Cross-section measurement of neutrino interactions in the TPC gas of the T2K off-axis near detector

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Lukas Koch, M.Sc.

aus

Düsseldorf

Berichter: Prof. Dr. Stefan Roth
Prof. Dr. Christopher Wiebusch

Tag der mündlichen Prüfung: 11.10.2018

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek verfügbar.

Abstract

Die TPCs des T2K Off-Axis-Nahdetektors sind insgesamt mit etwa $10\,\mathrm{m}^3$ einer Argon-basierten Gasmischung (95% [91%] Argon nach Volumen [Masse]) gefüllt. Da sie seit Beginn des T2K-Experiments dem Neutrinostrahl ausgesetzt sind, ist es möglich mit ihnen Neutrino-Wechselwirkungen mit dem TPC-Gas als aktivem Target zu untersuchen. In dieser Arbeit werden die Ergebnisse der Datennahmeperioden Zwei bis Vier vorgestellt. Dies entspricht 5.73×10^{20} Protonen auf dem Graphittarget des Beschleunigers, $\sim 15\%$ der aktuell bewilligten Datenmenge mit Myon-Neutrino-Strahl für das T2K-Experiment. Aktuelle Modelle sagen etwa 600 Wechselwirkungen von Myon-Neutrinos über geladene Ströme im Referenzvolumen der TPCs in diesem Datensatz voraus.

Neutrino-Ereignisse in den TPCs werden mithilfe der Rekonstruktionssoftware `TREx` selektiert. Dies geschieht in zwei Schritten: Zuerst wird das negative, in den TPCs startende Teilchen mit dem höchsten Impuls selektiert. Dieses definieren den Vertexkandidaten. Anschließend werden zusätzliche Schnitte angewandt, um den Untergrund an Ereignissen von außerhalb des Referenzvolumens zu reduzieren. Die Selektion erreicht eine Reinheit von $\sim 26\,\%$ und eine Effizienz von $\sim 32\,\%$.

Mit dieser Selektion wird eine Messung des Wirkungsquerschnitts von Myon-Neutrinos über geladene Ströme im TPC-Gas durchgeführt. Hierzu wird ein Ansatz mit Fokus auf der Detektorantwort-Matrix gewählt. Diese Matrix übersetzt Erwartungswerte im wahren Phasenraum, also Ereignisbeschreibungen mit ihren wahren Eigenschaften, zu Erwartungswerten im rekonstruierten Phasenraum, also Beschreibungen mit rekonstruierten Eigenschaften. Sie beschreibt also die Detektoreffizienz und -auflösung.

Systematische Detektorunsicherheiten werden ebenfalls von der Matrix abgedeckt. Entsprechend der Unsicherheiten werden mehrere Antwortmatrizen generiert. Jede Matrix entspricht einem möglichen wahren Detektor und produziert eigene Erwartungswerte für die Rekonstruktion. Die Erwartungswerte der verschiedenen Matrizen können einzeln mit den Daten verglichen und anschließend in eine marginale Likelihood kombiniert werden.

Wenn die Antwortmatrix so konstruiert wird, dass sie nicht vom verwendeten Interaktionsmodell abhängt, können mit ihr beliebige Interaktionsmodelle mit den Daten verglichen werden. Da die Matrizen die Detektorunsicherheiten beinhalten, müssen diese Unsicherheiten nicht für jedes Modell neu evaluiert werden. Werden die gemessenen Daten und die Detektorantwort-Matrix zusammen veröffentlicht, können auch Modellentwickler, die nicht im Detail mit dem Detektor vertraut sind, ihre Modelle mit den Daten vergleichen. Hierfür müssen sie nur die Erwartungswerte im wahren Phasenraum entsprechend variieren und mit der Matrix multiplizieren. Verglichen mit dem "klassischen" Ansatz, bei dem neue Modelle normalerweise in dedizierten Analysen innerhalb der Experimente getestet werden, ermöglicht dies einen deutlich schnelleren Entwicklungszyklus.

Mit diesem Ansatz wurden Anpassungen nomineller Templates der Ereignisgeneratoren Neut und Genie unternommen. Diese liefern Wirkungsquerschnitte pro Nukleon und Neutrinoenergie von $(0.42 \pm 0.15) \times 10^{-38}\,\mathrm{cm}^2/\mathrm{GeV}$ und $(1.02 \pm 0.24) \times 10^{-38}\,\mathrm{cm}^2/\mathrm{GeV}$.

Abstract

The TPCs of the T2K off-axis near detector contain in total about $10\,\mathrm{m}^3$ of an argon-based gas mixture (95% [91%] argon by volume [mass]) at atmospheric pressure. They have been exposed to the T2K neutrino beam for the full duration of T2K operation. It is thus possible to investigate neutrino interactions that happen inside the gas volume of the TPCs, using it as active target. This work presents the results obtained from T2K runs two through four. The data corresponds to 5.73×10^{20} protons on target, $\sim 15\%$ of the total muon-neutrino beam exposure expected by the end of the currently approved T2K programme. Current neutrino cross-section models predict about 600 charged-current muon-neutrino interaction events in the TPCs' fiducial volume in this data set.

Neutrino events in the TPCs are selected using the new TPC reconstruction software `TREx`. The selection works in two stages: First, the highest momentum, negatively charged particle that starts inside one of the TPCs is selected. This defines the vertex candidate. Then, additional cuts are applied to that candidate to reduce the out-of-fiducial-volume background. The selection reaches a purity of $\sim 26\,\%$ and an efficiency of $\sim 32\,\%$.

This selection is then used to perform a measurement of the charged-current muon-neutrino cross-section in the TPC gas. The response-matrix centred approach is used. This method is based on the detector response matrix, which translates expectation values for event counts in truth space, i.e. describing the true properties of the events, into expectation values in reco space, i.e. describing the events in reconstructed variables. In it contained are the information about detector efficiency and detector smearing.

Systematic detector uncertainties are also handled with the response matrix. A set of response matrices is generated according to the uncertainties in detector properties. Each matrix corresponds to one possible true detector performance and yields its own reco-space prediction for a given truth-space prediction. The different predictions can then be compared to the recorded data, and combined into a marginal likelihood.

If the construction of the response matrix is done with care (ensuring its physics model independence), it can be used for tests of arbitrary theories against the data very easily. Since the set of response matrices contains the knowledge about the systematic detector uncertainties, the systematics do not have to be re-evaluated for each model that is tested. If the data and response matrix are published together, the model tests can also be done by theorists and model builders that are not intimately familiar with the detector that recorded the data. All they have to do is vary the truth space expectation values according to their models and multiply them with the matrix. Compared with the "classic" approach, where new models usually require a new dedicated analysis within the experiments, this makes tests of new model ideas much easier.

A simple template fit with the nominal Neut and Genie models was done using this method. It yields a cross section per nucleon and scaled with average neutrino energy of $(0.42 \pm 0.15) \times 10^{-38}\,\mathrm{cm}^2/\mathrm{GeV}$ and $(1.02 \pm 0.24) \times 10^{-38}\,\mathrm{cm}^2/\mathrm{GeV}$ respectively.

Contents

List of Figures

List of Tables

1. Introduction

1.1. The T2K experiment

The T2K (Tokai to Kamioka) experiment is a long-baseline neutrino beam experiment in Japan. A high intensity (anti-) muon neutrino beam is created at the J-PARC facility in Tokai, Ibaraki prefecture, at the east coast of Japan. Protons are accelerated in a linear accelerator and synchrotron, and then fired on a graphite target with a repetition rate of one spill every 2 to 3 seconds[1] (see Fig. 1.1). Each spill consists of 8 bunches of protons with a bunch-to-bunch time separation of $\sim 580\,\mathrm{ns}$ and bunch widths of $\sim 60\,\mathrm{ns}$. The resulting nuclear interactions in the target produce charged pions. These are then charge-filtered and focussed into a 96 m long, evacuated decay pipe by three magnetic horns. There the pions decay to (mostly) muons and neutrinos. Because of the high momentum of the pions relative to the target frame, the muons and neutrinos are boosted in the direction of the pion beam. Muons are stopped by the material after the decay pipe, while the neutrinos travel on unhindered.[1]

This neutrino beam is aimed at the Super Kamiokande detector at a distance of $\sim 295\,\mathrm{km}$ (see Fig. 1.2). Super Kamiokande (SK) is a water Cherenkov detector with a fiducial mass of 22.5 t. It is located underground in the Mozumi Mine in the Gifu prefecture of Japan. Since SK is positioned slightly off-axis at 2.5 ° with respect to the neutrino beam centre, a very narrow neutrino energy distribution is achieved (see Fig. 1.3). The peak neutrino energy is at 600 MeV.

[1] The beam performance has been steadily increased.

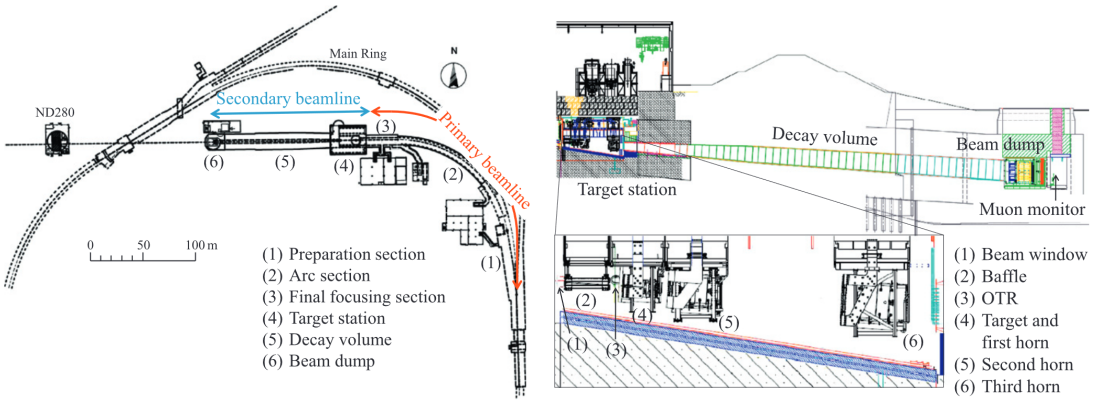


Figure 1.1.: The T2K neutrino beam line. [2]

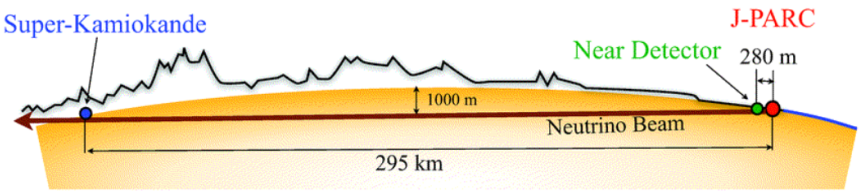


Figure 1.2.: The T2K experiment. [2]

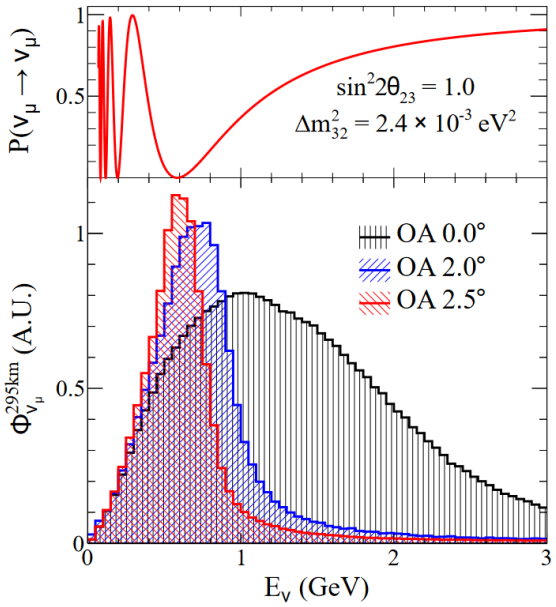


Figure 1.3.: Neutrino beam profiles at different off-axis angles. The Super-Kamiokande detector is at 2.5°. [3]

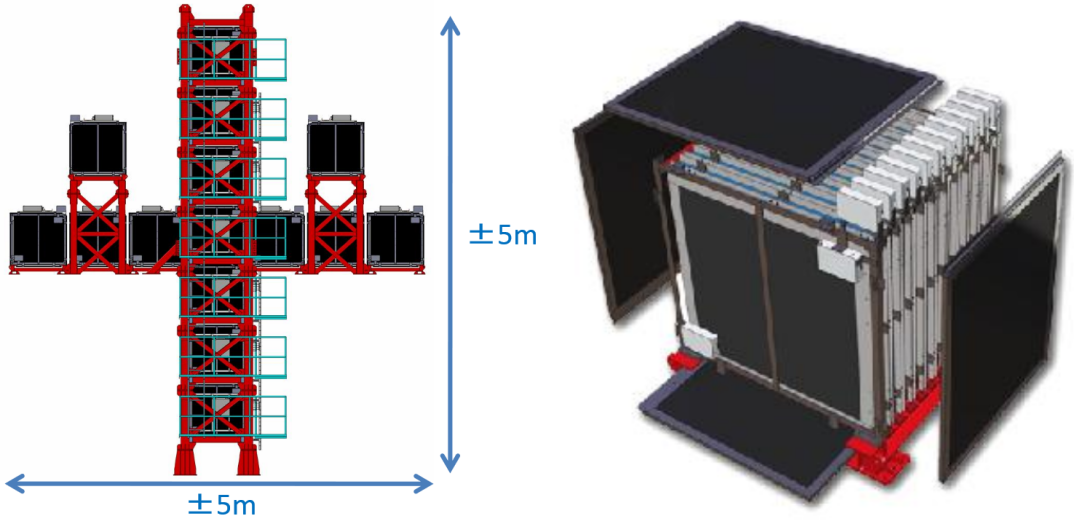


Figure 1.4.: The INGRID on-axis near detector. The 16 identical modules are centred on the nominal neutrino beam centre. Each module is a sandwich made of nine iron target plates and eleven scintillator trackers, surrounded by scintillator veto planes. [6]

T2K's main objective is the investigation of neutrino oscillations. Notable results include the world's first detection of electron-neutrino appearance from a muon-neutrino beam[4] and finding hints for normal neutrino mass ordering and maximal lepton sector CP violation in a joint neutrino and anti-neutrino beam analysis[5].

These measurements are only possible because of a precise characterisation of the unoscillated neutrino beam properties at the near detectors. They are located 280 m downstream of the graphite target at J-PARC. The on-axis detector INGRID consists of 14 detector modules of sandwiched iron plates and scintillator planes (see Fig. 1.4). It measures the actual neutrino beam position to with an accuracy better than 0.4 mrad, and monitors the total flux stability via event rate measurements[6].

The off-axis detector ND280 consists of several solid scintillator detectors and three large TPCs in a magnet yoke (see Fig. 1.5). It is positioned at the same off-axis angle as Super-Kamiokande, and thus investigating the same (unoscillated) neutrino energy spectrum that is relevant for the oscillation analyses. The Pi-Zero-Detector (P0D) is a scintillating bar tracker with added layers of lead and bags that can be filled with water as passive targets. The Time Projection Chambers (TPCs) are three large particle detectors based on gas ionisation. The Fine Grained Detectors (FGDs) are two scintillating bar trackers sandwiched between the three TPCs. The second (downstream) FGD also includes water layers as passive target for cross-section measurements. These three detector systems are surrounded by electromagnetic calorimeters (P0D-, Barrel-, and Downstream-ECALs). All of this is contained in the magnet yoke, which creates a magnetic field of $\sim 0.2\,\mathrm{T}$ in the detector. The magnetic yoke itself is instrumented with the Side Muon Range Detector (SMRD). It consists of scintillating tiles in-between the

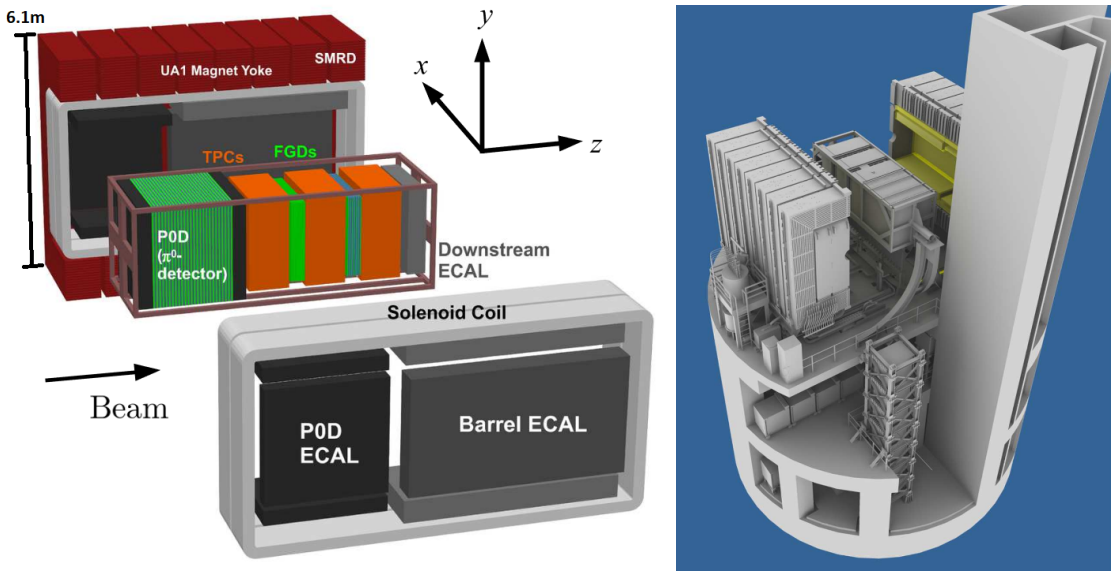


Figure 1.5.: The off-axis near detector ND280. It is located at the same off-axis angle as the Super-Kamiokande detector, above the INGRID on-axis detector. [2]

Table 1.1.: The ND280 sub-detectors.

	Resolution	Notes
P0D	33.0 mm	pitch of triangular scintillator bars in x and y layers interleaved with passive lead and water layers
FGD1	9.61 mm	square scintillator bars in x and y layers
FGD2	9.61 mm	alternating with passive water target layers
TPCs	7.0 mm × 9.8 mm	vertical × horizontal pad pitch on readout plane
P0D ECAL	40 mm × 10 mm	rectangular scintillator bars in x and y layers interleaved with passive lead layers
Barrel ECAL	40 mm × 10 mm	rectangular scintillator bars in x and y layer interleaved with passive lead layers
Downstream ECAL	40 mm × 10 mm	rectangular scintillator bars in x and y layers interleaved with passive lead layers
SMRD	875 mm × 167/175 mm	scintillator plates in the (horizontal/vertical) gaps of the magnet yoke

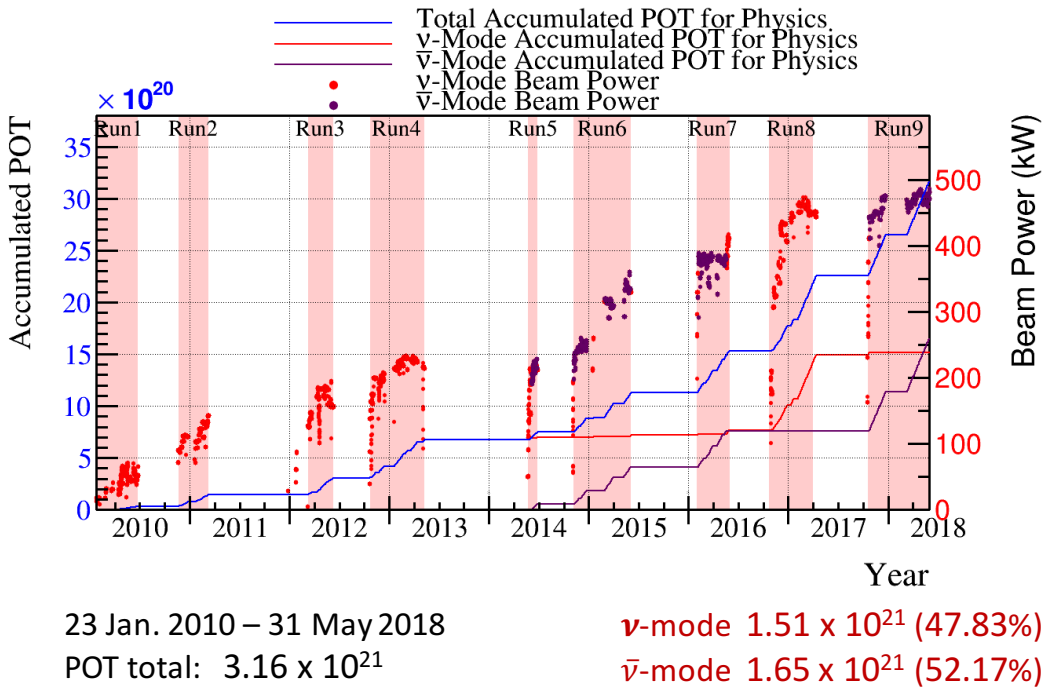


Figure 1.6.: Protons On Target at T2K [7]. The data set used in this analysis corresponds to Runs 2, 3, and 4. In total they make up about half of the presently available data in neutrino enhanced beam mode.

yoke's iron sheets. See Tab. 1.1 for an overview of the respective sub-detector resolutions.

The T2K beam can be operated in "neutrino enhanced" or "anti-neutrino enhanced" mode, by focussing either positive or negative pions into the decay pipe. The data analysed in this work corresponds to about half the currently available data in neutrino enhanced mode (see Fig. 1.6). A detailed description of the used data samples can be found in Section 2.1.

1.2. Nuclear effects

1.2.1. Why nuclear matter matters

Despite the narrow neutrino energy distribution, it is still necessary to reconstruct the neutrino energy on an event-by-event basis to achieve the highest possible sensitivity in oscillation analyses. At Super Kamiokande the energy reconstruction is done using event kinematics. The momenta and masses of reconstructed particles are combined to calculate the original (invisible) momentum and energy of the incoming neutrino.

In the simplest case of charged-current quasi-elastic (CCQE) interactions on free nucleons, the incoming neutrino converts to a charged lepton (e.g. a muon) and the nucleon recoils from the interaction as a whole (see Fig. 1.7). The energy transfer to the nucleon

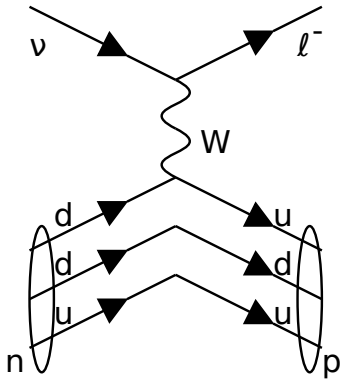


Figure 1.7: Charged-current quasi-elastic (CCQE) interaction. A neutrino weakly interacts with a (free) neutron, converting a down-quark into an up-quark. The quark stays bound in the nucleon (now a proton) after the interaction.

is usually smaller than the charged lepton energy, due to the nucleon's higher mass. The nucleon thus often stays below the detection threshold of the detector. Even so, the neutrino energy is almost[2] perfectly determined by the charged lepton kinematics, as the initial conditions of the nucleus and direction of the neutrino are known:

$$\boldsymbol{p}_\nu + \boldsymbol{m}_n = \boldsymbol{p}_l + \boldsymbol{p}_n,$$

$$\boldsymbol{p}_\nu = \begin{pmatrix} E_\nu \\ E_\nu \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{m}_n = \begin{pmatrix} m_n \\ 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{p}_l = \begin{pmatrix} E_l \\ p_l \\ p_t \\ 0 \end{pmatrix}, \boldsymbol{p}_n = \begin{pmatrix} \sqrt{p_n^2 + p_t^2 + m_n^2} \\ p_n \\ -p_t \\ 0 \end{pmatrix}.$$

Here p_t is the (balanced) momentum of the outgoing particles, transverse to the incoming neutrino direction. p_l and p_n are the longitudinal momenta of the charged lepton and outgoing nucleon respectively. This method of determining the neutrino energy breaks down if there are more invisible particles involved, e.g. pions from inelastic interactions. The average effect of such particles can be corrected, but an accurate event-by-event reconstruction of the energy becomes impossible.

Unfortunately there is currently no all-encompassing model that is able to describe the scattering of neutrinos off nuclei. The field can be broadly categorised in three (overlapping) areas of study (see Fig. 1.8): the initial state, the cross section, and final state interactions (FSI).

An overview of the current state of nuclear interaction modelling, event generation, and their impact on neutrino oscillation analyses is given in [8]. All these uncertainties propagate to the corrections that are necessary to reconstruct the neutrino energy from just the charged lepton kinematics. They are thus a direct source of systematic uncertainties in neutrino oscillation experiments such as T2K. Currently, their impact is in the order of $5\% - 10\%$

1.2.2. The initial state

The initial state describes the nucleus before the interaction with the neutrino. In the common *impulse approximation* neutrinos are assumed to interact with the nucleons

[2] Modulo thermal movement of the nucleon

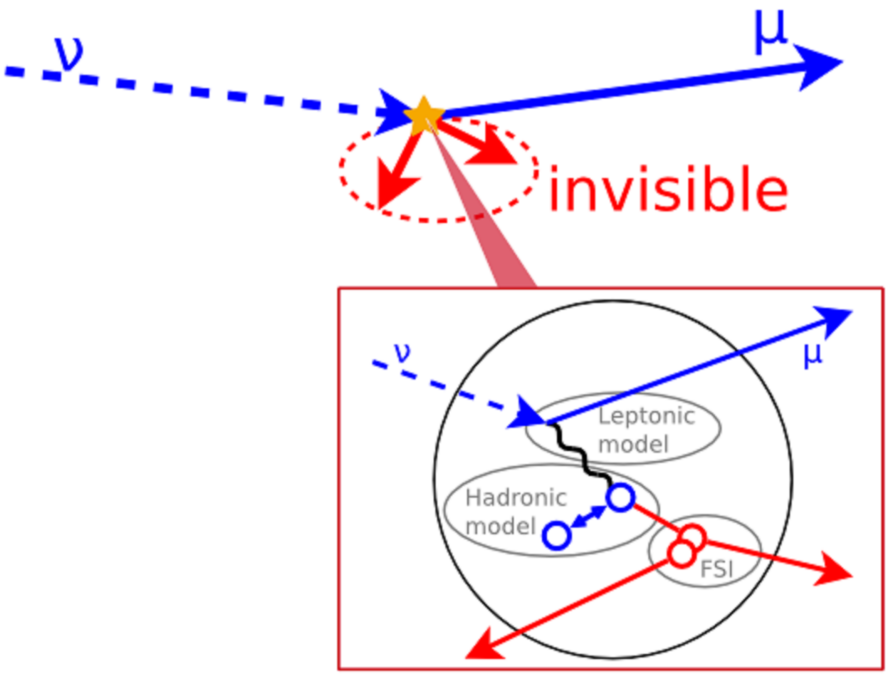


Figure 1.8.: Nuclear effects.

inside the nucleus, rather than the whole nucleus or the partons inside the nucleons. This is a useful approximation for "medium range" neutrino energies $\mathcal{O}(1\,\mathrm{GeV})$. The initial state is then described with a *spectral function*, which give the probability of encountering a nucleon with a given momentum and energy.

One such spectral function is the *relativistic Fermi gas* (RFG) model, where the nucleons are assumed to be non-interacting, populating a homogeneous, spherical nucleus. They fill up available states up to the Fermi momentum p_F and its corresponding energy E_F. The energy it takes to remove one nucleon from the nucleus is called binding energy E_B (see Fig. 1.9). The values of p_F and E_B can be deduced from electron scattering data [9].

An extensions of the RFG model is the *local Fermi gas* (LFG). In it, the global Fermi momentum p_F is replaced by a local one, dependent on the local density of protons/neutrons (see Fig. 1.10). This is a somewhat more realistic approach [11].

While both the RFG and LFG models provide spectral functions, usually only more complicated models that cannot be summarised in closed form are called *spectral function* (SF) models. These general models usually include the effects of nucleon-nucleon interactions and generalise the single-particle spectral function to a multi-particle one, see e.g. [12] and [13].

This kind of *initial state correlation* (ISC) can lead to final states with multiple nucleons and/or "holes" (empty momentum states left behind by excited nucleons) in them. Especially final states with two particles and two holes (2p2h) have recently been found

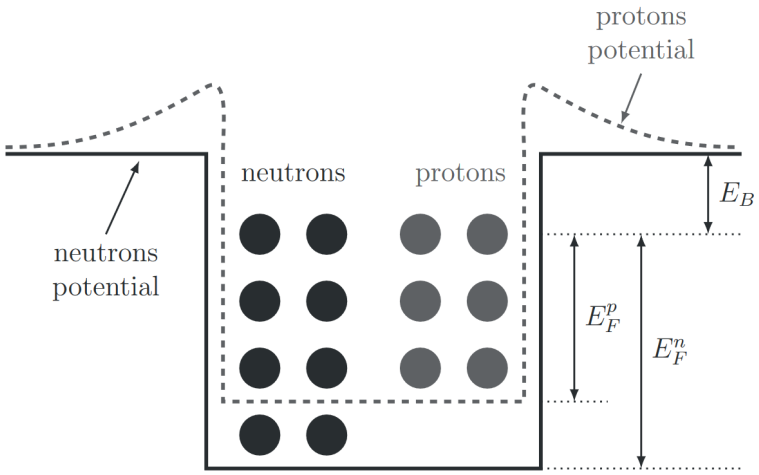


Figure 1.9.: Nuclear potential in the Fermi gas model [10], with the binding energy E_B, and the Fermi energies of protons and neutrons $E_F^{p,n}$.

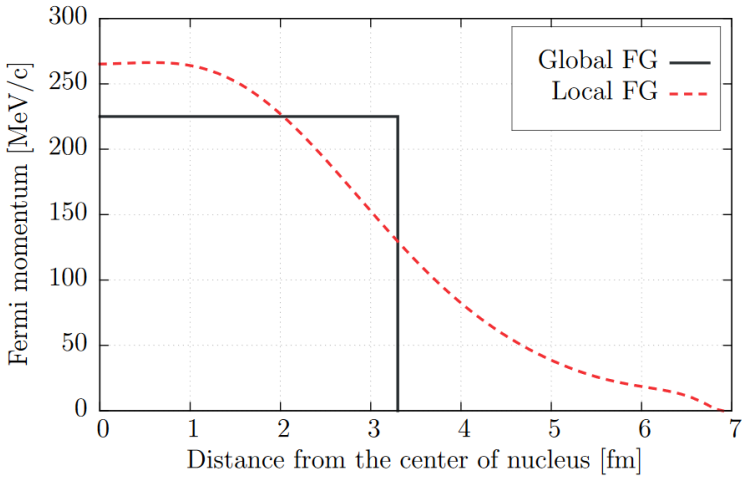


Figure 1.10.: Fermi momentum in the local Fermi gas model [10].

to play an important tole in describing differential neutrino cross-sectionsi, e.g. [14]. ISCs are not the only mechanism that leads to 2p2h final states though. *Final state correlations* (FSC) caused by the struck nucleon scattering off the spectator nucleons and *meson exchange currents* (MEC) also play a role (see e.g. [15]).

1.2.3. The cross section

Cross sections of neutrino interactions with nuclei at medium energies are dominated by four processes: (quasi-)elastic scattering, resonant pion production, coherent pion production, and deep inelastic scattering. In the *neutral current elastic* (NCEL) scattering process, the neutrinos scatter on the nucleons without producing any additional particles:

$$\nu + N \rightarrow \nu + N.$$

This is a true elastic process. The *charged current quasi-elastic* (CCQE) process is very similar, but because both the lepton and the nucleon change their mass in the interaction, it is not truly elastic:

$$\nu_l + n \rightarrow l^- + p, \quad \bar{\nu}_l + p \rightarrow l^+ + n.$$

The electroweak model of quantum field theory that governs this interaction is very well understood and allows very precise predictions of cross-sections. The uncertainties of this interaction stem from the fact that the nucleons are not point-like particles, but have an inner structure of partons. This structure enters the cross-section calculation as form factors, which have to be determined experimentally.

During *resonant pion production* (RES) the neutrino scatters off a nucleon, exciting it to a resonant state, which in turn decays to (mostly) a nucleon and a pion:

$$\nu_l + n/p \rightarrow l^- + \Delta^{+/++}, \quad \bar{\nu}_l + n/p \rightarrow l^+ + \Delta^{-/0}. \quad \overset{(-)}{\nu_l} + n/p \rightarrow \overset{(-)}{\nu_l} + \Delta^{0/+}.$$

Calculations of this cross section have to consider the effects of all possible resonances (up to a cut-off energy) [16] [17]. Just like in the NCEL and CCQE case, the inner structure of the nucleons must be considered in the form of form factors [18]. Recent developments of this cross section have combined the resonant channel with non-resonant interference terms[19] [20], leading to a much more realistic model [21]. This model is currently being implemented in the NEUT[22] and GENIE[23] event generators, which are used by most long-baseline neutrino beam experiments.

At the high energies, neutrinos can be modelled to interact with the nucleon constituents rather than the nucleons as a whole. This regime is called *deep inelastic scattering* (DIS). These interactions will generally "destroy" the nucleon and produce a hadronic shower of particles:

$$\overset{(-)}{\nu_l} + N \rightarrow l^{-(+)} + X, \quad \overset{(-)}{\nu_l} + N \rightarrow \overset{(-)}{\nu_l} + X.$$

At the highest energies, this leads to predictable cross sections (see Fig. 1.11), but it becomes hard at intermediate energies, when it overlaps with the resonant production region. Care has to be taken to avoid double-counting of processes.

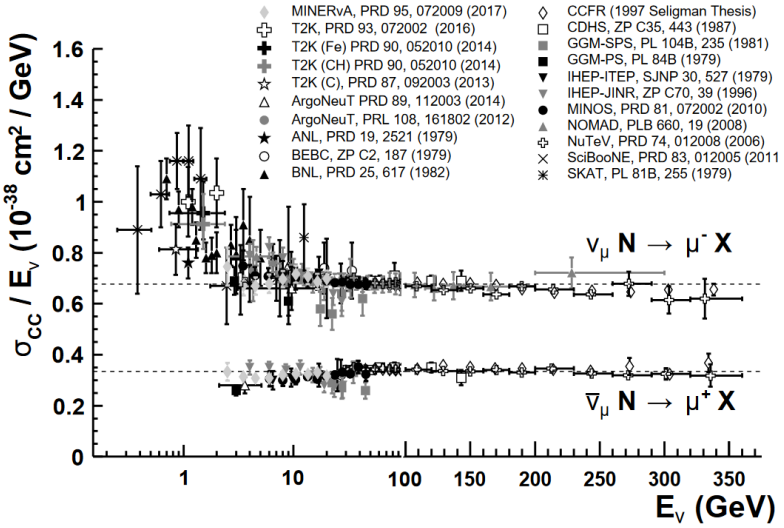


Figure 1.11.: World data of inclusive charged-current neutrino cross sections [24]. At high energies the total cross section becomes proportional to the neutrino energy E_ν. This is a feature of the scattering on the point-like quarks in the nucleons.

1.2.4. Final state interactions

The largest uncertainties stem from so called *final state interactions* (FSI), which is a collective term for effects induced by interactions of the final state particles with the nuclear matter on their way out of the nucleus. Most implementations of this process assume the final stat particles of the interaction to independently traverse the dense nuclear medium (cascade models). The GiBUU model[25] is an exception to this rule. It propagates the phase-space density of the final state particle species through a semi-classical mean field of the nucleus. The evolutions of the different particle species is explicitly coupled this mean field and explicit collision terms [26].

1.3. Neutrino gas interactions in the ND280 TPCs

At T2K these uncertainties are constrained with external cross-section data and by doing cross-section measurements with the near detectors ND280 and INGRID. The design of ND280 is primarily optimised to record neutrino interactions that occur in the solid state detectors (like the FGD and P0D). For these events, the TPCs are used in conjunction with the solid detectors to identify the particles that are produced in the vertex. This means that the detection thresholds of particles are determined by the properties of the solid detectors, e.g. the FGDs are sensitive to protons with a minimal momentum of $\sim 450\,\mathrm{MeV/c}$.

The detection thresholds of particles originating in the TPCs are much lower than that, since the TPCs are instrumented with a higher granularity and the density of the

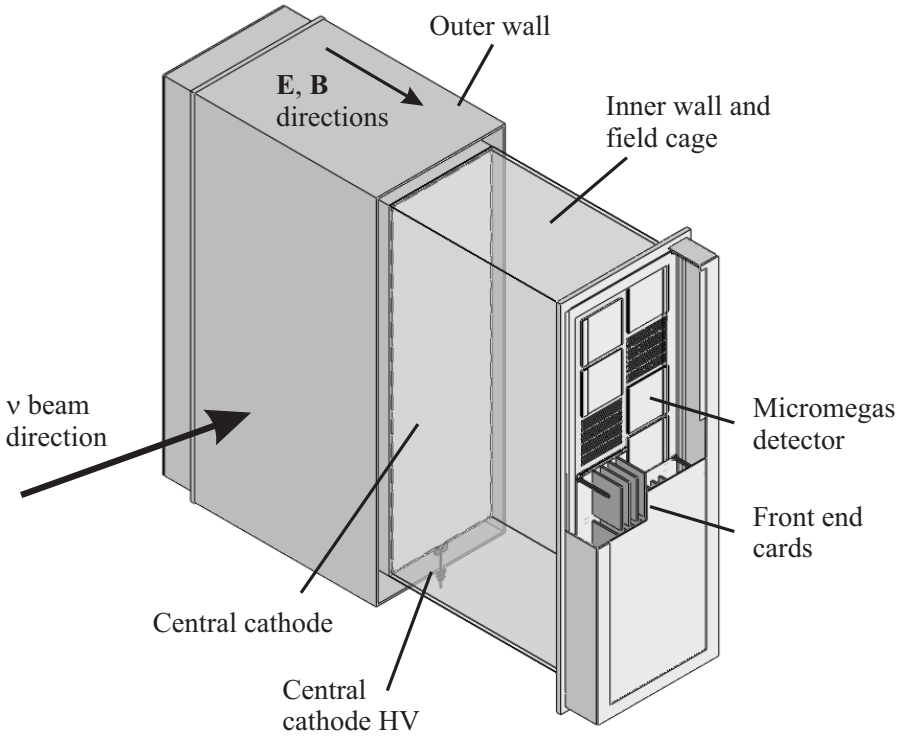


Figure 1.12.: The TPCs [27]. The inner wall and field cage's outer dimensions are $(1808 \times 2230 \times 854)$ mm (width × height × depth). The outer box is separated from the inner box by at least 68 mm of carbon dioxide for insulation of the −25 kV cathode voltage. The outer boxes measure $(2302 \times 2400 \times 974)$ mm. The inner box is operated at 0.4 mbar overpressure with respect to the carbon dioxide in the gap. The carbon dioxide is held at less than 5 mbar overpressure with respect to the surrounding atmosphere.

active material is much lower than that of the solid detectors. The ND280 TPCs consist of three large gas-filled volumes (total $\sim 10\,\mathrm{m}^3$), each divided in two halves by a central cathode (see Fig. 1.12). The anodes are instrumented with Micro Mesh Gaseous Structure (micromegas) modules. They are segmented into pads with a vertical (horizontal) pitch of 7.0 mm (9.8 mm)[27]. The nominal gas mixture by volume (by mass) is:

- Argon: 95% (91%)
- CF_4: 3% (6%)
- iC_4H_{10}: 2% (3%)

The gas is a factor ~ 1000 less dense than the plastics of the scintillators, leading to a corresponding increase in particle track lengths. The TPCs can efficiently reconstruct and identify protons down to a momentum of $\sim 200\,\mathrm{MeV/c}$ with the potential to push this limit down to $\sim 60\,\mathrm{MeV/c}$ with improvements of the reconstruction code for these low-momentum, high-energy-loss particles. At $60\,\mathrm{MeV/c}$ the protons have a predicted average track length in the TPCs of only about 5 cm, probably reaching the limit of the TPC resolution for tracking. But even without any dedicated efforts, the detection threshold is already much better than in the solid detectors, allowing to investigate the nuclear effects in previously inaccessible regions of the phase space.

The downside of the low target density is the low total target mass. The interaction rate in the TPC gas volume is also about a factor ~ 1000 lower than in the solid detectors. Nevertheless, since the ND280 TPCs have been exposed to the T2K neutrino beam since the beginning of the experiment, they have recorded a usable amount of neutrino interactions in the sensitive gas volume of the TPCs. The generator models (Neut[22] and Genie[23]) predict a signal event rate of about ~ 100 (~ 50) events per 1×10^{20} protons on target (POT) in the outer (inner) fiducial volume (see Section 2.2.1). This corresponds to ~ 600 (~ 300) events in the data analysed for this thesis (see Section 2.1).

The selection of those events is described in Chapter 2. Since the original reconstruction software's sole intended used was to track particles that enter the TPC from the solid detectors, it had no concept of vertices inside the TPCs. Therefore, it was necessary to extend the functionality of the software. This was implemented with the TPC Reconstruction Extension (`TREx` [28]), which is not only used for the gas interaction analysis, but will also replace the old TPC reconstruction software in future iterations of the solid detector analyses.

1.4. Event reconstruction in the TPCs

1.4.1. General operation

The process of measuring particles with a Time Projection Chamber can be divided in three stages:

1. Ionisation

2. Drift

3. Charge collection

If a charged particle traverses the gas inside the TPCs, it loses energy by ionising the gas molecules. This ionisation track is what is being measured in the TPC. The mean energy loss $\langle -\mathrm{d}E/\mathrm{d}x \rangle$ of heavy (i.e. not electrons) charged particles at intermediate momenta ($0.1 \lesssim \beta\gamma \lesssim 1000$) is described by the Bethe equation [24]:

$$\left\langle -\frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle = Kz^2 \frac{Z}{A} \frac{1}{\beta^2} \left[\frac{1}{2} \ln \frac{2m_e c^2 \beta^2 \gamma^2 W_{\max}}{I^2} - \beta^2 - \frac{\delta(\beta\gamma)}{2} \right].$$

Here z is the charge of the traversing particle, Z and A are the atomic number and molar mass of the stopping material (i.e. the gas in the TPCs), and K is the constant:

$$K = 4\pi N_A r_e^2 m_e c^2 \approx 0.307\,\mathrm{MeV\,mol^{-1}\,cm^2}.$$

I is the material's mean excitation energy ($\mathcal{O}(Z \times 10\,\mathrm{eV})$). $\delta(\beta\gamma)/2$ is the density effect correction that is needed at high energies because the stopping material gets polarised, effectively blocking interactions of the traversing particle's electric field. $W_{\max}$ is the maximum energy transfer in a single collision. For a traversing particle with mass M it is:

$$W_{\max} = \frac{2m_e c^2 \beta^2 \gamma^2}{1 + 2\gamma m_e/M + (m_e/M)^2}.$$

It is worth mentioning that the mean energy loss is very much driven by these unlikely, high-energy-loss interactions. The variation of energy deposited in a thin detector slice is well described by a Landau distribution. Its most probable value is much lower than the mean energy loss according to the Bethe equation (roughly half).

The electron-ion-pairs are separated by the applied drift field. In the T2K TPCs this field is 275 V/cm. The free electrons drift along the electric field lines towards the anode read-out plane, while the ions drift towards the central cathode. The electron drift velocity is 7.8 cm/μs, while the ions are about a factor 1000 slower. At T2K, only the electron signal is used for reconstruction.

At the anode, the electrons are recorded, providing a 2D projection of the track. The read-out technology employed at T2K is that of micro mesh gaseous structures (micromegas). They consist of a thin wire mesh $\sim 100\,\mu$m over sensitive rectangular copper pads. A high electric field between between the mesh and the pads causes the free electrons to ionise more gas molecules, leading to an electron cascade (see Fig. 1.13). The voltage applied between mesh and pads at T2K is 350 V, leading to a gas amplification factor of about 2000. The amplified charge is collected on the pads and digitised using an Application-Specific Integrated Circuit (ASIC) called "AFTER" [27].

1.4.2. Hit calibration

For a reliable track reconstruction and particle identification in the TPCs, the influence of changing environmental conditions has to be taken into account. The TPCs are operated

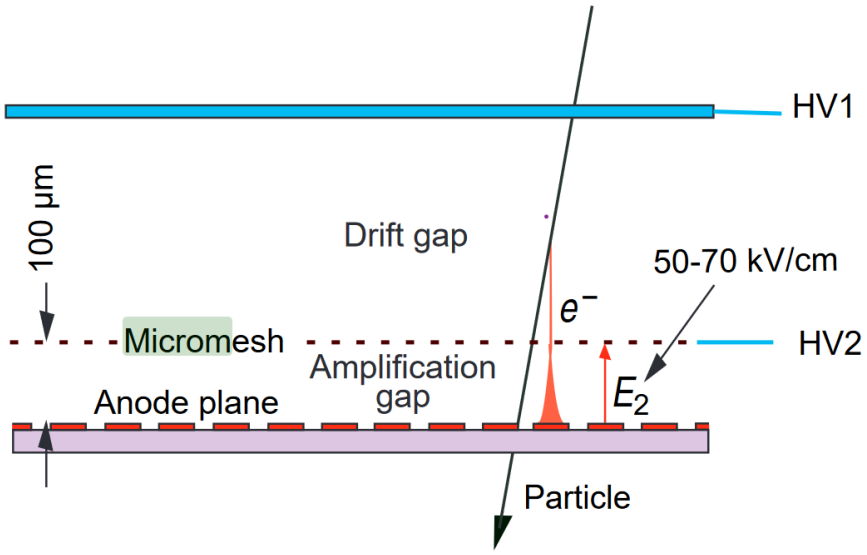


Figure 1.13.: Schematic drawing of a micromegas detector[24]. In the T2K TPCs the amplification voltage $HV2$ is 350 V and the drift "gap" is the drift volume with an electric field of 275 V/cm.

slightly above atmospheric pressure to avoid contamination of the drift volume with other gases. To reduce the load on the structural elements, this overpressure is kept constant, so the absolute pressure in the TPCs varies together with the atmospheric pressure outside the detector. There is no temperature regulation either, so both temperature and pressure depend on the weather conditions. This leads to changing gas densities and thus changing detector responses.

The specific energy loss of particles in the gas is proportional to the gas density. In the approximation of an ideal gas:

$$\frac{\mathrm{d}E}{\mathrm{d}x} \propto \rho \propto \frac{p}{T}.$$

This means the measured hit[3] charges Q have to be converted to a nominal charge Q_0 at a standard density to be comparable:

$$Q_0 = Q\frac{p_0}{T_0}\frac{T}{p},$$

with the pressure p_0 and the temperature T_0 of the chosen standard conditions.

The electron avalanche amplification in the micromegas also depends on the gas density, so it has to be corrected as well:

$$Q = Q'\frac{A_0}{A},$$

where Q' is the uncorrected measured charge and A is the amplification in the readout plane. In first approximation, the electric field in the micromegas is homogeneous. The

[3]The signal recorded on a single pad at the readout plane.

amplification is then an exponential process depending on the first Townsend coefficient α_T and the length of the amplification region d:

$$A = \exp(\alpha_T d).$$

The Townsend coefficient of a given gas mixture depends on the gas density and the electric field:

$$\alpha_T = \frac{p}{T} \mathrm{f}\left(E\frac{T}{p}\right),$$

with the function f only depending on the reduced electric field ET/p. Above the amplification threshold, it can be approximated very well with a linear function:

$$\mathrm{f}(ET/p) = a \cdot ET/p + b.$$

For the amplification that means:

$$\begin{aligned} A &= \exp\left(d\frac{p}{T}\left(a \cdot E\frac{T}{p} + b\right)\right) \\ &= \exp\left(adE + bd\frac{p}{T}\right). \end{aligned}$$

The correction then becomes:

$$\frac{A_0}{A} = \exp\left(bd\left(\frac{p_0}{T_0} - \frac{p}{T}\right)\right)$$

For small density changes, this can be approximated as a relative correction:

$$\frac{\Delta A}{A_0} = c\frac{\Delta(p/T)}{p_0/T_0}.$$

At the T2K TPCs the constant c is roughly -3.2, so for every increase of gas density of 1% the gain drops by 3.2%. It is measured in regular intervals in dedicated Gas Monitoring Chambers (GMCs).

