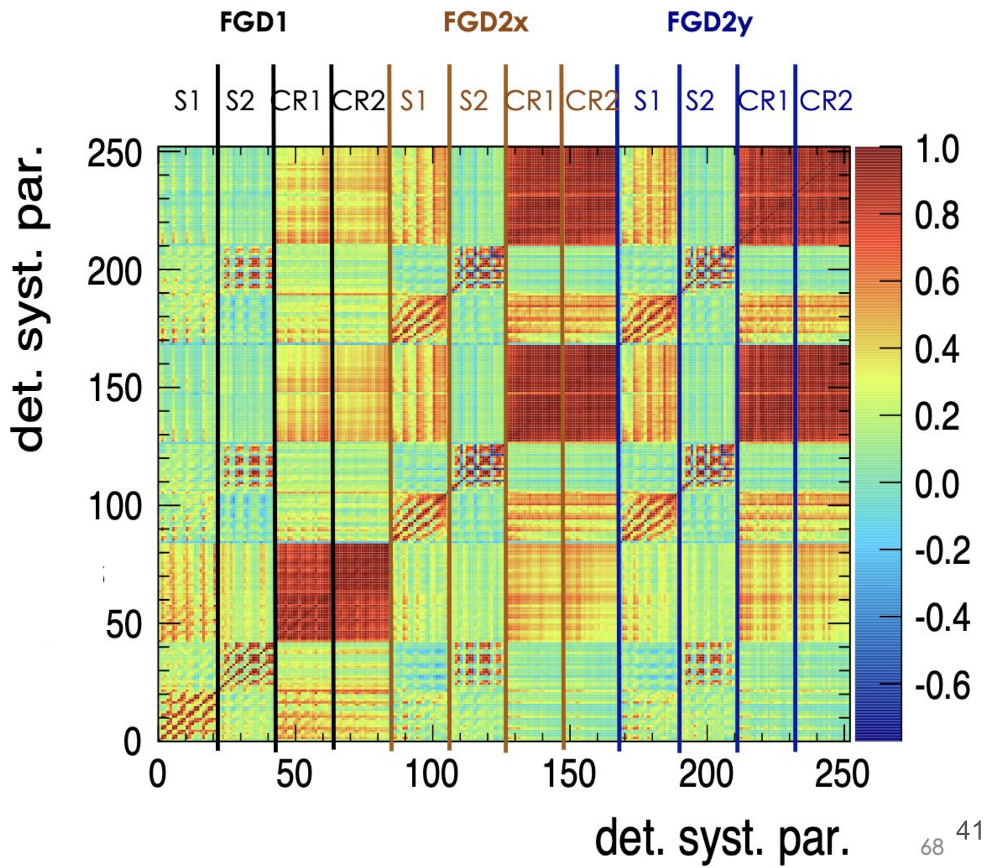
In T2K (cross-section) analyses, flux systematics are implemented as weight parameters that affect the true events depending on the neutrino energy and according to a prior covariance matrix provided by the beam experts

How we describe the DETECTOR systematics

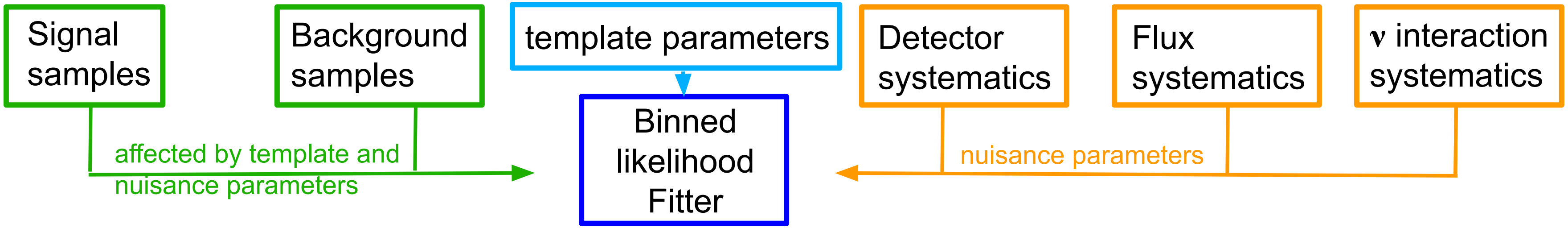
Taking as an example used inputs for the published O and C CC0pi analysis of T2K

In a similar way as for the flux, we provide a prior covariance matrix, by taking into account the event rate change in each reco bin and sample when varying all the detector systematics together.