The drift velocity v_d is a function of the reduced electric field ET/p. Since the T2K TPC are operated at the local maximum drift velocity, no immediate correction for temperature and pressure changes is needed. Instead, the drift velocity is monitored in the GMCs for changes due to varying gas compositions. The results of these measurements are used to update the drift velocity calibration about twice per day.

1.4.3. Reconstruction with `TREx`

The `TREx` reconstruction software is described in detail in [28]. This section will only present an overview over the most important concepts and terminology.

The reconstruction begins by grouping the recorded hits into *patterns* of connected, i.e. neighbouring, hits. The ends of particle tracks (either from the real ends of particle trajectories or from particles leaving the TPC) are identified by their extreme positions in the pattern.

Connections between these *path ends* are found using the A^* algorithm[29]. Hits along these connections are grouped into *paths*. Hits where the paths branch off are grouped into *junctions*. This ensures that hits assigned to a certain path likely only contain charge from a single particle. The hits of the junction can contain the charge information of multiple overlapping particles.

This algorithm is very good at finding vertices with three or more particles, even if two of them are produced back to back[4]. It would not, however, find a vertex with two particles, as there are no branching-off paths. These are found by a kink-finding algorithm that looks for a sharp angle in the paths. If one is found, an additional junction is inserted at that position. The kink-finding does *not* work well when the two particles are emitted back to back in the detector frame. Since the neutrino interaction events are usually boosted in the neutrino direction, this is not a big problem though.

The hits of horizontal (vertical) paths are grouped into vertical (horizontal) slices. These *clusters* of hits are used for a more precise determination of the particle position at that point. A fit of the charge distribution in each cluster yields a position with an uncertainty that can be smaller than the widths of the pads of the readout plane.

Lastly, helix *tracks* are fitted to the paths, or rather their hit clusters. The parameters of the best fit helix determine the reconstructed momentum of the track, and its direction. See Fig. 1.14 for an overview of the reconstructed objects.

1.5. Model-independent cross-section measurements

A big challenge for cross-section measurements is their explicit or implicit dependence on the very cross-section models they are supposed to constrain. The reason for this dependence is that, in general, the detector response to the event can vary considerably depending on the particular properties of the event. This can lead to very different efficiencies and purities of event selections if two theories predict very different coverages of the phase space. For example, all detectors have certain energy/momentum thresholds below which they are not sensitive to particles. If two theories (or a theory and reality for that matter) now predict very different fractions of events/particles below that threshold, the resulting efficiency of selecting the events will vary accordingly.

A lot of work is done on minimising or at least quantifying these effects. Strategies range from doing multi-dimensional differential cross-section measurements (to ensure a flat efficiency in each bin), to repeating the analysis with multiple theories and simply quoting how much the results depend on the used model. The former approach requires a lot of data to have a significant number of events in every bin, while the latter suffers from the uncertainty of whether all available models even cover reality at all.

The response-matrix-centred method described in this work aims to combine the model-independence of the multi-dimensional approach with the ability to work with low number of events of the naive model test. This is achieved by de-coupling the binning of the reconstructed events from the description of the events on the generator level.

[4] In fact, the A^*algorithm does not use the orientation of the hits. It only uses the local distances between neighbouring hits.

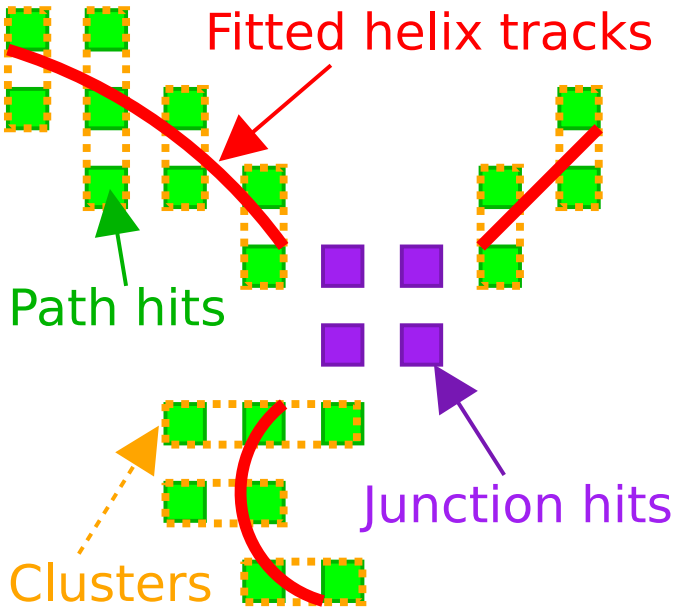


Figure 1.14.: `TREx` objects. Connected collections of hits in the TPCs are called a *pattern*. Patterns are further divided into *paths* – corresponding to single particles – and *junctions*, which cannot be attributed to a single particle. The hits within the paths are grouped in horizontal and vertical clusters, depending on the local angle of the path. Within each cluster, the horizontal (vertical) position is fitted assuming a Gaussian charge distribution about the true particle position. With these fitted positions, *helix tracks* are fitted to the paths using the maximum likelihood method.

The high dimensionality of variables is only needed in *truth space*, i.e. the description of the events on the generator level. The actual recorded data can be binned much coarser in *reco space*, i.e. with fewer reconstructed variables. The response matrix is the connecting piece between the two, describing how likely an event in a particular truth space bin is going to end up in any of the reco space bins.

If the truth binning is chosen carefully, the response matrix should be independent of any assumed physics model of the interactions. That is, different models can predict different truth space distributions, but the values of the response matrix elements do not depend on the model that is used to build the matrix[5]. The real data and response matrix can then be used with arbitrary models to calculate a likelihood and extract cross sections.

This is so far not different from the naive model testing method. The advantage of the response matrix approach is realised when considering the matrix and the raw data as the main result of the measurement. They are (ideally) independent of any model assumptions and can be used to test any new model or model improvement that will be developed in the future. Furthermore, if the raw data and response matrix are published, model developers can use them directly to test any changes or completely new models against old data. Compared to the classical approach, where the theories are thought up by theorists and then tested within the experimental groups in dedicated analyses, this reduces the time of the development cycle considerably. In fact, a lot of work has been spent to make old experimental results available for easy model tuning, for example with the NUISANCE framework[30]. Results obtained with the response-matrix-centred approach would be very easy to include in such global fits.

It might seem like a shortcut for lazy experimental physicist to simply publish the raw data and response matrix to leave the rest to the model builders. This is not the case though, since the construction of the response matrix requires exactly the same understanding of the detector and care to cover all systematics as a classical analysis. Also it is unlikely that any experimental group would publish the data and response matrix without also using them for their own model tests.

The detailed description of how to build the matrix and how to contain the knowledge about the systematic uncertainties in it is given in Chapter 3. The algorithms are implemented in a Python software library called `ReMU`, Response Matrix Utilities. It is intended to make the usage of the data and response matrix as easy as possible. More informations about the software and data formats are included in Section 3.2.4. Finally, the response matrix approach is applied to the neutrino gas interaction analysis and a likelihood template fit is performed to extract an inclusive charged-current cross section from the data. The results of this are presented in Section 3.5.

[5] Aside from statistical effects from the number of available simulated events in each truth bin.

2. Event selection

2.1. Data samples

Table 2.1 shows an overview of the used data samples. Different "productions" refer to different versions of the T2K software stack. The samples "TPC MC Neut" and "TPC MC Genie" are special MC runs for which only the TPC gas was simulated as target material for the neutrinos. They are high statistics signal samples and cannot be used for background studies. These samples only exist as reco-files from production 6B (i.e. without `TREx`), but they were re-processed with the same software version as was used in productions 6H/I. The same applies to the "Sand MC Neut" sample, which is a simulation of "sand muons", i.e. neutrino interactions in the ground material between the decay pipe and the detectors. The regular Monte Carlo samples only include interactions in the active and passive components of the detector. Production 6H and 6I only differ in some minor book-keeping details. The used software is virtually identical. The "Air" or "Water" in the sample names refers to the filling of the P0D's water bags (or lack thereof). Unless written otherwise, the Neut MC samples are used for plots and tables presented in this chapter.

2.2. The selection

2.2.1. Fiducial volume and signal definition

The central particle of in this selection is the negatively charged Minimum Ionising Particle (MIP) with the highest momentum (the *primary MIP*). We define MIPs as both muons and charged pions of any momentum. This definition is used, because the TPCs are not able to distinguish between pions and muons very well. The primary MIP is used to categorize the events by their true properties. For the reconstructed events, the *Highest Momentum MIP candidate* (HMM) is the basis for event identification and differential measurements. It is defined as the track with the highest momentum, which passes all cuts of Phase 1 of the selection (see Section 2.2.2).

One would like to select only charged-current scattering events of muon-neutrinos inside the TPCs, but that is not possible. We would be dealing with an intrinsic background of other events (e.g. electron-neutrino or neutral-current) that produce a negative pion as highest-momentum track. The number of these events would heavily depend on the cross-section models and fluxes involved. To avoid evaluating the selection with model dependent efficiencies and purities, we define the signal as all neutrino interactions inside the TPCs that produce a MIP as highest-momentum negative particle. The overwhelming majority of these events are charged-current muon-neutrino events and this selection

Table 2.1.: Data samples. Signal is defined without any momentum restrictions (see Section 2.2.1).

Sample name	POT	Number of signal events		Production
		track-FV	vertex-FV	
TPC MC Neut	-	57454	27880	6B+TREx
TPC MC Genie	-	63186	30837	6B+TREx
Sand MC Neut	11.13×10^{20}	0	0	6B+TREx
Run 2 Air MC Neut	9.23×10^{20}	927	461	6H
Run 2 Air MC Genie	9.56×10^{20}	836	428	6H
Run 2 Air Data	0.35×10^{20}	-	-	6I
Run 2 Water MC Neut	12.02×10^{20}	1251	630	6H
Run 2 Water MC Genie	12.83×10^{20}	1185	562	6H
Run 2 Water Data	0.43×10^{20}	-	-	6I
Run 3 Air MC Neut	30.79×10^{20}	3166	1520	6H
Run 3 Air MC Genie	32.78×10^{20}	2968	1456	6H
Run 3 Air Data	1.56×10^{20}	-	-	6I
Run 4 Air MC Neut	34.95×10^{20}	3592	1729	6H
Run 4 Air MC Genie	37.27×10^{20}	3207	1583	6H
Run 4 Air Data	1.76×10^{20}	-	-	6I
Run 4 Water MC Neut	34.96×10^{20}	3505	1680	6H
Run 4 Water MC Genie	37.24×10^{20}	3353	1637	6H
Run 4 Water Data	1.63×10^{20}	-	-	6I

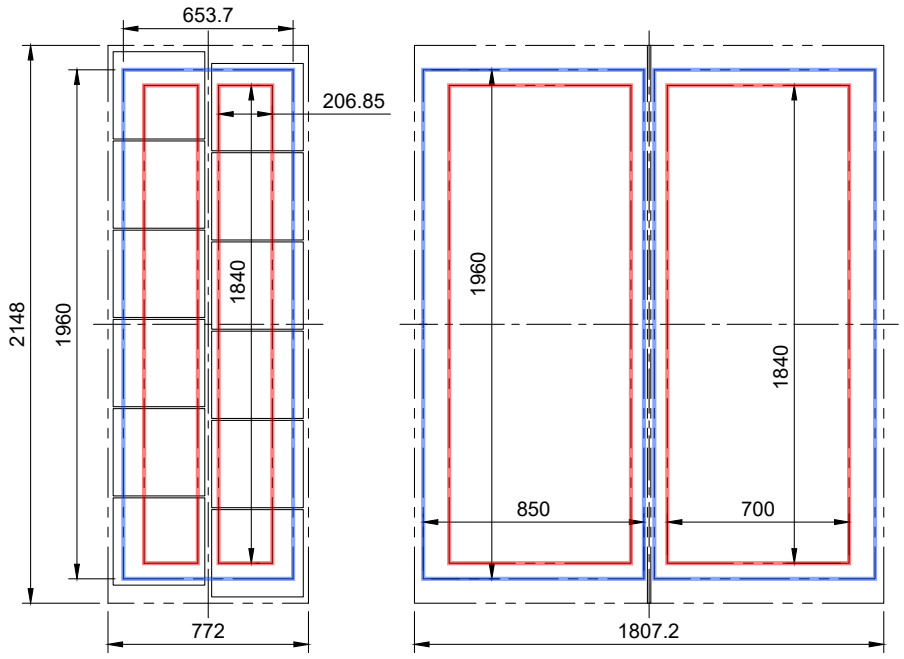


Figure 2.1.: Fiducial volume definitions. The track fiducial volume (blue) and vertex fiducial volume (red) within the TPCs' gas volume, as seen from the readout plane (left) and from the beam direction (right). The squares on the readout plane show the sensitive area of the micromegas modules. All measurements are in millimetres.

Table 2.2.: Signal definitions.

unrestricted, inner	no restriction on primary MIP momentum, vertex in vertex-FV
restricted, inner	primary MIP momentum $> 200\,\mathrm{MeV/c}$, vertex in vertex-FV
restricted, outer	primary MIP momentum $> 200\,\mathrm{MeV/c}$, vertex in track-FV

can then still be used to study them. Events with other particles as highest-momentum negative particle, and events with no negative particle at all are considered background.

There are two kinds of fiducial volumes (FV) defined (see Fig.2.1). The first is the *track fiducial volume*, or track-FV, which determines where the MIP candidates can originate to pass the cut described in Section 2.2.3.2. It consists of one cuboid per TPC-half, located in the instrumented region of the gas volume with some additional spacing to the TPCs' walls, anodes and cathodes. The total track-FV is $6.534\,\mathrm{m}^3$.

The second volume is the *vertex fiducial volume*, or vertex-FV. It is a proper subset of the track-FV and defines where the vertex candidates must be to pass the cut described in Section 2.2.4.3. Aside from an additional spacing to the TPC walls, anodes and cathode, it also excludes a volume over the vertical gap between the two micromegas columns in each TPC-half and some regions with faulty electronics during some of the data runs[1] (not shown in figure). The total vertex-FV is $3.188\,\mathrm{m}^3$.

In general, the track-FV defines the signal region for the neutrino interactions. If an interaction happens outside the track-FV it is considered out-of-fiducial-volume (OOFV) background, even if it happens within the TPC. This is *not* the case for some of the cut optimisation studies, where the fiducial volume is defined as the vertex-FV. This was done to ensure that the efficiencies were not unnecessarily influenced by the change of inner-to-outer volume ratio during the optimisation.

Since the reconstruction efficiency drops significantly for low-momentum particles (even without any momentum cuts), we define a phase-space-restricted signal. It includes only events where the primary MIP has a true momentum of $p > 200\,\mathrm{MeV/c}$. The motivation for this specific value is given in Section 2.2.4.1. Table 2.2 shows an overview of the different signal definitions.

2.2.2. Overview

The event selection works in two phases: First, all negative candidate tracks that start in the track fiducial volume are identified. The highest-momentum candidate then defines the vertex candidate and HMM track. In the second phase, cuts are applied on that single candidate to reduce the background:

- Phase 1 – Selecting the MIP candidate
  1. Event quality cut
  2. TPC multiplicity cut

[1] For simplicity's sake, we exclude these regions for all runs in both MC and data samples. They reduce the vertex-FV by $0.009\,\mathrm{m}^3$ from $3.197\,\mathrm{m}^3$.

3. Track postion cut
4. Track charge cut
5. Track momentum cut

- Phase 2 – Cuts on the MIP candidate
  6. MIP PID cut
  7. Track drift direction cut
  8. Vertex position cut
  9. T0 quality cut
  10. Likelihood match cut
  11. TPC cleanliness cut

The result is a selection of events in the fiducial volume, where the negative track with the highest momentum is a muon or pion.

2.2.3. Selecting the MIP candidate

2.2.3.1. Event quality and TPC activity

The first basic cut demands that the event passes the default ND280 event quality criteria. Then the event is checked for it containing any TPC tracks at all.

2.2.3.2. Track position and orientation

For most tracks that start in the TPCs, no reliable time-of-flight information is available[2]. The reconstructed orientation of the tracks, i.e. forward- or backward-going, is thus not reliable. That is why all tracks that start or end in the track-FV are considered as possible MIP candidates. All tracks that end in the track-FV are flipped (direction and charge), so they start in the TPC by definition and the selection algorithms do not have to consider the separate cases of ending or starting tracks. The tracks must not start *and* stop in the track-FV, as a MIP that is energetic enough to be identified (see below) would not be stopped by the TPC gas.

2.2.4. Cuts on the MIP candidate

2.2.4.1. MIP identification

Among the remaining tracks, the highest momentum one is chosen as *the* HMM candidate and defines the vertex candidate. The HMM is then checked for compatibility with the MIP hypothesis. The Particle Identification (PID) is based on the specific energy loss of charged particles in matter. When a particle traverses a medium it loses energy. That

[2] A valid time-of-flight measurements requires the track to cross at least two distant *scintillation* detectors.

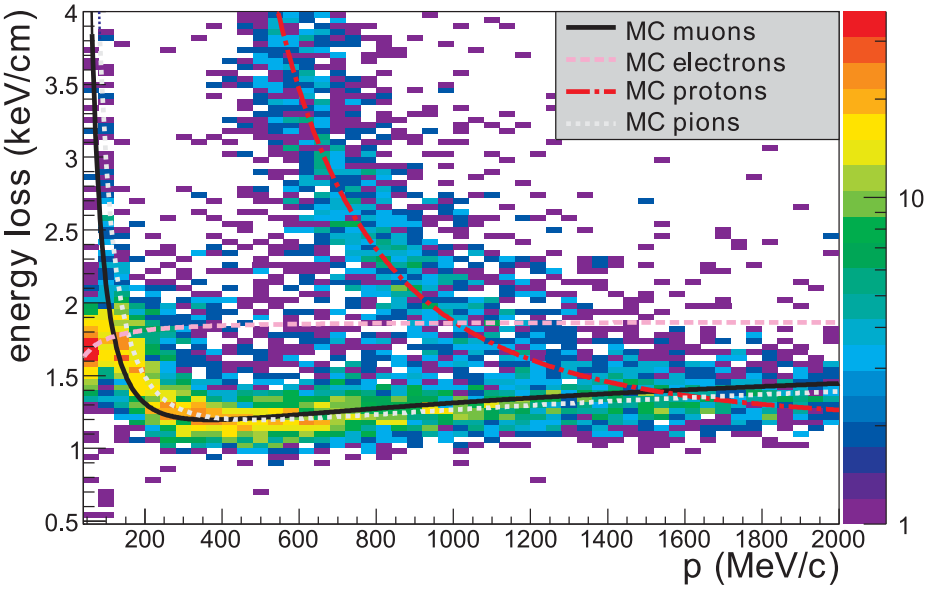


Figure 2.2.: Particle Identification. The expected and measured specific energy losses of different particles from neutrino interactions in the solid detectors [27].

energy loss depends on the particle's charge, mass and momentum (see Fig. 2.2). The TPCs can measure both the energy loss (i.e. the ionisation of the gas along the track) and the momentum (i.e. the curvature of the track in th magnetic field). The measured energy loss is compared to the expectations of four particle hypotheses: electron, muon, pion, and proton. Each hypotheses defines a "pull" P towards that hypothesis:

$$P_{e/\mu/\pi/p} = \frac{\mathrm{d}E/\mathrm{d}x|_{\text{measured}} - \mathrm{d}E/\mathrm{d}x(p)_{e/\mu/\pi/p}}{\sigma(\mathrm{d}E/\mathrm{d}x|_{\text{measured}} - \mathrm{d}E/\mathrm{d}x(p)_{e/\mu/\pi/p})},$$

with the expected uncertainty of the energy loss measurement $\sigma(\cdots)$. These pulls are combined into a likelihood of the particle being of a particular type:

$$L_{e/\mu/\pi/p} = \frac{\exp(P^2_{e/\mu/\pi/p})}{\sum_i \exp(P_i^2)}$$

Particles are assumed to be MIPs when they pass one of the following criteria:

- $(p > 500\,\mathrm{MeV/c}) \wedge (L_\mu > 0.05)$
- $((L_\mu + L_\pi)/(1 - L_p) > 0.8) \wedge (L_\mu > 0.05)$
- Reconstructed track reaches SMRD

The first two points correspond to the standard muon PID criteria also used for the solid detector analyses. As can be seen in Fig. 2.2, muons and pions cannot be distinguished well. The PID should thus be considered a MIP selection rather than a muon selection.

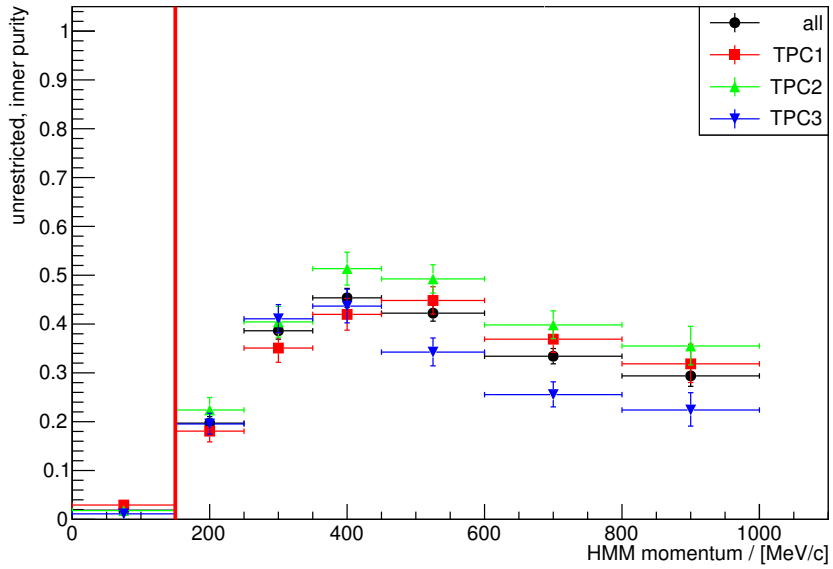


Figure 2.3.: Purity vs momentum. The cut at $p_{\text{reco}} = 150\,\text{MeV/c}$ is motivated by the influence on the efficiency as shown in Fig. 2.4. This plot uses a signal definition without restricted momentum phase space.

A reliable $\mathrm{d}E/\mathrm{d}x$ measurement requires sufficiently long tracks in the TPC. Interactions close to the edge of the TPC are unable to provide this. Fortunately, forward or backward going tracks can still be identified in the neighbouring TPCs. This is not possible for high-angle tracks. These can be recovered if they reach the SMRD. As the name implies (Side Muon Range Detector), it can be used to identify MIPs since most other particles should be stopped in the calorimeters.

Particle identification in the TPCs does not work reliably for particles with a momentum $p < 100\,\text{MeV/c}$. Figure 2.3 shows the achieved purities for different reconstructed momentum bins. The purity seems to increase steadily between 0 and 300 MeV/c. A cut at 300 MeV/c is not desirable though, because it directly cuts into interesting signal regions at low momenta. The effects of different cuts on the selection efficiency can be seen in Fig. 2.4. Even without any momentum cut, the efficiency drops around $p_{\text{reco}} = 150\,\text{MeV/c}$. This is where the energy loss predictions for MIPs and electrons cross. A seizable fraction of MIPs is discarded because they could as well have been an electron. Taking this "electron valley" into account, a cut at $p_{\text{reco}} = 150\,\text{MeV/c}$ seems the most natural choice. It ensures a sufficient efficiency above 200 MeV/c and removes the events left of the electron valley, which would count as background when the phase-space-restricted signal is defined as the region right of the valley.

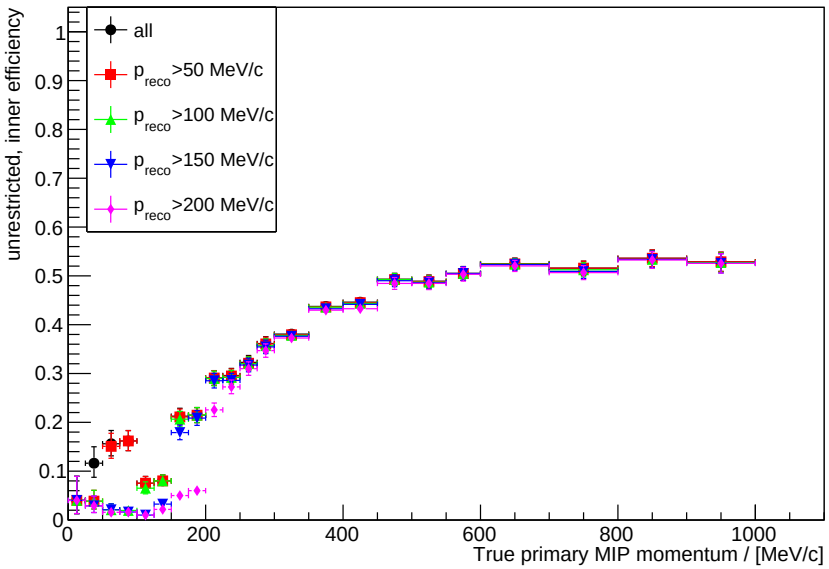


Figure 2.4.: Efficiency vs momentum in the TPC MC Neut sample. The effect of different cuts on the reconstructed momentum is shown. The dip in efficiency around $p_{\text{true}} = 150\,\text{MeV/c}$ appears, because the specific energy losses of MIPs and electrons are almost indistinguishable in that momentum region. This plot uses a signal definition without restricted momentum phase space.

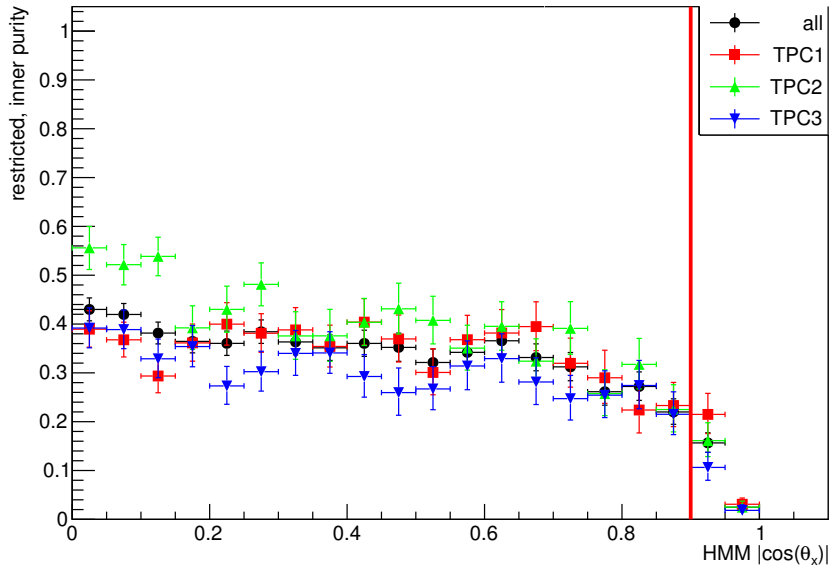


Figure 2.5.: Purity vs cosine of drift angle $\cos(\theta_x)$. Tracks parallel to the TPC drift direction cannot be reconstructed reliably. The contamination with background increases. MIP candidates with $|\cos(\theta_x)| > 0.9$ are discarded.

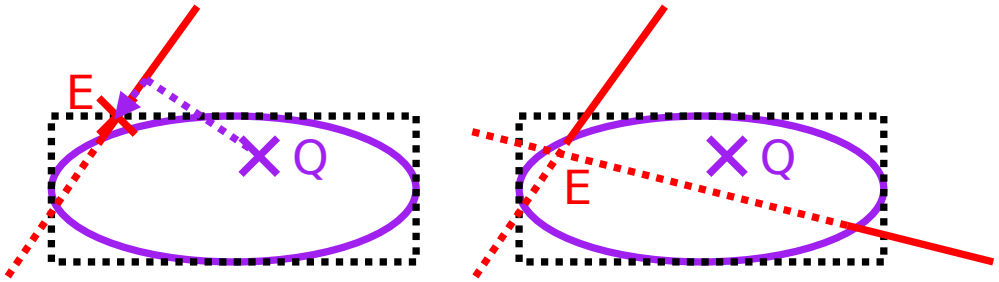
2.2.4.2. Drift direction consideration

The reconstruction in the TPCs does not work completely isotropically. Tracks that are close to parallel to the drift direction have a higher chance of being misreconstructed. To find a good cut value, the drift direction cut was disabled to see the efficiencies and purities for different angles (see Fig. 2.5 and Fig. A.1 in the appendix). In the final selection, MIP candidates with $|\cos(\theta_x)| > 0.9$, where θ_x is the angle between the track's start direction and the TPC drift direction, are removed.

2.2.4.3. Vertex position

Two cases have to be distinguished: In events with only one reconstructed track, the track's starting position is identical to the vertex position. In events with more than one reconstructed track, the position of the main track and the vertex are different. In this case, the vertex is defined by the junction, i.e. the collection of hits in the detector where multiple tracks meet.

The junction is an extended object, but the true vertex position is a single point in space. Without any additional information, the junctions position is reconstructed by calculating its center of charge, i.e. the charge-weighted average position of the constituent hits. This reconstructed position can vary considerably from the true vertex position, especially for asymmetric junctions. A more precise vertex position is determined by combining the information available about the junction and the tracks. If there is only



(a) (b)

Figure 2.6.: Vertex position extrapolation. If there is only one track with $p > 100\,\mathrm{MeV/c}$, the center of charge Q is projected onto that track. If the extrapolated position E falls outside of the junction extent, it is shifted along the track to the limiting side (a). If there are two or more tracks with $p > 100\,\mathrm{MeV/c}$, E is taken from the position of closest approach of the two highest momentum tracks (b).

one track with $p > 100\,\mathrm{MeV/c}$ (and one or more additional tracks below this threshold[3]), the center of charge of the junction is projected onto the linear extrapolation of that track and used as the vertex position (see Fig. 2.6(a)). Should that extrapolation fall outside of the limits of the junction extent in any of the three coordinate axes, the track is only extrapolated to the limiting side. If there are two or more tracks with $p > 100\,\mathrm{MeV/c}$, the vertex position is determined by the position of closest approach between the linear extrapolations of the two highest momentum tracks (see Fig. 2.6(b)).

An event passes the vertex fiducial volume cut, if the vertex position lies within the vertex-FV as described in Section 2.2.1. The vertex-FV is chosen to be smaller than the track-FV, because in general the vertex position and the HMM track starting position are different. Since one first has to find a MIP candidate to determine the vertex and its position, the larger track-FV ensures that no signal events are lost because of the distance between track and vertex.

The vertex-FV was optimised in a study of the efficiencies and purities at different positions in the detector with the actual vertex-FV cut disabled (see purities in Fig. 2.7 through Fig. 2.9, efficiencies in Fig. A.2 through Fig. A.4 in the appendix). For this study, the signal was defined as all MIP events within the *track-FV* rather than the vertex-FV. Cut values were chosen as to minimise edge effects while maximizing the total FV.

2.2.4.4. T0 quality cut

TPCs use timing information to reconstruct a track's position along the drift direction. The orientation of the track is reconstructed from the relative timing of the charge arriving at the read-out area. The absolute position of the track inside the TPC requires

[3]The acceptance threshold of 150 MeV/c only applies to the HMM candidate, not to the other tracks at the vertex.

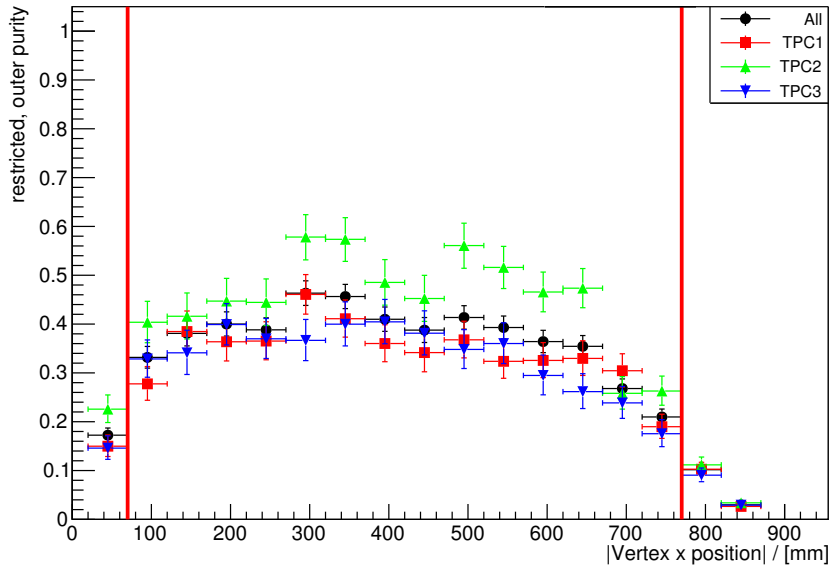


Figure 2.7.: Purity vs vertex x-position. Cut thresholds are shown as vertical red lines. The purity decreases close to the central cathodes ($|x| = 0$) and towards the micromegas readout planes at high $|x|$.

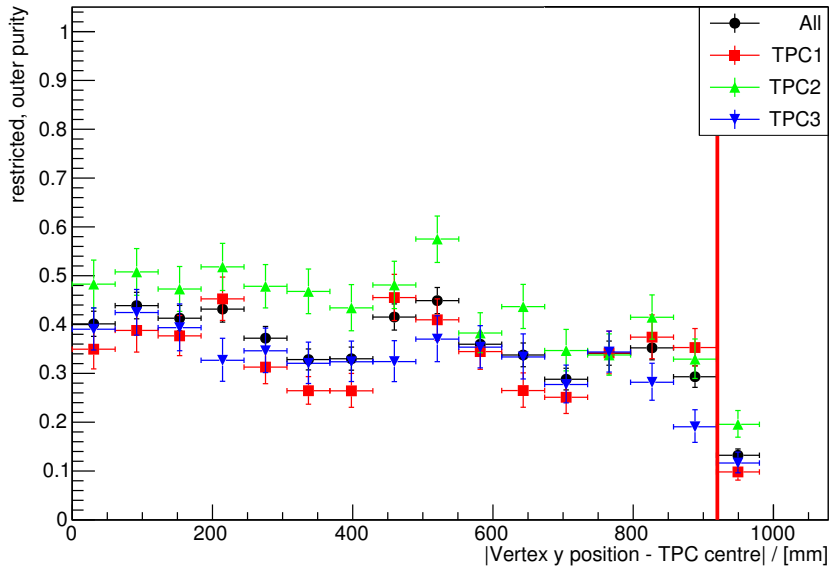


Figure 2.8.: Purity vs vertex y-position. Cut thresholds are shown as vertical red lines. The purity decreases towards the TPC walls at high $|y|$.

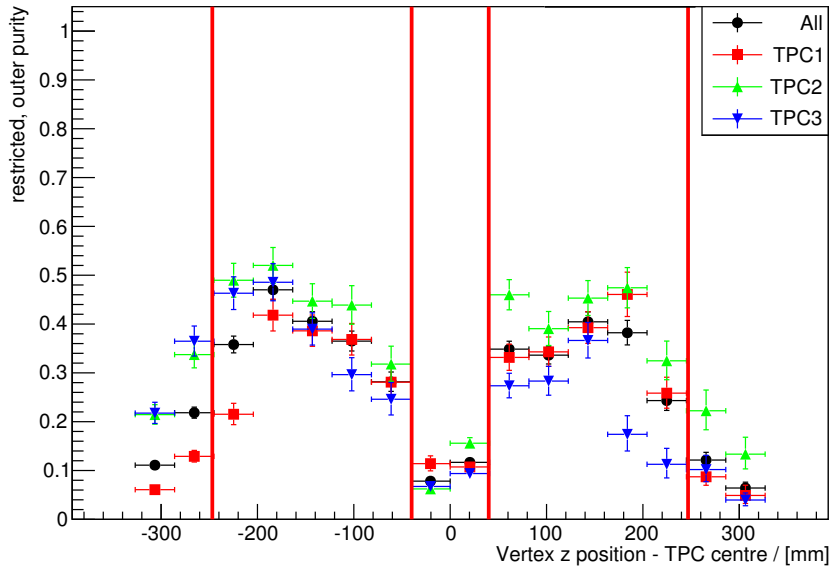


Figure 2.9.: Purity vs vertex z-position. Cut thresholds are shown as vertical red lines. The purity decreases around the vertical gap between the micromegas columns at $z = 0$ and towards the TPC walls at high $|z|$.

an external start time T0, though. This T0 is provided by the scintillating detectors around the TPCs.

A wrong T0 shifts a reconstructed track in the drift direction inside the TPCs. This can lead to tracks erroneously being reconstructed inside the fiducial volume. The T0 is reconstructed by matching the TPC tracks to the hits of the surrounding scintillation detectors (for details see [28]). If the T0 is correct, one can expect a particle track to actually include hits in the detector that provided the T0. If there is no such track connected to the vertex, the T0 is considered to be wrong and the event is discarded.

2.2.4.5. Likelihood match

`TREx` implements the following method to analyse whether two tracks are actually caused by the same particle [28]:

1. Fit two independent helices to the two paths (i.e. hit collections) and calculate the log-likelihood of the hit distribution under the two-track hypothesis, $\ln(L_{11} \cdot L_{22})$.

2. Extrapolate the two helices to the respective other track. This yields two likelihoods of the hit distributions under the one-track hypotheses, helix one extrapolated onto track two, L_{12}, and helix two extrapolated onto track one, L_{21}.

3. Compare the total likelihoods of the two-track hypothesis with the one-track hypotheses. Reject the one-track hypothesis if the log-likelihood difference is above a certain threshold $\Delta \ln(L) = \ln(L_{11} \cdot L_{22}) - \ln(\max(L_{11} \cdot L_{12}, L_{22} \cdot L_{21})) > c$.

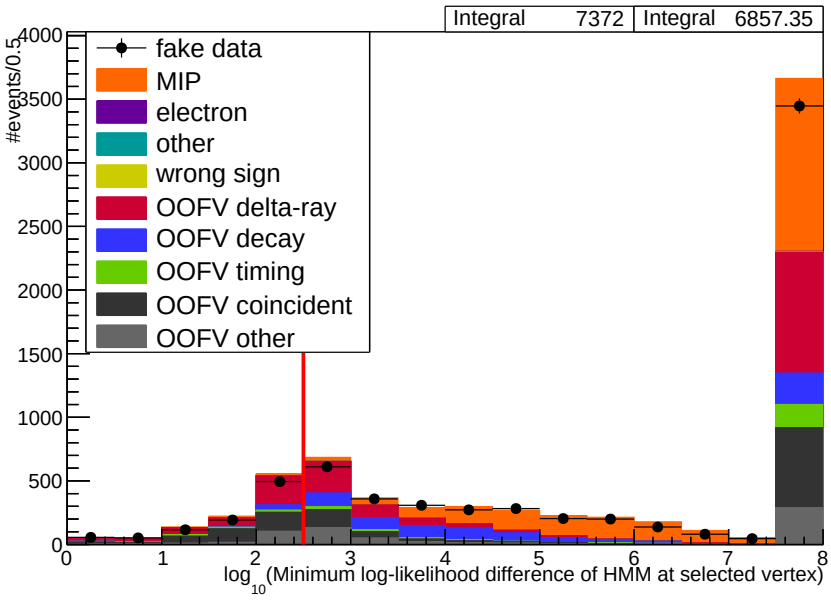


Figure 2.10.: Likelihood match cut. Events with $\log_{10}(\Delta \ln(L)) < 2.5$ are excluded for the gas interaction selection and selected for the delta-ray control sample (see Section 2.4.3). The stacked histogram shows Neut and the fake data is Genie. The Neut data is scaled to match Genie's POT.

At a true gas vertex, no two tracks should match each other except for the case of two particles being emitted back-to-back with identical momenta (in the detector frame) and opposite charges. Since the neutrino interaction events are usually boosted in the detector frame, this is a rare occurrence. For fake vertices caused by delta-rays on the other hand, the two "halves" of the through-going particle should match one another.

Therefore a cut on the minimum log-likelihood difference between the HMM and any of the other tracks at the vertex is applied (see Fig. 2.10). If any tracks are matched with a log-likelihood difference below $10^{2.5} = 316.2$, i.e. it seems like the two tracks belong to the same particle, the event is discarded. The cut value was chosen by disabling the cut and studying the purities for different likelihood match values (see Fig. 2.11). A more aggressive cut could lead to model dependent efficiencies, since the signal distribution of the likelihood differences directly depends on the model and its predicted kinematic distributions (see Fig. 2.12).

2.2.4.6. TPC cleanliness

A lot of out-of-fiducial-volume (OOFV) events are accompanied by other particles in the TPC that show no direct connection to the vertex. Additional activity in the vertex' TPC thus indicates that the event is caused by the "debris" of OOFV events instead of a true gas interaction event. To quantify this activity, the number of passing-by tracks is counted.

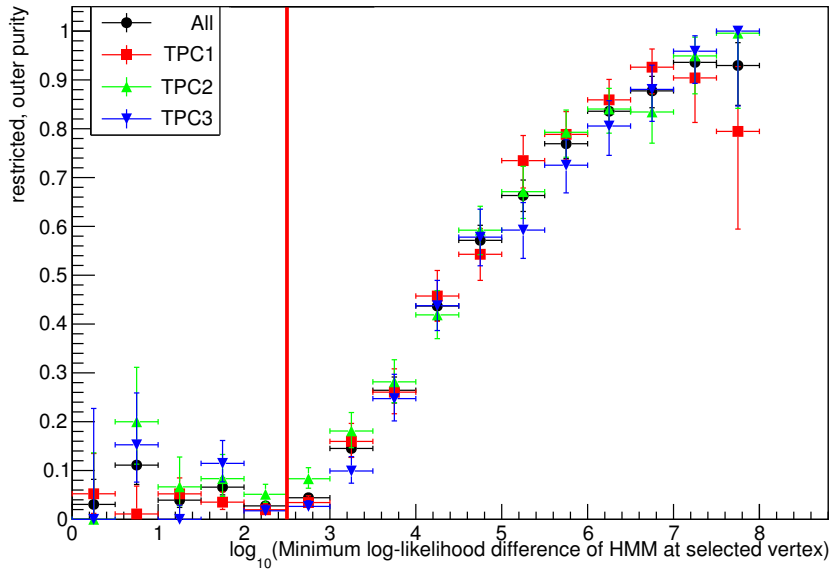


Figure 2.11.: Purity vs likelihood match.

A track is considered as passing-by the vertex if the following conditions are met (see Fig. 2.13):

- The track has hits in the same TPC as the vertex.
- The track's momentum is at least 100 MeV/c.
- One can define planes at the start and end position of the track, each perpendicular to the tracks direction at that point. For a passing-by track, the vertex has to lie in between these two planes, minus a safety margin of 100 mm. This safety margin prevents tracks from messy gas interaction vertices from vetoing their own vertex. The start and end positions do *not* have to lie inside the TPC.

The event is discarded if there is at least one passing-by track present (see Fig. 2.14).

2.3. Performance

2.3.1. Efficiency and purity

The cut performance is shown in Fig. 2.15. Using the restricted, outer (inner) signal definition, we achieve an efficiency of $\sim 17\%$ ($\sim 32\%$) and a purity of $\sim 26\%$. The purities and efficiencies for the different samples are summarised in Tab. 2.3. The efficiencies are slightly lower for Genie than for Neut, as can be seen in the high-statistics, TPC-only samples. The purities also seem to be lower in the Genie samples than in the Neut samples, but the difference is hardly statistically significant. The numbers *do* include

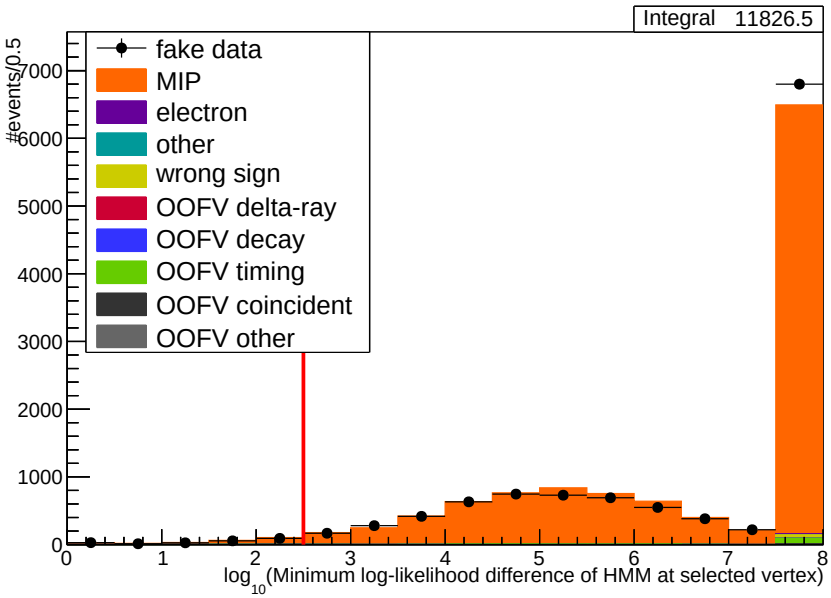


Figure 2.12.: Signal likelihood match. The distribution of signal events depends on the predictions of the generators. Neut (stacked histogram) and Genie (fake data) start to differ for values $\log_{10}(\Delta \ln(L)) > 5$, but it cannot be excluded that other models would show larger differences at lower values as well. The cut level of 2.5 is a conservative choice to avoid possible model dependencies. The Neut data is scaled to match Genie's total number of events, since there is no POT information in the signal samples.

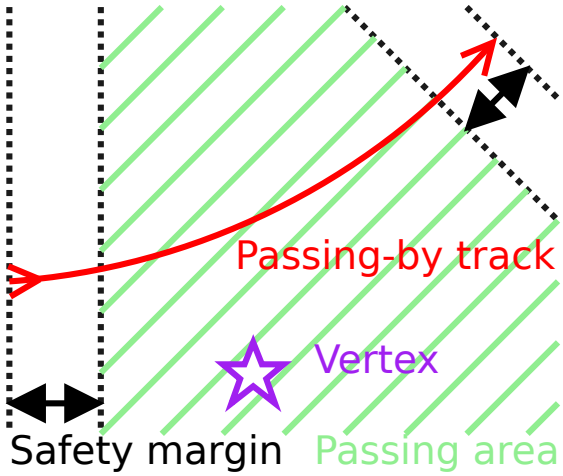


Figure 2.13.: Sketch of a passing-by track.

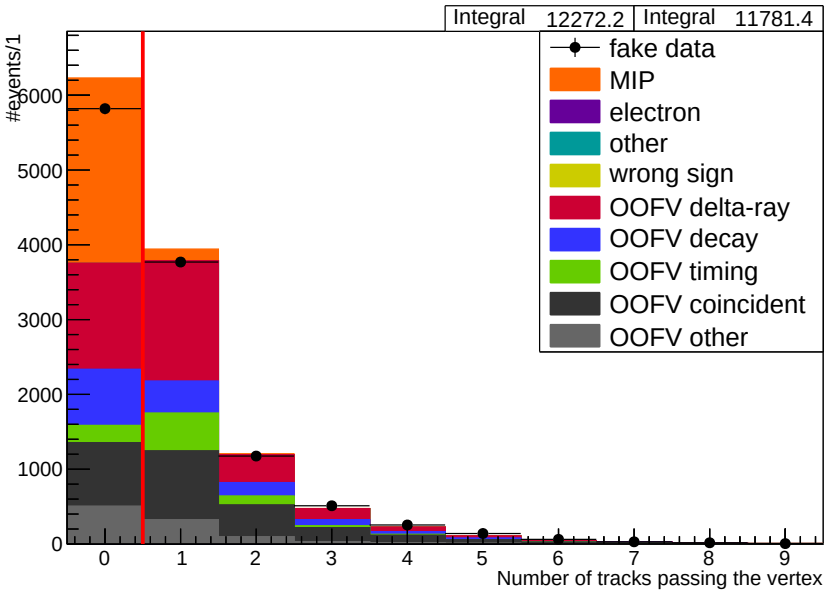


Figure 2.14.: Cut on number of passing-by tracks in the vertex-TPC. Events with at least one passing-by track in the same TPC as the vertex are discarded for the gas interaction selection. They are used for the TPC cleanliness control sample (see Section 2.4.1). The stacked histogram shows Neut and the fake data is Genie. The Neut data is scaled to match Genie's POT.

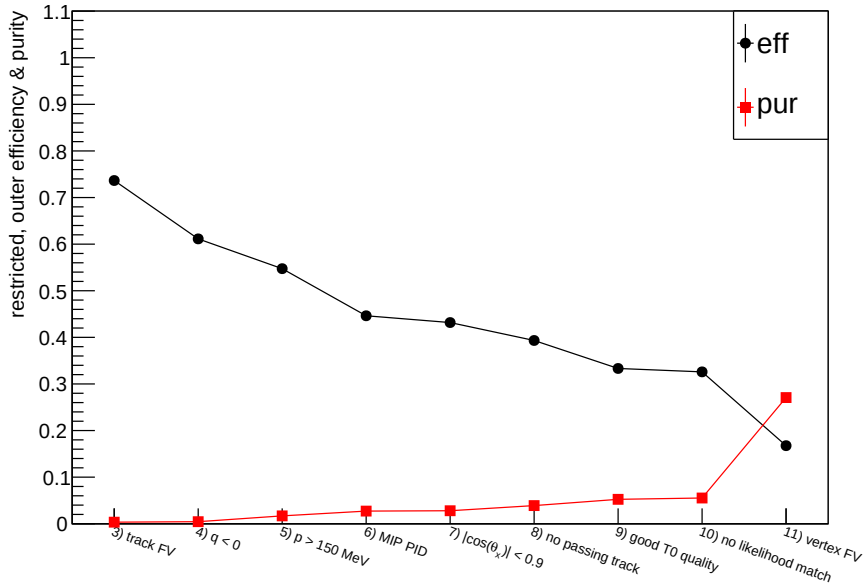


Figure 2.15.: Cut performance – outer FV. Cuts up to and including the momentum-cut (Phase 1) are done on a list of possible candidates tracks. An event is dropped only if no candidate remains after a cut. Later cuts (Phase 2) are applied to the single HMM and vertex candidate defined by the remaining candidate of Phase 1 with the highest momentum (see Section 2.2.2). The considerable drop in efficiency at the last cut is caused by the fact that the vertex-FV makes up only about 50% of the track-FV that defined the outer signal events.

Table 2.3.: Restricted efficiencies and purities for different samples. See Section 2.1 for sample descriptions. The quoted errors are statistical only. The values *do* include the effects of the ad hoc weight (see Section 2.5.1). The weights of the flux, target mass uncertainties/corrections were disabled for this evaluation, as they would affect true and reconstructed events in the same way, thus having no effect on the efficiency. See Section 2.5.2 for a discussion of the systematic uncertainties.

Sample	Efficiency / [%]		Purity / [%]
	inner	outer	outer
TPC MC Neut	32.23 ± 0.28	16.67 ± 0.16	-
TPC MC Genie	31.06 ± 0.27	16.39 ± 0.15	-
Run 2 Air MC Neut	34.7 ± 2.3	18.0 ± 1.3	26.1 ± 1.8
Run 2 Air MC Genie	32.2 ± 2.3	17.9 ± 1.4	29.7 ± 2.1
Run 2 Water MC Neut	32.7 ± 1.9	17.6 ± 1.1	31.7 ± 1.8
Run 2 Water MC Genie	34.7 ± 2.1	17.5 ± 1.1	29.0 ± 1.7
Run 3 Air MC Neut	31.9 ± 1.2	16.39 ± 0.68	28.8 ± 1.1
Run 3 Air MC Genie	31.9 ± 1.2	16.93 ± 0.70	26.9 ± 1.0
Run 4 Air MC Neut	32.4 ± 1.1	16.57 ± 0.63	28.5 ± 1.0
Run 4 Air MC Genie	31.0 ± 1.2	16.72 ± 0.67	25.30 ± 0.96
Run 4 Water MC Neut	32.7 ± 1.2	16.65 ± 0.64	28.1 ± 1.0
Run 4 Water MC Genie	32.6 ± 1.2	17.33 ± 0.67	27.50 ± 0.99
Total POT scaled Neut MC	32.57 ± 0.61	16.76 ± 0.34	27.08 ± 0.51
Total POT scaled Genie MC	32.14 ± 0.63	17.11 ± 0.36	25.61 ± 0.50

the effects of the "ad hoc" weight, which is introduced in Section 2.5.1. See Section 2.5.2 for a discussion of the total efficiency and purity in the final selection including systematic uncertainties.

2.3.2. Background categories

The selected background from inside the fiducial volume is very small. It consist of neutrino interactions that happened inside the track-FV, but which do *not* produce a MIP as highest momentum negative track. They can be further split into events with an electron as primary track, any other negative particle as primary track (very rare), and events with no negative track at all.

The overwhelming majority of background events stems from interactions that happened outside the fiducial volume (OOFV). Five categories of OOFV background have been identified (see Fig. 2.16):

- delta-ray induced
- timing related

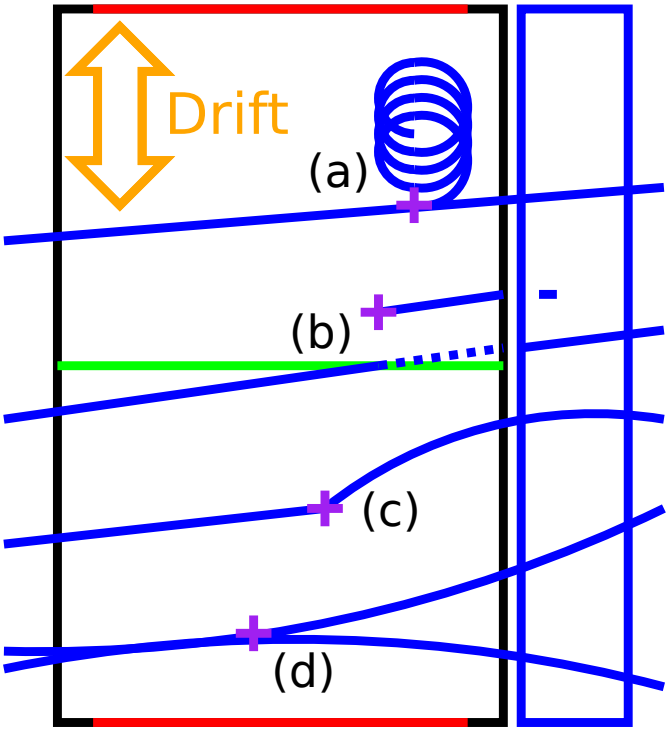


Figure 2.16.: Main OOFV background modes. A delta-ray emitted from a through going particle can be misidentified as vertex (a). If (a part of) a track is matched with a wrong hit for the T0 assignment, it is shifted along the TPC drift direction. This can lead to track ends erroneously being reconstructed in the fiducial volume (b). A particle decaying in the TPC can look like a 2-track vertex (c). Multiple particles entering the TPC coincidentally can confuse the reconstruction and cause fake vertices (d).

- particle decay/stop
- coincident particles
- other

The delta-ray induced background is caused by delta-ray junctions that are misidentified as gas interaction vertices. The timing related background is caused by a failure to determine the correct T0 of a track in the TPC. This causes either the main track to be shifted into the fiducial volume, or a broken track to be shifted away from the vertex candidate, preventing a match and thus rejection. The particle decay/stop background is caused by particles stopping in the TPC (and thus looking like a track starting in the FV) or decaying there, causing a fake vertex to be found.

Coincident particles can also confuse the reconstruction. If another particle enters the TPC within a distance of 5 cm or less from the HMM candidate, it is counted as coincident. All background that cannot be assigned to one of these four categories is counted as "other" background.

Detailed listings of the selection compositions can be found in the appendix: Table A.1 shows the event composition of the main selection and the control samples (see Section 2.4). Table A.2 shows the reaction modes of the events in the fiducial volume. Only between 3.5% and 4% of the selected events in the fiducial volume are not charged-current muon-neutrino events. Table A.3 shows the target nuclei of the different samples and Tab. A.4 the sources of the OOFV background. Please note that the numbers describing the OOFV background do *not* include the effects of the "ad hoc" weight described in Section 2.5.1.

2.3.3. Kinematic distributions

The momentum and angular distributions of the main selection can be seen in Fig. 2.17 and Fig. 2.18. Agreement between the two generators is very good. Figure 2.19 shows the reconstructed track multiplicity at the vertex. The multiplicity smearing matrix can be seen in Fig. 2.20. Even high multiplicities are reliably reconstructed as such.

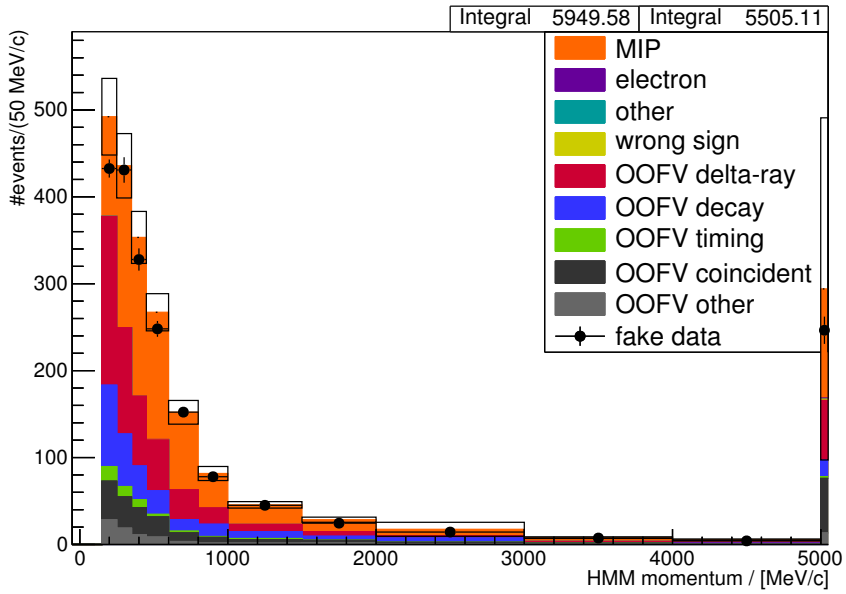


Figure 2.17.: Main selection momentum distribution. The stacked histogram shows Neut with total detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

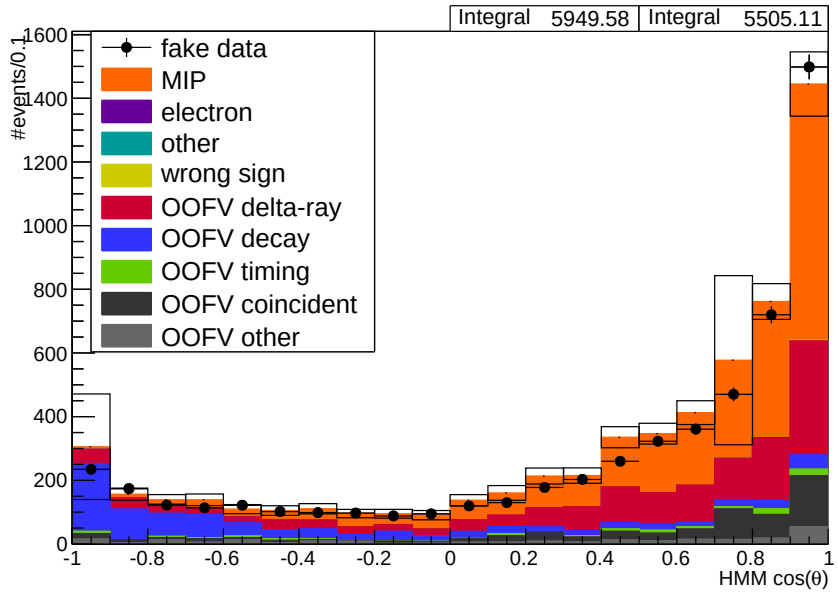


Figure 2.18.: Main selection angular distribution w.r.t the assumed neutrino direction. The stacked histogram shows Neut with total detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

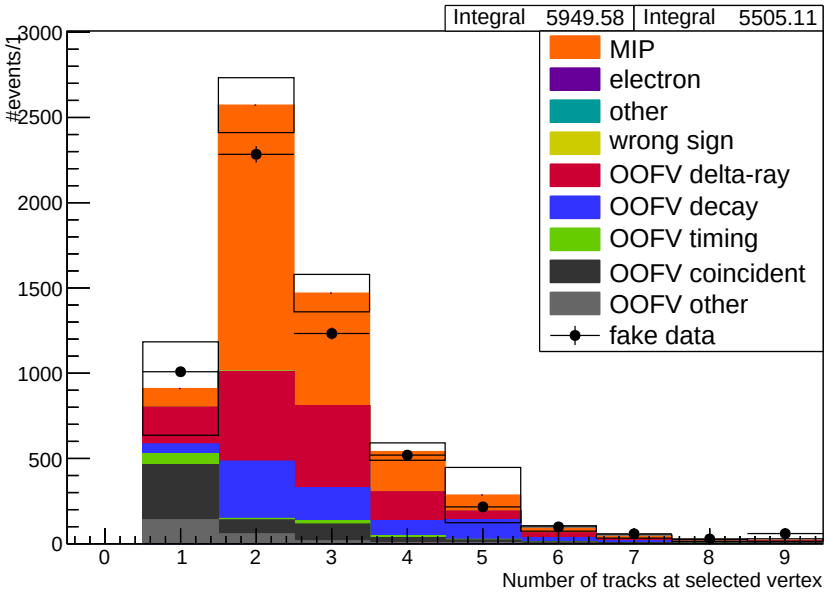


Figure 2.19.: Main selection track multiplicity. The stacked histogram shows Neut with total detector and flux uncertainties and the fake data is Genie with statistical errors. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

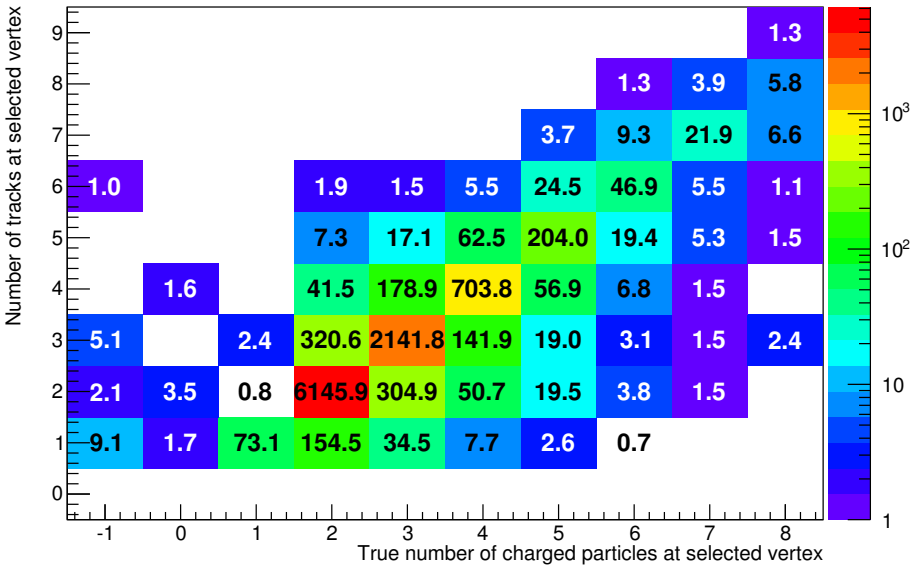


Figure 2.20.: Signal track multiplicity migration in the "TPC MC Neut" sample. A true multiplicity of -1 denotes events outside the track-FV.

2.4. Control samples

The amount of OOFV background in the final selection is very high. Since this is the first analysis of interactions in the TPCs ever, it is not clear how well that background is modelled by the detector simulation. To constrain the background in the main selection with data, three control samples are defined:

- The TPC cleanliness control sample
- The timing control sample
- The delta-ray control sample

Their compositions can be seen in Tab. A.1.

The control samples are disjoint[4] modifications of the main selection. Except for the changes described below, they use exactly the same cuts as listed in Section 2.2.2.

2.4.1. The TPC cleanliness control sample

The TPC cleanliness control sample (or Passing-by CS) is defined by inverting the TPC cleanliness cut (see Section 2.2.4.6), so there must be at least one track in the vertex' TPC passing by the vertex. It contains almost no signal events and its composition is closest to the compostition of the remaining background in the main selection (see Tab. A.1). The momentum distribution can be seen in Fig. 2.21, the $\cos(\theta)$ and multiplicity distributions in Fig. A.5 and Fig. A.6 in the appendix.

2.4.2. The timing control sample

The timing control sample (or Timing CS) is defined by inverting the T0 quality cut (see Section 2.2.4.4 and ignoring the TPC cleanliness cut (see Section 2.2.4.6). It contains almost no signal events and is enriched in out-of-fiducial-volume events that were caused by some sort of timing mismatch between the TPCs and surrounding detectors. The momentum distribution can be seen in Fig. 2.22, the $\cos(\theta)$ and multiplicity distributions in Fig. A.7 and Fig. A.8 in the appendix.

2.4.3. The delta-ray control sample

The delta-ray control sample (or Delta-ray CS) is created by inverting the likelihood-matching cut (see Section 2.2.4.5) and ignoring the T0 quality and TPC cleanliness cuts (see Section 2.2.4.4 and Section 2.2.4.6). It contains almost no signal events and is enriched in out-of-fiducial-volume events that were caused by delta-rays emitted from a through-going particle, or by coincidentally entering particles that look like a single particle entering the TPC. The momentum distribution can be seen in Fig. 2.23, the $\cos(\theta)$ and multiplicity distributions in Fig. A.9 and Fig. A.10 in the appendix.

[4] An event can only pass the selection criteria of either the main selection or one of the control samples. By construction, it is impossible for an event to pass two or more selections.

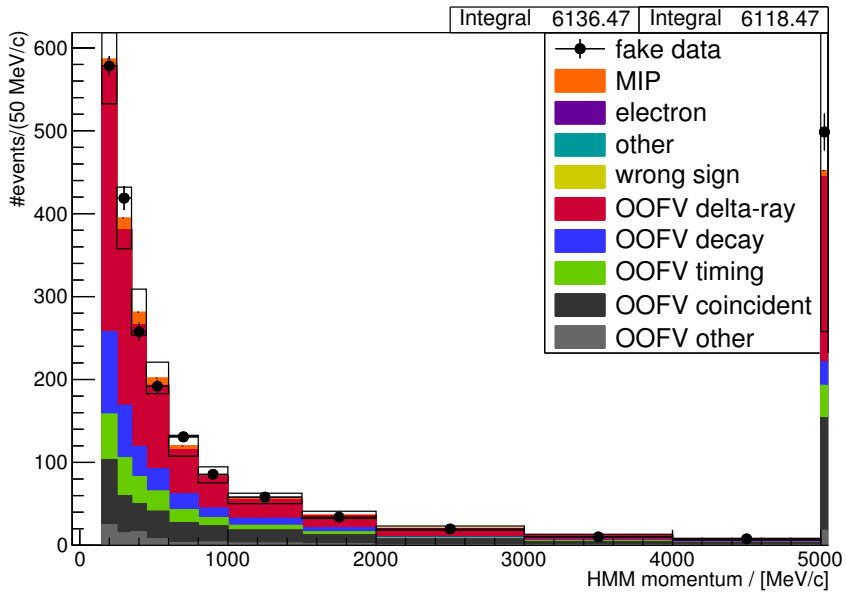


Figure 2.21.: Passing-by CS momentum distribution. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

2.5. Data results

2.5.1. Control samples

The selection has been developed "blind", i.e. using only Monte Carlo simulations to develop and optimise the cuts. Data unblinding, i.e. the process of looking at real data for the first time, is done in two steps: first with the control samples, and then – only after making sure the OOFV background is understood sufficiently well – with the main selection. Comparisons of real data and Monte Carlo momentum distributions in the control samples can be seen in Fig. 2.24 through Fig. 2.26, with angle and multiplicity distributions shown in Fig. A.11 through Fig. A.16. The stacked histogram consists of the Neut samples (including the Sand MC Neut sample), scaled to the data's POT values. The black points are the real data.

It is obvious that there is a huge difference between data and MC. The overall number of data events is larger by a factor of 2. This effect is much stronger for 1-track events than for n-track events, i.e. events with a proper junction as the vertex. A closer investigation of this difference shows that it not only depends on the track multiplicity (1-track vs n-track), but also on the main track's momentum, orientation on the micromegas plane, and x-position (see Tab. A.5 and Tab. A.6).

There are a few possible explanations for this data-MC discrepancy. It is possible that the MC is simply missing an important background process. Atmospheric muons

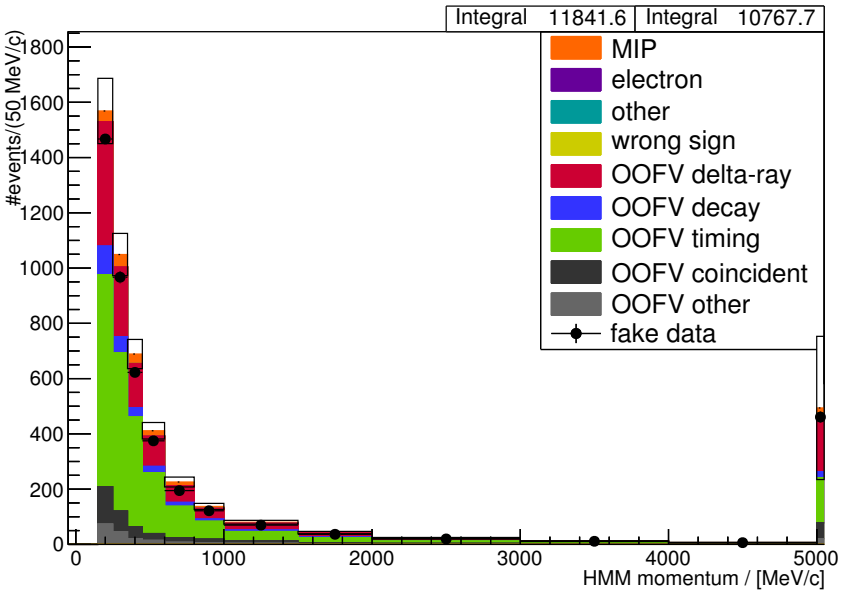


Figure 2.22.: Timing CS sample momentum distribution. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

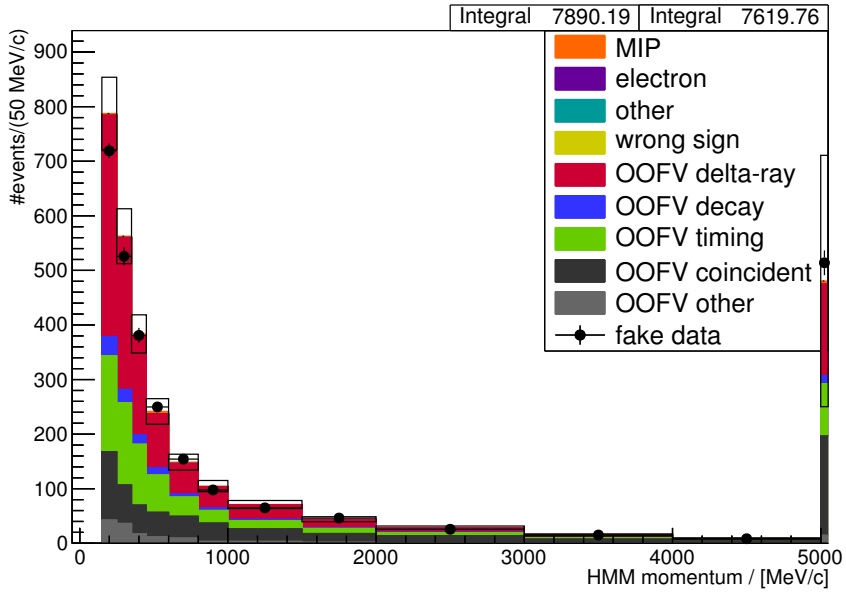


Figure 2.23.: Delta-ray CS momentum distribution. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

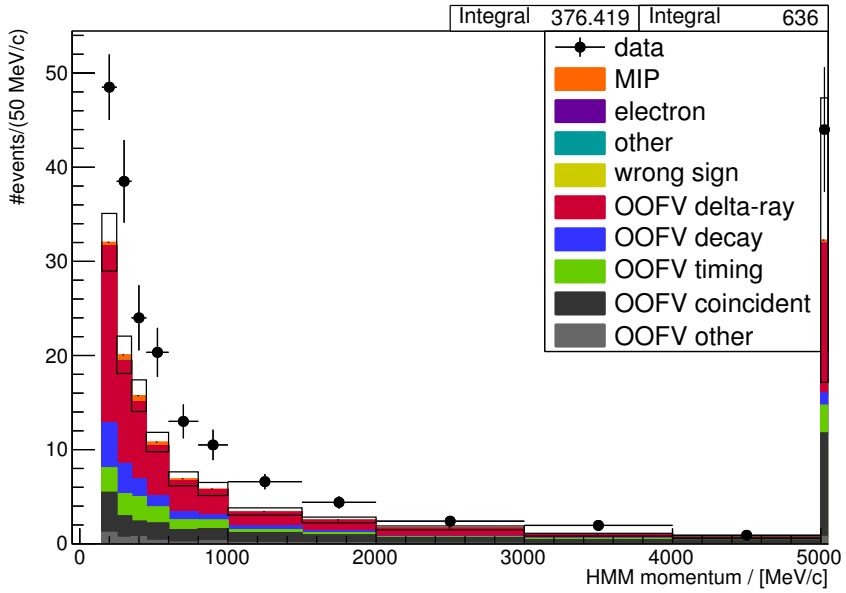


Figure 2.24.: Unblinding Passing-by CS momentum. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

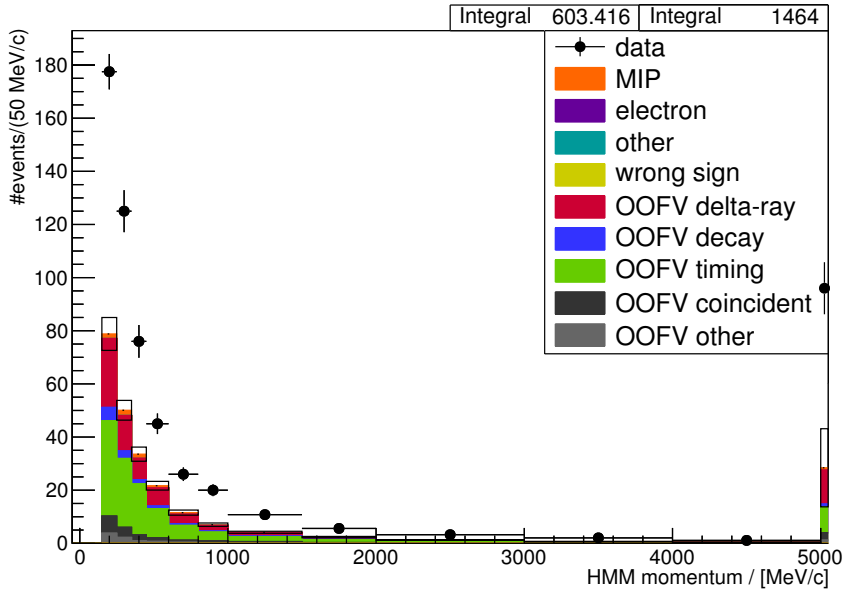


Figure 2.25.: Unblinding Timing CS momentum. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

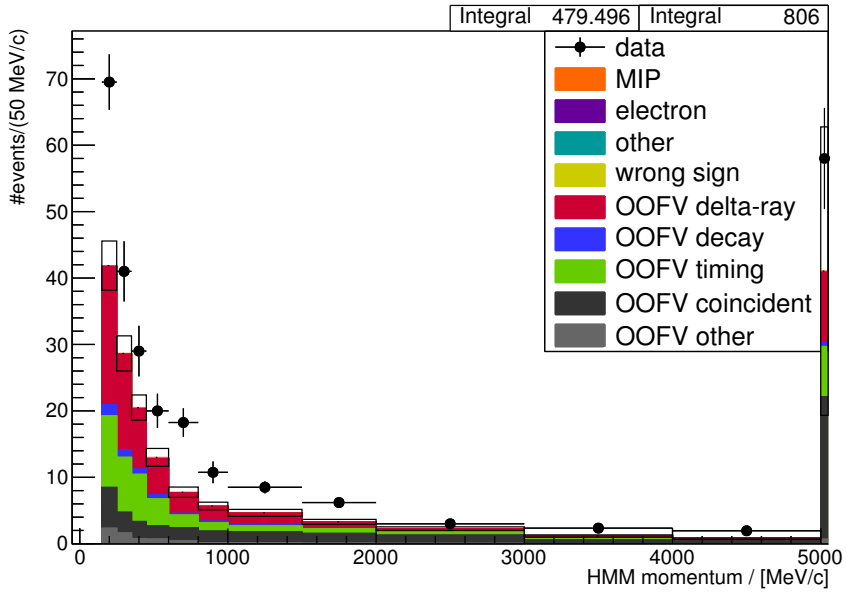


Figure 2.26.: Unblinding Delta-ray CS momentum. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

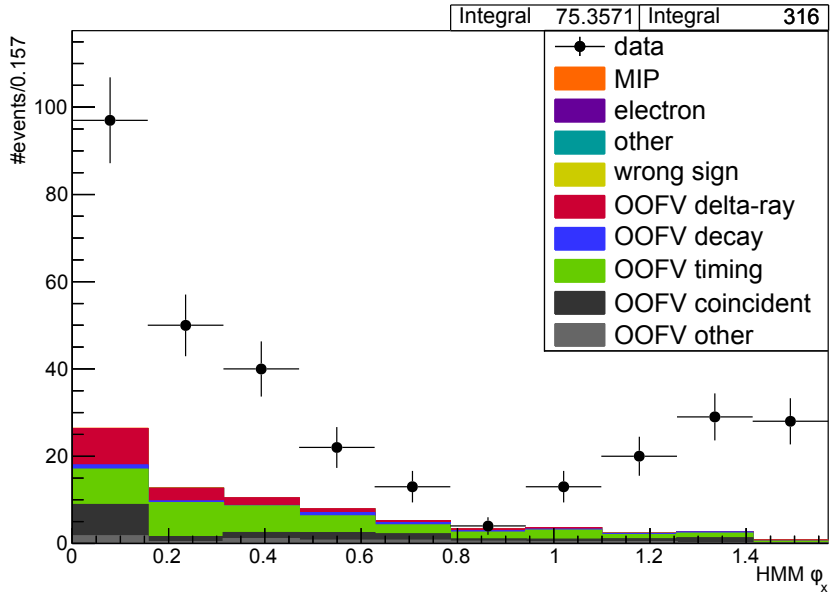


Figure 2.27.: Unblinding angle on micromegas plane. Combined control samples, 1-track events with momentum $> 1000\,\mathrm{MeV/c}$. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

come to mind, as those are completely missing from the MC samples. Estimates of the cosmic background contribution (see Section 2.6.3.2) make it seem unlikely that this is the explanation for this huge difference, though. Also, if the missing BG was primarily cosmic background, its absolute contribution would clearly peak at vertical tracks. This is not the case.

Instead, the observed dependencies point towards an incorrectly modelled track reconstruction failure in the TPCs. Especially high momentum tracks parallel to the pad rows or columns are missing. This becomes more pronounced for high momenta, i.e. very straight tracks (see Fig. 2.27). Together with the dependence on the drift distance in the TPCs (i.e. the x-position) and the fact that 1-track events are more affected than n-track events, it seems like the problem might be related to the widths of tracks on the micromegas plane. We know from previous studies, that the electron drift in the TPCs is not modelled perfectly and multiple corrections have been implemented already for Production 7 of the T2K software stack, including fixes to the diffusion parameters [**Diffusion**], field distortions, and micromegas alignment [**Distortion**].

On first glance, a TPC reconstruction uncertainty in the order of 100% seems unlikely, given that the TPCs have been used extensively in the FGD and P0D analyses and nothing of this magnitude has emerged before. The known uncertainties are actually in the right order of magnitude though. In all P0D and FGD analyses so far, only correctly reconstructed TPC tracks were of interest. In general, the TPC performance

has been very good and reconstruction efficiencies were high (very close to 100%) with low uncertainties in the order of 1% or less.

However, the OOFV background in this analysis is exactly caused by the few instances where things go wrong, as reconstruction failures of through-going particles can look like vertices in the fiducial volume. The selection reduces the number of events from all events that show TPC activity to the ones that end up in the final selections by a factor of ~ 500. This means an (absolute) uncertainty of 1% in the failure rate is more than enough to cover the observed differences:

$$\frac{\sigma(\text{failure rate})}{\text{failure rate}} \approx \frac{1\%}{1/500} = 5 > 2.$$

Similarly, the given implementations of the TPC systematics are focused on modifying correctly reconstructed tracks (e.g. by momentum smearing and efficiency weights) rather than dealing with how the reconstruction failures are affected. The present implementations can thus not cover the differences.

All of this is only conjecture, though, and more in-depth studies are needed to find the real reason of the data-MC discrepancy. Only then can the difference be fixed in the detector simulation. It might even be possible to prevent the reconstruction failure in the first place by improving the reconstruction software. This is not possible within the time limits of this first pass of the analysis. As a stop-gap solution for this problem, we decided to implement an "ad hoc weight" of the simulated OOFV background that makes it compatible with the real data in the control samples. The mean weights are binned according to Tab. A.5 and Tab. A.6, except for events with momenta $< 500\,\text{MeV/c}$. For those events the binning in ϕ_x was removed since the dependence seems to be weak. We assume an uncertainty of 50% of the applied weights, i.e. they are varied according to a normal distribution with expectation value $\mu = N_{\text{data}}/N_{\text{MC}}$ and standard deviation $\sigma = \mu/2$. Bins with no data events are assigned a mean weight of 1.0.

The results of these ad hoc weights can be seen in Fig. 2.28 through Fig. 2.30, and Fig. A.17 through Fig. A.22. Despite the weights not distinguishing between the different control samples, they are able to "fix" the data-MC difference in all three samples. This suggests that the OOFV background in the main selection is improved accordingly. The weight is only applied to OOFV events, not to events with the true vertex in the TPC fiducial volume. If the cause of the discrepancy really is a reconstruction failure in the order of 1%, its influence on the true signal events is already covered in the other detector systematics.

2.5.2. Main selection

Comparisons of real data and Monte Carlo distributions in the Main Selection can be seen in Fig. 2.31 through Fig. 2.33. The stacked histogram consists of the Neut samples, scaled to the data's POT values. Figure A.23 through Fig. A.25 in the appendix show the same data with Genie Monte Carlo as comparison. Since there is no sample of sand muons generated with Genie available, the Sand MC Neut sample is used in both cases (see Section 2.1).

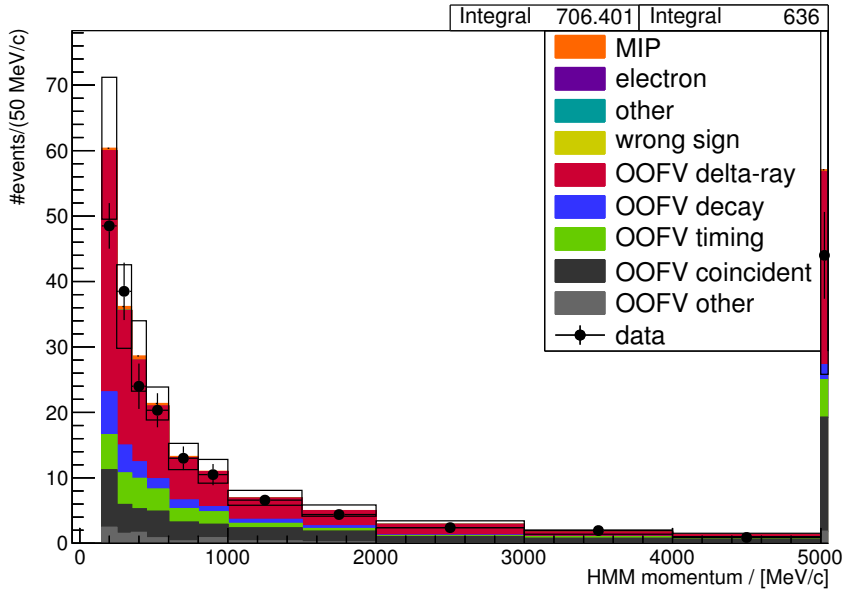


Figure 2.28.: Unblinding Passing-by CS momentum with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

Both Neut and Genie predict more overall events than present in data:

$$\frac{N_{\text{data}} - N_{\text{Neut}}}{N_{\text{Neut}}} = \frac{370 - 445.0}{445.0} = -16.9\%$$
$$\frac{N_{\text{data}} - N_{\text{Genie}}}{N_{\text{Genie}}} = \frac{370 - 421.1}{421.1} = -12.1\%.$$

This is completely compatible within the total systematic uncertainty of 17.7% (see Section 2.6.8) and the statistical uncertainty of $1/\sqrt{N_{\text{data}}} = 5.2\%$.

The restricted, outer MIP purities (see Section 2.2.1) in Neut and Genie are 26.9% and 25.4% respectively. The content of non MIP events in the fiducial volume is negligible, so the remaining events can be assumed to be OOFV. Assuming a worst case of 100% correlation between the ad hoc weights, the relative uncertainty on the number of background events is 50%:

$$N_{\text{OOFV,Neut}} = 325 \pm 163,$$
$$N_{\text{OOFV,Genie}} = 313 \pm 157.$$

All other error sources as well as the difference between the generators are negligible in comparison. An estimate of the purity uncertainty is obtained by varying the number of OOFV events but keeping the number of signal events constant:

$$\frac{N_{\text{MIP}}}{N_{\text{MIP}} + N_{\text{OOFV}}} = (26^{+15}_{-7})\%.$$

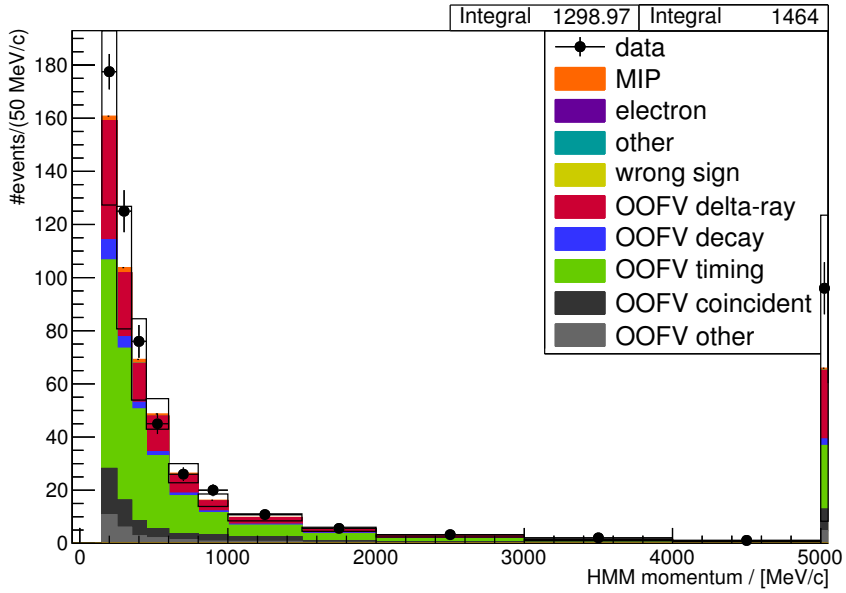


Figure 2.29.: Unblinding Timing CS momentum with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

Varying the number of OOFV events, but keeping the total number of events constant, yields a purity that is compatible with 0.

The expected restricted outer (inner) efficiencies in Neut and Genie are 16.8% (32.6%) and 17.1% (32.1%) respectively. All systematics that affect the reconstruction of signal events are well below the level of 10%[5]. We can thus conservatively apply this error to the efficiency:

$$\epsilon_{\text{outer}} = (17.0 \pm 1.7)\%,$$

$$\epsilon_{\text{inner}} = (32.4 \pm 3.2)\%.$$

The extraction of a CC_{inc} cross section from this selection is shown in Chapter 3.

2.5.3. Outlook

The main advantage of using the TPCs as active target is the low detection threshold compared to solid state detectors. This will be especially useful to investigate the kinematic distributions of "secondary" particles emitted from the vertex, i.e. particles other than the primary MIP. Due to time constraints, the systematic uncertainties of detecting and characterising these particles have not been sufficiently investigated for an inclusion in the analysis. Monte Carlo generator comparisons can show a few interesting venues of future development, though.

[5]Evaluated with the dedicated signal MC samples.

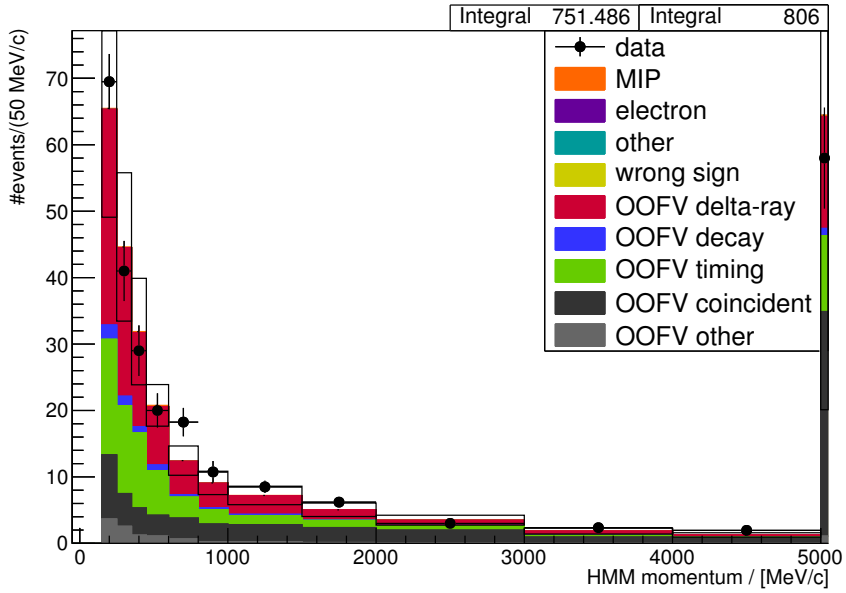


Figure 2.30.: Unblinding Delta-ray CS momentum with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

We look at "pure" proton events, which we define as events that yield exactly one identified MIP (from the main selection) and one or more identified protons. Tracks are identified as protons, if the TPC PID likelihood of being a proton is $> 90\%$. This simple criterion yields a very flat proton detection efficiency in the momentum range of 200 MeV/c up to 800 MeV/c (see Fig. 2.34). Above 800 MeV/c, the expected $\mathrm{d}E/\mathrm{d}x$ of protons crosses that of the other particles. Below 200 MeV/c, the high energy loss might lead to saturation effects influencing the reconstruction and PID.

Figure 2.35 shows a comparison of Neut and Genie predictions for the distribution of the largest proton momentum in these events. Interestingly, they start to show shape differences for momenta just below the sensitivity threshold of previous solid-state-detector analyses of ~ 450 MeV/c. Figure 2.36 shows the distribution for the combined control samples. Here Neut and Genie agree very well with each other. It is noteworthy that the purities of the "pure proton" selections are much higher than in the general selection. These selections thus promise to be quite useful despite their low number of events.

The ArgoNeuT collaboration has detected interesting events with back-to-back proton pairs in the laboratory frame [31]. Figure 2.37 shows the prediction of the angle between the two protons when demanding exactly two protons in the "pure" sample. Figure 2.38 shows the same for the combined control regions. Scaling the Monte Carlo POT to the amount of real data, both Neut and Genie predict only about 0.5 events with back-to-back protons ($\cos(\gamma) < -0.95$), so finding 3 or 4 of such events would already correspond

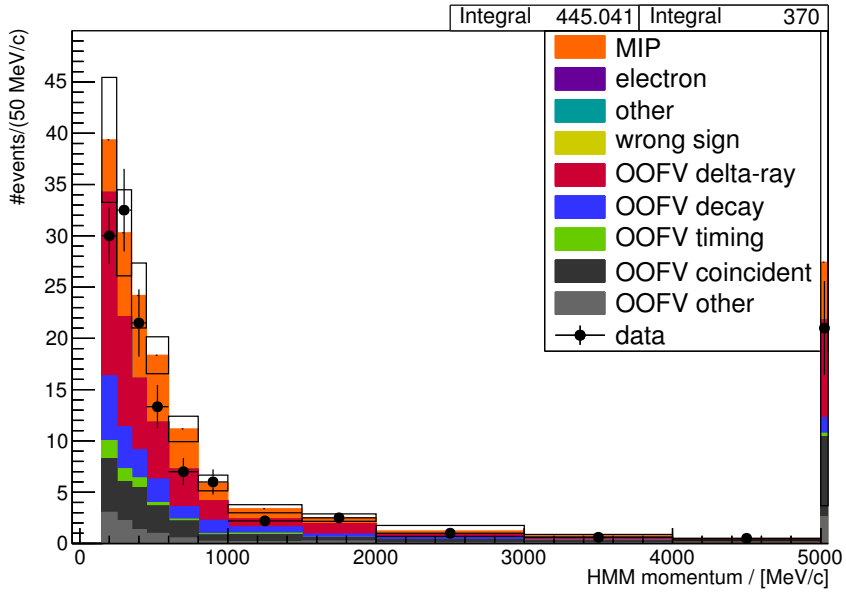


Figure 2.31.: Unblinding Main Selection momentum (vs Neut). The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

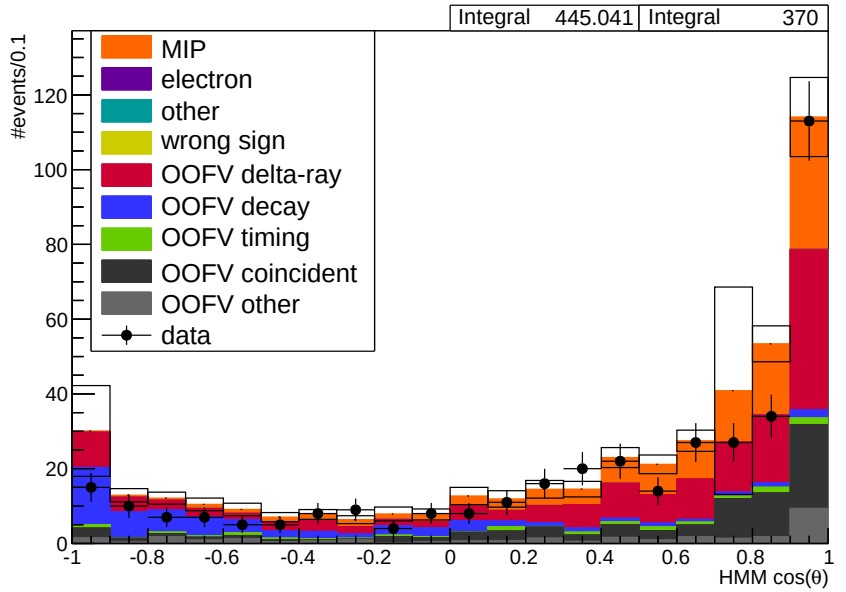


Figure 2.32.: Unblinding Main Selection angular distribution w.r.t. the assumed neutrino direction (vs Neut). The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

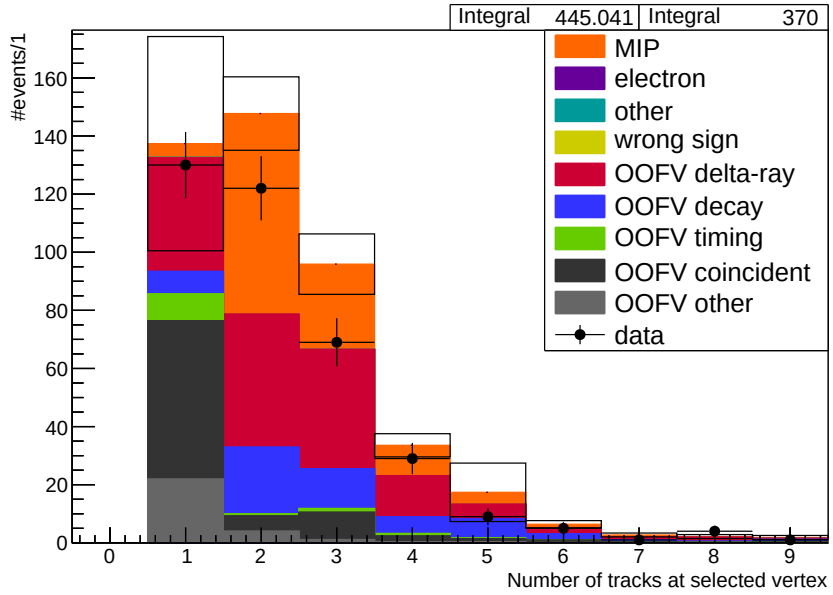


Figure 2.33.: Unblinding Main Selection multiplicity (vs Neut). The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

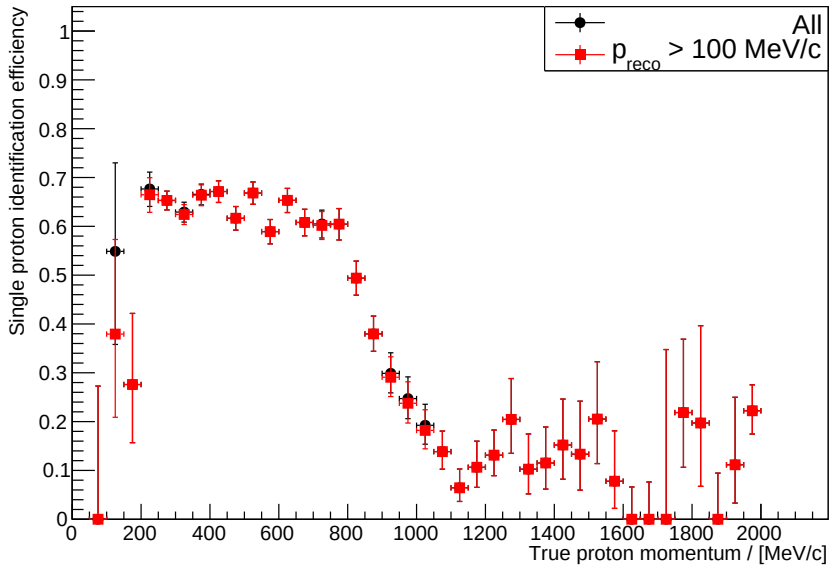


Figure 2.34.: Single proton efficiency in TPC MC Neut sample. The efficiency is calculated as the ratio of reconstructed events with exactly one true proton that have the proton successfully identified as such: N(truth: 1MIP+1p, reco: 1MIP+1p)/N(truth: 1MIP+1p, reco: 1MIP+1x)

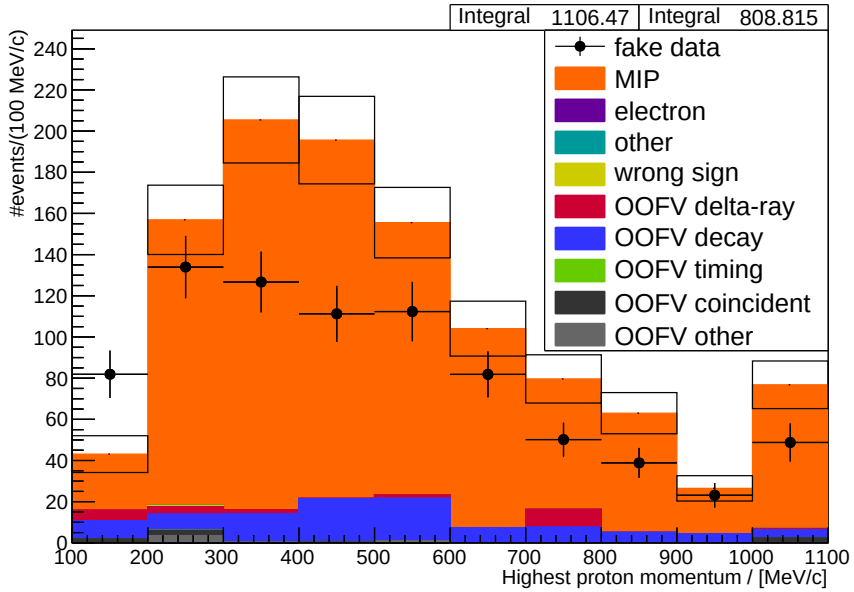


Figure 2.35.: Primary proton momentum. The stacked histogram shows Neut with total detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT.