Detector parameters act as weight parameters on each event depending on its reconstructed bin and samples and according to the input covariance matrix



Cross-section extraction



Adjust the **unconstrained template pars in the truth space** to match the MC and data distributions in the reco space (especially in the signal samples)

Adjust the constrained nuisance parameters to match the MC and data distributions ⇒ propagation of systematics uncertainties

Affect simultaneously the truth and the reco distributions ⇒ **unfolding** of the detector effects

Simultaneously on signal and background samples ⇒ **background subtraction**

By minimizing these two quantities with Minuit (Migrad/HESSE)

dominated by the effect of the template FREE parameters

$$-2\ln(\mathrm{L}^{\mathrm{stat}}) = \sum_{\mathrm{s}}^{\text{sub-samples}} \sum_{\mathrm{j}}^{\text{reco bins}} 2\left(\mathrm{N}_{\mathrm{j}}^{\mathrm{s}} - \mathrm{N}_{\mathrm{j}}^{\mathrm{s,\,obs}} + \mathrm{N}_{\mathrm{j}}^{\mathrm{s,\,obs}} \ln \frac{\mathrm{N}_{\mathrm{j}}^{\mathrm{s,\,obs}}}{\mathrm{N}_{\mathrm{j}}^{\mathrm{s}}}\right)$$

encodes all our prior knowledge of the systematics parameters

$$-2\ln(\mathrm{L}^{\mathrm{syst}}) = \sum_{\mathrm{p}} (\vec{\mathrm{p}} - \vec{\mathrm{p}}_{\mathrm{prior}})\,(\mathrm{V}_{\mathrm{cov}}^{\mathrm{syst}})^{-1}\,(\vec{\mathrm{p}} - \vec{\mathrm{p}}_{\mathrm{prior}})$$

Optional regularisation term that helps to smoothen anticorrelation between nearby ci

$$-2\ln\mathcal{L}_{reg} = \lambda \sum_{i}^{N-1} (c_i - c_{i+1})^2$$

see backup

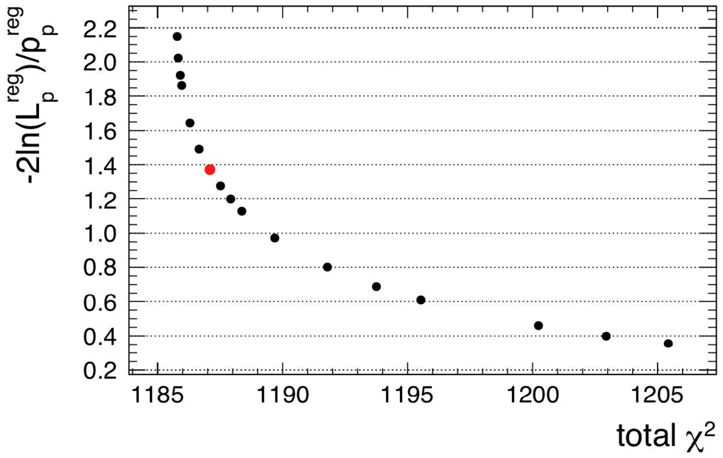
Regularisation

Known effect of the unfolding procedure: when the binning is fine wrt detector resolution, many combinations of true bins lead to the same set of reconstructed bins. The unfolded results appear to fluctuate around the truth, cursing a typical "zig-zaging" between near-by bins

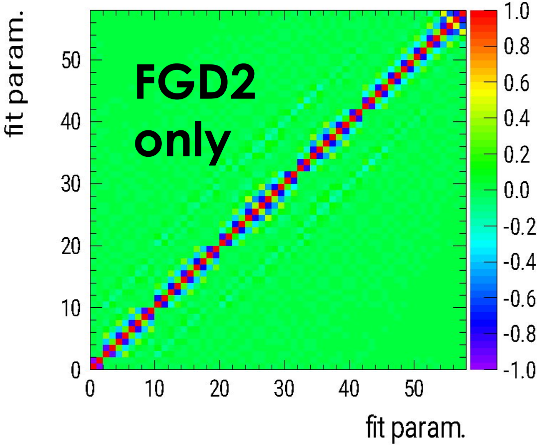
Introduce penalty terms in the likelihood to avoid too big bin-to-bin "oscillations". Can indeed bias the final result towards the input model ⇒ L-curve method to decide the "strenght" of the regularisation by balancing between the obtained smoothness against the introduced bias.