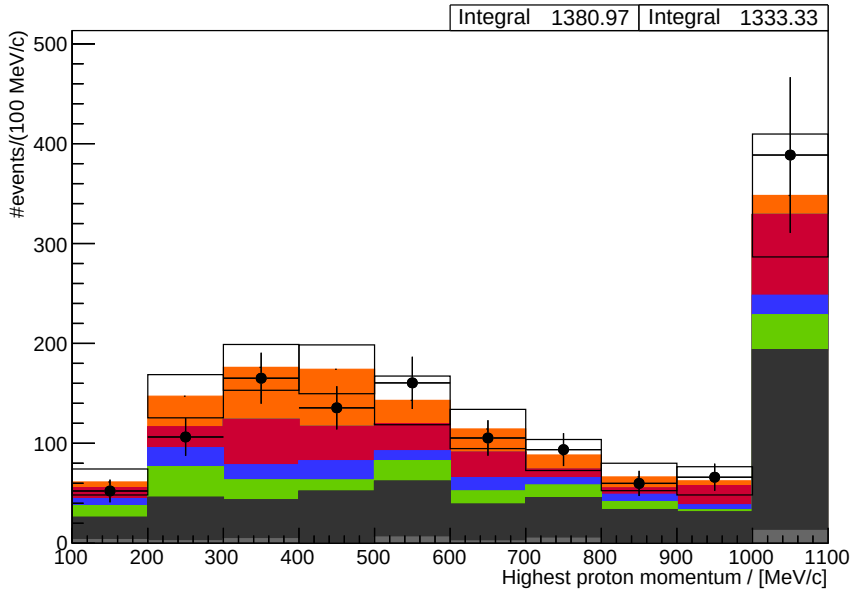


Figure 2.36.: Primary proton momentum in the combined control regions. The stacked histogram shows Neut with total detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT.

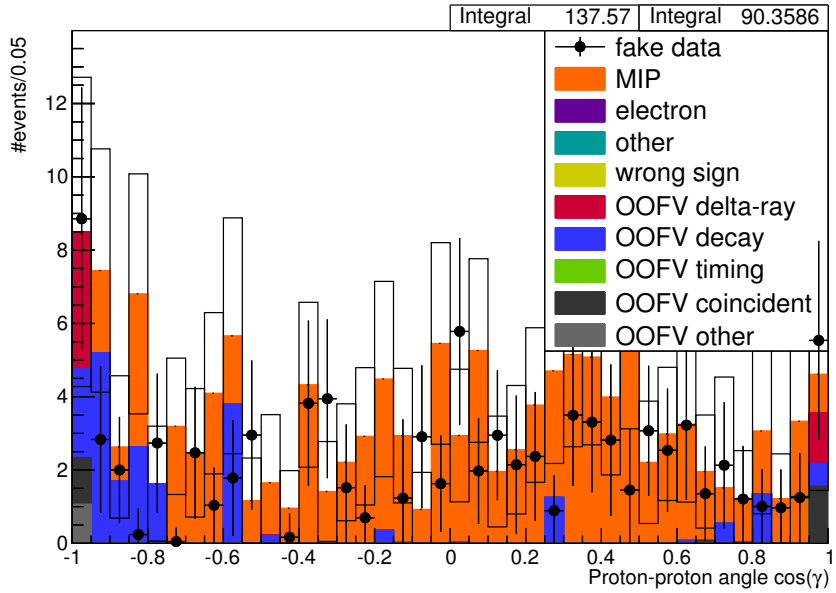


Figure 2.37.: Proton-proton angle. The stacked histogram shows Neut with total detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT.

to a (statistical) significance of about 3σ.

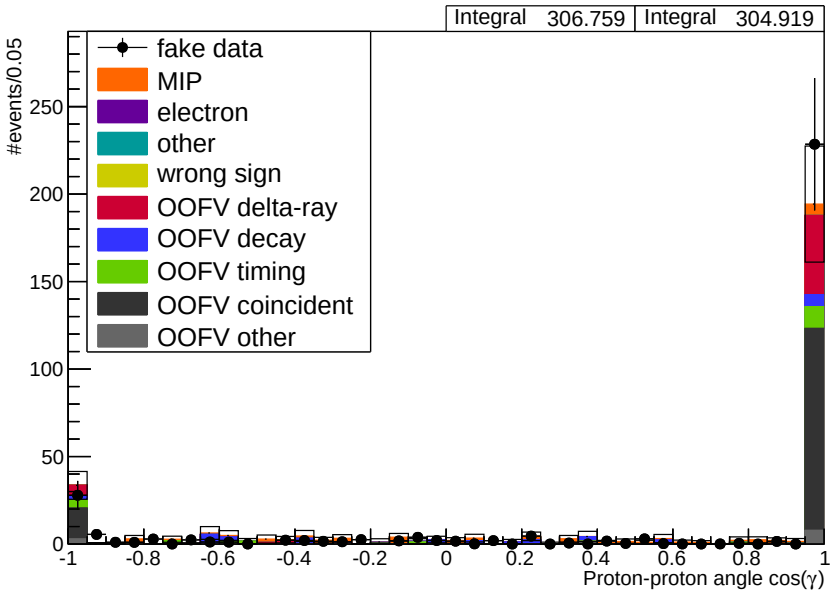


Figure 2.38.: Proton-proton angle in the combined control regions. The stacked histogram shows Neut with total detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT.

Table 2.4.: Weights applied as corrections for relative masses of gases in the TPC gas mixture.

Nucleus	Correction factor
Ar	0.9555
H	1.4450
C	1.5787
F	2.1133

2.6. Systematic uncertainties

Most of the systematic uncertainties that apply to this selection have been evaluated by Edward Larkin in his PhD thesis[28]. Because of their importance, they are repeated here in their entirety. Some systematics have been added and the implementation of others were modified to adapt to the evolving requirements of the analysis.

2.6.1. Corrections

Before systematic uncertainties are propagated, there are several corrections applied to either data or Monte Carlo to account for known and understood differences between real data and simulation. The gas mixture and proton range corrections have been newly developed for this analysis. All other corrections have been implemented for previous ND280 analyses and are used unchanged.

2.6.1.1. Gas mixture correction

During the course of the gas interaction analysis, a bug in the simulation of the gas mixture was discovered causing incorrect relative proportions of argon, isobutane and tetrafluoromethane leading to incorrectly simulated total target masses in the TPC gas volume. The gas mixture in the TPCs consists of Ar, CF_4, and iC_4H_{10} with the *volume* fractions of 95 %, 3 %, and 2 % respectively. The detector model used in Production 6 used *mass* fractions with the same values instead. To correct this, each simulated event with an interaction in the TPC gas volume is given a weight according to the true identity of the target nucleus in simulation. These corrections are shown in Tab. 2.4. There are substantial corrections to the relative masses of hydrogen, carbon and fluorine and a 5% reduction in the simulated mass of argon.

Additionally, the Monte Carlo gas density (and thus the target mass) must be matched to reality. The simulation assumes normal temperature and pressure (NTP) conditions, while the real temperature and pressure in the TPCs are subject to change due to the weather. The mean temperature and pressure in the TPCs was estimated from slow control data of the corresponding sensors (see Tab. 2.5). These numbers are then used to calculate the ratio of mean gas density in MC and real data:

$$\frac{p_{\mathrm{NTP}}}{p_{\mathrm{MIDAS}}}\frac{T_{\mathrm{MIDAS}}}{T_{\mathrm{NTP}}} = 0.994 \pm 0.026.$$

Table 2.5.: Mean temperature and pressure in MC and data. The range describes the yearly variation and the spread of the observed differences between sensors.

	T	$\pm\Delta T_{\text{range}}$	$\pm\Delta T_{\text{spread}}$		p	$\pm\Delta p_{\text{range}}$	$\pm\Delta p_{\text{spread}}$	
MC	(293.15 ±	-	± -	) K	(1013 ±	-	± -	) mbar
Data	(290.65 ±	2.5	± 3.0	) K	(1010 ±	20	± 10	) mbar

The mean density of the gas target is underestimated in the MC by a factor of 0.994. This is corrected by a weight applied to all MC neutrino interactions in the gas volume. The uncertainty of 2.6 % on this weight is propagated as a systematic error (see Section 2.6.3.1). This is a rough estimate of the average density. The correct way to evaluate it would involve calculating the POT-weighted mean temperature and pressure. The spread of the sensors could probably also be reduced by calibrating them correctly[6]. Since the uncertainty of this rough estimate is already well below the flux uncertainty, it was decided that this is not necessary at this point, though.

2.6.1.2. Proton range correction

A bug in the default photo absorption ionization model (PAI) used in the detector simulation causes low-momentum protons to deposit much less energy in the TPCs in Monte Carlo than in reality. A correction is applied to proton ranges at the analysis level to resolve this. The correction reduces the physical length of tracks by moving their end points closer to their starts. This also reduces the number of clusters the tracks are containing and requires correcting for the lower reconstruction efficiency for short tracks compared to long tracks.

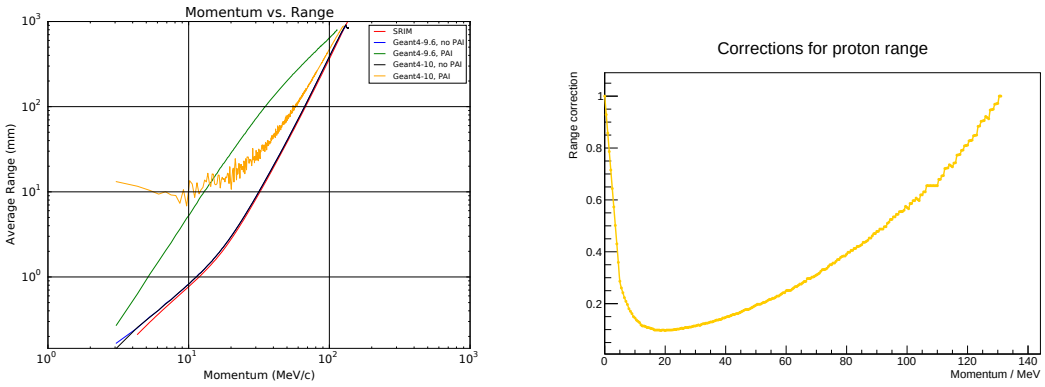
Figure 2.39 shows the differences between simulated and measured ranges for low energy protons in argon. Also presented are the corrections which result from them. This issue will be fixed with the upcoming Production 7 of the T2K software stack.

2.6.1.3. PID corrections

Since the PID determination is mostly unchanged from its pre-`TREx` implementation, the $\frac{dE}{dx}$ corrections used with the previous TPC reconstruction are still necessary. They correct well understood limitations in the hardware, calibration and reconstruction:

- Slight corrections for data (no more than a few percent) are applied to the measured $\frac{dE}{dx}$ in each TPC. These depend on the specific run and subrun range.
- Monte Carlo corrections of around 1% are applied to electrons. This is to account for a known overestimation of electron energy loss in simulation.
- Finally the expected $\frac{dE}{dx}$ values used to calculate pulls are corrected to account for unresolved inaccuracies introduced at the reconstruction stage.

[6] At the moment, the sensors are only used for ensuring that the detector is operating within design parameters and they do not need a precise calibration.



(a) (b)

Figure 2.39.: Proton ranges in different simulations (a) and the corrections which result from them (b). The corrections are derived by taking the ratio between the range in Geant 9.6 with the photo absorption ionisation model (PAI) and in the Stopping Range of Ions in Matter software (SRIM) [32], and enforcing no correction at 0 MeV/c or above 130 MeV/c. [28]

2.6.1.4. Momentum resolution corrections

As with the $\frac{dE}{dx}$ corrections, the momentum resolution corrections used with the previous TPC reconstruction are still used in this analysis. The corrections are quite large, providing an almost 40% smearing in inverse transverse momentum. They accounts for known differences in momentum resolution between data and Monte Carlo.

2.6.2. Propagation of systematic uncertainties

Systematic uncertainties used in the gas interaction analysis fall into three broad categories:

- Those which are unchanged since pre-`TREx` analyses
- Those which are in principle unchanged but require recomputation
- Those which are completely new for the gas interaction analysis

In the first case, some cross checks are performed to verify the validity of the old values of the uncertainties. In the other cases, the values are (re-)computed for the gas interaction analysis. The general implementation of the systematics is outlined here. It covers two classes of systematics: variations and weights (of which efficiency-like uncertainties are a special case). In any case, systematic variations are only applied to the Monte Carlo events. The real data is *not* modified.

Uncertainties based on weights only require the selection to be run once. A set of random weights within the systematic error in question is then calculated. Variation

based uncertainties require multiple passes of the selection with the analysis level variable in question varied many times within its systematic error. The variation in the final number of signal and background events in a given bin after the selection then gives the total systematic uncertainty for that bin. Unless otherwise stated all sources of uncertainty are assumed to be gaussian. Each drawing of detector properties and its subsequent weighting and varying of events is called a "toy". Every toy thus describes one possible true detector.

2.6.2.1. Variations

Variation uncertainties account for the uncertainty in some variable v, for example the track momentum. The selection is repeated multiple times, with this variable modified each time as

$$v_t = \mathrm{f}(v, p_t), \tag{2.1}$$

where v is the unmodified value, v_t the varied value of each toy simulation and p_t a parameter randomly drawn according to the uncertainty[7] about the parameter in question. The detailed properties of the function f depend on the nature of the variable and uncertainty.

2.6.2.2. Weights

Weight uncertainties can be propagated without modifying the underlying event. They simply modify the overall weight w of the event depending on some property of the event v and a randomly drawn parameter p_t,

$$w_t = w \cdot \mathrm{f}(v, p_t). \tag{2.2}$$

Again the details of the weight calculation depend on the specific systematic uncertainty.

2.6.2.3. Efficiencies

Many weight uncertainties, such as the probability of reconstructing a true particle, correspond to a chance for a particular event to be either completely accepted or completely rejected. In this case, the event property v is the information of whether the event was accepted (1) or not (0). The parameter is the assumed probability of accepting an event (the efficiency) in real data ϵ_{data}. The weight function is then:

$$\mathrm{f}(v, \epsilon_{\mathrm{data},t}) = \begin{cases} \epsilon_{\mathrm{data},t}/\epsilon_{\mathrm{MC}} & | \quad v = 1 \\ (1-\epsilon_{\mathrm{data},t})/(1-\epsilon_{\mathrm{MC}}) & | \quad v = 0, \end{cases}$$

where ϵ_{MC} is the efficiency in the Monte Carlo.

The uncertainty – and thus the distribution of $\epsilon_{\mathrm{data},t}$ – is ideally evaluated by the difference between efficiency in data and Monte Carlo in well understood control samples.

[7] The parameters are usually sampled from a normal or uniform distribution.

The tracking efficiency, for example, can be evaluated with a control sample of muons traversing the whole detector. It is selected by demanding that at least two of the three TPCs each contain exactly one track[8]. If those two tracks can be matched and extrapolated to the third TPC, it means that there should be a track there. By counting how often this third track exists, an efficiency can be calculated in both real data and Monte Carlo. For some gas interaction specific systematic uncertainties, control samples are unavailable and we have used comparisons between nominal and modified Monte Carlo instead.

2.6.3. Event level systematic uncertainties

2.6.3.1. Gas Monte Carlo systematics

Since the density of the gas target changes over time, we have to estimate a mean density and correct the simulated target mass (see Section 2.6.1.1). The uncertainty on this estimate is handled as a weight systematic of all events with true vertices inside the TPCs. The total uncertainty on the mean gas density is assumed to be 2.6 %.

2.6.3.2. Sand and cosmic muon background

It is currently not feasible to perform a robust evaluation of cosmic muon background due to practical constraints of time and data availability. To get an idea of the expected cosmic background we look to previous analyses. Those predicted total cosmic ray contamination far below 0.1%, with a total out-of-fiducial-volume background of about 5%. If we assume that the fraction of out-of-fiducial-volume background from cosmic rays is the same order as for other analyses', contamination is below 1% and thus neglected. This decision is also supported by the measured distribution of in-bunch-timing (see Fig. 2.40). Cosmic background events would lead to a uniformly distributed baseline. This is not observed.

We do not have a full set of simulated data for each run period for sand muons. Instead, we have just one sample, corresponding to 11.13×10^{20} POT (about two times higher than real data POT). We find 44.9 sand events passing the main selection (including ad hoc weights), which scales to a predicted 23.0 events in the real data. These selected sand events and their associated error are added as additional background in the analyses. The systematic uncertainty on sand muon flux is 10%.

2.6.3.3. Cosmic and sand muons as coincident veto tracks

Sand muons and coincident cosmic muons can cause a gas interaction event to be discarded, if they produce a passing-by veto track in the same TPC as the vertex. Previous analyses evaluated the probability of sand muon activity in TPC1 (the TPC closest to the graphite target) to be between 0.5 % and 1.4 % per bunch (see Tab. 2.6). The rate

[8] Plus some additional demands, that are not important for understanding the principle of the methodology.

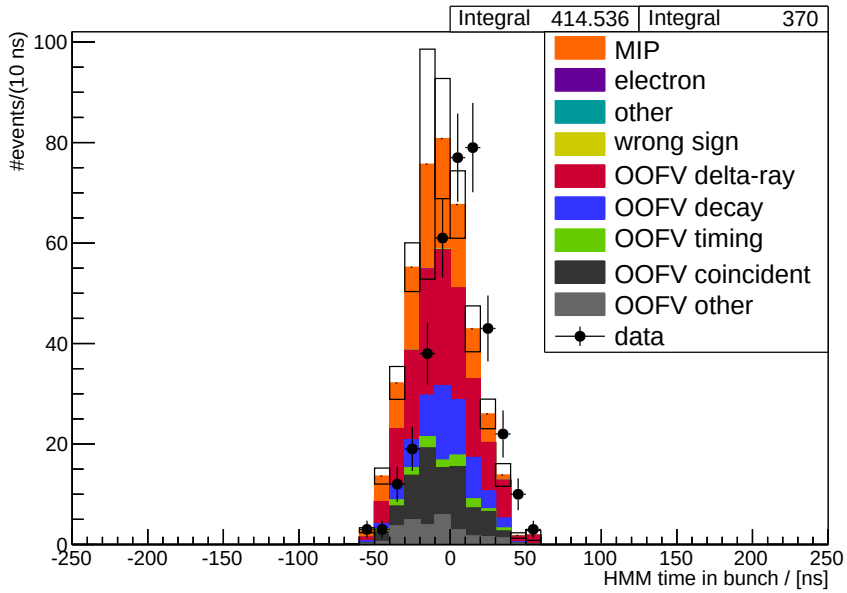


Figure 2.40.: In-bunch timing. Background from cosmic muons would be uniformly distributed within the bunch window ($\sim \pm 200\,\mathrm{ns}$). This is not observed.

in TPC2 and TPC3 will be close to that number[9]. The uncertainty propagation is implemented as a variation. Events are randomly assigned *virtual veto tracks* according to the varied sand muon probability. Events with virtual veto track will fail the veto track cut, even if no actual veto track is present.

The integrated flux of "hard" cosmic muons ($p > 350\,\mathrm{MeV/c}$) at sea level is $J_2 = (190 \pm 12)\,\mathrm{m^{-2}s^{-1}}$ [33]. Assuming that the sensitive area of a single TPC is less than $A = 4\,\mathrm{m^2}$ in any given direction, we can estimate an upper limit for the cosmic ray rate R_C in each TPC:

$$R_C < AJ_2 = (760 \pm 48)\,\mathrm{s^{-1}}.$$

The time window Δt for a beam spill is $< 8\,\mu\mathrm{s}$. The expected number of cosmic muons passing a TPC during a spill is thus:

$$N_C = R_C \Delta t < (6.08 \pm 0.38) \times 10^{-3}.$$

The rate of cosmic muons being associated with a given bunch in a spill will be even lower than that. Therefore, the effect of cosmic muons is neglected in this analysis.

2.6.3.4. Event pile up

The rate of selected events per spill is so low, that event pile up is entirely negligible.

[9] The relative change of muon flux between the TPCs from a possible vertex before the ND280 volume is in the order of $2 \times (1\,\mathrm{m}/100\,\mathrm{m}) = 2\%$.

Table 2.6.: Sand muon probabilities in TPC1. Numbers were taken from previous analyses. The uncertainties were estimated from the number of simulated events and the difference of total (sand + beam) MC and data occupancy.

Run period	Sand muon probability
Run1 Water	0.0051 ± 0.0010
Run2 Water	0.0080 ± 0.0011
Run2 Air	0.0099 ± 0.0014
Run3b Air	0.0096 ± 0.0010
Run3c Air	0.0108 ± 0.0015
Run4 Water	0.0120 ± 0.0010
Run4 Air	0.0138 ± 0.0010

2.6.3.5. Flux related systematic uncertainties

Flux related uncertainties are caused by different sources: from hadron production rates to the alignment of the beams and magnetic fields in the focussing horns. For the gas interaction analysis we use the same flux uncertainties and implementation thereof as in other ND280 analyses (see for example [34]).

As shown in Fig. 2.41, hadron interactions are the dominant contributor to the systematic uncertainty. This is constrained with data from NA61/SHINE [35]. Other sources of error are uncertainties over the precise profile and alignment of the proton beam, the magnetic field and physical alignment of the focusing horns and target. These errors are evaluated independently by varying the relevant quantity in simulation and determining its final effect on flux. All errors are ultimately combined into a covariance matrix binned in neutrino energy and type. Assuming a perfect correlation between the neutrino energy bins, the flux uncertainty is about 10%.

2.6.4. Standard TPC variation uncertainties

There is a large degree of overlap between the variation based uncertainties of the gas interaction analysis and those common to solid detector analyses. It was considered acceptable to reuse existing values for these uncertainties. This is justified by the close matching of kinematic distributions between `TRExFit` and the old reconstruction, and the fact that the underlying procedures for fitting and PID were not changed for `TRExFit`.

2.6.4.1. Momentum scale

Uncertainties in absolute momentum scale ultimately stem from uncertainties in the measurement and calibration of the ND280 B field. We use the same values as pre-`TRExFit` analyses. Altogether, the uncertainty on momentum scale is 0.57%. The scale uncertainty is implemented as a simple variation in absolute momentum for all tracks (partially) reconstructed in the TPC.

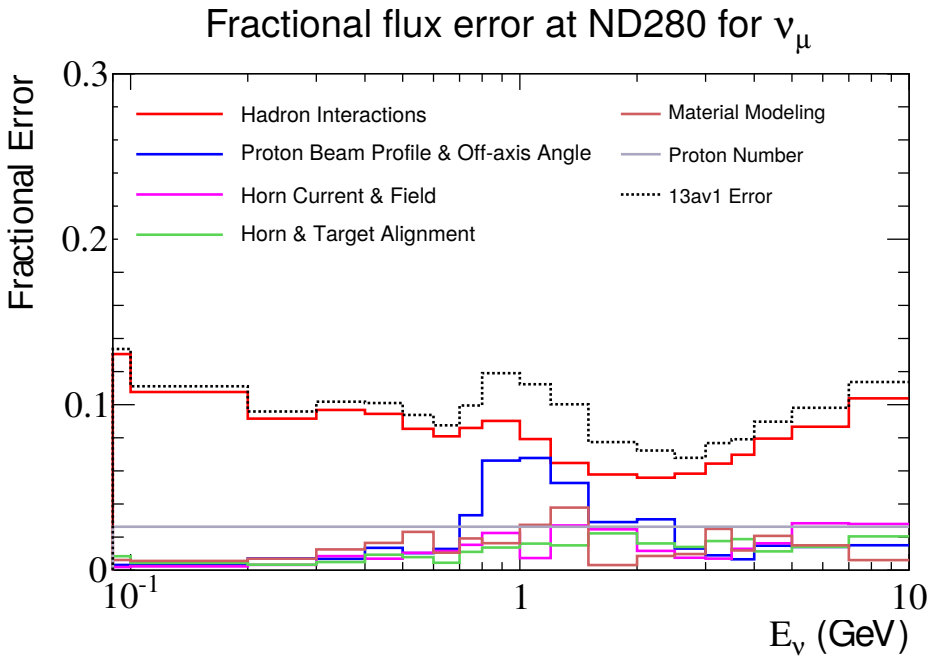


Figure 2.41.: T2K fractional flux uncertainties for ν_μ [36]. In this analysis the T2K flux tuning "13av1" is used. Its total fractional uncertainty is shown by the dashed line.

2.6.4.2. Momentum resolution

The distribution of the reconstructed momentum around the true value is not known precisely. The specific uncertainty depends on the track's x-coordinate, but generally is in the order of 10%.

The momentum resolution uncertainty is implemented as a variation in the difference $1/p_t - 1/p_t{}^{\text{true}}$ between reconstructed and true inverse transverse momentum. For each toy experiment each track's inverse transverse momentum error (i.e. the difference to the true value) is smeared by 10%. The new value is then used to calculate a new momentum.

2.6.4.3. B field distortion

This momentum variation is necessary because of our limited knowledge of distortions in the ND280 magnetic field, which ultimately stems from our ignorance of magnetic yoke properties. Unlike the momentum scale error, which accounts for uncertainty on the mean field, this handles our uncertainty on deviations from this mean field.

The main correction to B field distortions comes from a field map produced by direct measurements in the magnet. Additional corrections are derived from measurements using the TPC laser system, which can insert photo-electron clouds into the drift volume at known positions of the central cathode. The difference between nominal track momentum and that produced by evaluating these corrections ("refit momentum") provides the B field distortion systematics. Like the momentum scale uncertainty, this uncertainty is propagated as a variation in absolute momentum of the tracks.

2.6.4.4. PID

The PID uncertainties stem from the uncertainties in the $\frac{dE}{dx}$ measurements used to evaluate particle identity. We reuse the values from an earlier study, where this is evaluated through the data-Monte Carlo difference in $\frac{dE}{dx}$ between well understood control samples of electrons, protons and minimum ionizing particles (muons and pions) as a function of momentum.

This uncertainty is implemented as a variation in the $\frac{dE}{dx}$ value for each TPC segment of each charged track reconstructed in our TPCs. The exact variation is decided based on the true particle's identity and momentum. Electrons, protons and minimum ionizing particles use separate variations but the method is the same in each case. The reconstructed $\frac{dE}{dx}$ is varied according to this uncertainty.

2.6.5. Standard TPC efficiency uncertainties

There is a large degree of overlap between efficiency uncertainties required for the gas interaction analyses and those required for general ND280 tracker analyses. In most cases the overall systematic uncertainty is small. `TREx` and the previous reconstruction show very similar performances, justifying the reuse of values from pre-`TREx` analyses. Only the track reconstruction efficiency is recomputed, due to its sensitivity to changes in pattern recognition.

Table 2.7.: TPC track finding efficiencies for data and Monte Carlo.

TPC	Data efficiency	Monte Carlo efficiency
TPC1	$(99.6 \pm 1.0)\,\%$	$(98.9 \pm 1.6)\,\%$
TPC2	$(99.4 \pm 0.6)\,\%$	$(99.3 \pm 0.7)\,\%$
TPC3	$(99.1 \pm 2.0)\,\%$	$(99.0 \pm 1.0)\,\%$

Table 2.8.: TPC cluster finding efficiency systematic as a function of angle in the y–z plane. Extra Monte Carlo efficiency corresponds to $(\epsilon_{\mathrm{MC}} - \epsilon_{\mathrm{data}})\,/\epsilon_{\mathrm{MC}}$ where ϵ_{MC} is Monte Carlo efficiency and ϵ_{data} data.

$\cos\theta$	Extra Monte Carlo efficiency
0–0.5735	$(0.11 \pm 0.02)\,\%$
0.5735–1	$(0.07 \pm 0.01)\,\%$

2.6.5.1. Track reconstruction efficiency

This uncertainty covers our ignorance about the probability that a charged particle, which passes through the TPC's instrumented region, is successfully reconstructed as a track. The track efficiency contains both the pattern recognition and the likelihood fitting efficiencies. The methodology for computing and propagating this systematic is unchanged since pre-`TREx` analyses. Specific values have, however, been recomputed for this analysis. The efficiencies are tabulated in Tab. 2.7. The method for computing the systematic uncertainty uses conservative minimum efficiencies for each TPC in data and Monte Carlo. These come from the least efficient bin when efficiencies are binned in both angle ($0 < \cos(\theta) < 0.84$, $0.84 < \cos(\theta) < 0.9$, $0.9 < \cos(\theta) < 0.94$ or $0.94 < \cos(\theta) < 1$) and momentum ($0\,\mathrm{MeV/c} < p < 400\,\mathrm{MeV/c}$, $400\,\mathrm{MeV/c} < p < 500\,\mathrm{MeV/c}$, $500\,\mathrm{MeV/c} < p < 700\,\mathrm{MeV/c}$ or $700\,\mathrm{MeV/c}$ and over). Both data and Monte Carlo show high levels of efficiency which are equivalent within statistical fluctuations.

An efficiency weight is applied for true primary MIPs from gas interactions. In principle, the track efficiency should also affect the presence of veto tracks, but the dedicated sand-veto uncertainty propagation (see Section 2.6.3.3) is already varying the presence of veto tracks more strongly than the track efficiency would.

2.6.5.2. Cluster efficiency

The cluster efficiency accounts for the ignorance of the probability of missing one or more horizontal or vertical clusters in an otherwise reconstructed track. The values used for this systematic uncertainty come from work done prior to the introduction of `TREx`. In this study, the cluster efficiency was compared between data and Monte Carlo for well understood samples of through-going muons in the horizontal and vertical directions.

Results from this study are shown in Tab. 2.8. The uncertainty weights are applied to the main track of the selection. For the gas interaction analysis they are entirely

negligible.

2.6.5.3. Charge ID uncertainty

Charge ID uncertainties account for the chance of a particle being misidentified as its antiparticle due to incorrectly reconstructed charge. This can be caused by both local mis-ID (a track segment being assigned the wrong charge) and global mis-ID (a global track being assigned different charge to its best local segment). It depends on the number of TPC segments in a track and their relative charge sign. Tracks with one, two and three segments require different errors as do tracks where the charge sign disagrees between segments.

The efficiency of finding the right charge is derived from a relatively complex parametrisation in momentum fitting error. Since the momentum fit is in principle unchanged in `TREx`, we reuse the values and propagation methodology used in previous tracker analyses. The charge ID uncertainty is propagated as an efficiency based on the primary MIP candidate track. The precise amount by which the weights are varied depends on the candidate track's momentum fit error, the number of TPC segments and their charges relative to each other and relative to the global track's charge. These systematic uncertainties contribute a final uncertainty in the gas interaction analysis of less than 1%.

2.6.6. Gas interaction specific uncertainties

Many of the uncertainties inherent in a gas interaction analysis have not been computed before, either because they are only relevant to vertices within the TPC, or because they didn't have a significant effect in other analyses. We evaluate these from scratch.

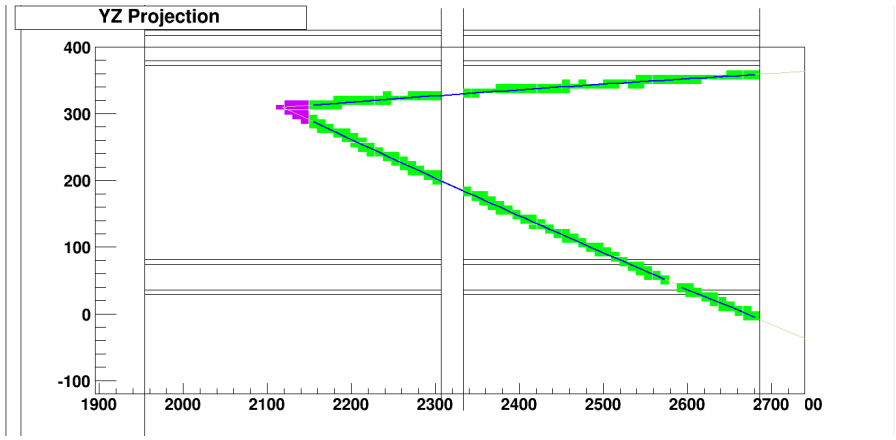
Due to a lack of control samples for gas interactions, many of these systematic uncertainties rely on studies using only Monte Carlo data. Since it cannot be assumed that these precisely describe the behaviour of real data, they are used to compute conservative upper limits on the uncertainties.

2.6.6.1. Vertexing uncertainties

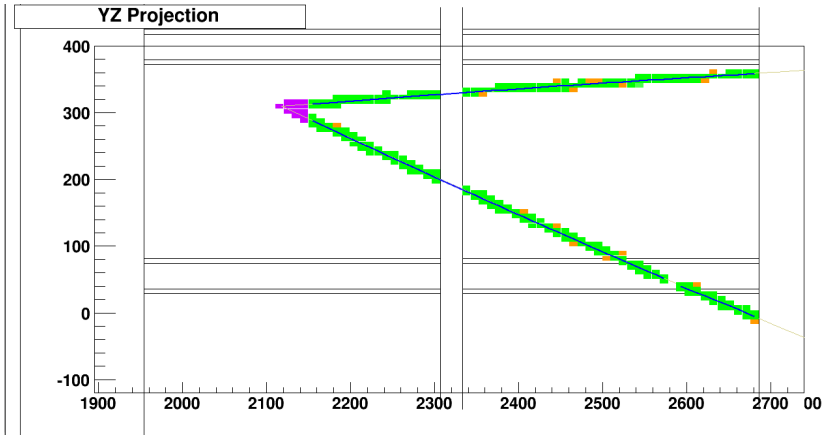
Vertexing uncertainties address our ignorance about the probability of correctly identifying vertices and their properties within the TPC. Since reconstruction differs substantially when it comes to identifying vertices in one-, two- and multi-track events, the parameters describing this uncertainty are binned in charged particle multiplicity. For two-track interactions vertices are identified through kink finding and in the case of multiple tracks they are identified through the `TREx` junction finding algorithm.

Due to lack of a viable control sample in data and Monte Carlo, a Monte Carlo only study was used. The nominal simulation was compared with a noisy sample, in which the variation in simulated pedestal noise was doubled and the zero suppression threshold[10] halved to create extra noise hits. As illustrated in Fig. 2.42, even such a large change in

[10]Only signals above the zero suppresion threshold are saved as hits.



(a)



(b)

Figure 2.42.: Normal (a) and extra-noisy (b) version of the same gas interaction event. Extra noise hits are highlighted in orange. [28]

Table 2.9.: Summary of vertex resolutions: Gaussian fits to difference of reconstructed and true vertex position. These studies use the *charge based* vertex position.

Two-track vertex resolution	x / [mm]	y / [mm]	z / [mm]
Nominal mean	-0.286 ± 0.037	-2.51 ± 0.040	1.445 ± 0.059
Nominal sigma	4.143 ± 0.049	5.650 ± 0.041	6.332 ± 0.051
Noisy mean	-1.320 ± 0.084	-2.839 ± 0.052	3.275 ± 0.079
Noisy sigma	8.14 ± 0.11	6.642 ± 0.058	7.583 ± 0.066
Multi-track vertex resolution	x / [mm]	y / [mm]	z / [mm]
Nominal mean	-1.463 ± 0.080	-3.813 ± 0.066	8.65 ± 0.11
Nominal sigma	8.32 ± 0.10	9.538 ± 0.070	14.14 ± 0.11
Noisy mean	-2.23 ± 0.12	-3.935 ± 0.080	9.99 ± 0.14
Noisy sigma	10.56 ± 0.15	10.332 ± 0.088	15.59 ± 0.13
"Hairy" resolution (see Sec. 2.6.6.3)	x / [mm]	y / [mm]	z / [mm]
Nominal mean	-0.723 ± 0.040	-3.097 ± 0.037	4.716 ± 0.069
Nominal sigma	5.870 ± 0.060	7.417 ± 0.042	10.119 ± 0.066
Hairy mean	-1.361 ± 0.075	-3.887 ± 0.090	7.26 ± 0.15
Hairy sigma	11.12 ± 0.12	16.63 ± 0.11	21.53 ± 0.15

simulation results only in a slight broadening of tracks, but the difference between the two can be taken as an estimate of our uncertainty of real data performance. Three values were checked for the purpose of quantifying their uncertainty: primary track efficiency, vertex resolution, and secondary charged particle efficiency.

The primary track efficiency is the probability of reconstructing the primary track originating within the TPC's instrumented region. Results of the two Monte Carlo samples are shown in Tab. 2.10. The one-track case is ignored since the general track efficiency uncertainty already covers it.

The vertex resolution describes the precision with which a vertex's position can be identified. To evaluate this, we calculated the difference between true and reconstructed *charge based* vertex positions in each of the three dimensions and fitted normal distributions to the histograms. The results are summarised in Tab. 2.9. In the noisy sample, the resolution is up to a factor two worse than in the nominal Monte Carlo sample. This is a conservative estimate for the systematic uncertainty; the extrapolated position is much more accurate than the charge based one. Vertices with exactly one charged particle do not have a junction, so this evaluation cannot be done for them. We assume the real ratio for those events is somewhere between one (i.e. the nominal case) and the (noisy) two-track case.

The secondary charged particle efficiency is the probability of reconstructing a secondary (i.e. not the primary MIP) charged particle emerging from a vertex. This is heavily dependent on the length of each track, which requires a correction for protons at the analysis level (see Section 2.6.1.2). The results are listed in Tab. 2.10. They are dominated by statistical uncertainties in evaluating them, i.e. the differences between

Table 2.10.: Summary of vertex efficiencies.

Systematic		Nominal MC	Noisy MC
Main track efficiency	Two-track	0.8149 ± 0.0025	0.8043 ± 0.0023
	Multi-track	0.8281 ± 0.0027	0.8240 ± 0.0025
Secondary particle efficiency	0 mm – 50 mm tracks	0.080 ± 0.018	0.071 ± 0.015
	50 mm – 100 mm	0.457 ± 0.024	0.430 ± 0.021
	100 mm – 200 mm	0.745 ± 0.019	0.744 ± 0.017
	200 mm – 500 mm	0.8338 ± 0.0077	0.8248 ± 0.0069
	500 mm – 1000 mm	0.9257 ± 0.0054	0.9238 ± 0.0049
	1000 mm or more	0.9622 ± 0.0058	0.9602 ± 0.0052

the normal and noisy sample could still be a statistical fluctuation.

The main track efficiency uncertainty is propagated in the same way as the general TPC track efficiency uncertainty. The only difference is the binning in track multiplicity. The uncertainty in resolution is accounted for by varying the positions of the vertices and track starting positions in our toy experiments with a scaling factor s.

$$x^t_{\text{reco}} = x_{\text{true}} + s \cdot (x_{\text{reco}} - x_{\text{true}})$$

We cannot assume the noisy MC sample to represent the true data behaviour, so we add the nominal to noisy difference to the uncertainty of the evaluation. This means the scaling factor is randomly drawn from a normal distribution centered around the evaluated noisy/nominal ratio r with a standard deviation that includes the data-MC difference.

$$s \sim \text{Norm}(\mu = r, \sigma = (r-1) + \sigma(r))$$

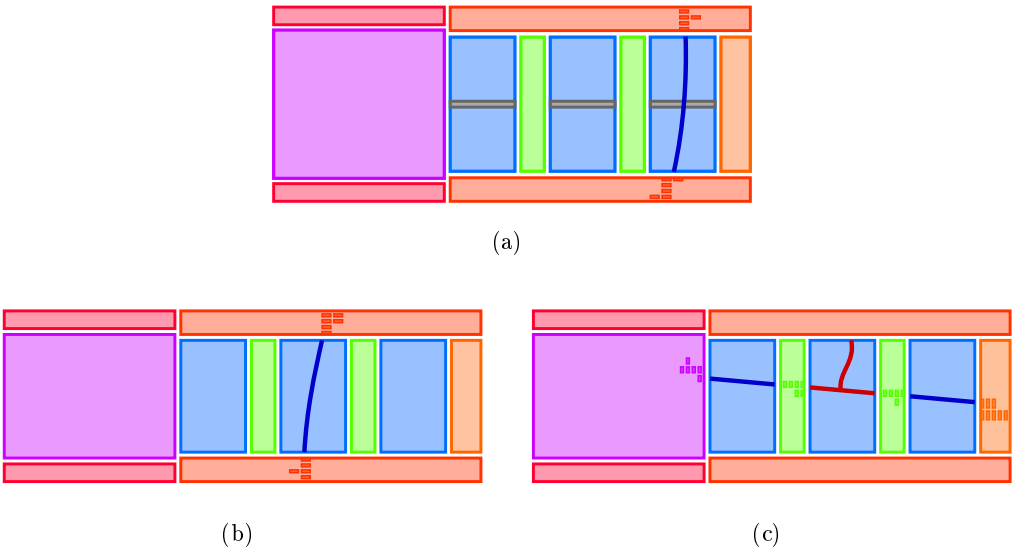
$$r = \sigma^x_{\text{noisy}} / \sigma^x_{\text{nominal}}$$

The secondary particle multiplicity is propagated as a combined efficiency weight, i.e. the product of the single efficiency weights of all secondary tracks.

2.6.6.2. T0 determination

The efficiency uncertainty for the T0 determination represents the uncertainty on the probability of a T0 being found from a source (P0D, FGD or ECal) through which the track passes. The uncertainty is computed by comparing the efficiencies of each source in data and Monte Carlo.

"Clean" events with through-going particles are selected. These are events where a unique single track pattern in a given TPC covers its entire length in x, y or z depending on the detector being checked. Hits are also required in the nearest layers of the two detectors to either side of the TPC. For z tracks these are either the P0D and FGD1, FGD1 and FGD2 or FGD2 and the downstream ECal. For x and y tracks they are the two side ECals and the top and bottom ECals respectively. The categorisation of these candidate events is illustrated in Fig. 2.43.



(a)

(b) (c)

Figure 2.43.: T0 candidates in x (a), y (b) and z (c). The red pattern shown in (c) is not a candidate since it is not a unique single track pattern. [28]

Table 2.11.: Data and Monte Carlo T0 efficiencies for different sources based on samples of through-going tracks.

T0 source	MC efficiency	Data efficiency
P0D	0.920 ± 0.017	0.931 ± 0.020
FGD1	0.936 ± 0.012	0.951 ± 0.015
FGD2	0.968 ± 0.011	0.981 ± 0.015
Downstream ECal	0.967 ± 0.015	0.959 ± 0.021
Top/Bottom ECal	0.776 ± 0.096	0.795 ± 0.143
Side ECals	0.741 ± 0.023	0.917 ± 0.025

The resulting T0 efficiencies are shown in Tab. 2.11. The uncertainties (statistical and data-MC difference) for top, bottom and side ECals are notably higher than those from other sources. This is thought to be a result of the methodology used for evaluating the uncertainty for these detectors, e.g. not accounting for ECAL noise hits that could be mistaken as proper T0 source. A more robust study could produce lower uncertainties but on our current time scale these conservative values are considered acceptable.

The systematic uncertainty is propagated as a variation. According to the assumed efficiency differences in each toy, tracks that do (not) pass the T0 cut are flagged for a forced failure (success) of the cut. Additionally, the reconstructed tracks with a forced T0 success are shifted along the drift direction according to the difference between reconstructed and true T0. The T0 source used for this variation depends on whether the track passes the T0 cut or not. If it does, the detector that provided the reconstructed T0 is used. If the track fails the cut, the detector is chosen from the last detector before the TPC and the first detector after the TPC according to their priorities in the T0-finding algorithm of TREx:

1. FGDs
2. Downstream ECAL
3. POD
4. Side ECALs
5. Top/Bottom ECAL

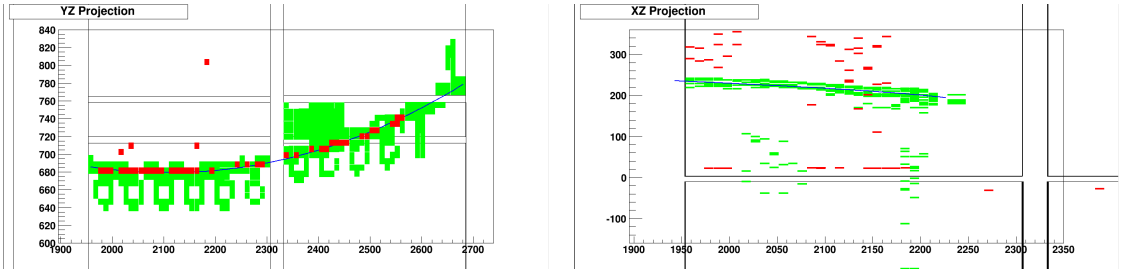
The SMRD is ignored in this case, since a true track coming from there would have to traverse the ECAL before reaching the TPCs. In the case of tracks starting in the TPC, the first detector after the TPC is used.

2.6.6.3. Hairy track systematic uncertainties

One of the major issues identified during the development of `TREx` and this analysis was the existence of "hairy tracks", i.e. low momentum proton tracks with large amounts of noise surrounding them. Although reconstruction procedures were implemented to resolve them, the observed differences between data and Monte Carlo necessitate a large systematic uncertainty.

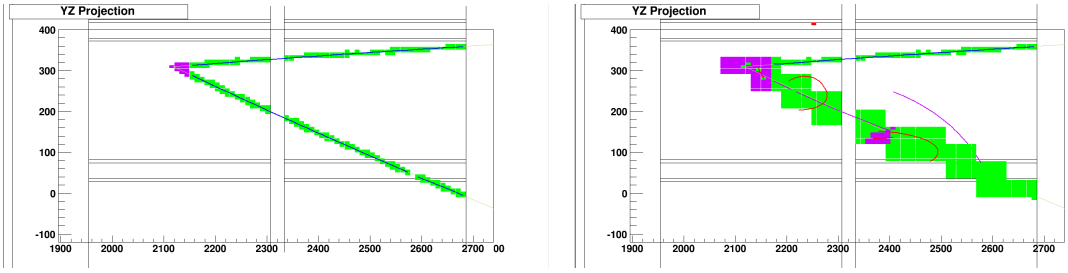
As seen in Fig. 2.44 the `TREx` reconstruction handles hairy topologies fairly well. In most of the worst cases, It can reconstruct a hairy particle's trajectory with good accuracy. Unfortunately, no viable control samples are available for these events. Some proton samples were used while improving and testing the reconstruction of hairy tracks, but these are low in statistics and do not allow us to check tracks starting inside the TPC. Tracks originating in the TPC are expected to be highly sensitive to the presence of hairy tracks. For these reasons a Monte Carlo only study was used.

To compute our uncertainty, we compare the reconstruction efficiency and resolution between a nominal Monte Carlo sample and an artificially hairy sample. This sample



(a) (b)

Figure 2.44.: Two hairy events: a "chainsaw" event (a) and a "low charge shadow" event (b). In each case green and red blocks represent used and unused hits respectively and blue lines represent reconstructed tracks (the single blue line in (a) is partially obscured by unused hits). [28]



(a) (b)

Figure 2.45.: Normal (a) and artificially hairy (b) version of the same gas interaction event. [28]

is created by filling every ASIC electronics unit[11] which the lowest energy proton track passed through with low charge hits. As can be seen in Fig. 2.45, this sample produces far worse reconstruction performance than any hairy event observed in data (compare with Fig. 2.44). Since the goal of the sample is to produce a conservative, "worst case scenario" uncertainty covering any unexpected drop in performance of `TREx` in the presence of hairy gas interactions, this was considered appropriate.

Results for the track efficiency are shown in Tab. 2.12. They demonstrate a surprisingly good reconstruction efficiency for our artificially hairy sample, with only about 7% of primary tracks being lost in the presence of a highly hairy proton track. Since the artificial hairy sample performs worse than the actual hairy events in real data, we characterize the efficiency uncertainty as a normal distribution with the central value halfway between nominal and hairy Monte Carlo. Hairy tracks will also affect the efficiency of secondary track reconstruction, i.e. the multiplicity. We assign the efficiency loss for secondary

[11] The pads of the micromegas readout modules are clustered in rectangles. Each rectangle is read out by a single Application Specific Integrated Circuit (ASIC) module.

Table 2.12.: Hairy track efficiency

	Track reconstruction efficiency
nominal sample	0.8131 ± 0.0017
"hairy" sample	0.7456 ± 0.0020

tracks the same uncertainties as for the primary tracks. It is handled as a separate systematic uncertainty though, so the primary and secondary track efficiencies can be varied independently in the toy simulations.

The hairy track vertex resolution systematic is evaluated analogously to the regular vertex resolution systematic in Section 2.6.6.1. Results are shown in Tab. 2.9. The hairy sample's resolution is about a factor two worse than the nominal sample. Again, this is a conservative estimate for the uncertainty, as the extrapolation based position performs much better than the charge based one. We thus model our uncertainty on the scaling factor s as a uniform distribution between nominal and hairy Monte Carlo.

The efficiency and resolution systematics for hairy tracks were propagated like their analogues for general vertex systematic uncertainties. The difference is that they are only implemented for vertices featuring at least one proton below 500 MeV/c. The additional noise surrounding low-momentum protons is probably caused by the very high specific energy loss of those protons, so this cut-off was established because we know to a reasonable level of certainty that protons above this threshold do not produce hairy events. Further restrictions on events for which hairy systematic uncertainties are considered may be applied in future iterations of the analyses, since we know from experience that a lot of protons below 500 MeV/c are not hairy. The actual threshold seems to be lower than that.

2.6.6.4. Track matching

Particles that cannot be identified in the vertex's TPC, can still be identified as MIPs by TPC-PID in the other TPCs or by reaching the SMRD. This means that the track matching efficiency between the TPCs and between TPCs and SMRD has an influence on the overall reconstruction efficiency. Thus the differences of this matching between data and MC must be evaluated and covered by a systematic weight.

The TPC-SMRD-matching efficiency was evaluated and implemented for an FGD analysis of events including neutral pions. We re-use this implementation. The TPC-TPC-matching uncertainty is extrapolated from the FGD-TPC-matching uncertainties of previous analyses. In those analyses, the systematic uncertainty was only applied to FGD tracks with two or less hits, since data and MC agree very well for longer FGD tracks. For TPC-TPC-matching the FGD-tracks would have to be much longer than that, so one could argue that we can neglect this uncertainty. Unfortunately we cannot be sure that the vertex being in the TPC rather than the FGD does not change this behaviour.

We thus decided to apply the FGD-TPC-matching uncertainty of 2-hit FGD tracks to

Table 2.13.: OOFV rate uncertainties.

OOFV source	rate uncertainty
P0D / FGDs	5.1%
ECALs	11.6%
SMRD	4.9%
other	13.6%

all primary MIPs that cross at least two TPCs, regardless of the number of FGD hits. Since the TPC-TPC-matching is actually a TPC-FGD-TPC-matching, we apply it in quadrature. The TPC-TPC-matching only matters if the track length in the first TPC is not sufficient for the PID. To be more precise, only the track length perpendicular to the drift direction in the TPCs matters, since this determines the possible number of hits on the readout plane for the $\mathrm{d}E/\mathrm{d}x$ measurement. That is why we do not apply the weight to particles with true perpendicular track lengths over 40 cm. This is a conservative threshold, including more tracks than is strictly necessary.

2.6.7. Out-of-fiducial-volume uncertainties

Background in the gas interaction analysis comes almost exclusively from tracks entering the TPC from other detectors. The uncertainty in this background is split into two sources:

- A *rate uncertainty* on the total number of tracks entering from outside the TPC
- A *reconstruction uncertainty* on the probability for a track entering from outside the TPC to be reconstructed as a vertex within the TPC

Rate uncertainties come from uncertainties in the number of neutrino interactions in the detector material surrounding our TPCs. Reconstruction uncertainties are the uncertainties on the probability for a side-entering track to be shifted into the fiducial volume by bad T0 or the probability of a through-going track to be broken.

2.6.7.1. Rate uncertainties

All of the rate uncertainties we use have been evaluated for previous FGD analyses. They reflect our incomplete knowledge of the mass and neutrino-interaction cross sections of the material surrounding the TPCs. A summary is shown in Tab. 2.13. Unlike in the FGD analyses, we cannot assume the FGD-part of the out-of-fiducial-volume background to be small. That is why we apply the same uncertainty as for the P0D. The total effect of the OOFV rate uncertainties is lower than the single values, as they are treated as uncorrelated and partially cancel out.

Table 2.14.: Junction rates.

	Junction probability
Monte Carlo	0.1187 ± 0.0009
Data	0.1217 ± 0.0009

2.6.7.2. Delta-ray rate uncertainty

The delta-ray rate uncertainty is the uncertainty regarding the rate of delta-rays emitted by particles traversing the TPCs. To compute it, a study was performed using the same sample of through-going muons as for the track reconstruction efficiency (Section 2.6.5.1). Through-going tracks were checked in both data and Monte Carlo for the presence of reconstructed junctions. These usually stem from delta-rays in the TPC. The number of tracks with at least one reconstructed junction was compared between data and Monte Carlo. As can be seen in Tab.2.14 there is good agreement between the two. An efficiency weight is applied to all out-of-fiducial-volume MIPs that pass through the active volume of the TPCs.

2.6.7.3. Ad hoc weights

Unblinding the control regions has unveiled a large difference between the data and Monte Carlo distributions of OOFV events. This difference is "fixed" with an ad hoc weight as described in Section 2.5.1. Since we do not know the exact cause of the discrepancy at this point, an uncertainty of 50% is assumed on these weights. This is the largest source of uncertainty in the selection.

2.6.8. Summary of systematic uncertainties

The systematic uncertainties are shown in Tab. 2.15. The dominating systematics uncertainty is the ad hoc weight at 10.0% (see Section 2.5.1), followed by the flux uncertainty at a level of about 4.5 %. Please note that these numbers are lower than the actual uncertainty, due to missing correlations of the systematics' parameters in the used software. A simple "worst case" correlation estimate yields an uncertainty of $\sim$ 10% for the flux weight and $\sim$ 35% for the ad hoc weight (50% realtive uncertainty on $\sim$ 70% background events in the main selection).

The largest "non-standard" source of detector systematics affecting the signal events are hairy secondary tracks. It is obvious that more effort should be put into understanding those hairy events and the TPC performance in general. The effect of the TPC vertex resolution is also relatively high. This might be caused by the high surface area to volume ratio of the fiducial volumes.

Table 2.15.: Systematic uncertainties overview. Average differential errors for the main selection. Maximum correlation approximation in parentheses.

All weights	13.1%	
Flux weight	4.5%	($\sim 10\%$)
Hairy multiplicity	2.9%	
Hairy track eff.	1.0%	
Gas MC	1.3%	
OOFV	3.1%	
Ad hoc weights	10.0%	($\sim 35\%$)
Rest	1.3%	
All variations	6.6%	
Momentum resolution	3.7%	
TPC PID	0.9%	
B-field distortion	3.4%	
TPC vertex resolution	2.0%	
Sand veto	0.8%	
TPC T0 eff.	2.7%	
Hairy track vertex resol.	2.8%	
Momentum scale	1.8%	
Total	17.7%	

3. Cross-section analysis

This chapter describes the fundamentals of the response-matrix-centred approach to cross-section measurements. The mathematical foundations are explained in Section 3.1 and the implementation for the measurement in the ND280 TPCs can be found in Section 3.2. Section 3.3 shows an example of how to use the approach to do a very simple model fit, producing a model dependent, phase-space-restricted, inclusive, charged-current cross-section measurement of muon neutrinos on the T2K TPC gas mixture.

The response-matrix-centred approach is a way of presenting cross-section measurements (or any other kind of counting experiment) in a way that tries to be as model-independent as possible. Its main philosophy can be summarised in three main points:

1. There is a linear relationship between "true" physics expectation values and expected number of measured events.
2. Our knowledge of that relationship is imperfect.
3. The data is the data is the data.

The linear relationship mentioned in the first point is the response matrix. It describes how likely it is to count an event that happened in the detector (efficiency) and in which reconstructed bin it is probably going to end up, i.e. what the reconstructed properties of the event will be (smearing). We know the elements of this matrix only to a certain precision. They are subject to uncertainties of evaluating them.

The actually measured data on the other hand is the only thing we can be 100% sure about. It consists of exact numbers, and systematic or even statistical errors only apply if one interprets the actual data as expectation values for future measurements. E.g. if we do a cross-section measurement and measure 16 events of a certain type, we measure *exactly* 16 events, not something between 12 and 20. Once we try to predict future repetitions of the experiment, we have to interpret this number as measurement of the expectation value, so we get an uncertainty on that: the expectation value is 16 ± 4. In general, there is no one-to-one correspondence between data and the physics variables we are interested in, so the response matrix must be used to translate between the two. Possible ways to do this are described in Section 3.1.

The main result of any measurement presented in this way consists of the *raw* reconstructed data (without any systematic errors) and the response matrix including all uncertainties on the matrix elements. These two objects are everything that is needed to test arbitrary physics models against the data in a consistent way. The tests then produce "traditional" results in truth space.

3.1. Measurement Strategy

3.1.1. Aims

Cross-section measurements are often used to constrain parameters of a given interaction model. This usually means that assumptions of that model are built into the analysis, and published measurements can lose their applicability when the underlying models change. The data then has an implicitly limited "shelf life".

Ideally, a measurement should remain useful not only for the current interaction model, but also for all possible future models. This can be achieved if:

- Arbitrary models can be checked for compatibility with the published data.
- The publication contains all tools and information to do this.
- These tools do not depend on the currently favoured model.

All of this is possible with the response-matrix-centred approach.

3.1.2. The detector model

We categorise all events by their true properties and sort them into a set of truth bins. For Monte Carlo data, these properties are directly accessible, while for real data they remain hidden. Selected events are also binned according to their reconstructed properties. The Poisson expectation value for the number of events in the j-th truth bin is μ_j. It is determined by the underlying physics models and the integrated neutrino flux.

If an event happens in truth bin j, it has a certain probability $P(j \to i)$ to be selected and reconstructed in the i-th reconstruction bin. This probability can be calculated from the MC sample:

$$P(j \to i) = \lim_{N \to \infty} \frac{N(\text{truth} = j, \text{reco} = i)}{N(\text{truth} = j)}.$$

It should depend *only* on the detector properties and *not* on the interaction model. This can be achieved by choosing an appropriate binning in truth space (see Section 3.2.2). The expectation value for the i-th reconstruction bin is then

$$\nu_i = \sum_j P(j \to i)\mu_j.$$

This can be expressed as a matrix product[1]

$$\nu_i = R_{ij}\mu_j,$$

where R is the detector response matrix. Please note that this matrix models both the selection efficiency and reconstruction smearing.

Since we need to know the truth information, R can only be built from MC samples. Unfortunately the simulated detector does not mirror the real detector perfectly. The

[1] Using Einstein notation.

differences are parametrised in a set of systematic uncertainties (see Section 3.4). Their effect on the response matrix is evaluated by producing lots of "toy simulations", in which the same dataset is processed, but the detector properties are sampled from their uncertainty distribution. This yields a set of N_{toy} response matrices R^t, each describing one possible true detector:

$$\nu_i^t = R_{ij}^t \mu_j.$$

3.1.3. The likelihood

One way to measure the compatibility of a given hypothesis and the measured data is the likelihood L. For a discrete counting experiment, it describes the probability of getting exactly the measured result $\boldsymbol{n}$, given the tested hypothesis $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = P(\boldsymbol{n}|\boldsymbol{\theta}).$$

In our framework, the hypothesis is described by the expectation values of the truth bins $\boldsymbol{\mu}$:

$$L(\boldsymbol{\mu}) = P(\boldsymbol{n}|\boldsymbol{\mu}).$$

We can expand this expression to explicitly include the possibility of different detector responses[2]:

$$P(\boldsymbol{n}|\boldsymbol{\mu}) = \sum_R P(\boldsymbol{n}|\boldsymbol{\mu}, R) P(R),$$

where the sum is over all possible detectors and their probabilities to be true. This is impractical, but we can replace the infinite sum with the random sample of toy detectors R^t:

$$P(\boldsymbol{n}|\boldsymbol{\mu}) = \frac{1}{N_{\text{toy}}} \sum_t P(\boldsymbol{n}|\boldsymbol{\mu}, R^t).$$

The sample is drawn from the uncertainty distributions of the detector properties, so more-probable matrices will appear more often than unlikely ones. Within the set of matrices, each one is equally likely.

The remaining probability term is just that of a multi-bin Poisson counting experiment:

$$\begin{aligned} P(\boldsymbol{n}|\boldsymbol{\mu}, R^t) &= P_{\text{Poisson}}(\boldsymbol{n}|\boldsymbol{\nu} = R^t \cdot \boldsymbol{\mu}) \\ &= \prod_i \frac{(R_{ij}^t \mu_j)^{n_i}}{n_i!} \exp(-R_{ij}^t \mu_j) \end{aligned}$$

So ultimately the total marginal likelihood of a tested hypothesis, given the measured data, is

$$L(\boldsymbol{\mu}) = P(\boldsymbol{n}|\boldsymbol{\mu}) = \frac{1}{N_{\text{toy}}} \sum_t \prod_i \frac{(R_{ij}^t \mu_j)^{n_i}}{n_i!} \exp(-R_{ij}^t \mu_j).$$

[2] Choosing a Bayesian approach for brevity.

Alternatively one can also choose to use the profile likelihood

$$L_{\text{profile}}(\boldsymbol{\mu}) = \max_t \prod_i \frac{(R^t_{ij}\mu_j)^{n_i}}{n_i!} \exp(-R^t_{ij}\mu_j),$$

which just selects the toy migration matrix with the highest resulting likelihood.

The profile likelihood is only useful for strictly constrained response matrices, i.e. with hard bounds on all parameters. If a parameter of the matrix is distributed without strict limits[3] and the maximum likelihood is achieved for very extreme matrices, the achieved likelihood will depend a lot on the number of toy matrices. The more matrices are sampled from the unlimited distribution, the more extreme the most extreme matrix will become. If, on the other hand, all parameters are sampled from bounded distributions[4], the extremeness of the most extreme matrix will tend to a limiting value instead of rising towards infinity with the number of toy matrices.

3.1.4. Absolute maximum likelihood

A *simple hypothesis* is completely characterised by the vector of truth expectation values $\boldsymbol{\mu}$. It has no free parameters. Each expectation value must be a non-negative real number, $\mu_j \in \mathbb{R}_{\geq 0}$. This defines the set of all conceivable hypotheses Ω:

$$\Omega = \mathbb{R}^d_{\geq 0},$$

where $d = \dim(\boldsymbol{\mu})$ is the number of truth bins. We can thus define a maximum likelihood hypothesis $\boldsymbol{\mu}_{\max\, L}$ such that

$$L_{\max}(\Omega) = L(\boldsymbol{\mu}_{\max\, L}) = \max_{\boldsymbol{\mu}\in\Omega} L(\boldsymbol{\mu}).$$

This hypothesis and its likelihood value can then be used as a baseline to compare other hypotheses to it.

3.1.5. Likelihood ratio testing

The agreement between the data and any given hypothesis can be evaluated with the likelihood ratio λ:

$$\lambda(\boldsymbol{\mu}) = \frac{L(\boldsymbol{\mu})}{L_{\max}(\Omega)}.$$

By construction, this value is in the range $[0, 1]$. A high value shows good agreement, while low values indicate disagreement.

According to the *Neyman–Pearson lemma* [37], a hypothesis test using λ as a test statistic is the most powerful[5] test possible. So we define a critical value η and reject hypotheses where $\lambda < \eta$. The choice of η depends on the desired significance[6] of the test

[3] E.g. with a normal distribution, which is not bounded in either direction.
[4] E.g. uniform distributions.
[5] The power of a test $1 - \beta$ describes the probability of rejecting a false hypothesis.
[6] The significance of a test α describes the probability of rejecting a true hypothesis.

α and the expected distribution of the likelihood ratio $f(\lambda)$ given that $\boldsymbol{\mu}$ is true. It must be chosen such that

$$P(\lambda < \eta|\boldsymbol{\mu}) = \int_0^\eta f(\lambda|\boldsymbol{\mu})\,\mathrm{d}\lambda \overset{!}{\leq} \alpha.$$

The distribution of λ must be evaluated for each tested hypothesis separately, for example by doing a sufficient number of MC experiments. The critical value is thus a function of the hypothesis $\eta(\boldsymbol{\mu})$.

This can be rectified by using the likelihood ratio p-value as test statistic directly. It is the probability of measuring a likelihood "worse" than the actually measured one λ_0, assuming that the tested hypothesis is true:

$$p_\lambda(\boldsymbol{\mu}) = P(\lambda < \lambda_0|\boldsymbol{\mu}) = \int_0^{\lambda_0} f(\lambda|\boldsymbol{\mu})\,\mathrm{d}\lambda.$$

By construction, this value is uniformly distributed for the true hypothesis, so the critical value is just the significance, and hypotheses are rejected if

$$p_\lambda(\boldsymbol{\mu}) < \alpha.$$

3.1.6. Composite hypotheses

Often a tested hypothesis will have some free (nuisance) parameters. Those are called *composite hypotheses*. They will define the truth expectation values $\boldsymbol{\mu}$ as a function of these free parameters $\boldsymbol{\mu}(\boldsymbol{\theta})$, with the number of free parameters $d' = \dim(\boldsymbol{\theta}) < d$. The possible values of $\boldsymbol{\theta}$ define the set of *simple* hypotheses Θ, which is a subset of all conceivable hypotheses:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) \in \Theta \subset \Omega \quad \Leftrightarrow \quad \boldsymbol{\theta} \in \omega,$$

where ω is the set of allowed values of $\boldsymbol{\theta}$. For example, if all parameters are unrestricted real values, we have

$$\omega = \mathbb{R}^{d'}.$$

We consider a composite hypothesis *true* if it contains the true simple hypothesis $\boldsymbol{\mu}_{\mathrm{true}}$, and *false* otherwise.

Again, we would like to test hypotheses with the highest possible power at a given significance. To reject a composite hypothesis Θ, we must reject all contained simple hypotheses $\boldsymbol{\mu}(\boldsymbol{\theta})$:

$$\lambda(\boldsymbol{\mu}) < \eta(\boldsymbol{\mu}) \quad \forall \quad \boldsymbol{\mu} \in \Theta.$$

As an approximation with lower than ideal power, we can consider the maximum likelihood and minimum critical value:

$$\lambda_{\max}(\Theta) = \max_{\boldsymbol{\mu}\in\Theta} \lambda(\boldsymbol{\mu}) \overset{!}{<} \eta_{\min}(\Theta) = \min_{\boldsymbol{\mu}} \eta(\boldsymbol{\mu}).$$

Depending on the variation of $\eta(\boldsymbol{\mu})$ within Θ[7], the significance of the test will be less than or equal to the nominal value α.

[7] In the limit of large sample sets, the distribution of $\lambda_{\max}$ will approach a χ^2-distribution [38] and an exact value for η can be chosen accordingly. Unfortunately this does not apply in this selection.

To increase the power of the test, we can use the p-values of the likelihood ratios directly. With this test statistic, the critical value is identical for all simple hypotheses, and we reject a composite hypothesis if

$$p_\lambda(\boldsymbol{\mu}) < \alpha \quad \forall \quad \boldsymbol{\mu} \in \Omega.$$

That means we can exclude a composite hypothesis by checking whether

$$p_{\max}(\Theta) = \max_{\boldsymbol{\mu}\in\Theta} p_\lambda(\boldsymbol{\mu}) < \alpha$$

with maximum power. This is called the "supremum method" [39] and not computationally harder than finding $\eta_{\min}$. In both cases $f(\lambda|\boldsymbol{\mu})$ has to be calculated for each evaluation in the minimisation/maximisation process.

3.1.7. Parameter estimation

If a composite theory Θ is not rejected, one might want to quote a set of "best fit" parameters and/or a range of allowed values, i.e. confidence intervals. The maximum likelihood point estimator for the parameters $\hat{\boldsymbol{\theta}}$ is straight forward. It is the set of parameters that produce the highest likelihood:

$$L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})) = L_{\max}(\Theta).$$

Confidence intervals for the parameters can be calculated by rejecting part of the possible parameter space analogously to the general composite hypothesis test in Section 3.1.6. For this, we split the parameters into interesting parameters $\boldsymbol{\theta}$, where we want to quote the intervals, and nuisance parameters $\boldsymbol{\phi}$. We then interpret the set of all $\boldsymbol{\mu}(\boldsymbol{\theta}, \boldsymbol{\phi})$ with a fixed $\boldsymbol{\theta}$ as a new composite hypothesis $\Theta(\boldsymbol{\theta})$:

$$\boldsymbol{\mu}(\boldsymbol{\theta}, \phi) \in \Theta(\boldsymbol{\theta}) \subset \Theta \quad \Leftrightarrow \quad \boldsymbol{\phi} \in \Phi(\boldsymbol{\theta}),$$

where $\Phi(\boldsymbol{\theta})$ is the set of allowed values of $\boldsymbol{\phi}$ given a specific $\boldsymbol{\theta}$. Now we can exclude values of $\boldsymbol{\theta}$ by checking whether

$$p_{\max}(\Theta(\boldsymbol{\theta})) < \alpha.$$

Those values of $\boldsymbol{\theta}$ that have not been rejected define the confidence region.

It might be useful to construct confidence intervals for parameters of composite hypotheses whether or not they have been excluded. In these cases, one is usually only interested in the allowed parameter range within the context of the analysed hypotheses. This can easily be achieved by replacing the absolute maximum likelihood $L_{\max}(\Omega)$ with the maximum likelihood of the hypothesis $L_{\max}(\Theta)$. The likelihood ratio is then

$$\lambda(\boldsymbol{\mu}) = \frac{L(\boldsymbol{\mu})}{L_{\max}(\Theta)},$$

and the construction of the confidence interval only ever compares the nested hypothesis $\Theta(\boldsymbol{\theta})$ directly with the enveloping hypothesis Θ. This can reduce the number of parameters considerably, as no evaluation of the absolute maximum likelihood needs to be done.

3.1.8. Profile plug-in p-values

Even when only comparing two hypotheses with moderate number of parameters, finding $p_{\max}(\Theta(\boldsymbol{\theta}))$ is a computationally intensive task. Calculating the p-value for a single $\boldsymbol{\mu}$ takes the generation of $\mathcal{O}(100)$ toy data sets from the reco predictions of that hypothesis, and then maximising the likelihoods of both compared composite hypotheses for each data set. Maximising the p-value with a typical optimisation algorithm means that it has to be evaluated at least $\mathcal{O}(10000)$ times, depending on the difficulty of finding the global(!) likelihood maxima. This quickly escalates into millions upon millions necessary fits and a corresponding demand of computing power.

A drastic reduction can be achieved when using the "profile plug-in" p-value instead of the maximum p-value. Instead of maximising the p-value over all possible hypotheses $\boldsymbol{\mu} \in \Theta$, one only evaluates the p-value of the most likely hypothesis $\hat{\boldsymbol{\mu}}$:

$$p_{\text{plug}}(\Theta) = p_\lambda(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})),$$

or in the context of parameter estimation:

$$p_{\text{plug}}(\Theta(\boldsymbol{\theta})) = p_\lambda(\boldsymbol{\mu}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}})).$$

Here $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\phi}}$ are the maximum likelihood estimates of the (nuisance) parameters:

$$L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})) = L_{\max}(\Theta),$$

$$L(\boldsymbol{\mu}(\boldsymbol{\theta}, \hat{\boldsymbol{\phi}})) = L_{\max}(\Theta(\boldsymbol{\theta})).$$

The calculation of this value requires only a single optimisation of the likelihood. The p-value itself is then computed with toy data assuming the truth of the estimate.

This is called "profile plug-in" p-value as we plug-in the profile maximum likelihood estimate for the nuisance values as an estimate for the distribution of the likelihood ratios of the true hypothesis. The method has certain advantages over other approximation methods [40], but it is still an approximation. It is thus important to check the coverage properties of any analysis using this method.

3.1.9. Bayesian posterior sampling

The exact Frequentist approaches described above need a prohibitive amount of computing power when the number of parameters of the tested composite hypothesis is large. Those models are better handled by a Bayesian approach. Using a Marcov Chain Monte Carlo (MCMC) method, it is relatively easy to sample parameter sets $\boldsymbol{\theta}$ from the posterior probability

$$P(\boldsymbol{\theta}|\boldsymbol{n}) \propto L(\boldsymbol{\mu}(\boldsymbol{\theta}))\, P(\boldsymbol{\theta}),$$

with the prior probability $P(\boldsymbol{\theta})$. These sets can then be used to infer information about the parameters, e.g. point estimates or credible intervals, and to compare different hypotheses with one another.

Since the information about the absolute probability of the hypothesis is lost in the MCMC, we propose using the *Posterior distribution of the Likelihood Ratio* (PLR) to infer the data preference of one model over another. The PLR is defined as the posterior probability of the likelihood ratio of the compared hypotheses being below or equal to a certain threshold value:

$$\mathrm{PLR}_{\Theta_0,\Theta_1}(\boldsymbol{n},\zeta) = P\left(\frac{L(\boldsymbol{\mu}(\boldsymbol{\theta}_0))}{L(\boldsymbol{\mu}(\boldsymbol{\theta}_1))} \leq \zeta \,\middle|\, \boldsymbol{\theta}_0 \sim P(\boldsymbol{\theta}_0|\boldsymbol{n}), \boldsymbol{\theta}_1 \sim P(\boldsymbol{\theta}_1|\boldsymbol{n})\right),$$

with $\boldsymbol{\mu}(\boldsymbol{\theta}_0) \in \Theta_0$ and $\boldsymbol{\mu}(\boldsymbol{\theta}_1) \in \Theta_1$ the (completely independent) parametrisations of the tested hypotheses. If the threshold value ζ is set to 1, the PLR is equivalent to a Frequentist p-value under certain circumstances [41]. But even when this is not the case, the interpretation is straight forward: $\mathrm{PLR}_{\Theta_0,\Theta_1}(\boldsymbol{n},\zeta=1)$ is the posterior probability of the data being more likely under Θ_1 than under Θ_0.

As with all Bayesian analyses, it is important to choose suitable priors $P(\boldsymbol{\theta})$. There is no single "correct" way to do this, but we recommend using a *Jeffreys prior* [42]. Its main advantage is that its probability density – and especially the posterior probability density resulting from using this prior – is *invariant* under variable transformations. This means that the results of the analysis do *not* depend on the particular parametrisation of $\boldsymbol{\mu}(\boldsymbol{\theta})$. A drawback of Jeffreys priors is that they are not necessarily proper, i.e. they cannot be normalised. This is not a problem here though, as long as the posterior is well defined.

The Bayesian approach treats all unknown parameters equal. It is therefore natural to also include the detector uncertainties in the MCMC sampling. We simply treat the detector toy index as additional (nuisance) parameter of the model. The posterior probability thus also includes information about how likely or unlikely the different toy detectors are:

$$\begin{aligned} P(\boldsymbol{\theta},t|\boldsymbol{n}) &\propto L^t(\boldsymbol{\mu}(\boldsymbol{\theta}))\, P(\boldsymbol{\theta},t) \\ &= P(\boldsymbol{n}|\boldsymbol{\mu}(\boldsymbol{\theta}),R^t)\, P(\boldsymbol{\theta},t). \end{aligned}$$

3.1.10. Closure tests

All methods described here depend on the model-independence of the response matrix, so this needs to be ensured with dedicated tests. For the Bayesian analysis one also needs to make sure of the convergence of the MCMC samples. To this effect, one should apply two checks (see Fig. 3.1). We use two different event generators to build the detector response matrices and demand that their results are compatible (see 3.2.3). When doing the MCMC analyses one should generate multiple MCMC posterior samples with different start values and check them for compatibility. If the MCMCs have converged, the differences should be statistical only.

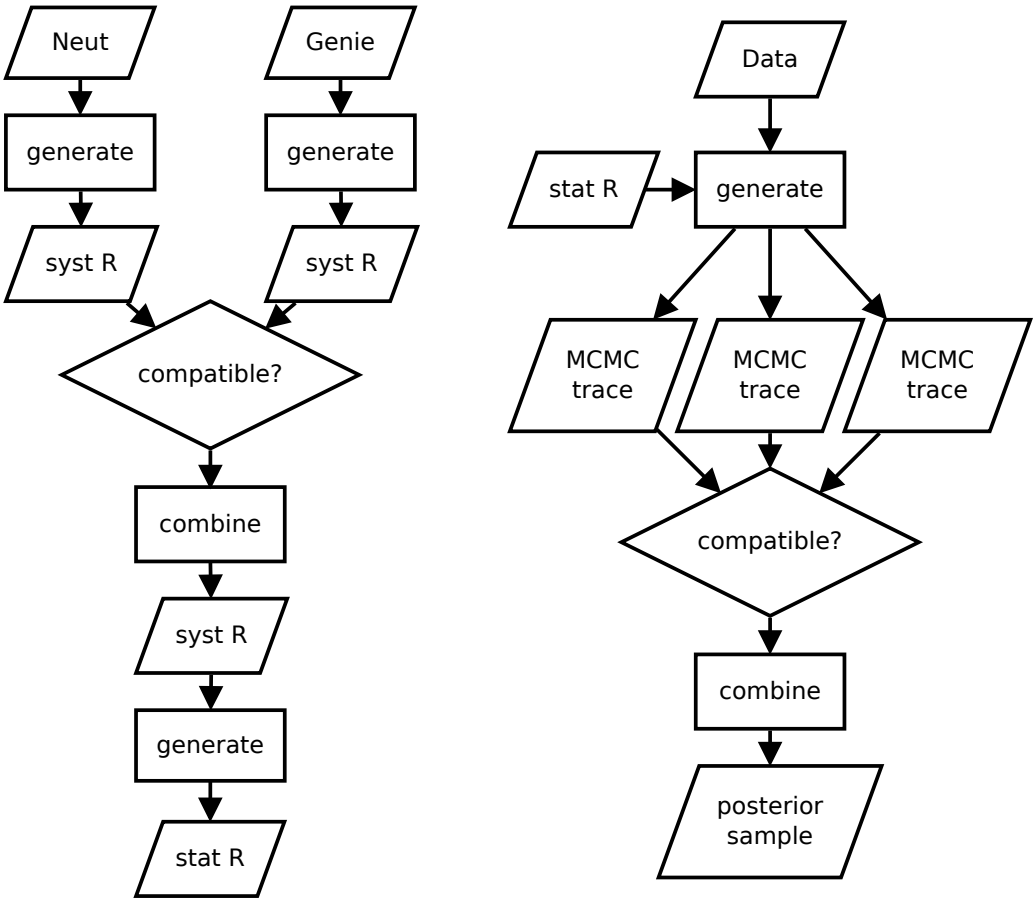


Figure 3.1.: Data flow. The two neutrino event generators Neut and Genie are used to build systematically varied response matrices (see Section 3.2.1). If the true matrix is independent of the generator model, the matrices should be compatible within the detector variations and they are combined to increase the available MC statistics. Based on these systematic response matrices, the statistical variations are evaluated. In the case of a Bayesian analysis, the data is then used together with the statistically varied matrices to generate multiple MCMC posterior samples. If the MCMCs converged, the posterior samples should only differ within statistical fluctuations. In this case, they can be combined into the final a-posteriori distribution of the hypothesis parameters.

3.2. Implementation

3.2.1. Building the detector response matrices

The detector response matrix is built from Monte Carlo simulations. Events are first categorised by their truth information and assigned a truth bin number j. It is possible that events do not get assigned a truth bin (see Section 3.2.2). Those events are ignored for the build of the response matrix. All events with a truth bin number are then categorised by their reconstructed information. Events that end up being selected by the gas interaction selection[8] get assigned a reco bin number i.

The probability for an event in truth bin j ending up in reco bin i is

$$P(j \to i) = \lim_{N_j \to \infty} \frac{N_{ij}}{N_j},$$

as defined in Section 3.1. Here N_j is the number of events in truth bin j, including the events that do not get assigned a reco bin, and N_{ij} the number of events in truth bin j and reco bin i. Since the number of Monte Carlo events is limited by the available computing resources, this value can only be approximated:

$$R_{ij} = \frac{N_{ij}}{N_j}.$$

The simulated detector does not reproduce the behaviour of the real one perfectly. We parametrise the estimated difference as a set of systematic uncertainties that get propagated as weights and variations in the selection (see Section 3.4). The systematic parameters are sampled from their assumed distributions (normal or uniform) and the events are weighted and varied accordingly. We call each sampling of the parameter space a toy simulation. Each toy simulation t yields its own response matrix

$$R^t_{ij} = \frac{W^t_{ij}}{W_j} = \frac{N^t_{ij} w^t_{ij}}{N_j w_j},$$

where W_j and W^t_{ij} are the sum of weights, and w_j and w_{ij} the average weights of all events in the respective bins. Since the detector variations do not affect the events on the generator level, the number of event in the truth bin N_j is not affected by the toys.

The number of MC events is limited, so the values of R^t_{ij} will also suffer from statistical variations from the true MC value. These fluctuations are not represented in the systematic toys. We estimate the statistical uncertainties in a "Bayesian inspired" three-step process.

The first two uncertainties stem from the multinomial sampling of N^t_{ij}. For the purpose of statistical error estimation, we split the multinomial process in two parts:

- A binomial chance of being reconstructed at all (i.e. efficiency) ϵ^t_j
- A multinomial probability of ending up in a certain reco bin (i.e. smearing) p^t_{ij}

[8] The main selection or one of the control samples, see Chapter 2.

$$\epsilon_j^t = \lim_{N_j \to \infty} \frac{\sum_i N_{ij}^t}{N_j}.$$

$$p_{ij}^t = \lim_{N_j \to \infty} \frac{N_{ij}^t}{\epsilon_j^t N_j}.$$

$$\epsilon_j^t \cdot p_{ij}^t = \lim_{N_j \to \infty} \frac{N_{ij}^t}{N_j}.$$

We do not know the true values of these parameters, so we can treat them as Bayesian random variables.

If we assume a beta distribution[9] as a prior for the distribution of ϵ_j^t, we can use the simulated number of events directly to update the parameters of the prior, β'_{*j} and $\beta'_{\dagger j}$, to get the parameters of the posterior[10]:

$$\epsilon_j^t \sim \mathrm{Beta}(\beta_{*j}^t, \beta_{\dagger j}^t),$$

$$\beta_{*j}^t = \beta'_{*j} + \sum_i N_{ij}^t = \alpha'_j + N_{*j}^t,$$

$$\beta_{\dagger j}^t = \beta'_{\dagger j} + (N_j - \sum_i N_{ij}^t) = \beta'_j + N_{\dagger j}^t.$$

Here N_{*j}^t is the number of selected and $N_{\dagger j}^t$ the number of "lost", i.e. not selected, events in truth bin j.

We can do the same for the smearing uncertainty if we assume a Dirichlet distribution[11] as a prior for the distribution of p_{ij}^t. Again we can use the simulated number of events directly to update the parameters:

$$\boldsymbol{p}_j^t \sim \mathrm{Dir}(\boldsymbol{\alpha}_j^t),$$

$$\alpha_{ij}^t = \alpha'_{ij} + N_{ij}^t.$$

The variances of the resulting posterior distributions are

$$\sigma^2(\epsilon_j) = \frac{\beta_{*j}^t \beta_{\dagger j}^t}{(\beta_{*j}^t + \beta_{\dagger j}^t)^2(\beta_{*j}^t + \beta_{\dagger j}^t + 1)},$$

$$\sigma^2(p_{ij}^t) = \frac{\alpha_{ij}^t(\sum_{i' \neq i} \alpha_{i'j}^t)}{(\sum_{i'} \alpha_{i'j}^t)^2((\sum_{i'} \alpha_{i'j}^t) + 1)},$$

and the expectation values

$$\hat{\epsilon}_j^t = \frac{\beta_{*j}^t}{\beta_{*j}^t + \beta_{\dagger j}^t} = \frac{\beta_{*j}^t}{N_j + \beta'_{*j} + \beta'_{\dagger j}},$$

[9] The beta distribution is the conjugate prior for binomial distributed likelihoods. See [43].

[10] Usually the parameters of the beta function are denoted as α and β. To avoid confusion with the parameters of the Dirichlet distribution α_i, we decided to use β_* and $\beta_\dagger$ respectively instead.

[11] The Dirichlet distribution is the conjugate prior for multinomial distributed likelihoods. See [43].

$$\hat{p}^t_{ij} = \frac{\alpha^t_{ij}}{\sum_{i'} \alpha^t_{i'j}}.$$

As prior parameters we set

$$\beta'_{*j} = \beta'_{\dagger j} = 1$$

and

$$\alpha'_{ij} = \min(1, 3^{N_{\text{reco variables}}}/N_{\text{reco bins}}).$$

This choice of prior parameters ensures that the overall reconstruction (in)efficiency of the truth bins is uniformly distributed apriori. It also means that the prior assumes that the reconstruction probabilities are concentrated on a few reco bins (~ 3 per reco variable), while being completely agnostic about *which* reco bins those are.[12] The resulting variances of the posterior distributions are consistent with the standard frequentist approach in the limit of high statistics. Especially the binomial case correspond to an experiment where we added a "pseudo-observation" of two simulated events, of which one was successfully reconstructed, to the actual data.

The third step is to evaluate the statistical uncertainty of the weight correction. The true weight correction

$$m^t_{ij} = \lim_{N_j \to \infty} \frac{w^t_{ij}}{w_j}$$

is estimated from the sum of of weights:

$$\hat{m}^t_{ij} = \frac{w^t_{ij}}{w_j} = \frac{W^t_{ij}/N^t_{ij}}{W_j/N_j}.$$

For the purpose of the variance estimation, we treat w^t_{ij} *independently* from ϵ^t and p^t_{ij} as arithmetic means of samples with given sizes.

We apply the usual standard error of the mean formula for each average weight. The sample variance is estimated from the sum of squared weights. To be able to estimate variances even for bins with only one entry, we add a pseudo-observation event with an expected weight of 1:

$$\sigma^2(w^t_{ij}) = \left(\left(\frac{V^t_{ij} + 1^2}{N^t_{ij} + 1} \right) - \left(\frac{W^t_{ij} + 1}{N^t_{ij} + 1} \right)^2 \right) \frac{1}{N^t_{ij} + 1},$$

where V^t_{ij} are the sums of the squared weights in the respective bins. The pseudo-observation represents our prior knowledge of the weights and has no effect in the limit of high statistics. The variance of the weight correction is then

$$\sigma^2(m^t_{ij}) = \frac{\sigma^2(w^t_{ij})}{(w_j)^2} + \left(\frac{w^t_{ij}}{(w_j)^2} \right)^2 \sigma^2(w_j).$$

[12] Dirichlet distributions with $\alpha < 1$ favour "extreme" sets of p, where most of the probability is concentrated in few categories, over flat sets, where the probability is more uniformly distributed. The corresponding reco bins do *not* have to be contiguous.

All that is left now, is to combine the variances of the multinomial sampling and the weight correction:

$$\sigma^2_{\text{MC stat}}(R^t_{ij}) = (\hat{\epsilon}^t_j \hat{m}^t_{ij})^2 \sigma^2(p^t_{ij}) + (\hat{\epsilon}^t_j \hat{p}^t_{ij})^2 \sigma^2(m^t_{ij}) + (\hat{p}^t_{ij} \hat{m}^t_{ij})^2 \sigma^2(\epsilon^t_j)$$

If the statistical variance is much smaller than the systematic detector variation,

$$\sigma^2_{\text{MC stat}}(R^t_{ij}) \ll \sigma^2_{\text{syst}}(R_{ij}) \approx \frac{1}{N_{\text{toy}} - 1} \sum_t (R^t_{ij} - \bar{R}_{ij})^2,$$

for all toy experiments, we can neglect it. In practice there will almost certainly be bins where this is not the case, e.g. (elmost) empty matrix elements.

To deal with these non-negligible statistical uncertainties, we generate random toy matrices from every systematic toy matrix according to the three step process described above: First we draw a set of efficiencies and multinomial probabilities from the posterior beta/Dirichlet distributions, and then we modify these with weight factors calculated from normal distributed mean weights.

$$\epsilon^{t*}_j \sim \text{Beta}(\beta^t_{*j}, \beta^t_{\dagger j}),$$

$$\boldsymbol{p}^{t*}_j \sim \text{Dir}(\boldsymbol{\alpha}^t_j),$$

$$w^{t*}_{ij} \sim \text{Norm}(w^t_{ij}, \sigma^2(w^t_{ij})),$$

$$R^{t*}_{ij} = \frac{w^{t*}_{ij}}{w^{t*}_j} \epsilon^{t*}_j p^{t*}_{ij}.$$

These toy matrices are then handled just like the systematic toy matrices.

To further limit the influence of the statistical uncertainties, we constrain the truth bin expectation values to the number of simulated events:

$$\mu_j \overset{!}{<} N_j.$$

Hypotheses that predict more events in a given truth bin than were simulated are outside the testable scope of the response matrix. If the tested hypotheses (e.g. in a likelihood fit or Bayesian posterior sampling) are close to this limit, it could lead to model dependence of the results. Therefore it is necessary to check whether this is the case.

3.2.2. Binning

3.2.2.1. General considerations

The central object of this analysis is the detector response matrix R_{ij}. Its properties depend first and foremost on the chosen binning in truth and reco space. The binning has to balance the following (contradictory) aims:

- Ensure the independence of the interaction model.

Table 3.1.: Binning variables.

Truth variables	Reco variables
Event type	Successful sample
Primary MIP $\cos(\theta)$	HMM $\cos(\theta)$
Primary MIP momentum	Multiplicity
Angular separation	
(Charged multiplicity)	

- Minimise the influence of statistical errors.
- Maximise the separation power, i.e. resolution.

The following sections describe the general methodology of choosing the binning. The actually chosen binning for the CC_{inc} measurement is shown in Section B.1.

3.2.2.2. Choosing the variables to bin in

The response matrix can only be model-independent if it is binned in the right variables. Variables close to the actual observables are more suited than those that describe the event in a more fundamental way, which have to be inferred from the measurement. For example, the lepton momentum is a good variable, as the detector can directly measure it. The neutrino energy on the other hand is a bad choice, because the translation of neutrino energy to observables in the detector depends on the physics model (FSI, etc).

But even when binning in direct observables only, one has to take care not to introduce hidden model dependencies. The distribution of events in variables that we do not bin in, might still have an effect on the detector performance. If different models predict different angular distributions, which in turn change the detector efficiency, binning only in the muon momentum will *not* be model independent. One has to bin in *all* truth variables that affect the detector performance to be truly model independent.

Aside from detector performance considerations, one of course also has to bin in the variables of interest. The reco binning is dictated by the physics goals of the measurement. Again it is important to choose variables as close to the actual observables as possible. If a variable of interest is the function of other more basic observables, a binning in those observables would be less susceptible to hidden model dependencies. Unfortunately the number of events per bin decreases exponentially with the number of binning dimensions. Our choice of binning variables for the template fit analysis is summarised in Tab. 3.1 and described in detail in Section 3.2.2.5 and onwards.

3.2.2.3. Bin widths

As seen in Section 3.2.1, the efficiencies of the truth bins are estimated using a Bayesian approach that adds two pseudo-measurements as prior information. In order to not be biased too much towards that prior, we would like the actual number of events per truth

bin to be larger than that. The average number of events per truth bin is a good measure for this. We demand that $\text{mean}(N_j) \overset{!}{>} 50$.

To maximise the number of events per bin, one could choose a very wide binning. This can lead to model dependences though, if the detector performance varies considerably within a truth bin. Since model independence is a primary goal of this analysis, this defines an upper limit for the truth bin sizes.

We estimate the response variation within one bin ΔR_{ij} from the variation between neighbouring bins:

$$\Delta R_{ij} = \max_{j'} |R_{ij} - R_{ij'}|,$$

with the neighbouring bins j'. Ideally, one would like this variation to be not much higher[13] than the uncertainties on the matrix elements:

$$\Delta' R_{ij} = \max_{j'} \left| \frac{R_{ij} - R_{ij'}}{\sqrt{\sigma^2(R_{ij}) + \sigma^2(R_{ij'})}} \right| \overset{!}{<} l,$$

with the normalised in-bin variation $\Delta' R_{ij}$ and a limit $l \sim \mathcal{O}(5)$. Unfortunately this aim is contradictory to the need to fill each truth bin with sufficiently many events to reduce the influence of the priors (see above). Also, small scale variations might be hidden within the bins, so care has to be taken on a variable by variable base to optimise the binning.

If the detector response is sufficiently flat, the truth bin widths should be adjusted to include the necessary MC statistics. Other than that, they should be made as small as possible. The reco bin width is mostly dictated by the physics goals of the analysis and MC statistics. If the truth binning is chosen as to ensure model independence, no reco binning will introduce additional model dependence. A fine reco binning might expose model dependencies, but the cause is solely in the truth binning. Conversely, a coarse binning can hide dependencies, so one should aim for as fine a binning as MC and dara statistics permit. This also ensures the best hypothesis testing power. Reco bins should not be finer than their truth counterparts.