$$-2\ln(L_p^{reg}) = p_p^{reg} \sum_{k}^{\theta \text{ true bins -1}} \left(\sum_{i}^{p_\mu \text{bins in } \theta \text{ bin } k} \left[(c_i - c_{i+1})^2 + (o_i - o_{i+1})^2 \right] \right)$$

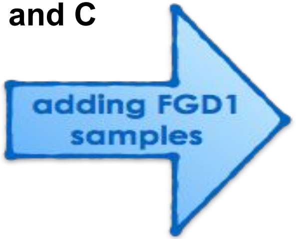
$$-2\ln(L_{\cos\theta}^{reg}) = p_\theta^{reg} \sum_{k}^{\theta \text{ true bins -1}} \left[(\bar{c}_k - \bar{c}_{k+1})^2 + (\bar{o}_k - \bar{o}_{k+1})^2 \right]$$

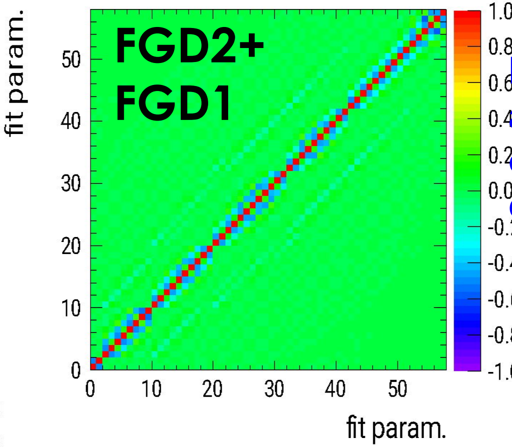


The effect of the regularisation

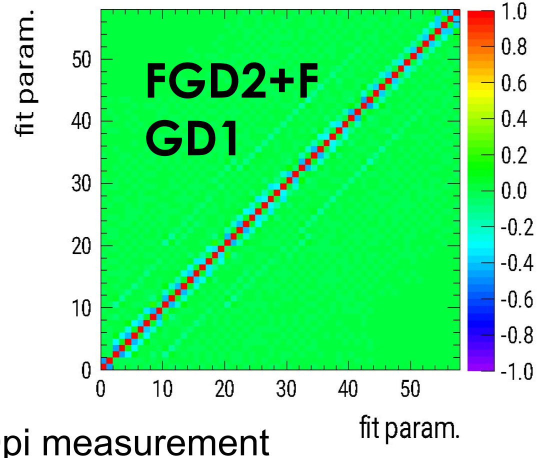


Strong anti-correlation between adjacent bins and O and C





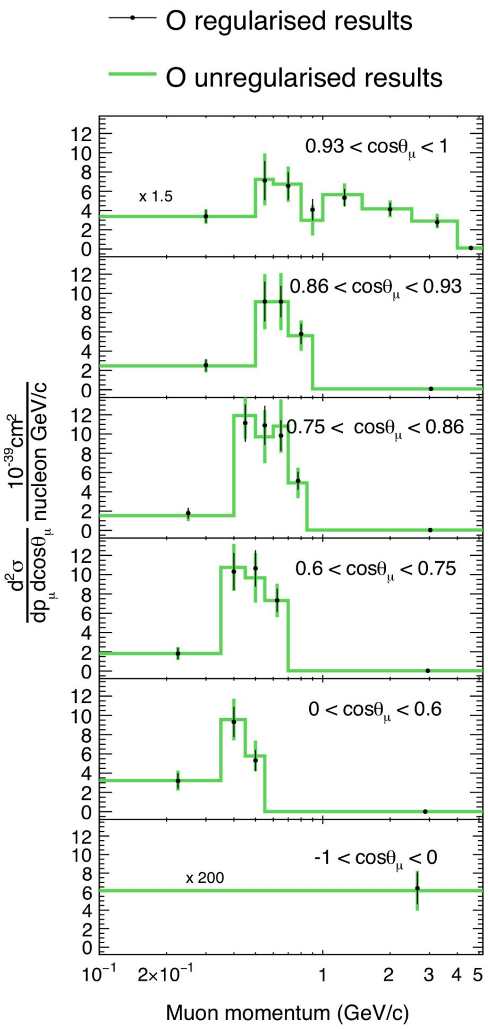
Reduced anti-correlation between c_i and o_i