We deploy the following algorithm to decide on the final bin widths:

1. Set reco binning according to expected statistics and physics goals.
2. Set truth binning very fine.
3. Merge truth bins until $\text{mean}(N_j) \overset{!}{>} 50$
   a) Set limit for in-bin variation l.
   b) Merge neighbouring bins with lowest number of entries until limit is reached.
   c) Merge neighbouring bins with lowest in-bin variation $\Delta' R_{ij}$ until limit is reached.

[13] Ideally one would like the variation within the bins to be lower than the statistical uncertainty, but if there is no actual in-bin variation, the statistical uncertainty will dominate this estimate.

d) If necessary, increase l and repeat.

4. Fine-tune binning by hand.

After this, the resulting matrix is checked for sufficient model-independence in the closure tests (see Section 3.2.3).

3.2.2.4. Empty bins

The Monte Carlo samples used to generate the response matrix use a physics model $\boldsymbol{\mu}'$. In that model, certain areas of the truth phase space are very unlikely to be realised and the corresponding truth bins will not be filled with a sufficient number of events during the response matrix construction. This means that we have not enough information about how the detector would react to these kinds of events. Ideally one would like to build the response matrix with simulation data that covers all possible phase space, but this is computationally difficult.

Since we cannot predict how those events behave in the detector, we remove those bins from the vector of truth expectation values $\boldsymbol{\mu}$. This is equivalent to setting those expectation values to 0 in all considered hypotheses, and reduces the dimensionality of Ω. The events that would have been assigned to these bins are now ignored. This means that we *cannot* test hypotheses that predict any events in these bins.

There might also be reconstruction bins that never get filled during the construction of the response matrix. The expectation value in those bins will be close[14] to 0 for all possible hypotheses in Ω. Finding events in these bins would necessitate further investigation and possibly the generation of more Monte Carlo data.

To judge how well the simulated data covers the real measurement and tested hypotheses, we can compare the number of simulated events to the number of measured/predicted events:

$$\xi_{\text{reco},i}(\boldsymbol{n}) = \max_t \frac{n_i}{N_i^t}$$

$$\xi_{\text{truth},j}(\boldsymbol{\mu}) = \frac{\mu_j}{N_j}.$$

Numbers close to or above one indicate that the simulated phase space is not sufficient and should be extended. More specifically, $\boldsymbol{\xi}_{\text{truth}}(\boldsymbol{\mu})$ indicates how well the given hypothesis $\boldsymbol{\mu}$ is covered by the simulation, while $\boldsymbol{\xi}_{\text{reco}}(\boldsymbol{n})$ shows whether the actual measurement is covered at all.

3.2.2.5. Event categories

True events fall into eight categories: four *in-fiducial-volume* (FV) categories and four *out-of-fiducial-volume* (OOFV) categories. The gas interaction selection categorises FV events by the highest-momentum, negatively-charged particle:

[14] It will not be exactly 0 due to the generation of statistically varied matrices as described in Section 3.2.1.

Table 3.2.: Reconstructed event categorisation by control samples.

No likelihood match?	Passes T0-quality?	No passing-by veto?	Selection branch
Yes	Yes	Yes	Main selection
Yes	Yes	No	Passing-by CS
Yes	No	*	Timing CS
No	*	*	Delta ray CS

- negative MIP (muon or pion)
- electron
- negative other
- no negative particle

The OOFV background is categorised by failure mode:

- delta-ray induced
- timing related
- particles stopping or decaying in the TPC
- multiple particles entering the TPC in coincidence
- other

The last two modes are actually combined into a single OOFV category in the matrix, as the difference between them is already encoded in the forward separation variable (see Section 3.2.2.7).

To constrain the number of OOFV events, the selection defines three control samples. They are created by inverting the T0-quality, passing-by veto and likelihood-match cuts. Table 3.2 shows the definitions of the samples. This means a total of four reco category bins, one for each sample.

Events in the OOFV categories are conceptually different from the FV events. Even their truth binning is defined with (some) reconstructed information, i.e. that something in the reconstruction went wrong. There is no pure truth-space definition as there is no true OOFV event unless it is actually reconstructed. It is not possible to define a useful efficiency for these events, so we modify the definition of the denominator in the response matrix for OOFV events slightly:

$$R_{ij} = \frac{N_{ij}}{N'_j} \quad \forall \quad j \in \text{OOFV},$$

with the total number of *selected* events N'_j. The efficiency of the these "nuisance bins" is fixed[15], reducing the number of free parameters in the description of the response matrix. This means that the respective expectation values μ_j do not describe the absolute number of potential OOFV events that happened in the detector, but the number of actually selected OOFV events. The background from events in the fiducial volume, e.g. neutral-current events, is not affected by this.

3.2.2.6. Primary MIP & HMM kinematics

The main idea of the selection is to find muon tracks that start in the TPCs' fiducial volume. The selection performance depends on the muon direction with respect to the neutrino beam and momentum, so we have to bin in these variables.

Actually, the selection is not looking for muons, but for *Minimum Ionizing Particles* (MIPs), since the detector cannot distinguish well between muons and pions. It selects the negative *Highest Momentum MIP candidate* (HMM) starting in the TPCs, and the reco binning is done in the momentum and angle of this particle.

To reflect this as closely as possible in the truth binning, we define a *primary MIP*. It is the highest momentum negative muon or pion for FV events. For OOFV events, it is the particle that has been misidentified as HMM, regardless of charge. Truth binning is done in the variables of this particle.

Since the expected data statistics are low, we decided to do a very coarse reco binning for the first analysis. There are only two bins in reconstructed $\cos(\theta)$, with bin-edges at $-1, 0, 1$, distinguishing only between forward and backward going events. We decided to bin in $\cos(\theta)$, because the OOFV background is more isotropic than the expected signal. It thus gives us some separation power between the two. We do *not* bin in the reconstructed momentum, since the expected momentum distributions are very similar.

3.2.2.7. Particle separation

The (mis-)reconstruction probability heavily depends on how well the main particle is "separated" from other particles in the event. As a measure for this, we define the forward and backward separation $s_{\text{fwd/bwd}}$. For high-momentum particles it simply corresponds to the closest angle between the particle direction and its neighbours (see Fig. 3.2).

For very-low-momentum particles, their limited range must be taken into account. The overlap x between two straight, long tracks can be calculated from their separation angle and a separation scale y (see Fig. 3.3):

$$\tan(\theta) = y/x.$$

The overlap of low-momentum particles is limited by their range, which corresponds to a minimum separation:

$$s = \max(\theta, \arctan(y/x)).$$

[15] For a single toy matrix, the efficiency will always be 1. To make things consistent between multiple toy matrices with different numbers of selected BG events, N'_j is defined using the sum of selected BG events in *all* toy selections. This ensures that all toy matrices reproduce the BG distributions they were created with.

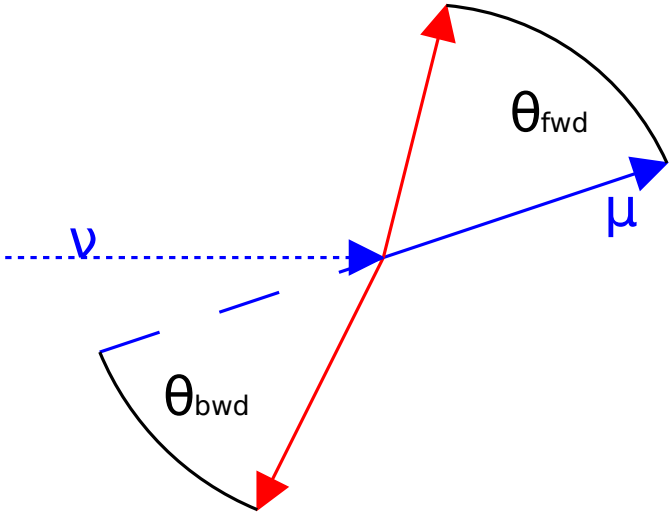


Figure 3.2.: Angular separation in forward and backward direction.

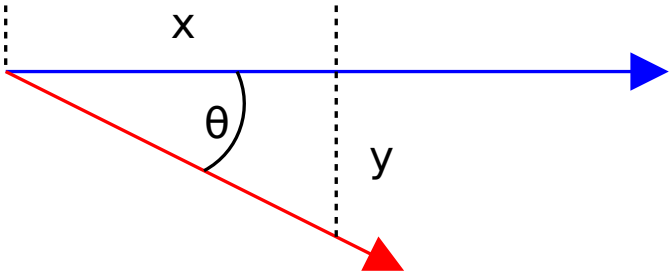


Figure 3.3.: Overlap between two tracks. After the overlap distance x, the two tracks are separated by more than the separation scale y.

We set the separation scale to $y = 5\,\mathrm{cm}$. This corresponds to more then five times the size of the pads of the readout plane, so the tracks should be fairly easy to separate at that distance. The particle range in the TPCs can be estimated from its mass and momentum (see Fig. 3.4):

$$r = (20.4\,\mathrm{cm/(MeVc)} \cdot p^3/m^2).$$

We use this approximation instead of the MC truth information, because external users of the response matrix (i.e. non-T2K-members), must be able to generate all necessary information to fill the truth bins with their predictions. The particle momentum and mass are usually available, while more involved calculations of ranges in different kinds of matter might not be. This leads to a range corrected separation of

$$s = \max(\theta, \arctan(5\,\mathrm{cm}/(20.4\,\mathrm{cm/(MeVc)} \cdot p^3/m^2)).$$

In OOFV events, the true separation of the main track at the true vertex outside the TPC is not of interest to the reconstructed vertex in the TPC. We thus change the

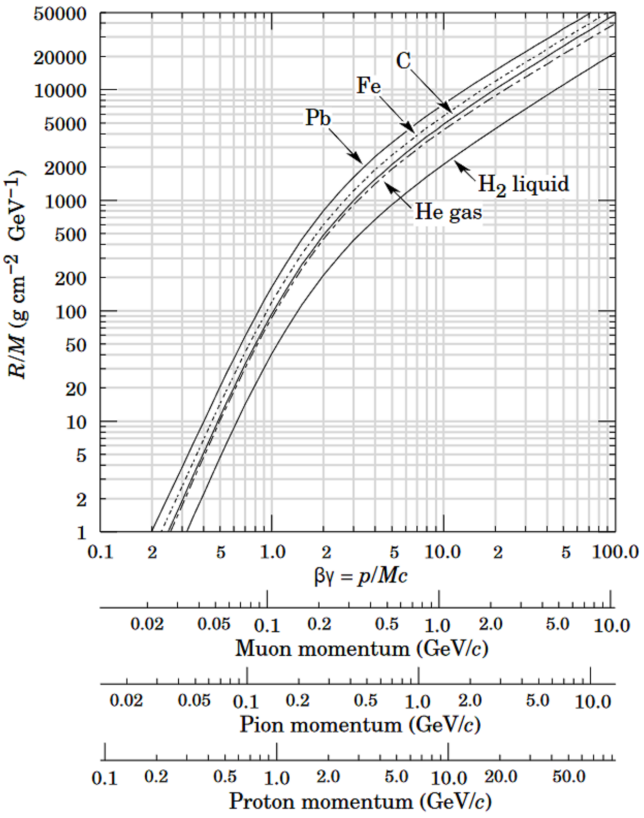


Figure 3.4.: Estimation of particle range from its mass and momentum [24]. In the low-momentum region ($p/m < 1$), x/m is approximately proportional to $(p/m)^3$, so we can estimate $x = k \cdot (p^3/m^2)$. The value of the constant k can be evaluated from simulations of protons in argon gas.

meaning of the separation variables in these cases. Their forward separation is defined as the angle between the main track and the closest other particle at the position where the true main particle first enters the TPC. This enables the response matrix to treat collinearly entering particles different from isolated particles. If there is no other particle entering the TPC within a 5 cm radius, the forward separation of that OOFV event is assigned a negative default value. This allows the backward separation to encode the information of whether an event is coincident OOFV background or "other" (see Section 3.2.2.5).

The backward separation of OOFV events is just the *reconstructed* backward separation of the reconstructed vertex. Like the OOFV categories, it is based on reconstructed information and offers an additional hint as to how the event's reconstruction failed. If an event is reconstructed with a 1-track vertex, i.e. there are no other particles except for the main one, the backward separation is assigned a negative default value. The OOFV separations do not use any range corrections.

3.2.2.8. Particle multiplicities

The track multiplicity directly affects the performance of the detector, since the vertex finding algorithm works very differently between one-, two-, and multi-track events. This information can actually be encoded in the separation and does not need its own variable. In one-track events, neither forward nor backward separation are defined and can be set to a negative default value. In two-track events, the secondary particle is either oriented in forward or in backward direction with respect to the main track. The respective separation is defined, while the other is set to the negative default value. In three-track events we can fill both the forward and the backward separation with values > 0. If all secondary tracks are oriented in the same direction, the other separation will get a value $> \pi/2$. This multiplicity encoding is summarised in Fig. 3.7. We employ this round-about encoding of the multiplicity to reduce the number truth variables we have to bin in. The remaining difference in reconstruction performance between the different multiplicities above three is much smaller than the variation within the angular separation (see Fig. 3.5 and Fig. 3.6).

The purity of the selected events is very different between 1-track and n-track events, giving the multiplicity discrimination power to distinguish between signal and background. The expected number of real data events is too low and the multiplicity systematics have not been tested sufficiently enough to do a finer binning in the n-track category. There is one bin for 1-track events and one bin for everything else.

3.2.2.9. Other variables

In the future it would be interesting to extend the number of binned variables in reco and truth space. Especially the properties of secondary particles (e.g. protons, pions) are of interest here, as they offer a window into nuclear effects. The TPCs could use their superior sensitivity to low-momentum protons to make very interesting measurements. Due to time constraints, this could not be achieved in this first analysis.

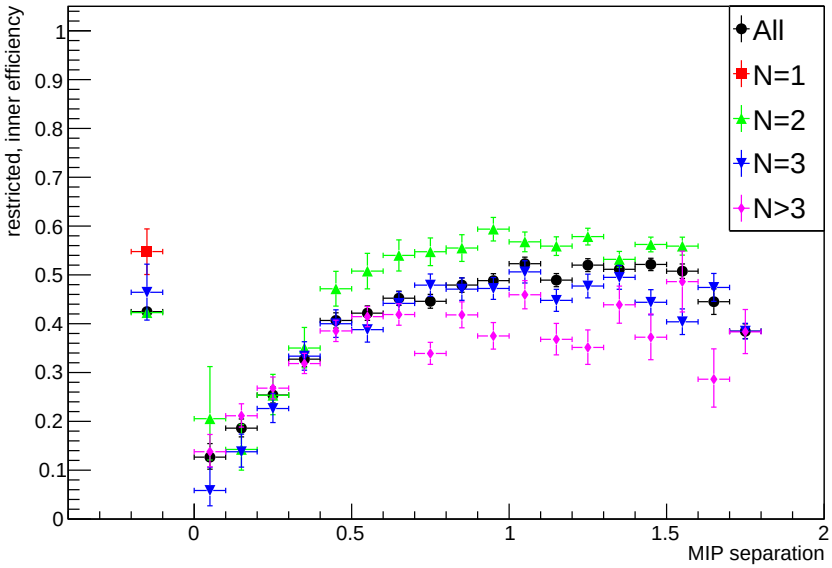


Figure 3.5.: Efficiency vs. forward particle separation for different charged-particle multiplicities.

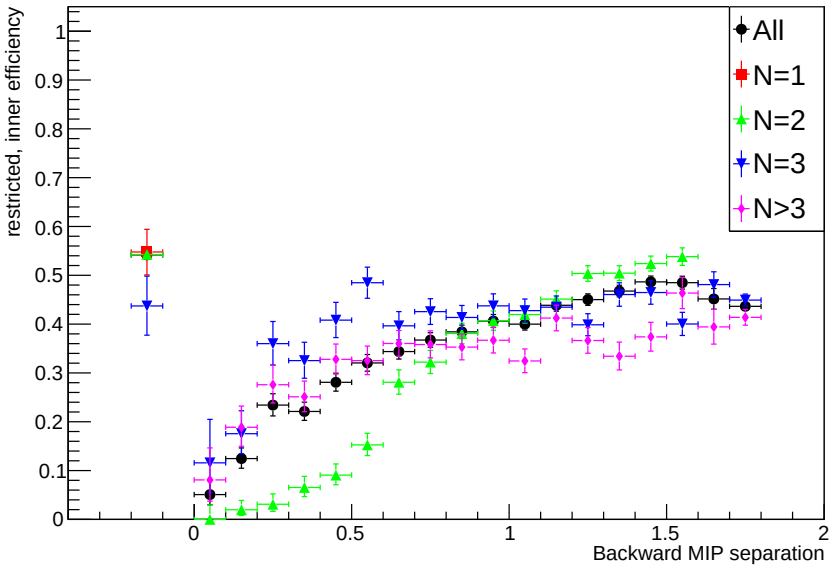


Figure 3.6.: Efficiency vs. backward particle separation for different charged-particle multiplicities.

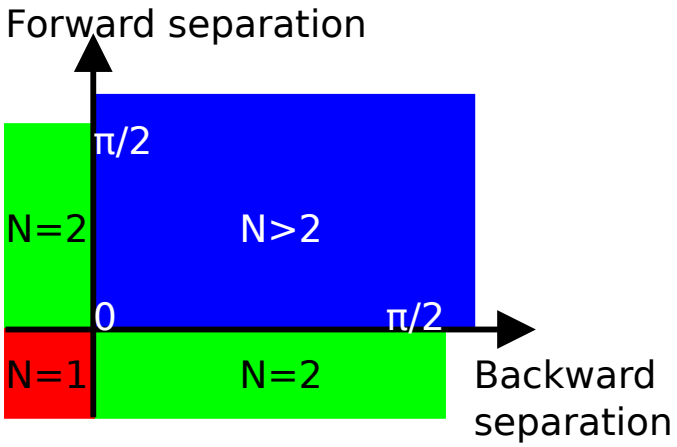


Figure 3.7.: Encoding of multiplicity in the separation.

3.2.3. Matrix tests

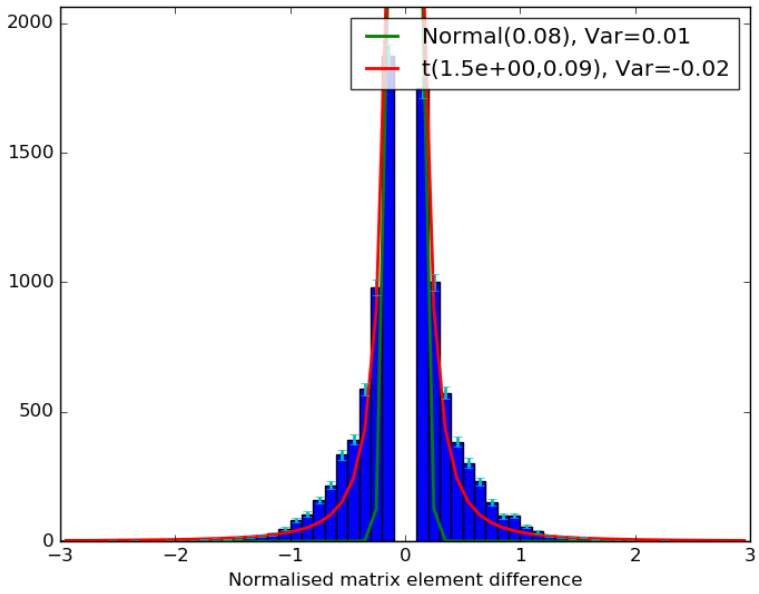
It is important to check whether the generated response matrix is actually as model-independent as assumed. The most straight-forward approach is to simply compare the matrix elements of matrices that were generated with different Monte Carlo samples, and check whether they agree within the assumed statistical uncertainties (see Section 3.2.1). Figure 3.8 shows this for matrices generated with Neut and Genie. They show a very good agreement within the statistical uncertainties. Additionally we check whether the two matrices produce compatible reco-space predictions given the same truth-space input (see Fig. 3.9). Again the differences are compatible within matrix uncertainties.

Care has to be taken when comparing matrices that were generated with very different sample sizes. The matrix construction uses certain prior assumptions or "fake data points" as starting points. When increasing the size of used data samples, this leads to the matrix elements shifting away from the prior values towards the real ones. This means the element difference between matrixes built with two samples of different size does *not* have to be cenetered at 0, even if the simulated physics of the samples is exactly the same. Any bias in element difference should still lie within the uncertainty of the elements, though.

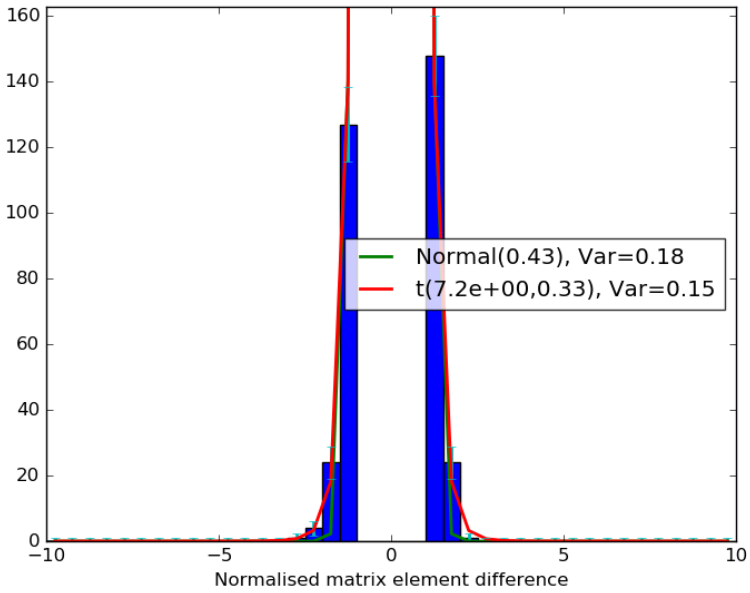
Since the matrices generated with the two generators are compatible, we use both Genie and Neut MC to generate the single nominal response matrix that is used in the analysis. The tests described above can only exclude a model-dependency to the level of the difference between the used generators, i.e. Neut and Genie, though. To check whether the matrix is also good enough to cover models that lie outside the scope of the generators, we varied the models with an arbitrary "crazy" weighting function

$$w = 0.5 + \left(\exp\left(\frac{-p}{600\,\mathrm{MeV/c}}\right) + 0.5|\cos(\theta)|\right) \cdot \frac{N_{\mathrm{charged}}}{2},$$

and compared the resulting matrix to the nominal one in the same way as done before. The results of this comparison can be seen in Fig. B.1 and B.2 in the appendix. The

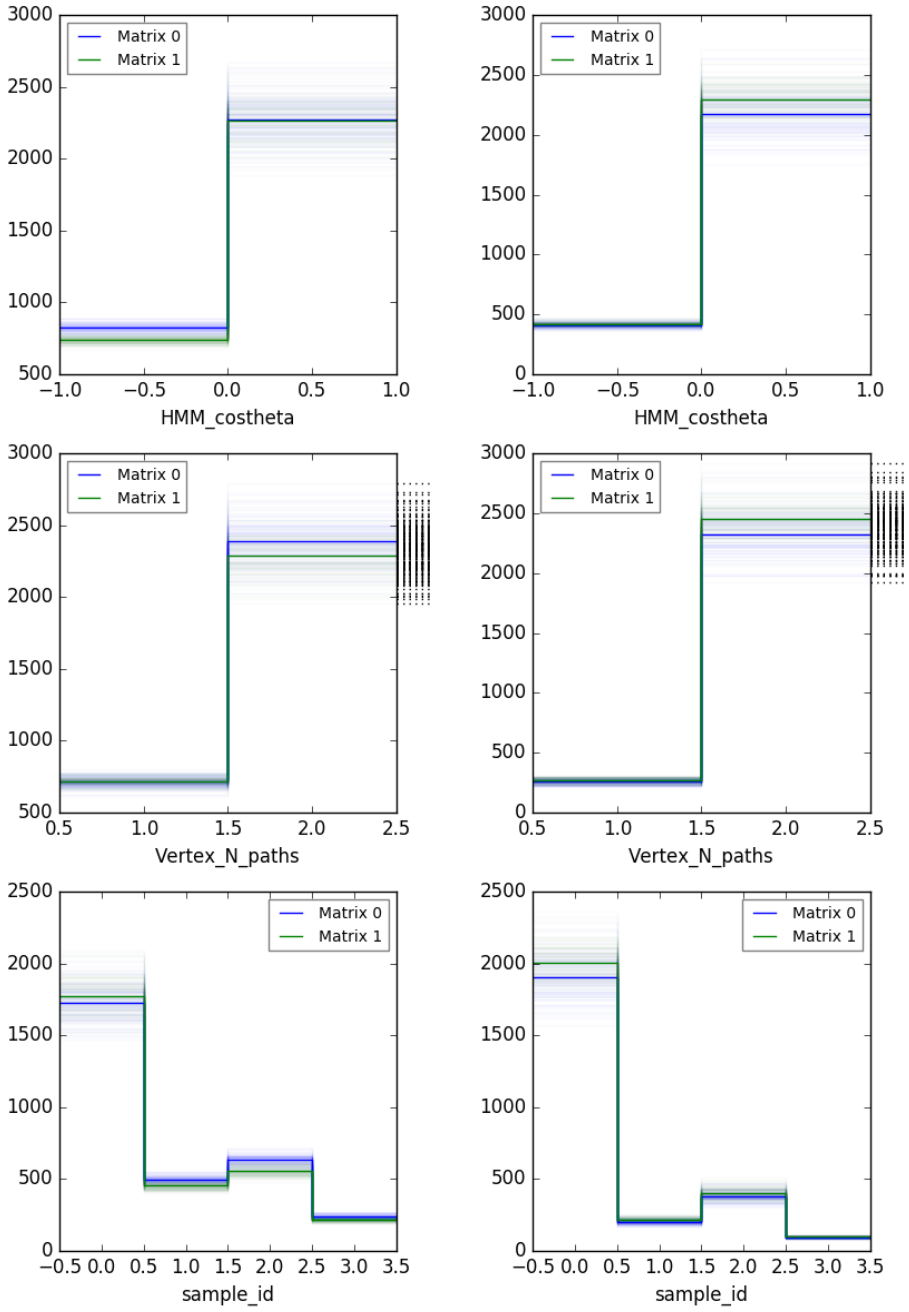


(a) Ignoring differences < 0.1



(b) Ignoring differences < 1.0

Figure 3.8.: Matrix element comparison Genie-Neut. Shown are the difference of matrix element expectation values, normalised by the statistical uncertainty (see Section 3.2.1): $(\hat{R}_{ij}^{\text{Genie}} - \hat{R}_{ij}^{\text{Neut}})/\sqrt{\sigma^2(R_{ij}^{\text{Genie}}) - \sigma^2(R_{ij}^{\text{Neut}})}$. A normal and a student's t distribution are fitted to the histogram to estimate the variance of the distribution. (a) and (b) show the same data, but with different thresholds for the minimum difference.



(a) Genie truth

(b) Neut truth

Figure 3.9.: Forward-folding test. Neut and Genie truth data are folded through the response matrices (one generated with Genie and one with Neut data). Each semi-transparent line corresponds to one toy-matrix, and the opaque lines to the average matrix. The predictions differ, but are still compatible within the uncertainties of the matrices. The dotted lines outside the axes indicate that the neighbouring bin is an overflow bin.

Table 3.3.: Software versions.

Reconstruction	ND280 software stack	Production 6H/I (v11r37)
	(including `TREx`)	v2r27p3
Selection	Highland2	v2r23
	trexAnalysis	v1r1
Response matrix	`ReMU`	0.2.0

weighted matrix is compatible with the nominal one, suggesting that it is sufficiently model-independent.

3.2.4. Software and data format

The software versions used for the reconstruction, selection and analysis are listed in Tab. 3.3. Reconstruction and selection are handled by software that is intended to be used inside the T2K collaboration. Data formats are based on the ROOT analysis framework developed at CERN [44] and also very specific to the experiment. The final output of the selection is a ROOT file with the reconstructed properties of each selected event. For Monte Carlo events, the truth information is also contained.

The ultimate goal of the response-matrix-centred approach is to enable people who are not intimately familiar with the experiment to compare neutrino event generators with the measured data. To this end, it was decided to develop the software that deals with the response matrix independently from the internal T2K software. The result of this effort is the Response Matrix Utilities framework `ReMU` [45].

`ReMU` is written in pure Python and thus able to run on any system that supports the scripting language. Numerical calculations are handled by the NumPy [46], SciPy [47], and PyMC [48] packages to take advantage of the performance gains of compiled code. `ReMU`'s source code is publicly available on the code-sharing platform GitHub [49], and releases of the software are distributed via the Python Package Index (PyPI) [50]. This means, installing the framework on systems supporting PyPI can be done with a single command:

```
pip install remu
```

Data is stored and exchanged with standard file formats. The binning of the response matrix is saved in YAML files [51], a human text format that is both human readable and easy to parse by machines. The response matrix is saved as a binary NumPy file. To save disk and RAM space, the matrix is saved as a "sparse" matrix, i.e. only bins that were filled during the matrix creation are saved. The information which bins were filled (and how many events were simulated in each) is saved in another binary NumPy file. The data itself (reco or truth space) can either be provided as binned histograms with binary NumPy files, or event-by-event with Comma-Separated Values (CSV) files. Since ROOT files are a de-facto standard for data-exchange in many particle-physics experiments, an extension to allow ROOT files as input is planned in the future.

A publication following the response-matrix centred approach would include at least these elements:

- Response matrix binning in reco space ("reco-binning.yml")
- Response matrix binning in truth space ("truth-binning.yml")
- The systematically and statistically varied sparse response matrices ("response.npy")
- A truth space histogram of how many events were simulated in each bin ("generator-truth.npy")
- Reco histogram of data ("data.npy") or CSV file of reco properties of all data events ("data.csv")
- Truth space background templates (background.npy)

Users of the publication could then provide their own signal predictions to calculate likelihoods. `ReMU` provides many functions to make this as easy as possible. This includes the definition of composite hypotheses and the likelihood maximisation over their parameter spaces. An example of a very simple analysis is given in Listing 1 in the appendix.

3.3. Template fit

3.3.1. Approach

As a first measurement to demonstrate the feasibility of the response matrix approach, we chose to do a simple template fit. For this we divide the Monte Carlo data into twelve categories, based on the reaction mode and neutrino type in the TPC fiducial volume, and the background type for out-of-fiducial-volume events:

- Charged current (CC) ν_μ events in the fiducial volume (FV)
- Neutral current (NC) events (any neutrino flavour) in the FV
- Charged current $\bar{\nu}_\mu$ events in the FV
- Charged current $\nu_e/\bar{\nu}_e$ events in the FV
- Timing related Out Of Fiducial Volume (OOFV) events (1-track)
- Timing related Out Of Fiducial Volume (OOFV) events (n-track)
- Delta-ray OOFV events (1-track)
- Delta-ray OOFV events (n-track)
- Particle decay OOFV events (1-track)
- Particle decay OOFV events (n-track)

- Other OOFV events (1-track)
- Other OOFV events (n-track)

The OOFV categories correspond to the ones defined in Section 3.2.2.5, but with the coincident particle background included in the "OOFV other" template. The FV categories do *not* correspond to the bins of the response matrix. Each of the FV templates will thus contain events from multiple event category bins. All used templates are depicted in Section B.2.

The unblinding of the control regions revealed a huge difference between data and MC predictions of the OOFV events. This was "fixed" with an "ad hoc" weight, which introduces the biggest single systematic detector uncertainty. There is a huge difference between the weights of events with multiple tracks and those with only one reconstructed track at the vertex. To give the fitter some power to vary the shape of the background distributions and the strength of this correction, the OOFV templates are split into 1-track and n-track events.

The templates for each category are normalised and then combined linearly to form the composite hypothesis that can be fitted to the data.

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \theta_0 \boldsymbol{\mu}^0 + \theta_1 \boldsymbol{\mu}^1 + \cdots,$$

where $\boldsymbol{\theta} \in \mathbb{R}_+^{12}$ is the vector of template weights and $\boldsymbol{\mu}^k$ are the normalised templates. The parameter of interest is the template weight of the FV ν_μ CC template θ_0. The other parameters θ_{1-11} are nuisance parameters that describe the composition of background events. We build a confidence interval for the parameter of interest as described in Section 3.1.8 with both a set of Neut and Genie templates.

It is worth emphasising that this very simple template fit does *not* vary any cross-section model parameters. The result will be a *model-dependent*, phase-space-restricted $\mathrm{CC_{inc}}$ cross section.

3.3.2. Final tests

Some final tests were done to confirm that the analysis method performs as intended. To check the maximum likelihood fitter, we used the full MC reco data to generate toy data-sets according to a Poisson distribution

$$\boldsymbol{n}^{\mathrm{toy}} \sim \mathrm{Poisson}(\boldsymbol{n}^{\mathrm{MC}}/f),$$

with the scaling factor $f > 100$, and the reconstructed distributions $\boldsymbol{n}^{\mathrm{toy/MC}}$ in the toy data-sets and full MC data-set respectively. The scaling factor needs to be this high, so the Poisson variation due to the original MC statistics is small compared to the variation of the toy data sets. With this condition, one knows the expectation value of true signal, i.e. FV ν_μ CC, events in the data, and can check how the fit results are distributed around that value. This was done in all combinations of generator data and generator templates. The toy data sets generated in this way are called A-sets.

Since the total number of events simulated in this test is lower than the number of real data events recorded[16], the same check is performed with toy data generated from the MC truth folded through the average response matrix:

$$\boldsymbol{n}^{\text{toy}} \sim \text{Poisson}(\overline{R} \cdot \boldsymbol{\mu}^{\text{MC}} / f^{\text{POT}}).$$

Here the scaling factor $f^{\text{POT}} \sim 20$ scales the MC statistics to the expected numbers in the real data. The truth expectation values $\boldsymbol{\mu}^{\text{MC}}$ are technically also just random samples from a Poisson distributed sample obtained by the event generator. The true expectation values of the model are only reached in the limit of an infinitively large MC sample. In this context the Poisson sampled MC expectation values are used as the definitive truth model that differs from the "real" generator model by the statistical differences. By folding them through the matrix directly, the corresponding reco expectation values are know precisely[17] and do not get influenced by the uncertainty of the MC statistics. The toy data sets generated in this way are called B-sets.

Figure 3.10 and Fig. 3.11 show the distribution of the maximum-likelihood-fit signal-template-weight values obtained from the toy data sets. There is a bias towards lower values in the A-sets. This might be caused by the fact that the expectation value in this study is close to 0 and the overall event numbers are low. In the B-sets (i.e. folded samples with realistic event numbers), the fit results are distributed around the expectation value. There are some slight biases when fitting the Neut data with the Genie templates and vice versa. This is no problem, as the results of the fits are expected to be model-dependent.

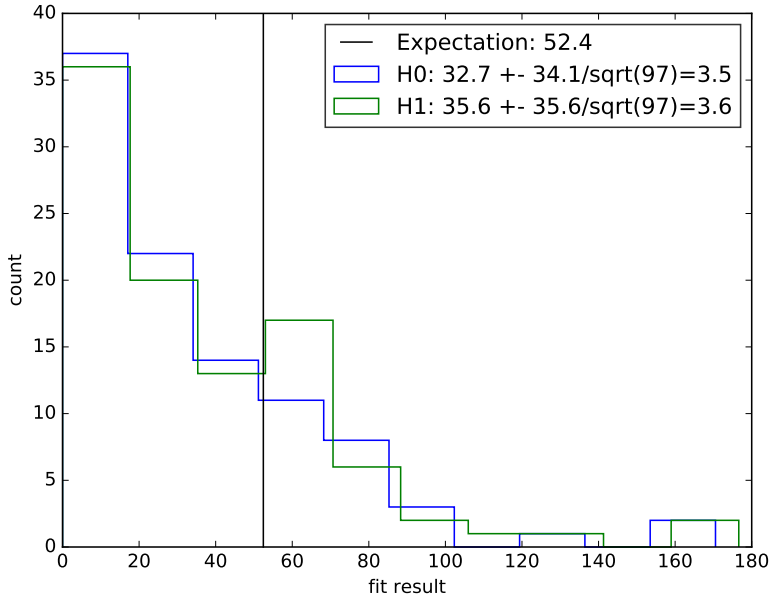
Another important benchmark is the distribution of the plug-in p-value (as described in Section 3.1.8) for the known, true expectation values in the toy data sets. It should be uniformly distributed. The cumulative distributions for the A-sets are shown in Fig. 3.12. The low p-values appear more often than they should. Using them directly as exclusion limits would thus lead to under-coverage of the confidence interval. This might be a consequence of the low signal expectation value, just like the bias in the maximum likelihood fit results. The profile plug-in p-value calculation cannot be employed here without additional corrections. The distributions for the B-sets are shown in Fig. 3.13. They show the expected behaviour, proving the profile plug-in p-value to be suitable for the amount of data analysed here.

3.4. Systematic uncertainties

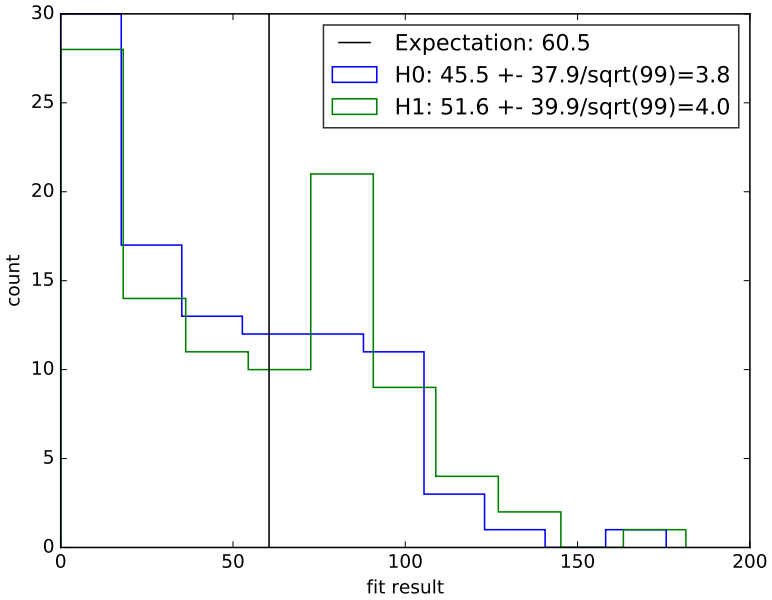
The systematic uncertainties of the selection are summarised in Tab. 2.15 The biggest uncertainties stem from the "ad hoc" and flux weights. These numbers refer to the total influence on the main selection. Their effect on the matrix construction (see Section 3.2.1) and template fit (see Section 3.3) will be different.

For example, the flux uncertainty affects the reconstructed events in the same way as the true events, so its influence on the matrix construction should be much smaller[18]

[16] Our MC sample is only a factor ~ 20 larger than the real data, not the required ~ 100.
[17] At least for a single given response matrix.
[18] Assuming a perfectly model-independent matrix, the systematic should have no effect at all. Since

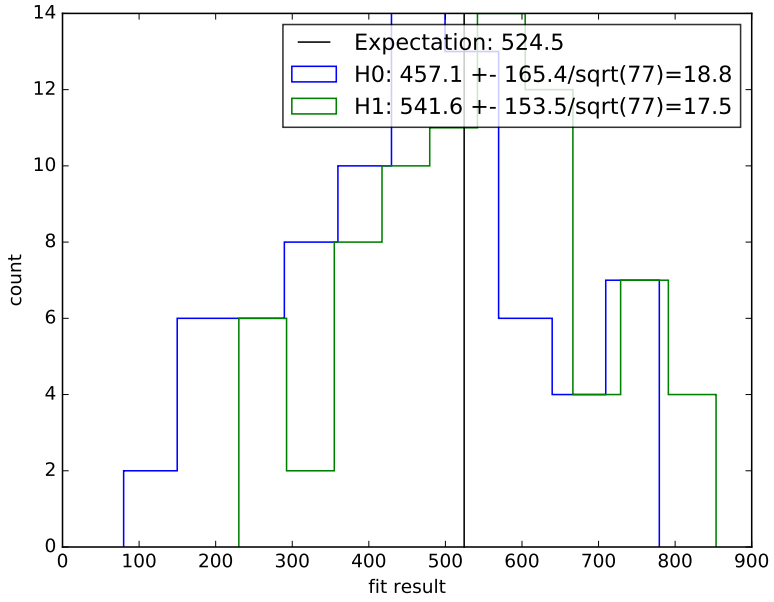


(a) Genie toy data

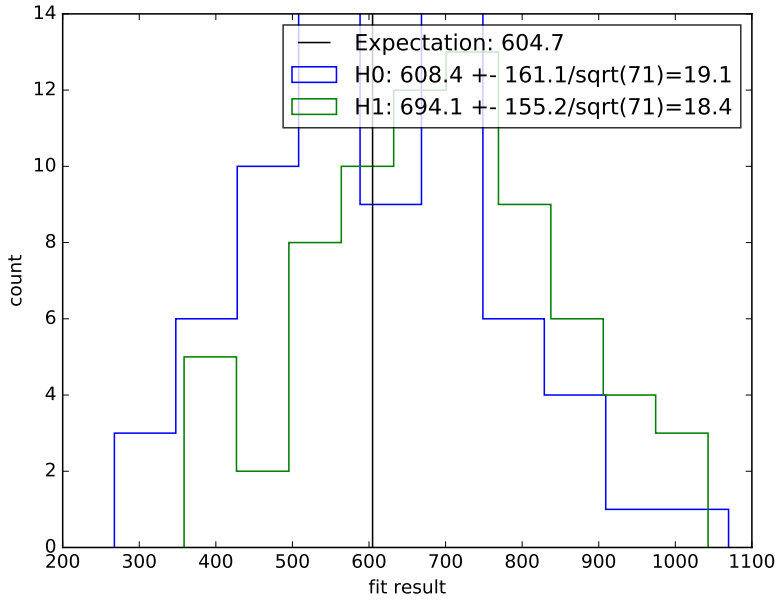


(b) Neut toy data

Figure 3.10.: Test of maximum likelihood estimate in the A-sets of toy data. The maximum likelihood estimates for the expectation value of signal events were calculated for 98 (100) randomly generated Genie (Neut) toy data sets. This was done for both Neut and Genie toy data and both Neut (H0) and Genie (H1) templates. The arithmetic mean and error on the mean are displayed in the legend. The true expectation value is shown in black.

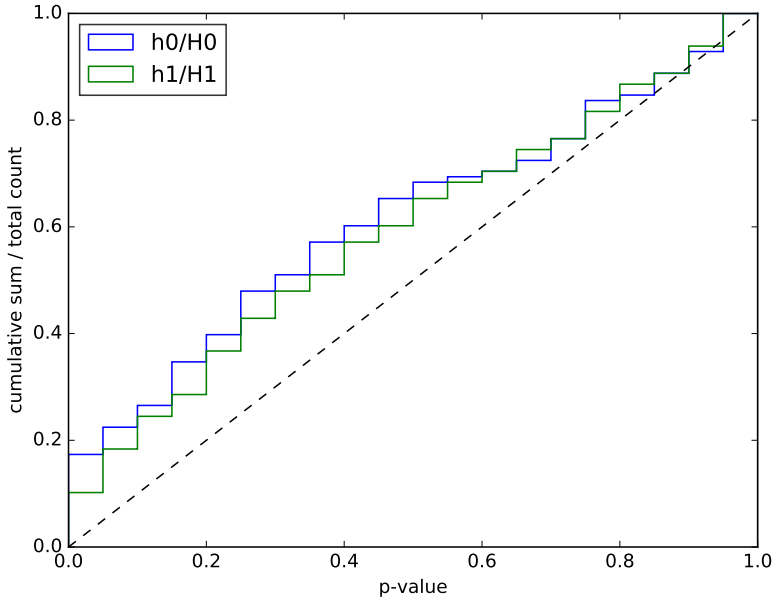


(a) Folded Genie toy data

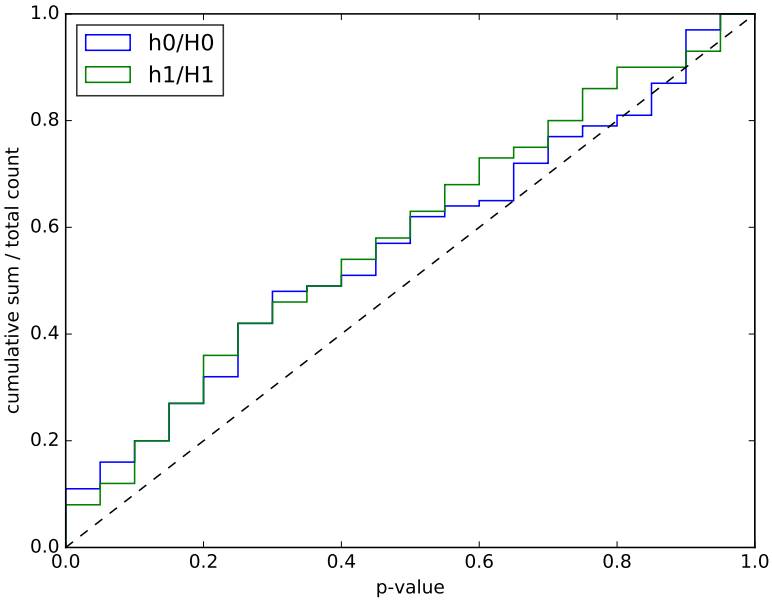


(b) Folded Neut toy data

Figure 3.11.: Test of maximum likelihood estimate in the (folded) B-sets of toy data. The maximum likelihood estimates for the expectation value of signal events were calculated for 78 (72) randomly generated Genie (Neut) toy data sets. This was done for both Neut and Genie toy data and both Neut (H0) and Genie (H1) templates. The arithmetic mean and error on the mean are displayed in the legend. The true expectation value is shown in black.

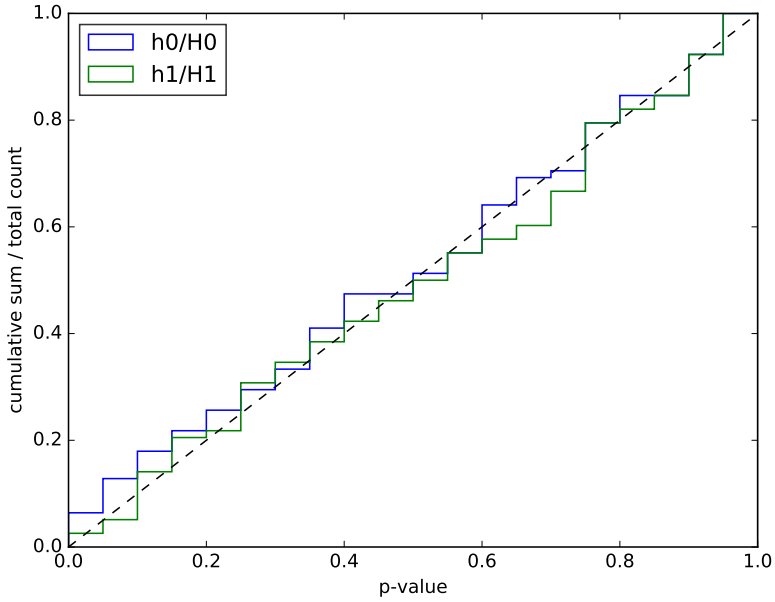


(a) Genie toy data

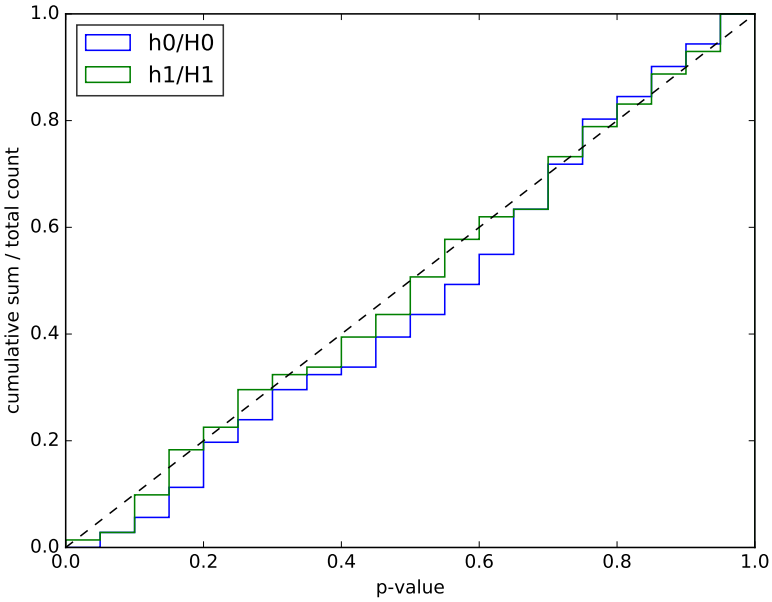


(b) Neut toy data

Figure 3.12.: Test of p-value calculation in the A-sets of toy data. The p-values of the true signal expectation value were calculated for 98 (100) randomly generated Genie (Neut) toy data sets. This was done for both Neut and Genie toy data and both Neut (H0) and Genie (H1) templates. The basis of the p-value calculation is the likelihood ratio of free template hypotheses (H0/1 with 12 free parameters) and the hypotheses with fixed signal weight (h0/1).