Reduced anti-correlation c_i-c_{i+1} and o_i-o_{i+1} !

even parameters: o_i,
odd parameters: c_i,
58 parameters in total

Example from the published O/C CC0pi measurement



Regularisation

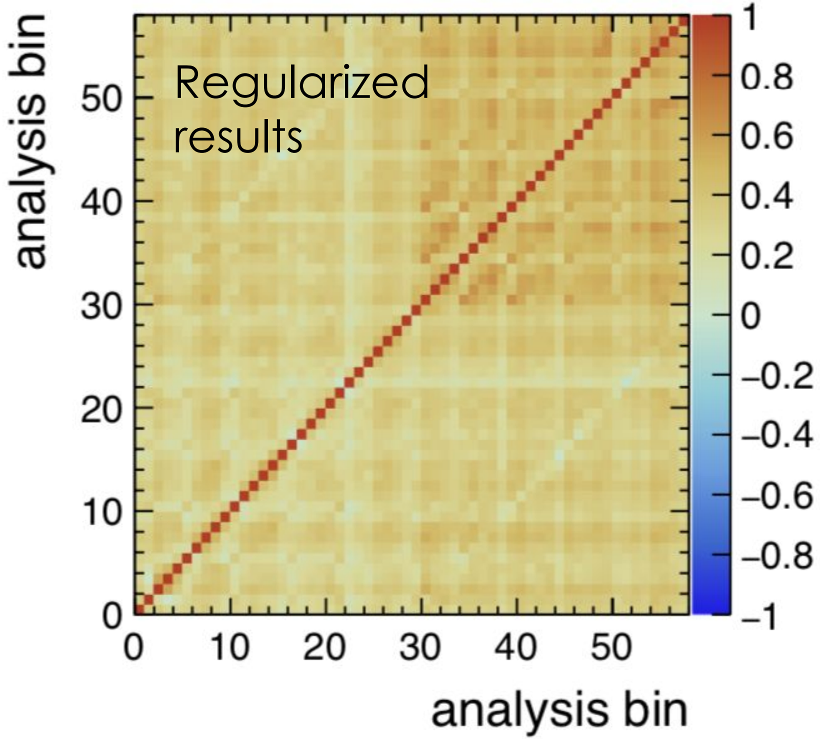
Regularised and un-regularised results perfectly compatible

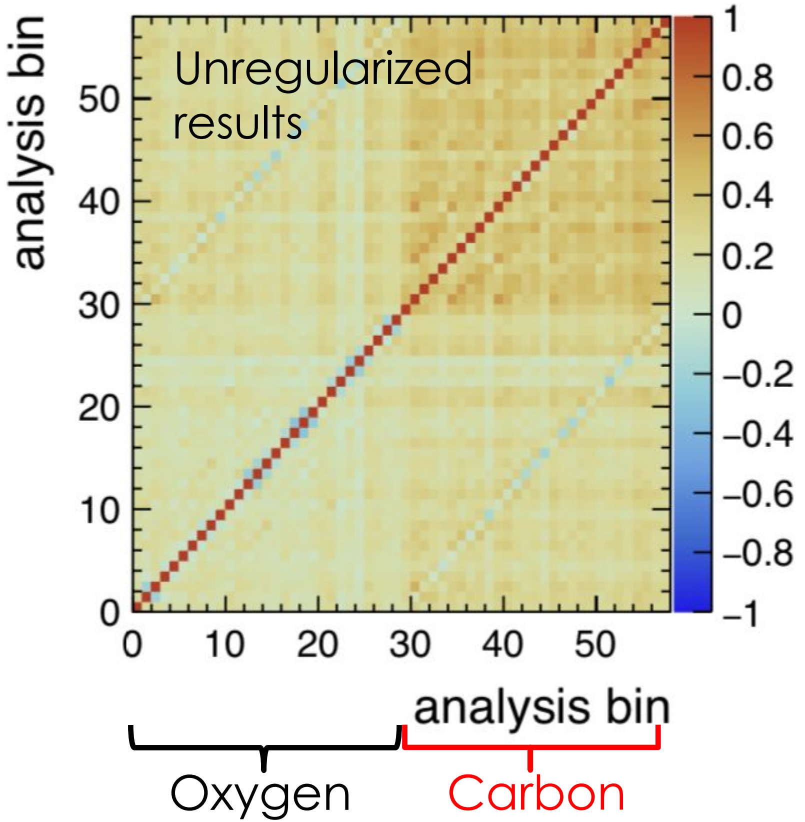
however regularised results are less zig-zaging and in principle easier to interpret “by-eye”

Anyway a proper estimation of the data/model agreement still needs the use of the post fit covariance matrix

In other word, applying a resularisaiton corresponds to select a subset of degenerate models without large fluctuations between neighboring bins. The penalty term enforces this extra constraint based on the assumption that the large fluctuations are unphysical, but this may also bias the results as a whole. The full, unreguarized result is best, but regularization can be used to choose the exact fit point used in published plots, since any of the degenerate results in the set are essentially equivalent even if they have small differences in the chisq.

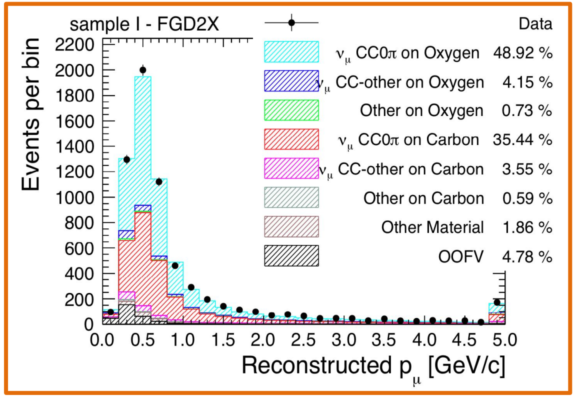
Covariance and correlation matrices





From the reco to the truth space

Reco space

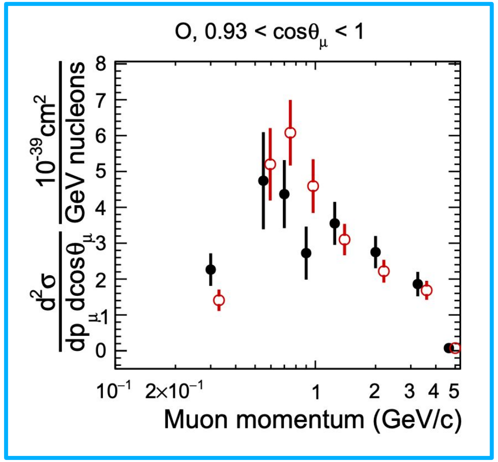


$$N_j^{reco} = \sum_i^{true\ bins} S_{ij} N_i^{true}$$

$$N_i^{true} = \sum_i^{reco\ bins} U_{ij} N_j^{reco}$$

simplified version assuming no background

Truth space



We dispose of the number of selected events in terms of reconstructed variables

efficiency correction

background subtraction

unfolding of detector effects

We want to extract the total number of signal events in terms of true variables

The binned likelihood fitter

Then we perform a binned likelihood fit where we try to minimize:

$$
\begin{aligned}
\chi^2 &= \chi^2_{stat} + \chi^2_{syst} + \chi^2_{reg,\,p} + \chi^2_{reg,\,\theta} \\
&= \boxed{\sum_{j}^{reco\ bins} 2\left(N_j^{MC} - N_j^{obs} + N_j^{obs} \ln \frac{N_j^{obs}}{N_j}\right)} + \boxed{\chi^2_{syst}} + \boxed{\chi^2_{reg,\,p} + \chi^2_{reg,\,\theta},}
\end{aligned}
$$

Key features of the fitter

- **Unconstrained fit (or template) parameters** that estimate the signal in each bin ⇒ minimize the model dependence!
- Detector, flux and theoretical interaction parameters included as **nuisance parameters** ⇒ reduce the systematics!
- **Data-driven regularization** ⇒ minimize the anti-correlation between adjacent bins (optional)!