(a) Genie toy data



(b) Neut toy data

Figure 3.13.: Test of p-value calculation in the (folded) B-sets of toy data. The p-values of the true signal expectation value were calculated for 78 (72) randomly generated Genie (Neut) toy data sets. This was done for both Neut and Genie toy data and both Neut (H0) and Genie (H1) templates. The basis of the p-value calculation is the likelihood ratio of free template hypotheses (H0/1 with 12 free parameters) and the hypotheses with fixed signal weight (h0/1).

than the worst case 10%. For the template fit, only the shape of the OOFV distributions matters, since the integral of each template is normalised to 1 anyway. The systematics' effect on the overall quantity of OOFV background in the selection, i.e. the normalisation, has no influence.

The final influence of the uncertainties depends on the details of the tested models and measurement method (likelihood template fit, MCMC, etc.). They will naturally be handled by the methods described in Section 3.1, leading to the construction of confidence/credible intervals that reflect our knowledge of the detector correctly. It is difficult to disentangle the single influences in these methods though, so one should use the numbers given here as an indication which improvements of statistical uncertainty would benefit the gas interaction analysis the most.

3.5. Results

A detailed description of the response matrix used for the analysis can be found in Section B.1. The calculation of the confidence interval with the real data was performed as described in Section 3.3 with both the nominal Neut and Genie model templates. Additionally, the fit was done with a "crazy" template that was created by using both the Neut and Genie MC events, and weighting them with the "crazy" weight as described in Section 3.2.3. The result of the fits can be seen in Fig. 3.14 and Fig. 3.15. All template weights are well within the acceptable range for the response matrix to be valid (see Section 3.2.1). The resulting expectation values for the number of true charged-current events in the real data are (maximum likelihood value and 1σ confidence interval):

$$\theta_{\text{Neut}} = 271\,^{+98}_{-83},$$

$$\theta_{\text{Genie}} = 653\,^{+119}_{-105},$$

$$\theta_{\text{Crazy}} = 305\,^{+88}_{-96}.$$

These have to be converted to a cross section that is comparable to other data.

For this, the values have to be divided by the integrated neutrino flux Φ and the number of targets N:

$$\sigma = \frac{\theta}{N\Phi}.$$

The total flux of the analysed data is shown in Fig. 3.16 and yields:

$$\Phi = 11.07 \times 10^{12}\,\text{cm}^{-2}.$$

The target mass m is simply the fiducial volume V multiplied with the average[19] density of the drift gas ρ:

$$m = \rho V.$$

any real matrix will have some model dependence left, we include these systematics when building the matrix anyway.

[19] The gas density in the TPCs changes over time due to atmospheric pressure and temperature changes.

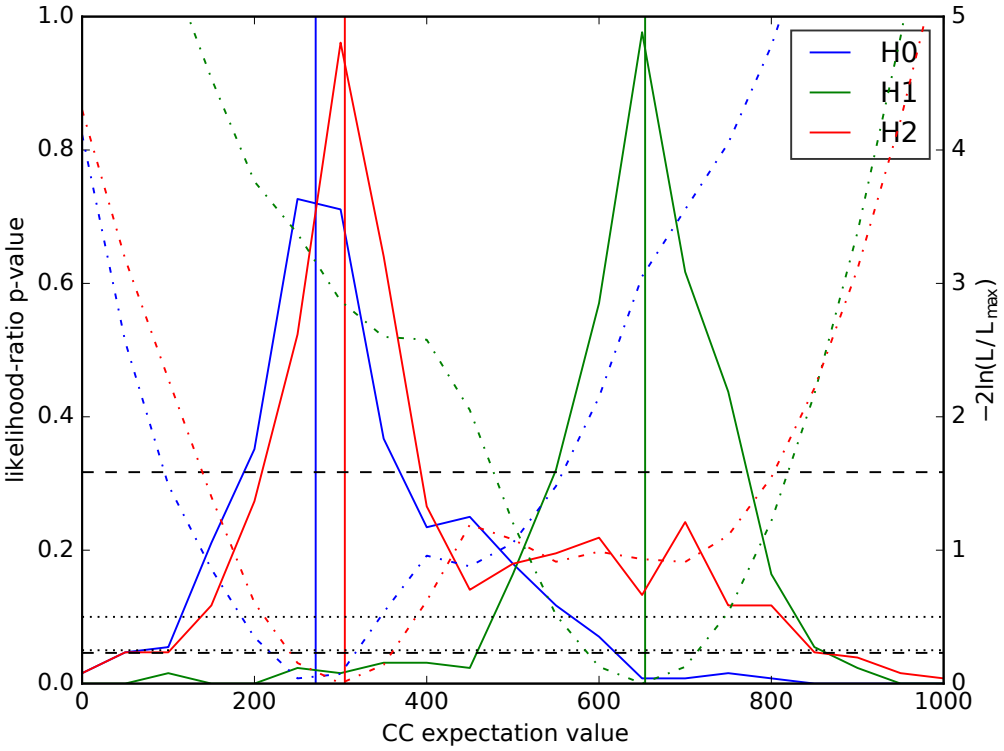


Figure 3.14.: Template fit results. Likelihood ratio p-values and profile likelihood ratios for the three performed template fits. H0: Neut templates, H1: Genie templates, H2: "Crazy" templates. Confidence intervals can be constructed from these values by including areas over the threshold p-value. The thresholds for 1σ, 2σ, 90%, and 95% confidence levels are shown. The maximum likelihood values are shown as vertical lines. The fits are done independently and the results are only valid within the confines of the template models.

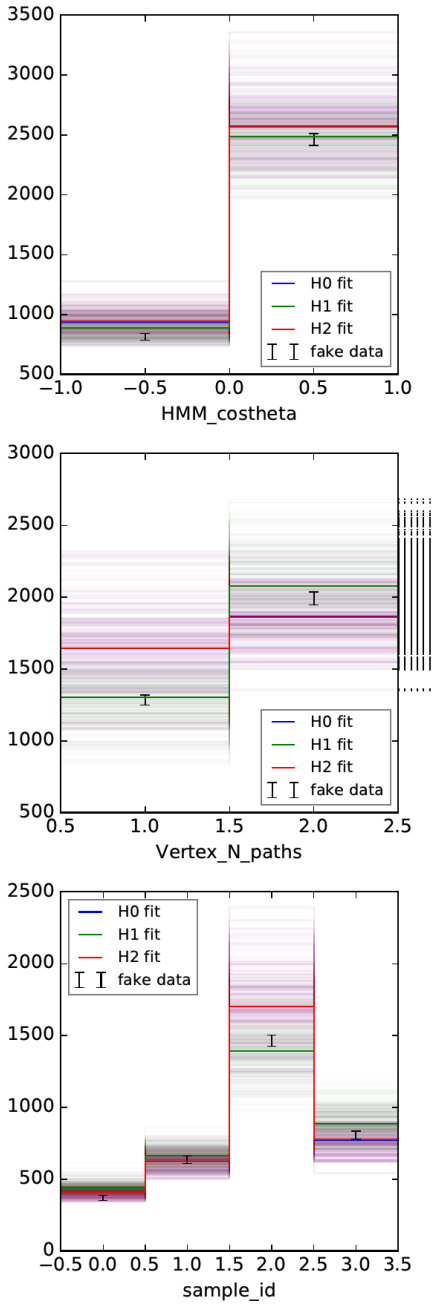


Figure 3.15.: Reco space fit results. Each plot is a projection of the data onto that variable. The maximum likelihood estimates of the three hypotheses (H0: Neut templates, H1: Genie templates, H2: "Crazy" templates) are folded through the response matrix. Each semi-transparent line is the folding result of one toy matrix. The solid lines show the result of the average matrix. The data is shown as points with $\sqrt{N}$ error bands. The dotted lines outside the axes indicate that the neighbouring bin is an overflow bin.

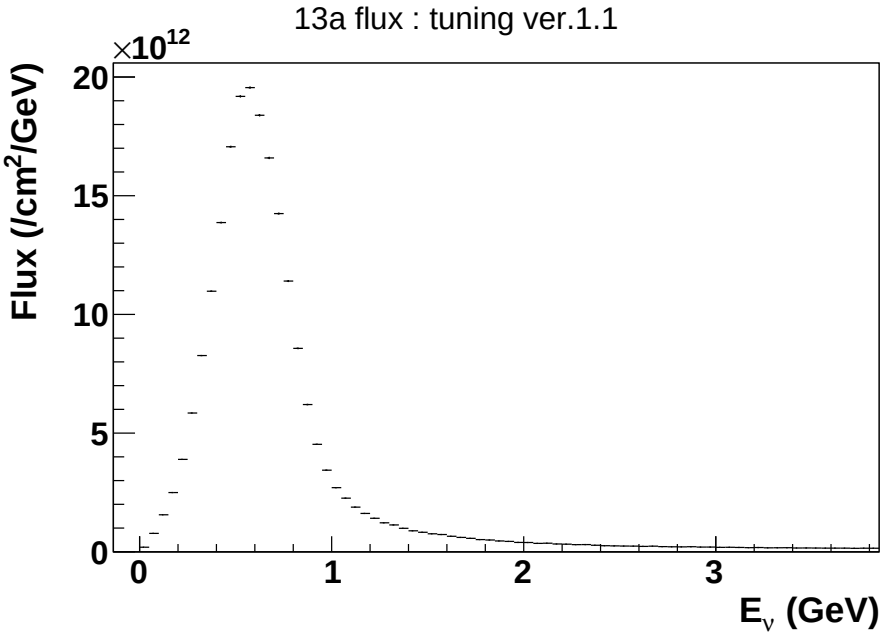


Figure 3.16.: Muon neutrino flux corresponding to the 5.73×10^{20} POT of the analysed data. The total flux adds up to $\Phi = 11.07 \times 10^{12}\,\mathrm{cm}^{-2}$. The average neutrino energy is $\overline{E}_\nu = 0.842\,\mathrm{GeV}$ and the central 68% of neutrinos lie in the interval between 0.405 GeV and 0.978 GeV. The peak energy is at 600 MeV.

The total fiducial volume of the TPCs is $V = 6.534\,\mathrm{m}^3$. The mixtures' density at NTP ($T = 293.15\,\mathrm{K}$, $p = 1013\,\mathrm{mbar}$) is $\rho_{\mathrm{NTP}} = 1.738\,\mathrm{kg/m}^3$, and the average ($T = 290.65\,\mathrm{K}$, $p = 1010\,\mathrm{mbar}$) density $\rho = 1.748\,\mathrm{kg/m}^3$. So the target mass is:

$$m = 11.4\,\mathrm{kg}.$$

The number of targets is not well defined since the target material is a mixture of different gases. The most robust definition would be by amount of substance, i.e. in mole:

$$N_{\mathrm{mol}} = m/M,$$

with the molar mass $M = x_{\mathrm{Ar}}M_{\mathrm{Ar}} + x_{\mathrm{CF4}}M_{\mathrm{CF4}} + x_{\mathrm{iC4H10}}M_{\mathrm{iC4H10}} = 41.77\,\mathrm{g/mol}$. The molar fractions of the constituent gases are $x_{\mathrm{Ar}} = 0.949$, $x_{\mathrm{CF4}} = 0.030$, and $x_{\mathrm{iC4H10}} = 0.021$. They differ slightly from the volume fractions due to especially isobutane being far from an ideal gas. The molar masses of the constituents are $M_{\mathrm{Ar}} = 39.95\,\mathrm{g/mol}$, $M_{\mathrm{CF4}} = 88.00\,\mathrm{g/mol}$, and $M_{\mathrm{iC4H10}} = 58.12\,\mathrm{g/mol}$. So the amount of target substance is:

$$N_{\mathrm{mol}} = 273\,\mathrm{mol}.$$

Charged-current cross sections are often quoted per nucleon or per neutron. The number of nucleons N_N can be calculated from the total target mass and the unified atomic mass u:

$$N_N = m/u = 6.87 \times 10^{27}.$$

The number of target protons N_p is:

$$N_p = N_{\mathrm{mol}}N_A(18x_{\mathrm{Ar}} + (6 + 4 \cdot 9)x_{\mathrm{CF4}} + (4 \cdot 6 + 10 \cdot 1)x_{\mathrm{iC4H10}}) = 3.13 \times 10^{27},$$

with the Avogadro constant N_A. So the number of target neutrons N_n is:

$$N_n = N_N - N_p = 3.74 \times 10^{27}.$$

As can be seen in Chapter 2, the selection is not particularly sensitive to momenta below $200\,\mathrm{MeV/c}$. Claiming to measure these would thus be bad practice. Instead, we decide to ignore the events below this threshold in our quoted cross section. The ratio of events above the threshold in each of the signal templates is

$$\epsilon = \sum_{p>200\,\mathrm{MeV/c}} \mu_j^0,$$

$$\epsilon_{\mathrm{Neut}} = 0.907,$$

$$\epsilon_{\mathrm{Genie}} = 0.908,$$

$$\epsilon_{\mathrm{Crazy}} = 0.873,$$

and can directly be applied to the measured cross section:

$$\sigma' = \epsilon\sigma = \frac{\epsilon\theta}{N\Phi}.$$

This means the final cross section result is a *model-dependent*, inclusive measurement with *restricted phase space*. The results presented here do *not* vary the model parameters in any way. Model uncertainties must be handled in separate analyses, e.g. by varying the templates according to the uncertainties and evaluating the influence on the cross-section results. These analyses can use the response matrix used in this analysis as-is, without having to re-evaluate the detector systematics for the different models.

To calculate an uncertainty on the cross section within the constraints of the model templates, one can apply normal error propagation:

$$\sigma^2(\sigma') = \left(\frac{\theta}{N\Phi}\right)^2 \sigma^2(\epsilon) + \left(\frac{\epsilon}{N\Phi}\right)^2 \sigma^2(\theta) + \left(\frac{\epsilon\theta}{N\Phi^2}\right)^2 \sigma^2(\Phi) + \left(\frac{\epsilon\theta}{N^2\Phi}\right)^2 \sigma^2(N).$$

The error on ϵ can be calculated from the MC statistics, the errors on m and Φ are given by the ND280 systematics, and the error on θ is evaluated from the 68 % confidence interval, using the larger of the asymmetric errors. A summary of the results and all error sources is given in Tab. 3.4.

The final uncertainties on the cross sections are in the order of 21% to 37%. Considering only the main selection with ~ 400 events in total and ~ 100 signal events, the purely statistical error would be $\sqrt{400}/100 = 20\%$. Detector systematics are in the order of 17%[20] and the flux uncertainty of 10%. These two increase this idealised error to $\sim 28\%$. This assumes a perfect knowledge of the number of background events. The fact that this corresponds (roughly) to the final uncertainty shows that the background constriction using the control regions is working very well.

The maximum likelihood estimates of the Genie model differs from the other two models by about 2σ. Both the Neut and Genie fit show dips in their likelihood ratio towards the other's maximum likelihood value (see Fig. 3.14). The "crazy" model even has a local minimum close to Genies global minimum. Its 95% confidence interval covers the results of both the Neut and the Genie fit. Overall, the results of all three models show compatible results. It should be noted though, that the results of one model are *not* able to exclude other models. The results are only valid within the tested model space, i.e. assuming that the model is correct and that the parameter space is wide enough to cover reality. The maximum likelihoods of the three template fits are comparable. Only the Neut templates seem to yield a slightly better fit to the data than the other two models.

This is the first measurement of a neutrino cross section on (mostly) argon at that neutrino energy. This makes comparisons with existing data difficult. The closest comparable measurement comes from the ArgoNeuT collaboration [52]. They measured the inclusive charged-current cross section per nucleon on argon at an average neutrino energy of 9.5 GeV to be

$$\frac{\sigma_{\text{ArgoNeuT}}}{E_\nu} = (0.66 \pm_{\text{stat}} 0.03 \pm_{\text{syst}} 0.08) \times 10^{-38}\,\text{cm}^2/\text{GeV}.$$

[20] The effect of the systematics on the template fit results are lower than on the absolute numbers of the selection. The fitter only uses the shape of the templates and can even vary the shape of the OOFV background to a certain degree.

Table 3.4.: Likelihood fit results.

General	
Target mass	$m = (11.4 \pm 0.3)\,\mathrm{kg}$
Amount of substance	$N_{\mathrm{mol}} = (203 \pm 5)\,\mathrm{mol}$
Molar composition	Ar: 0.95, CF4: 0.03, iC4H10: 0.02
Number of nucleons	$N_N = (6.87 \pm 0.18) \times 10^{27}$
Number of neutrons	$N_n = (3.74 \pm 0.10) \times 10^{27}$
Integrated muon neutrino flux	$\Phi = (11.1 \pm 1.1) \times 10^{12}\,\mathrm{cm}^{-2}$
Mean muon neutrino energy	$\overline{E}_\nu = 0.842\,\mathrm{GeV}$
Central 68% neutrino energy quantile	$(0.405, 0.978)\,\mathrm{GeV}$
Neut template fit	
Maximum log-likelihood	-74.72
$\mathrm{CC_{inc}}$ template weight	$\theta = 271 \pm 98$
Sensitive fraction ($p_{\mathrm{MIP,true}} > 200\,\mathrm{MeV/c}$)	$\epsilon = 0.907 \pm 0.009$
$\mathrm{CC_{inc}}$ cross section by mass	$\sigma_{\mathrm{m}} = (2.14 \pm 0.80) \times 10^{-12}\,\mathrm{cm^2 kg^{-1}}$
Sensitive $\mathrm{CC_{inc}}$ cross section by mass	$\sigma'_{\mathrm{m}} = (1.94 \pm 0.66) \times 10^{-12}\,\mathrm{cm^2 kg^{-1}}$
Molar $\mathrm{CC_{inc}}$ cross section	$\sigma_{\mathrm{mol}} = (0.120 \pm 0.045) \times 10^{-12}\,\mathrm{cm^2 mol^{-1}}$
Sensitive molar $\mathrm{CC_{inc}}$ cross section	$\sigma'_{\mathrm{mol}} = (0.109 \pm 0.037) \times 10^{-12}\,\mathrm{cm^2 mol^{-1}}$
$\mathrm{CC_{inc}}$ cross section per nucleon	$\sigma_{\mathrm{N}} = (3.56 \pm 1.33) \times 10^{-39}\,\mathrm{cm^2/nucleon}$
Sensitive $\mathrm{CC_{inc}}$ cross section per nucleon	$\sigma'_{\mathrm{N}} = (3.22 \pm 1.11) \times 10^{-39}\,\mathrm{cm^2/nucleon}$
Genie template fit	
Maximum log-likelihood	-76.04
$\mathrm{CC_{inc}}$ template weight	$\theta = 654 \pm 119$
Sensitive fraction ($p_{\mathrm{MIP,true}} > 200\,\mathrm{MeV/c}$)	$\epsilon = 0.908 \pm 0.009$
$\mathrm{CC_{inc}}$ cross section by mass	$\sigma_{\mathrm{m}} = (5.16 \pm 1.08) \times 10^{-12}\,\mathrm{cm^2 kg^{-1}}$
Sensitive $\mathrm{CC_{inc}}$ cross section by mass	$\sigma'_{\mathrm{m}} = (4.69 \pm 0.91) \times 10^{-12}\,\mathrm{cm^2 kg^{-1}}$
Molar $\mathrm{CC_{inc}}$ cross section	$\sigma_{\mathrm{mol}} = (0.290 \pm 0.060) \times 10^{-12}\,\mathrm{cm^2 mol^{-1}}$
Sensitive molar $\mathrm{CC_{inc}}$ cross section	$\sigma'_{\mathrm{mol}} = (0.263 \pm 0.051) \times 10^{-12}\,\mathrm{cm^2 mol^{-1}}$
$\mathrm{CC_{inc}}$ cross section per nucleon	$\sigma_{\mathrm{N}} = (8.57 \pm 1.79) \times 10^{-39}\,\mathrm{cm^2/nucleon}$
Sensitive $\mathrm{CC_{inc}}$ cross section per nucleon	$\sigma'_{\mathrm{N}} = (7.78 \pm 1.51) \times 10^{-39}\,\mathrm{cm^2/nucleon}$
Crazy template fit	
Maximum log-likelihood	-76.10
$\mathrm{CC_{inc}}$ template weight	$\theta = 305 \pm 96$
Sensitive fraction ($p_{\mathrm{MIP,true}} > 200\,\mathrm{MeV/c}$)	$\epsilon = 0.873 \pm 0.009$
$\mathrm{CC_{inc}}$ cross section by mass	$\sigma_{\mathrm{m}} = (2.41 \pm 0.80) \times 10^{-12}\,\mathrm{cm^2 kg^{-1}}$
Sensitive $\mathrm{CC_{inc}}$ cross section by mass	$\sigma'_{\mathrm{m}} = (2.10 \pm 0.62) \times 10^{-12}\,\mathrm{cm^2 kg^{-1}}$
Molar $\mathrm{CC_{inc}}$ cross section	$\sigma_{\mathrm{mol}} = (0.135 \pm 0.045) \times 10^{-12}\,\mathrm{cm^2 mol^{-1}}$
Sensitive molar $\mathrm{CC_{inc}}$ cross section	$\sigma'_{\mathrm{mol}} = (0.118 \pm 0.035) \times 10^{-12}\,\mathrm{cm^2 mol^{-1}}$
$\mathrm{CC_{inc}}$ cross section per nucleon	$\sigma_{\mathrm{N}} = (4.00 \pm 1.33) \times 10^{-39}\,\mathrm{cm^2/nucleon}$
Sensitive $\mathrm{CC_{inc}}$ cross section per nucleon	$\sigma'_{\mathrm{N}} = (3.49 \pm 1.03) \times 10^{-39}\,\mathrm{cm^2/nucleon}$

As the weak interaction cross section should be approximately proportional to the neutrino energy, one can compare these numbers to our measurements. The mean neutrino energy in the T2K beam is 0.842 GeV. Using the not-phase-space-restricted nucleon cross sections from Tab. 3.4, the corresponding values for the three fits are:

$$\frac{\sigma_{\text{Neut fit}}}{E_\nu} = (0.42 \pm 0.16) \times 10^{-38}\,\text{cm}^2/\text{GeV},$$

$$\frac{\sigma_{\text{Genie fit}}}{E_\nu} = (1.02 \pm 0.21) \times 10^{-38}\,\text{cm}^2/\text{GeV},$$

$$\frac{\sigma_{\text{Crazy fit}}}{E_\nu} = (0.48 \pm 0.16) \times 10^{-38}\,\text{cm}^2/\text{GeV}.$$

The values are completely compatible within the uncertainties, despite the naive extrapolation to lower neutrino energies. At this level of uncertainty it is also not necessary to include an "argon-to-isoscalar" correction[21] as has been done for the ArgoNeuT result in the order of 3%.

This simple comparison of energy-scaled cross sections ignores the considerable spread of the neutrino energies in both beams. More stringent treatments need to take the actual beam fluxes into account (see Fig. 3.16). The "most correct" results of this analysis, i.e. the ones involving the least assumptions, are the sensitive molar and mass cross sections: σ'_{mol} and σ'_{m} (see Tab. 3.4). They describe the sensitive absolute cross section of muon neutrinos in the T2K beam on the T2K gas mixture, assuming the fitted models. The cross section per nucleon σ'_N implies that the cross section is equal for all nucleons. This is obviously not the case. The need for "target-to-isoscalar" corrections in precision measurements underlines this.

[21] Corrects for the fact that the target does not have the same number of protons and neutrons. For this measurement it would be more correct to call it a "T2K-gas-to-isoscalar" correction.

4. Conclusions

We performed a selection of neutrino interactions in the ND280 TPCs. Events where the highest-momentum, negatively-charged particle is MIP-like (i.e. a muon or a pion) can be identified with a purity of 26% and an efficiency of 32%. The overwhelming majority of these events are charged-current muon-neutrino interactions with argon, which makes up 91% of the target mass.

Virtually all background events in the selection are out-of-fiducial-volume events. We could identify five failure modes that cause the events to be reconstructed as vertices in the fiducial volume. The amount of background in the main selection is constrained by three control regions that invert different cuts of the main selection.

The selection was developed "blind", i.e. only with Monte Carlo events. The unblinding of the control regions revealed considerable differences between the real data and Monte Carlo predictions. Further investigations suggest that this is caused by a faulty detector simulation. The differences could be "fixed" by applying binned "ad hoc" weights. Since the exact nature of the simulation fault remains unknown, the uncertainty on these weights is assumed to be 50%. It is the largest source of systematic uncertainties for the selection. Unblinding the main selection showed a good agreement between expected and predicted distributions (with included ad hoc weights).

A cross-section measurement was extracted from the selection with the response-matrix-centred approach. Its focus lies on constructing a response matrix that contains all needed information about the detector performance and the related uncertainties. Once such a matrix is available, it is relatively easy to test various physics models against the data, without having to re-evaluate the detector systematics for each model. This will be useful for the NUISANCE model tuning framework[30], for example.

The actual cross-section extraction from the data was done by performing a template fit with templates of three models: the Neut and Genie event generators, as well as a combination of the two with an arbitrary weighting function added. Since the detector efficiency drops considerably below MIP momenta of 200 MeV/c, we decided to calculate a sensitive cross section, i.e. the cross section of events above that threshold. As expected, the three model template fits to the data yield slightly different results, with values ranging from $(1.94 \pm 0.64) \times 10^{-12}\,\mathrm{cm^2 kg^{-1}}$ (Neut) to $(4.69 \pm 1.02) \times 10^{-12}\,\mathrm{cm^2 kg^{-1}}$ (Genie). Considering the asymmetric likelihood ratio profiles and confidence intervals, they are compatible with one another.

Systematic and statistical uncertainties of the measurement are roughly equal. The statistical uncertainty will be reduced by including more data sets in the selection. The Runs 2-4, which were used in this work, make up only about 15% of the projected total amount of muon-neutrino data that will have been collected at the end of the T2K

experiment[1]. The most important detector systematic is the ad hoc weight. There is a clear need to understand in detail what causes the data-MC differences observed in the OOFV background, and to fix the detector simulation to represent the real data performance more closely.

This is the first time a cross-section measurement with an exhaustive treatment of the systematic uncertainties has been performed in the T2K TPCs. The total inclusive measurement is of limited use to constrain nuclear interaction models. It is an important milestone on the path to more specialised measurements that take advantage of the TPCs superior sensitivity to low momentum particles, though. Future iterations of the selection and analysis will include these particles and offer unprecedented insights into the nuclear effects of neutrino interactions with argon. Measurements of the anti-muon-neutrino cross section will also be possible.

[1] Assuming an equal amount of muon-neutrino mode and anti-muon-neutrino mode beam. A currently discussed extension of the T2K programme called "T2K Phase 2" would more than double the amount of data. It would include an upgrade of the near detector, though.

Bibliography

[1] K. Suzuki et al. "Measurement of the muon beam direction and muon flux for the T2K neutrino experiment". *Progress of Theoretical and Experimental Physics* 2015.5 (2015), p. 053C01. DOI: 10.1093/ptep/ptv054.

[2] K. Abe et al. "The T2K experiment". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 659.1 (2011), pp. 106–135. DOI: 10.1016/j.nima.2011.06.067.

[3] K. Abe et al. "T2K neutrino flux prediction". *Physical Review D* 87.1 (2013). DOI: 10.1103/physrevd.87.012001.

[4] K. Abe et al. "Observation of Electron Neutrino Appearance in a Muon Neutrino Beam". *Physical Review Letters* 112.6 (2014). DOI: 10.1103/physrevlett.112.061802.

[5] K. Abe et al. "Combined Analysis of Neutrino and Antineutrino Oscillations at T2K". *Physical Review Letters* 118.15 (2017). DOI: 10.1103/physrevlett.118.151801.

[6] K. Abe et al. "Measurements of the T2K neutrino beam properties using the INGRID on-axis near detector". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 694 (2012), pp. 211–223. DOI: 10.1016/j.nima.2012.03.023.

[7] T2K beam group. *Accumulated POT for all runs*. June 2018.

[8] L. Alvarez-Ruso et al. "NuSTEC white paper: status and challenges of neutrino–nucleus scattering". *Progress in Particle and Nuclear Physics* 100 (2018), pp. 1–68. DOI: 10.1016/j.ppnp.2018.01.006.

[9] E. J. Moniz et al. "Nuclear Fermi Momenta from Quasielastic Electron Scattering". *Physical Review Letters* 26.8 (1971), pp. 445–448. DOI: 10.1103/physrevlett.26.445.

[10] Tomasz Golan. "Modeling nuclear effects in NuWro Monte Carlo neutrino event generator". PhD thesis. PhD thesis, University of Wroclaw, 2014. URL: http://neutrino.ift.uni.wroc.pl/files/phd_tomasz_golan.pdf.

[11] L Alvarez-Ruso, Y Hayato, and J Nieves. "Progress and open questions in the physics of neutrino cross sections at intermediate energies". *New Journal of Physics* 16.7 (2014), p. 075015. DOI: 10.1088/1367-2630/16/7/075015. URL: https://doi.org/10.1088%2F1367-2630%2F16%2F7%2F075015.

[12] Omar Benhar and Adelchi Fabrocini. "Two-nucleon spectral function in infinite nuclear matter". *Physical Review C* 62.3 (2000). DOI: 10.1103/physrevc.62.034304.

[13] A. Bodek, M. E. Christy, and B. Coopersmith. "Effective spectral function for quasielastic scattering on nuclei". *The European Physical Journal C* 74.10 (2014). DOI: 10.1140/epjc/s10052-014-3091-0.

[14] R. Gran et al. "Antineutrino Charged-Current Reactions on Hydrocarbon with Low Momentum Transfer". *Physical Review Letters* 120.22 (2018). DOI: 10.1103/physrevlett.120.221805.

[15] Omar Benhar, Alessandro Lovato, and Noemi Rocco. "Contribution of two-particle–two-hole final states to the nuclear response". *Physical Review C* 92.2 (2015). DOI: 10.1103/physrevc.92.024602.

[16] Dieter Rein and Lalit M Sehgal. "Neutrino-excitation of baryon resonances and single pion production". *Annals of Physics* 133.1 (1981), pp. 79–153. DOI: 10.1016/0003-4916(81)90242-6.

[17] Ch. Berger and L. M. Sehgal. "Lepton mass effects in single pion production by neutrinos". *Physical Review D* 76.11 (2007). DOI: 10.1103/physrevd.76.113004.

[18] Krzysztof M. Graczyk and Jan T. Sobczyk. "Form factors in the quark resonance model". *Physical Review D* 77.5 (2008). DOI: 10.1103/physrevd.77.053001.

[19] D. Rein. "Angular distribution in neutrino-induced single pion production processes". *Zeitschrift für Physik C Particles and Fields* 35.1 (1987), pp. 43–64. DOI: 10.1007/bf01561054.

[20] E. Hernández, J. Nieves, and M. Valverde. "Weak pion production off the nucleon". *Physical Review D* 76.3 (2007). DOI: 10.1103/physrevd.76.033005.

[21] M. Kabirnezhad. "Single pion production in neutrino-nucleon interactions". *Physical Review D* 97.1 (2018). DOI: 10.1103/physrevd.97.013002.

[22] Yoshinari Hayato. "A neutrino interaction simulation program library NEUT". *Acta Phys. Polon.* B40 (2009), pp. 2477–2489.

[23] C. Andreopoulos et al. "The GENIE Neutrino Monte Carlo Generator". *Nucl. Instrum. Meth. A* 614 (2010), pp. 87–104. DOI: 10.1016/j.nima.2009.12.009.

[24] Particle Data Group. "Review of Particle Physics". *Chinese Physics C* 40.10 (2016), p. 100001. DOI: 10.1088/1674-1137/40/10/100001.

[25] O. Buss et al. "Transport-theoretical description of nuclear reactions". *Physics Reports* 512.1-2 (2012), pp. 1–124. DOI: 10.1016/j.physrep.2011.12.001.

[26] O. Lalakulich, K. Gallmeister, and U. Mosel. "Neutrino nucleus reactions within the GiBUU model". *Journal of Physics: Conference Series* 408 (2013), p. 012053. DOI: 10.1088/1742-6596/408/1/012053.

[27] N. Abgrall et al. "Time projection chambers for the T2K near detectors". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 637.1 (2011), pp. 25–46. DOI: 10.1016/j.nima.2011.02.036.

[28] Edward Larkin. "A cross section measurement from neutrino interactions on argon gas". PhD thesis. University of Warwick, Department of Physics, 2017. URL: http://webcat.warwick.ac.uk/record=b3066130~S15.

[29] Peter Hart, Nils Nilsson, and Bertram Raphael. "A Formal Basis for the Heuristic Determination of Minimum Cost Paths". *IEEE Transactions on Systems Science and Cybernetics* 4.2 (1968), pp. 100–107. DOI: 10.1109/tssc.1968.300136.

[30] P. Stowell et al. "NUISANCE: a neutrino cross-section generator tuning and comparison framework". *Journal of Instrumentation* 12.01 (2017), P01016–P01016. DOI: 10.1088/1748-0221/12/01/p01016.

[31] R. Acciarri et al. "Detection of back-to-back proton pairs in charged-current neutrino interactions with the ArgoNeuT detector in the NuMI low energy beam line". *Physical Review D* 90.1 (2014). DOI: 10.1103/physrevd.90.012008.

[32] James F. Ziegler, M.D. Ziegler, and J.P. Biersack. "SRIM – The stopping and range of ions in matter (2010)". *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 268.11-12 (2010), pp. 1818–1823. DOI: 10.1016/j.nimb.2010.02.091.

[33] O C Allkofer et al. "The absolute cosmic ray flux at sea level". *Journal of Physics G: Nuclear Physics* 1.6 (1975), pp. L51–L52. DOI: 10.1088/0305-4616/1/6/005.

[34] K. Abe et al. "Measurement of the ν_μ charged current quasielastic cross section on carbon with the T2K on-axis neutrino beam". *Physical Review D* 91.11 (2015). DOI: 10.1103/physrevd.91.112002.

[35] N. Abgrall et al. "Pion emission from the T2K replica target: Method, results and application". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 701 (2013), pp. 99–114. DOI: 10.1016/j.nima.2012.10.079.

[36] T2K beam group. *Flux error*.

[37] J. Neyman and E. S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 231.694-706 (1933), pp. 289–337. DOI: 10.1098/rsta.1933.0009.

[38] S. S. Wilks. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". *The Annals of Mathematical Statistics* 9.1 (1938), pp. 60–62. DOI: 10.1214/aoms/1177732360.

[39] Luv Demortier. *P Values and Nuisance Parameters*. eng. 2008. DOI: 10.5170/cern-2008-001.23.

[40] Paul Kabaila and Chris J. Lloyd. "A computable confidence upper limit from discrete data with good coverage properties". *Statistics & Probability Letters* 47.2 (2000), pp. 189–198. DOI: 10.1016/s0167-7152(99)00156-x.

[41] I. Smith and A. Ferrari. "Generalizations related to hypothesis testing with the Posterior distribution of the Likelihood Ratio". *arXiv preprint* (June 4, 2014). arXiv: 1406.1023v1 [physics.data-an].

[42] H. Jeffreys. "An Invariant Form for the Prior Probability in Estimation Problems". *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 186.1007 (1946), pp. 453–461. DOI: 10.1098/rspa.1946.0056.

[43] Daniel Fink. *A Compendium of Conjugate Priors*. Tech. rep. Environmental Statistics Group, Department of Biology, Montana State University, Jan. 1997.

[44] Rene Brun and Fons Rademakers. "ROOT — An object oriented data analysis framework". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 389.1-2 (1997), pp. 81–86. DOI: 10.1016/s0168-9002(97)00048-x.

[45] Lukas Koch. *ReMU - Response Matrix Utilities*. en. 2018. DOI: 10.5281/zenodo.1217572.

[46] Travis E. Oliphant. *Guide to NumPy*. USA: Trelgol Publishing, 2006.

[47] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001–. URL: http://www.scipy.org/.

[48] C. Fonnesbeck et al. *PyMC: Bayesian Stochastic Modelling in Python*. URL: http://github.com/pymc-devs/pymc.

[49] *GitHub*. 2018. URL: http://github.com/.

[50] *PyPI - the Python Package Index*. 2018. URL: http://pypi.python.org/pypi.

[51] *YAML: YAML Ain't Markup Language*. 2018. URL: http://yaml.org/.

[52] R. Acciarri et al. "Measurements of inclusive muon neutrino and antineutrino charged current differential cross sections on argon in the NuMI antineutrino beam". *Physical Review D* 89.11 (2014). DOI: 10.1103/physrevd.89.112003.

A. Event selection appendix

Table A.1.: Selection composition. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

		Main Selection		Passing-by CS	
		Neut	Genie	Neut	Genie
In FV	MIP	38.27%	36.50%	2.56%	2.58%
	electron	0.06%	0.12%	0.00%	0.01%
	other	0.00%	0.06%	0.00%	0.01%
	wrong sign	0.06%	0.12%	0.02%	0.41%
Out of FV	delta-ray	22.56%	21.66%	36.60%	35.52%
	timing	3.95%	4.04%	10.26%	8.62%
	decay	13.02%	15.40%	13.09%	15.53%
	coincident	13.87%	13.69%	29.45%	29.75%
	other	8.20%	8.40%	8.02%	7.57%
		Timing CS		Delta-ray CS	
		Neut	Genie	Neut	Genie
In FV	MIP	4.28%	3.57%	0.84%	0.78%
	electron	0.00%	0.01%	0.01%	0.01%
	other	0.00%	0.03%	0.00%	0.00%
	wrong sign	0.05%	0.00%	0.00%	0.00%
Out of FV	delta-ray	19.45%	19.68%	30.80%	30.25%
	timing	41.23%	39.96%	21.59%	18.52%
	decay	10.96%	11.99%	4.14%	4.99%
	coincident	16.01%	16.67%	32.94%	35.82%
	other	8.00%	8.09%	9.69%	9.62%

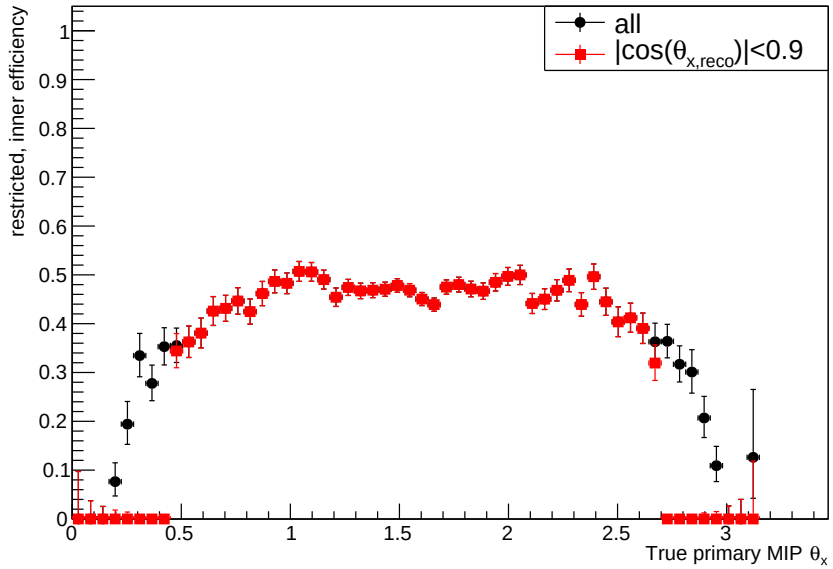


Figure A.1.: Efficiency vs drift angle θ_x in the TPC MC Neut sample. Tracks parallel to the TPC drift direction cannot be reconstructed reliably. The selection efficiency drops. The effect of the θ_x cut is shown.

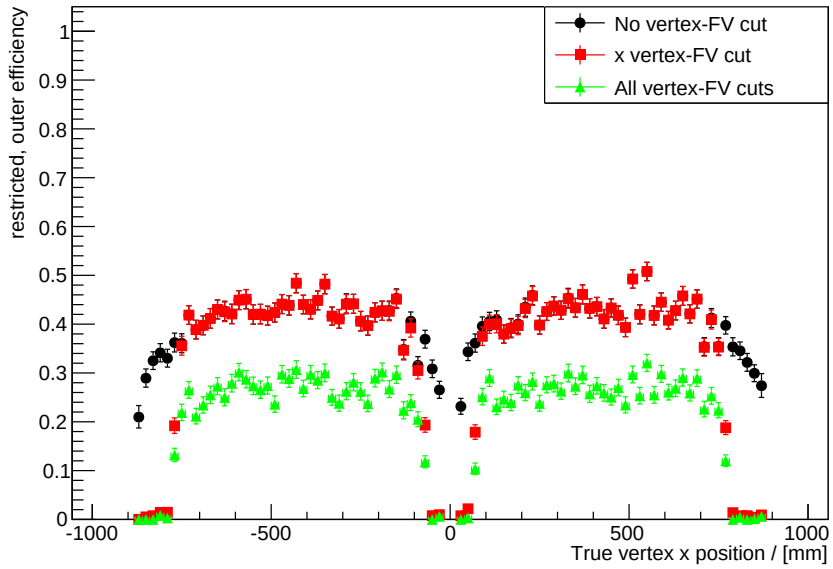


Figure A.2.: Efficiency vs vertex x-position in the TPC MC Neut sample. The effect of the vertex-FV cuts can be seen. The central cathode is located at $x = 0$.

Table A.2.: In-FV reactions in main selection.

	Neut				
	MIP	electron	other	wrong sign	all In FV
CCQE	61.01%	0.00%	-	0.00%	60.81%
2p2h	3.13%	0.00%	-	0.00%	3.13%
RES	19.89%	0.00%	-	0.00%	19.82%
DIS	11.53%	0.00%	-	0.00%	11.49%
COH	1.18%	0.00%	-	0.00%	1.18%
NC	2.64%	0.00%	-	67.16%	2.74%
CC-$\bar{\nu}_\mu$	0.50%	0.00%	-	32.84%	0.55%
CC-ν_e, CC-$\bar{\nu}_e$	0.11%	100.00%	-	0.00%	0.27%
other	0.00%	0.00%	-	0.00%	0.00%
	Genie				
	MIP	electron	other	wrong sign	all In FV
CCQE	61.16%	0.00%	0.00%	0.00%	60.65%
2p2h	0.00%	0.00%	0.00%	0.00%	0.00%
RES	25.95%	0.00%	0.00%	0.00%	25.73%
DIS	8.78%	0.00%	30.77%	0.00%	8.76%
COH	0.64%	0.00%	0.00%	0.00%	0.63%
NC	2.81%	0.00%	69.23%	48.69%	3.07%
CC-$\bar{\nu}_\mu$	0.60%	0.00%	0.00%	46.30%	0.75%
CC-ν_e, CC-$\bar{\nu}_e$	0.06%	100.00%	0.00%	5.00%	0.40%
other	0.00%	0.00%	0.00%	0.00%	0.00%

Table A.3.: Target elements. From the Neut MC. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

	Main Selection		Passing-by CS		Timing CS		Delta-ray CS	
	InFV	OOFV	InFV	OOFV	InFV	OOFV	InFV	OOFV
Argon	91.10%	0.16%	92.55%	0.09%	91.73%	0.43%	95.44%	0.03%
Fluorine	6.01%	0.00%	2.24%	0.03%	5.57%	0.00%	2.68%	0.01%
Carbon	2.72%	27.95%	5.20%	38.11%	2.32%	26.61%	1.45%	34.48%
Oxygen	0.00%	4.11%	0.00%	5.85%	0.00%	5.63%	0.00%	5.05%
Hydrogen	0.16%	1.77%	0.01%	2.46%	0.01%	1.68%	0.44%	1.74%
Aluminium	0.00%	16.18%	0.00%	13.29%	0.00%	16.43%	0.00%	14.98%
Iron	0.00%	18.90%	0.00%	13.07%	0.00%	20.35%	0.00%	13.53%
Copper	0.00%	1.85%	0.00%	2.70%	0.00%	1.62%	0.00%	2.46%
Lead	0.00%	25.84%	0.00%	21.40%	0.00%	22.82%	0.00%	24.80%
other	0.02%	3.25%	0.00%	3.01%	0.38%	4.42%	0.00%	2.92%

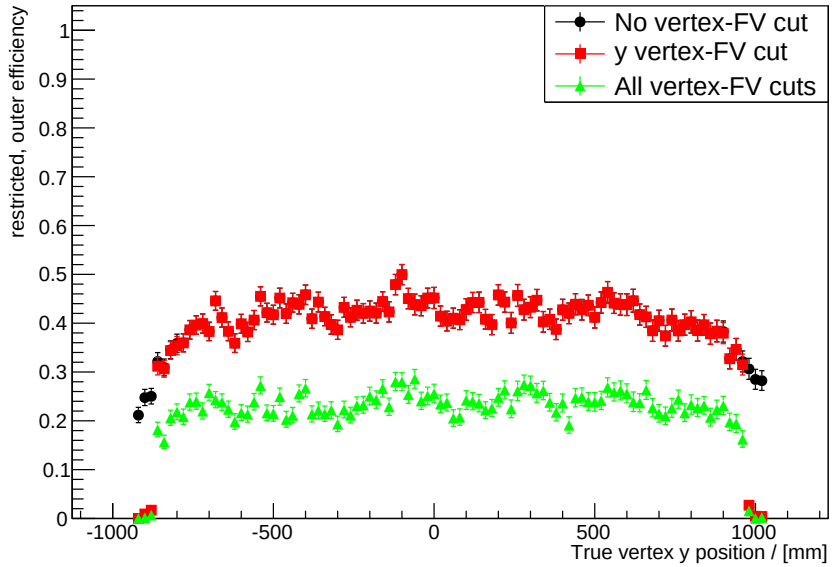


Figure A.3.: Efficiency vs vertex y-position in the TPC MC Neut sample. The effect of the vertex-FV cuts can be seen.

Table A.4.: OOFV background source. From the Neut MC. The TPC can be a source of OOFV background, because the fiducial volume is just a sub-set of the full TPC volumes. Also, particles that get reconstructed in the wrong bunch are counted as timing-related OOFV background, including events that spacially happened in the track-FV. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

	Main Selection	Passing-by CS	Timing CS	Delta-ray CS
TPC1	3.34%	2.66%	4.56%	2.90%
TPC2	3.43%	2.27%	4.92%	2.17%
TPC3	2.48%	1.45%	5.08%	2.10%
FGD1	3.98%	8.40%	3.58%	5.77%
FGD2	3.32%	5.60%	3.12%	4.21%
DsECAL	0.89%	0.28%	0.85%	0.81%
BrECAL	27.52%	14.09%	22.57%	18.20%
P0DECAL	6.38%	4.40%	7.30%	4.48%
P0D	24.39%	44.27%	22.05%	40.98%
SMRD	16.08%	9.94%	17.60%	10.48%
other	8.19%	6.62%	8.36%	7.89%

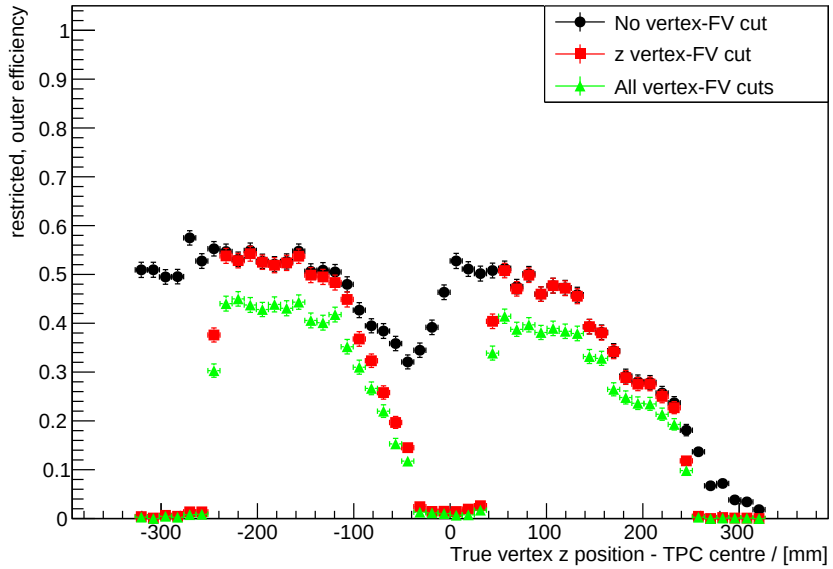


Figure A.4.: Efficiency vs vertex z-position in the TPC MC Neut sample. The effect of the vertex-FV cuts can be seen. Efficiency drops towards the downstream end of the fiducial volume. Because most particles are emitted in forward direction, chances increase that the particle tracks are not long enough for a proper PID in the first TPC the particle sees. Those particles can still be identified in the other TPCs further downstream, but shorter tracks also reduce the probability of a successful matching of the tracks in the different TPCs. The same applies for the drop before the vertical gap in the centre of the TPCs ($z = 0$). The closer the particle is created before the gap, the lower is the probability to match the hits before the gap to the hits behind the gap.