Cross section extraction

In a reconstructed ($\cos\theta_\mu$, p_μ) bin we have N_j reco events:

$$
\begin{aligned}
N_{\boxed{j}} = \sum_{\boxed{i}}^{\text{true bins}} \Bigg[\boxed{c_i} \Bigg(\sum_{\boxed{s}}^{\text{CC}0\pi\text{ true int.}} \Big(N_i^{MC,s-C} \prod_{\boxed{b}}^{\text{model syst.}} \boxed{w(b)_j^{s-C}} \Big) \Bigg) \\
+ \boxed{o_i} \Bigg(\sum_{\boxed{s}}^{\text{CC}0\pi\text{ true int.}} \Big(N_i^{MC,\ s-O} \prod_{\boxed{b}}^{\text{model syst.}} \boxed{w(b)_j^{s-O}} \Big) \Bigg) \\
+ \sum_{\boxed{k}}^{\text{bkg true int.}} N_i^{MC\ \text{bkg}\ k} \prod_{\boxed{a}}^{\text{model syst.}} \boxed{w(a)_j^k} \Bigg] \boxed{t_{ij}^{det}} \boxed{r_j^{det}} \sum_{\boxed{n}}^{E_\nu} w_n^j \boxed{f_n}
\end{aligned}
$$

Fit par for Carbon

Resp. function

reco bin

true bin

True ν sig. interactions

Interaction pars affecting CC0π-C

Fit par for Oxygen

Model pars affecting CC0π-O

Detector pars

flux pars

True ν bkg interactions

Model pars affecting bkg

Transfer matrix

True E_ν bins

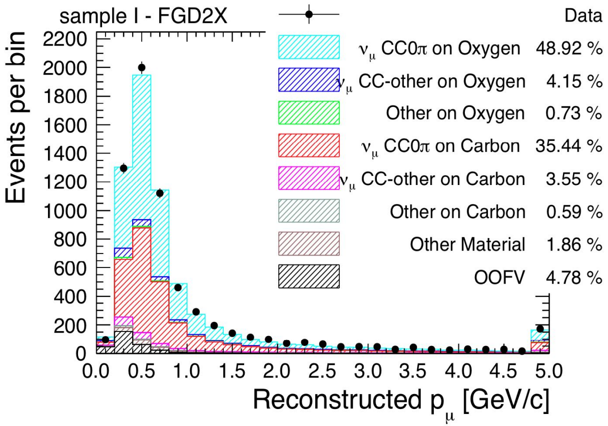
The Barlow-Beeston likelihood

$$-2\ln\mathcal{L}_{stat} = \chi^2_{\mathrm{stat}} = \sum_{j}^{\mathrm{bins}} 2\left(\beta_j N_j^{\mathrm{exp}} - N_j^{\mathrm{obs}} + N_j^{\mathrm{obs}}\ln\frac{N_j^{\mathrm{obs}}}{\beta_j N_j^{\mathrm{exp}}} + \frac{(\beta_j - 1)^2}{2\sigma_j^2}\right)$$

From the reco to the truth space

(taking as an example the [Oxygen and Carbon CC0pi measurement from T2K]())

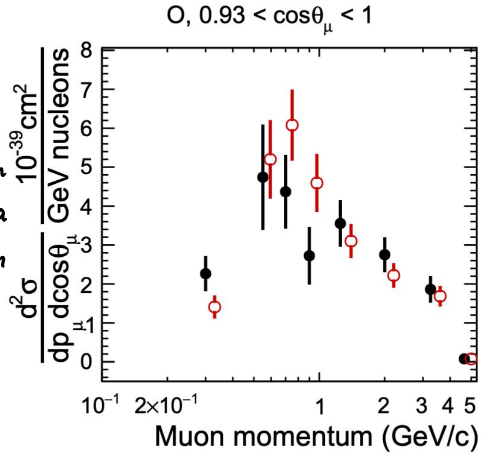
Reco space



Just an example in 1D and for one sample

Truth space

Just an example of a 2D bin



We dispose of the number of selected events in terms of reconstructed variables

efficiency correction

background subtraction

unfolding of detector effects

We want to extract the total number of signal events in terms of true variables