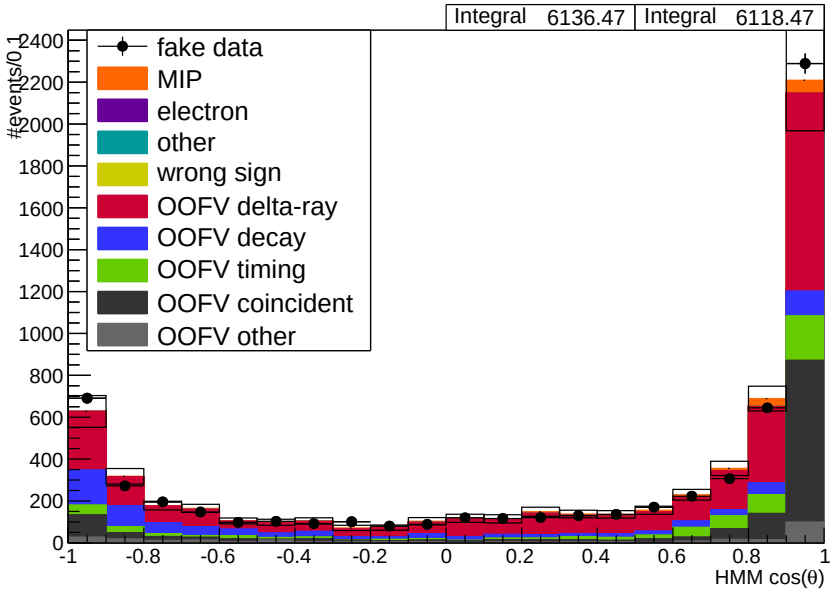


Figure A.5.: Passing-by CS angular distribution. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

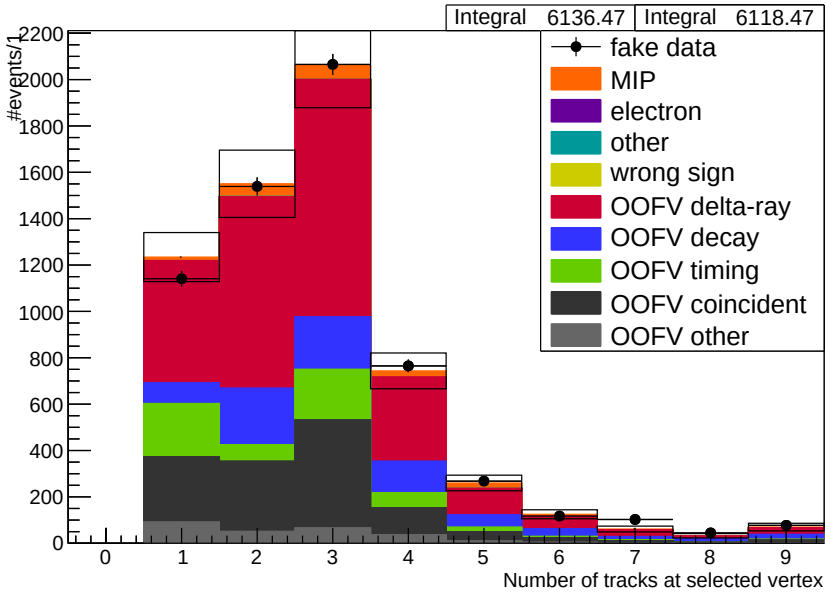


Figure A.6.: Passing-by CS track multiplicity. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

Table A.5.: Missing 1-track background in combined control regions. $|\phi_x|$ is the angle on the MicroMeGaS plane, with 0 being horizontal.

|x| position	momentum	$\|\phi_x\|$	data events	MC events	data - MC	data / MC
0 - 200	150 - 500	0.00 - 0.32	11	4.0	7.0	2.8
-	-	0.32 - 0.64	20	6.2	13.8	3.2
-	-	0.64 - 0.96	26	8.6	17.4	3.0
-	-	0.96 - 1.28	28	7.1	20.9	3.9
-	-	1.28 - 1.60	18	4.3	13.7	4.2
-	500 - 1000	0.00 - 0.32	19	3.5	15.5	5.4
-	-	0.32 - 0.64	5	3.6	1.4	1.4
-	-	0.64 - 0.96	4	4.9	-0.9	0.8
-	-	0.96 - 1.28	5	3.0	2.0	1.7
-	-	1.28 - 1.60	0	0.9	-0.9	0.0
-	1000 - ...	0.00 - 0.32	33	5.9	27.1	5.6
-	-	0.32 - 0.64	12	3.6	8.4	3.3
-	-	0.64 - 0.96	6	3.5	2.5	1.7
-	-	0.96 - 1.28	6	1.7	4.3	3.5
-	-	1.28 - 1.60	11	1.3	9.7	8.5
200 - 400	150 - 500	0.00 - 0.32	26	5.3	20.7	4.9
-	-	0.32 - 0.64	19	4.6	14.4	4.1
-	-	0.64 - 0.96	17	2.9	14.1	5.9
-	-	0.96 - 1.28	24	5.1	18.9	4.7
-	-	1.28 - 1.60	24	3.5	20.5	6.9
-	500 - 1000	0.00 - 0.32	18	5.0	13.0	3.6
-	-	0.32 - 0.64	4	2.0	2.0	2.0
-	-	0.64 - 0.96	7	0.8	6.2	8.8
-	-	0.96 - 1.28	10	1.9	8.1	5.3
-	-	1.28 - 1.60	6	0.4	5.6	15.0
-	1000 - ...	0.00 - 0.32	49	7.2	41.8	6.8
-	-	0.32 - 0.64	15	2.1	12.9	7.1
-	-	0.64 - 0.96	4	0.7	3.3	5.7
-	-	0.96 - 1.28	8	0.5	7.5	16.0
-	-	1.28 - 1.60	15	0.3	14.7	50.0
400 - 600	150 - 500	0.00 - 0.32	33	6.8	26.2	4.9
-	-	0.32 - 0.64	12	4.4	7.6	2.7
-	-	0.64 - 0.96	20	4.2	15.8	4.8
-	-	0.96 - 1.28	17	4.4	12.6	3.9
-	-	1.28 - 1.60	17	3.9	13.1	4.4
-	500 - 1000	0.00 - 0.32	19	2.9	16.1	6.6
-	-	0.32 - 0.64	11	1.7	9.3	6.5
-	-	0.64 - 0.96	3	1.2	1.8	2.5
-	-	0.96 - 1.28	4	1.0	3.0	4.0
-	-	1.28 - 1.60	7	0.3	6.7	23.3
-	1000 - ...	0.00 - 0.32	21	6.1	14.9	3.4
-	-	0.32 - 0.64	8	2.1	5.9	3.8
-	-	0.64 - 0.96	5	1.1	3.9	4.5
-	-	0.96 - 1.28	7	0.8	6.2	8.8
-	-	1.28 - 1.60	10	0.4	9.6	25.0
600 - 800	150 - 500	0.00 - 0.32	48	21.2	26.8	2.3
-	-	0.32 - 0.64	62	33.0	29.0	1.9
-	-	0.64 - 0.96	61	31.4	29.6	1.9
-	-	0.96 - 1.28	68	40.8	27.2	1.7
-	-	1.28 - 1.60	48	24.4	23.6	2.0
-	500 - 1000	0.00 - 0.32	32	13.7	18.3	2.3
-	-	0.32 - 0.64	38	15.2	22.8	2.5
-	-	0.64 - 0.96	20	7.3	12.7	2.7
-	-	0.96 - 1.28	15	4.8	10.2	3.1
-	-	1.28 - 1.60	12	2.4	9.6	5.0
-	1000 - ...	0.00 - 0.32	45	20.3	24.7	2.2
-	-	0.32 - 0.64	27	10.5	16.5	2.6
-	-	0.64 - 0.96	3	3.3	-0.3	0.9
-	-	0.96 - 1.28	13	2.9	10.1	4.5
-	-	1.28 - 1.60	18	1.1	16.9	16.4

Table A.6.: Missing n-track background in combined control regions. $|\phi_x|$ is the angle on the MicroMeGaS plane, with 0 being horizontal.

|x| position	momentum	$\|\phi_x\|$	data events	MC events	data - MC	data / MC
0 - 200	150 - 500	0.00 - 0.32	36	31.9	4.1	1.1
-	-	0.32 - 0.64	36	21.0	15.0	1.7
-	-	0.64 - 0.96	21	11.3	9.7	1.9
-	-	0.96 - 1.28	18	12.2	5.8	1.5
-	-	1.28 - 1.60	13	11.6	1.4	1.1
-	500 - 1000	0.00 - 0.32	27	20.1	6.9	1.3
-	-	0.32 - 0.64	13	9.2	3.8	1.4
-	-	0.64 - 0.96	6	4.2	1.8	1.4
-	-	0.96 - 1.28	6	2.7	3.3	2.2
-	-	1.28 - 1.60	2	1.0	1.0	2.0
-	1000 - ...	0.00 - 0.32	93	70.5	22.5	1.3
-	-	0.32 - 0.64	16	10.6	5.4	1.5
-	-	0.64 - 0.96	3	3.1	-0.1	1.0
-	-	0.96 - 1.28	2	1.9	0.1	1.1
-	-	1.28 - 1.60	2	0.9	1.1	2.2
200 - 400	150 - 500	0.00 - 0.32	51	43.5	7.5	1.2
-	-	0.32 - 0.64	40	27.0	13.0	1.5
-	-	0.64 - 0.96	35	16.0	19.0	2.2
-	-	0.96 - 1.28	29	20.5	8.5	1.4
-	-	1.28 - 1.60	35	17.7	17.3	2.0
-	500 - 1000	0.00 - 0.32	42	28.3	13.7	1.5
-	-	0.32 - 0.64	23	13.6	9.4	1.7
-	-	0.64 - 0.96	10	4.8	5.2	2.1
-	-	0.96 - 1.28	3	2.2	0.8	1.4
-	-	1.28 - 1.60	3	3.3	-0.3	0.9
-	1000 - ...	0.00 - 0.32	175	90.7	84.3	1.9
-	-	0.32 - 0.64	22	15.6	6.4	1.4
-	-	0.64 - 0.96	7	3.6	3.4	1.9
-	-	0.96 - 1.28	4	2.2	1.8	1.8
-	-	1.28 - 1.60	8	1.7	6.3	4.7
400 - 600	150 - 500	0.00 - 0.32	68	44.9	23.1	1.5
-	-	0.32 - 0.64	42	24.8	17.2	1.7
-	-	0.64 - 0.96	36	17.7	18.3	2.0
-	-	0.96 - 1.28	25	19.6	5.4	1.3
-	-	1.28 - 1.60	44	19.4	24.6	2.3
-	500 - 1000	0.00 - 0.32	51	31.2	19.8	1.6
-	-	0.32 - 0.64	25	10.2	14.8	2.5
-	-	0.64 - 0.96	16	5.2	10.8	3.1
-	-	0.96 - 1.28	5	2.5	2.5	2.0
-	-	1.28 - 1.60	2	1.4	0.6	1.4
-	1000 - ...	0.00 - 0.32	143	95.4	47.6	1.5
-	-	0.32 - 0.64	25	16.8	8.2	1.5
-	-	0.64 - 0.96	5	4.4	0.6	1.1
-	-	0.96 - 1.28	1	2.5	-1.5	0.4
-	-	1.28 - 1.60	7	1.0	6.0	7.0
600 - 800	150 - 500	0.00 - 0.32	70	49.6	20.4	1.4
-	-	0.32 - 0.64	56	28.3	27.7	2.0
-	-	0.64 - 0.96	26	19.8	6.2	1.3
-	-	0.96 - 1.28	37	20.8	16.2	1.8
-	-	1.28 - 1.60	29	19.3	9.7	1.5
-	500 - 1000	0.00 - 0.32	53	35.3	17.7	1.5
-	-	0.32 - 0.64	24	15.0	9.0	1.6
-	-	0.64 - 0.96	8	5.5	2.5	1.5
-	-	0.96 - 1.28	3	3.3	-0.3	0.9
-	-	1.28 - 1.60	6	2.1	3.9	2.9
-	1000 - ...	0.00 - 0.32	157	87.3	69.7	1.8
-	-	0.32 - 0.64	26	18.0	8.0	1.4
-	-	0.64 - 0.96	5	3.8	1.2	1.3
-	-	0.96 - 1.28	3	2.9	0.1	1.0
-	-	1.28 - 1.60	1	1.2	-0.2	0.8

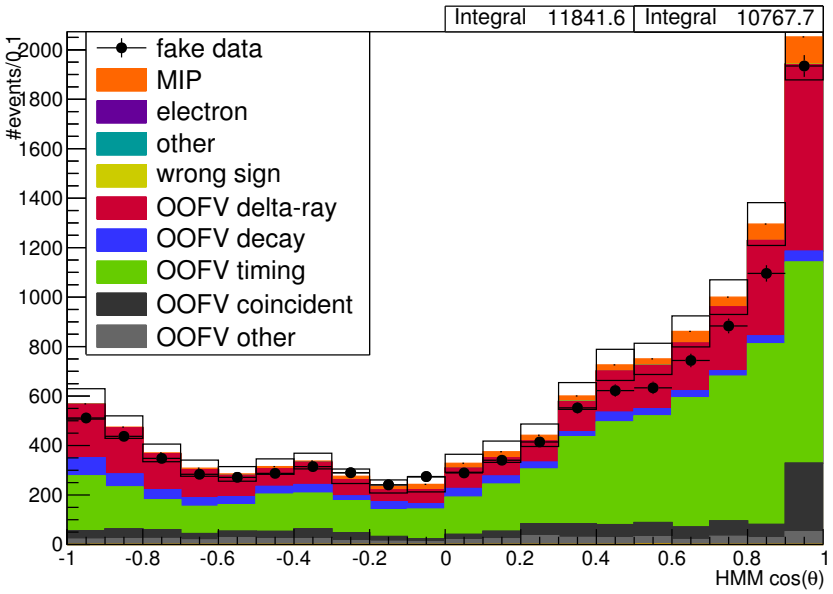


Figure A.7.: Timing CS angular distribution. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

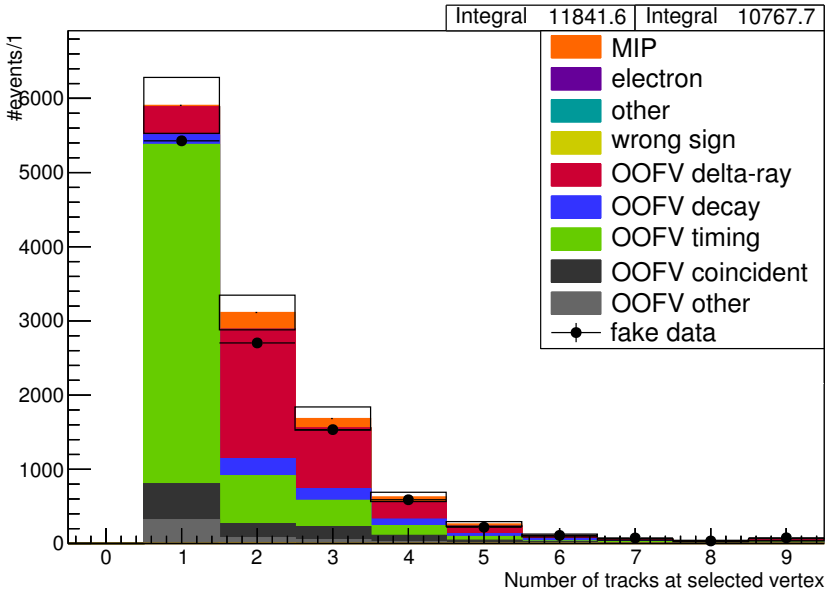


Figure A.8.: Timing CS track multiplicity. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

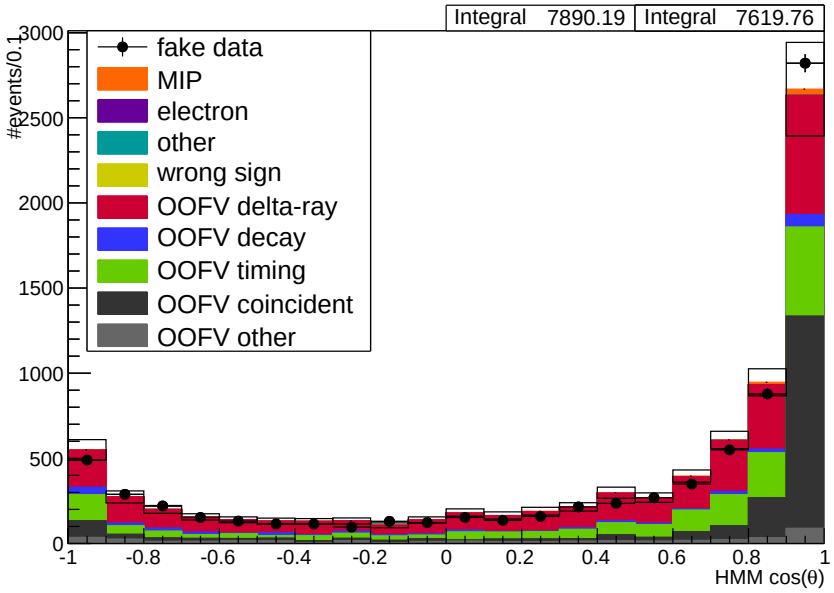


Figure A.9.: Delta-ray CS angular distribution. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

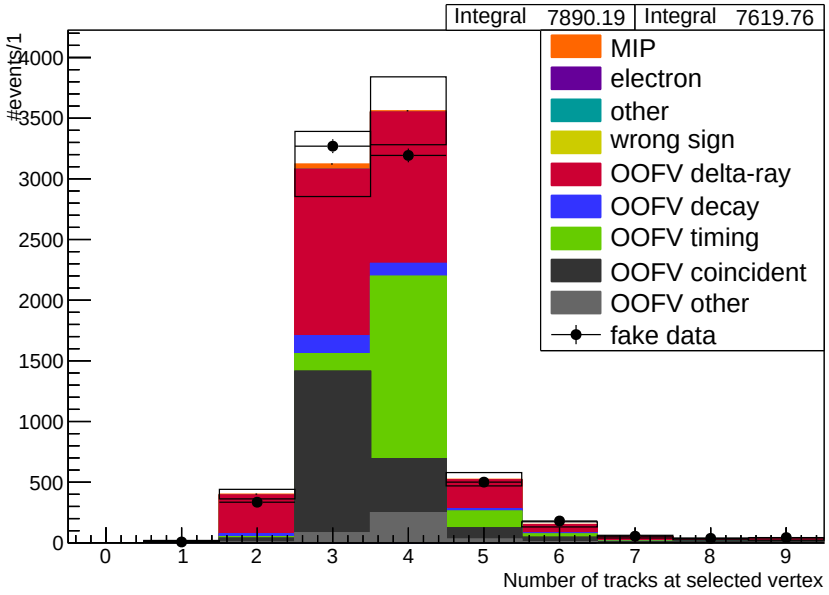


Figure A.10.: Delta-ray CS track multiplicity. The stacked histogram shows Neut with systematic detector and flux uncertainties and the fake data is Genie with statistical errors. The Neut data is scaled to match Genie's POT. The values do *not* include the effects of the ad hoc weight (see Section 2.5.1).

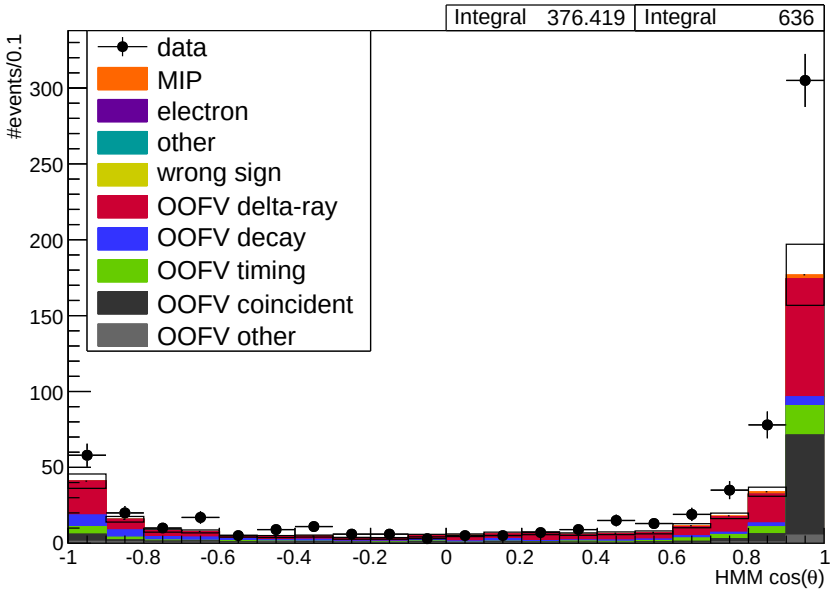


Figure A.11.: Unblinding Passing-by CS angular distribution. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

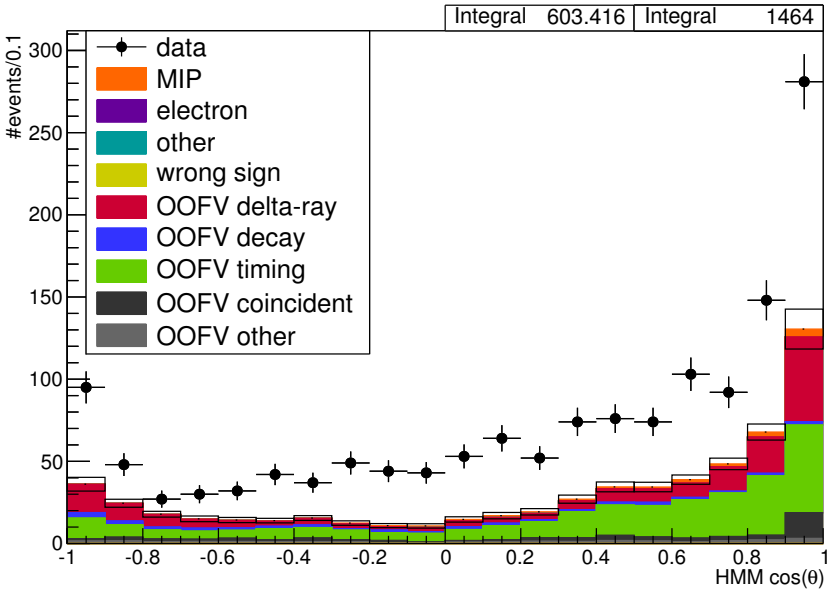


Figure A.12.: Unblinding Timing CS angular distribution. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

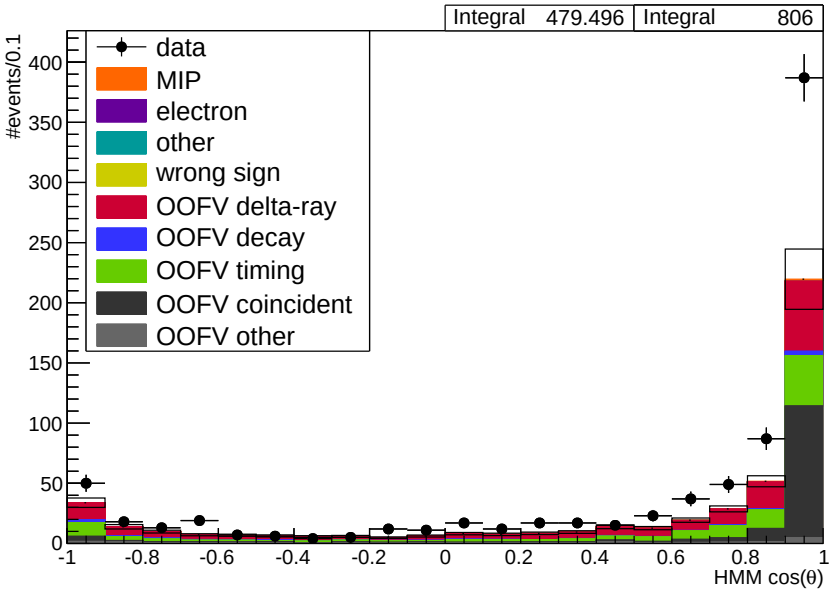


Figure A.13.: Unblinding Delta-ray CS angular distribution. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

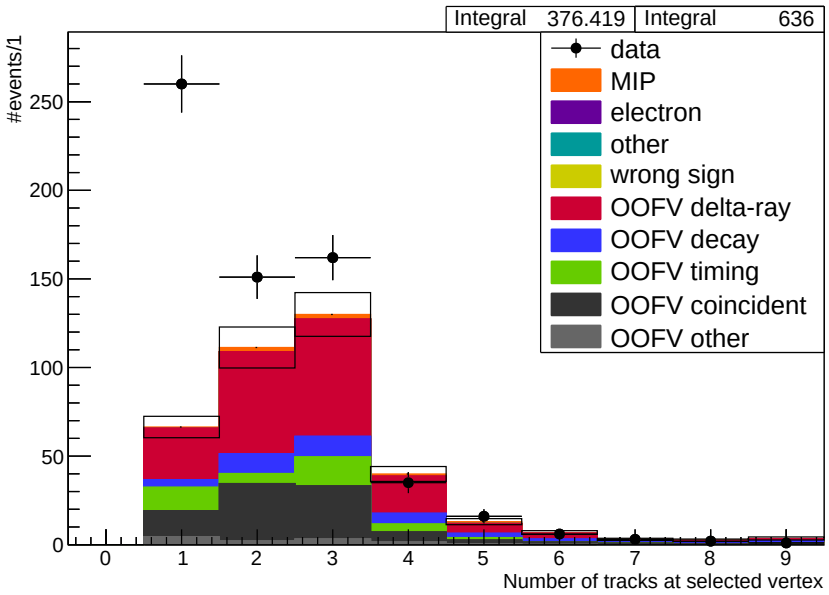


Figure A.14.: Unblinding Passing-by CS multiplicity. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

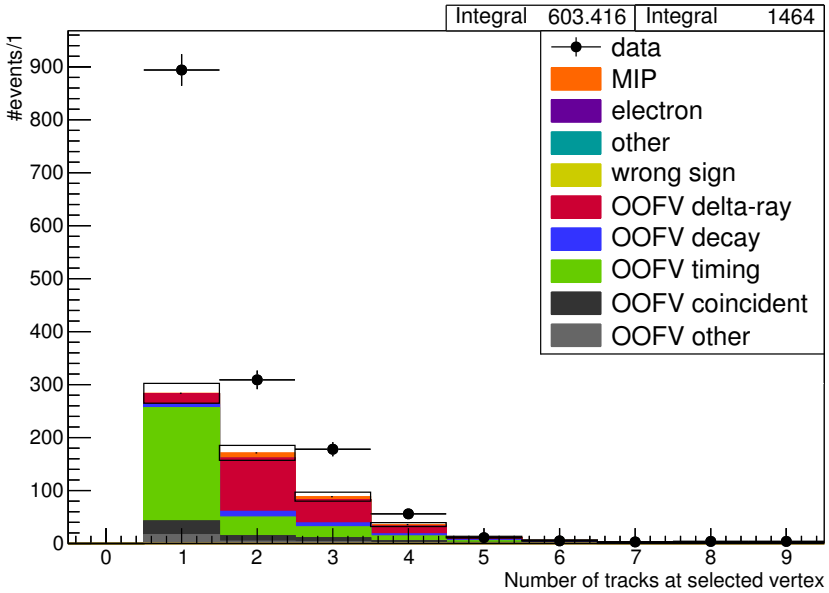


Figure A.15.: Unblinding Timing CS multiplicity. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

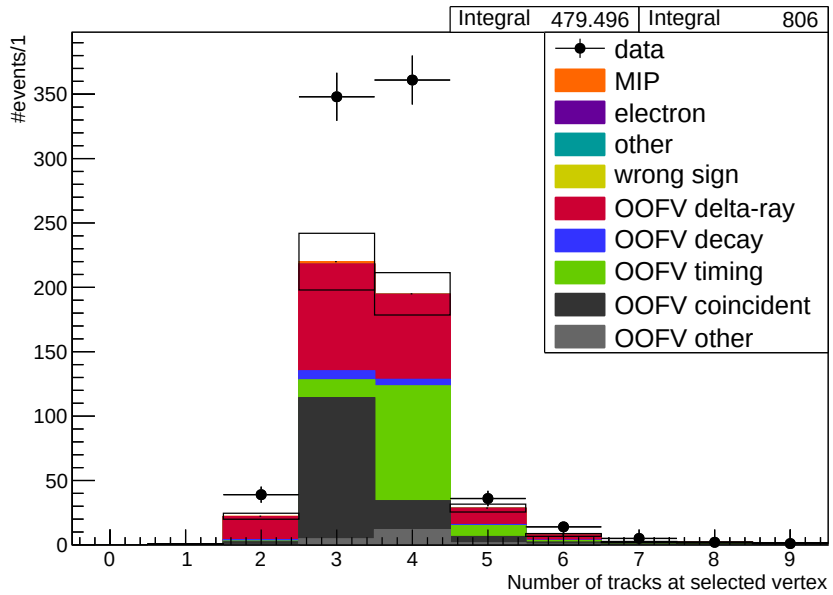


Figure A.16.: Unblinding Delta-ray CS multiplicity. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

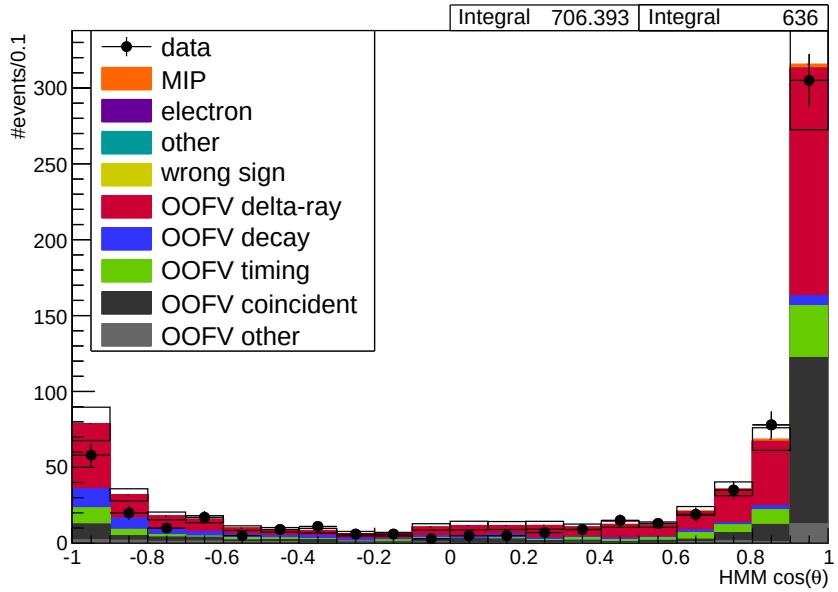


Figure A.17.: Unblinding Passing-by CS angular distribution with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

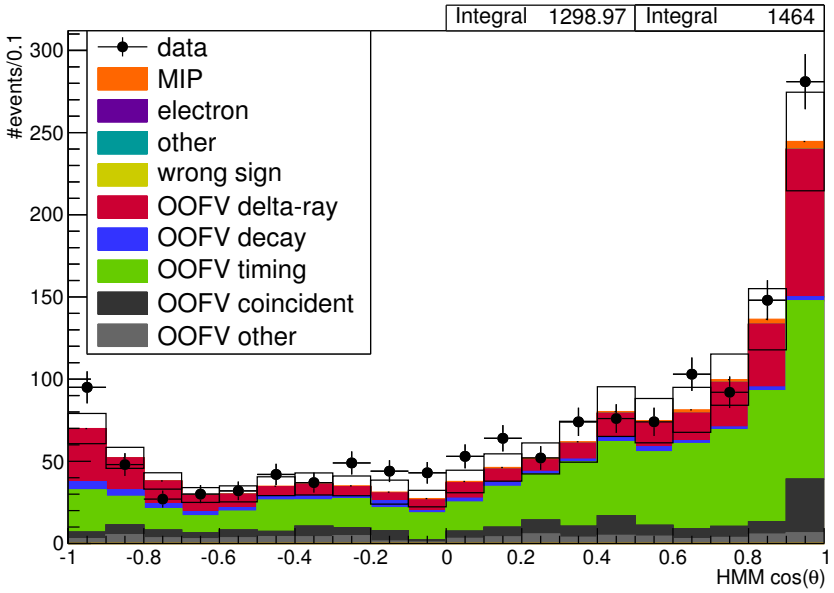


Figure A.18.: Unblinding Timing CS angular distribution with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

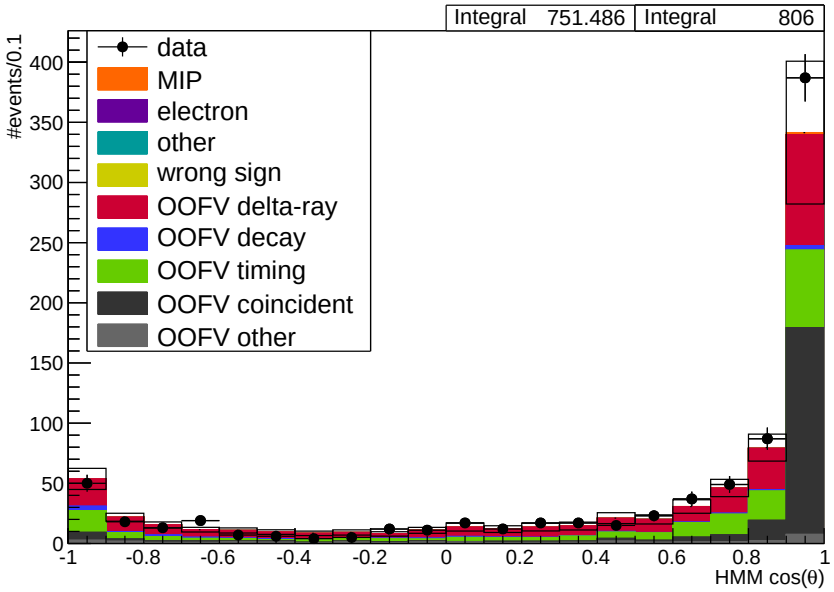


Figure A.19.: Unblinding Delta-ray CS angular distribution with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

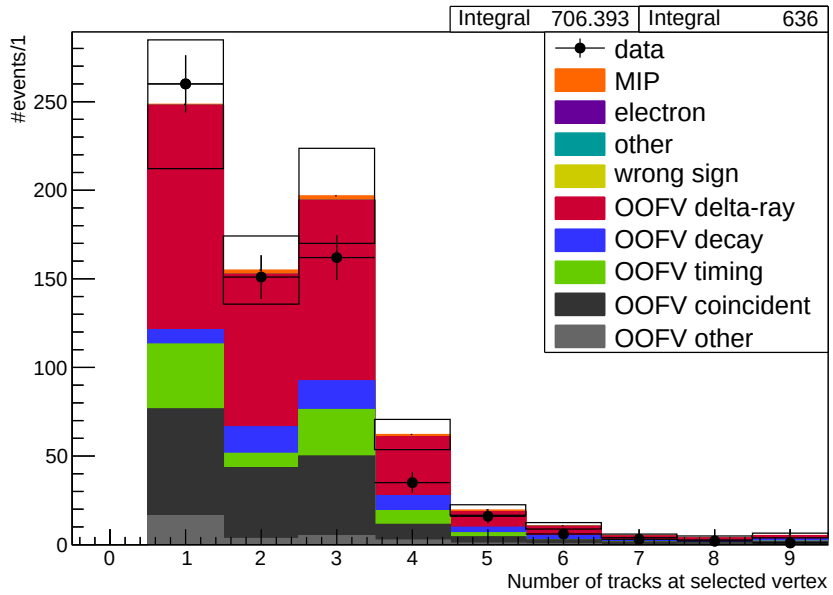


Figure A.20.: Unblinding Passing-by CS multiplicity with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

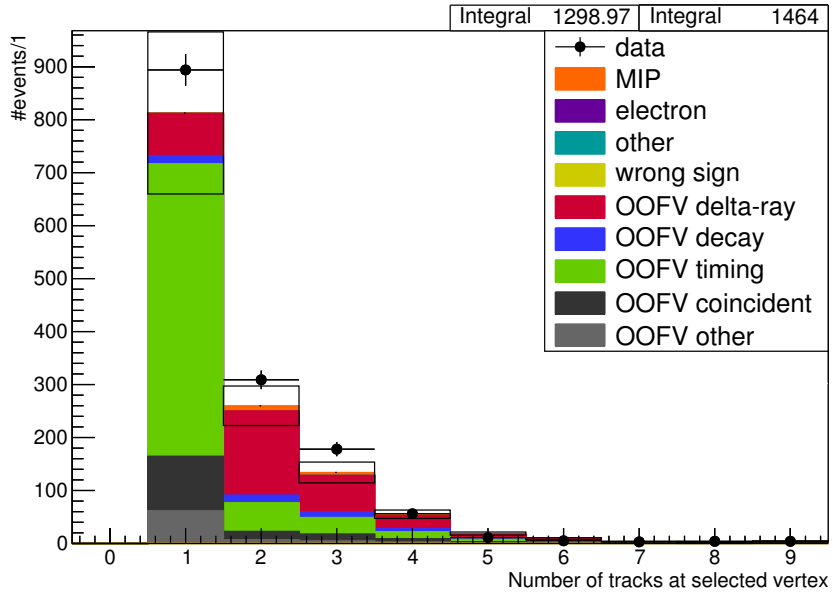


Figure A.21.: Unblinding Timing CS multiplicity with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

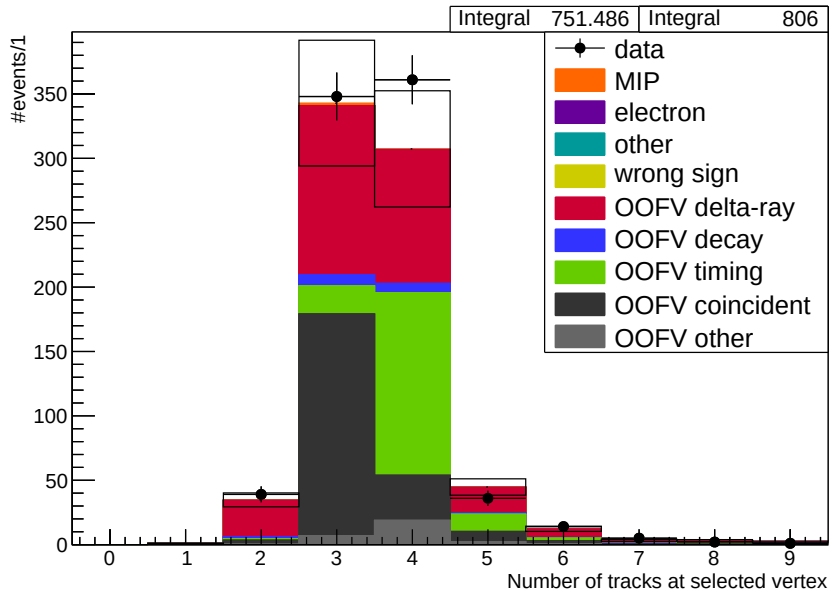


Figure A.22.: Unblinding Delta-ray CS multiplicity with ad hoc weights. The stacked histogram shows Neut with total detector and flux uncertainties and the data is shown with statistical errors. The Neut data is scaled to match the real data's POT.

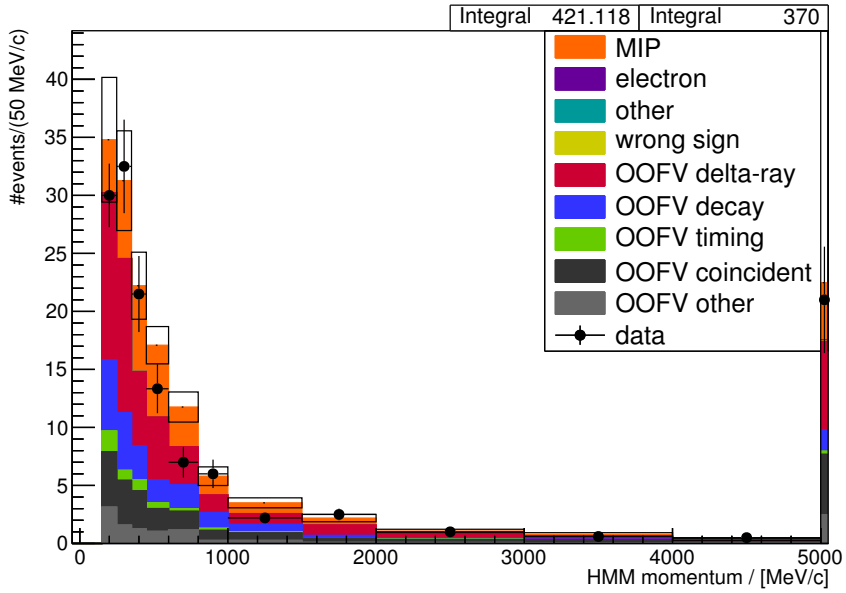


Figure A.23.: Unblinding Main Selection momentum (vs Genie). The stacked histogram shows Genie with total detector and flux uncertainties and the data is shown with statistical errors. The Genie data is scaled to match the real data's POT.

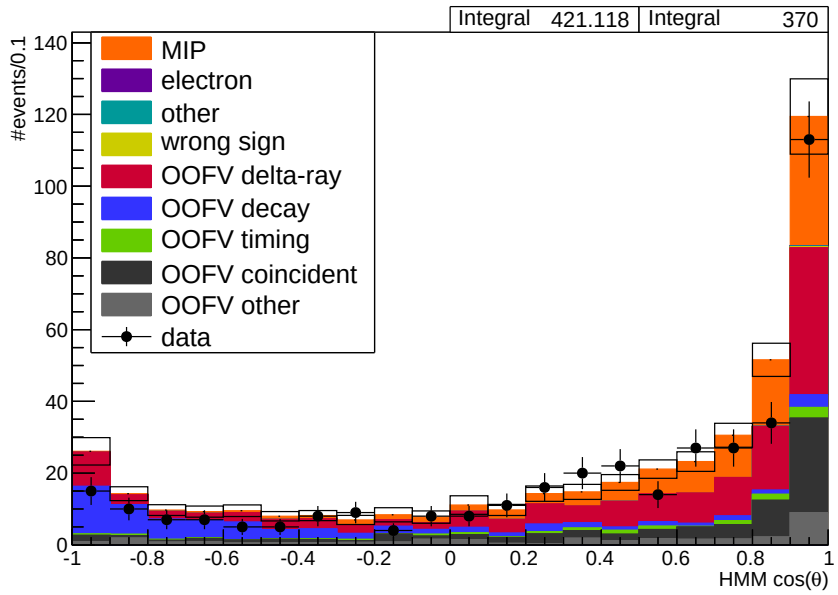


Figure A.24.: Unblinding Main Selection angular distribution (vs Genie). The stacked histogram shows Genie with total detector and flux uncertainties and the data is shown with statistical errors. The Genie data is scaled to match the real data's POT.

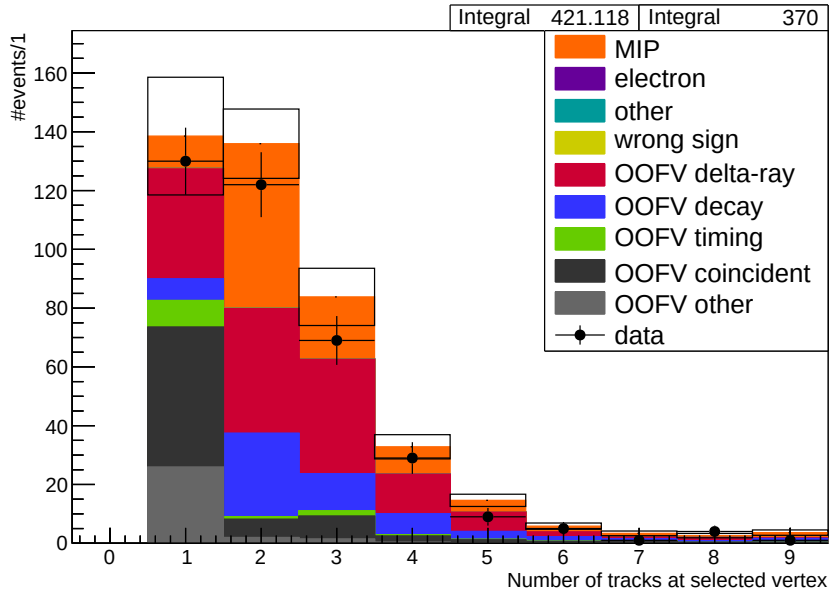
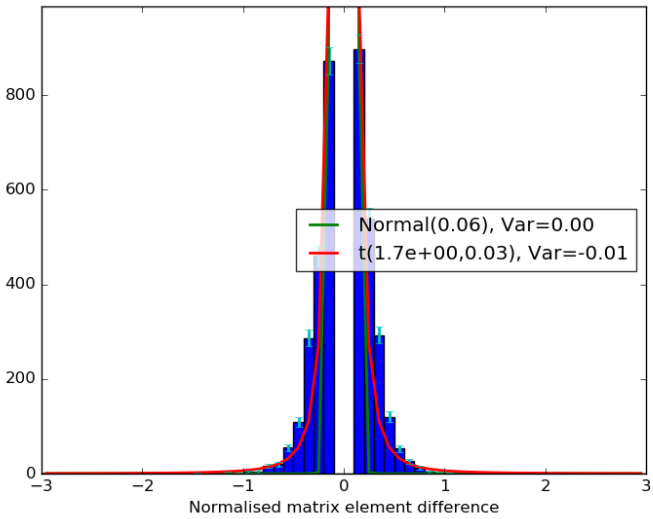
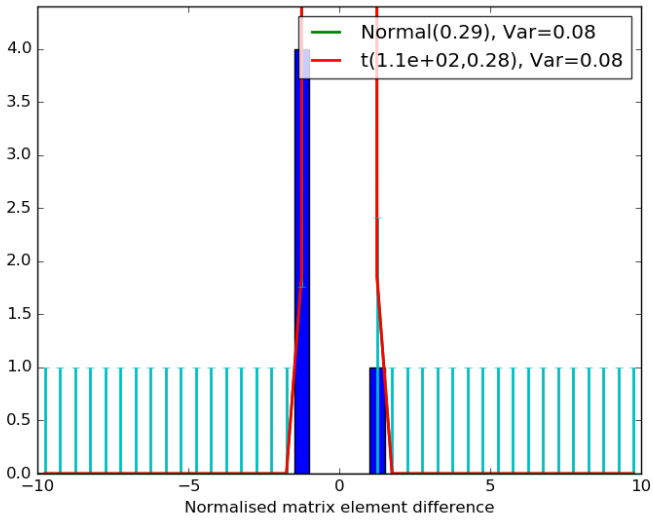


Figure A.25.: Unblinding Main Selection multiplicity (vs Genie). The stacked histogram shows Genie with total detector and flux uncertainties and the data is shown with statistical errors. The Genie data is scaled to match the real data's POT.

B. Cross-section analysis appendix

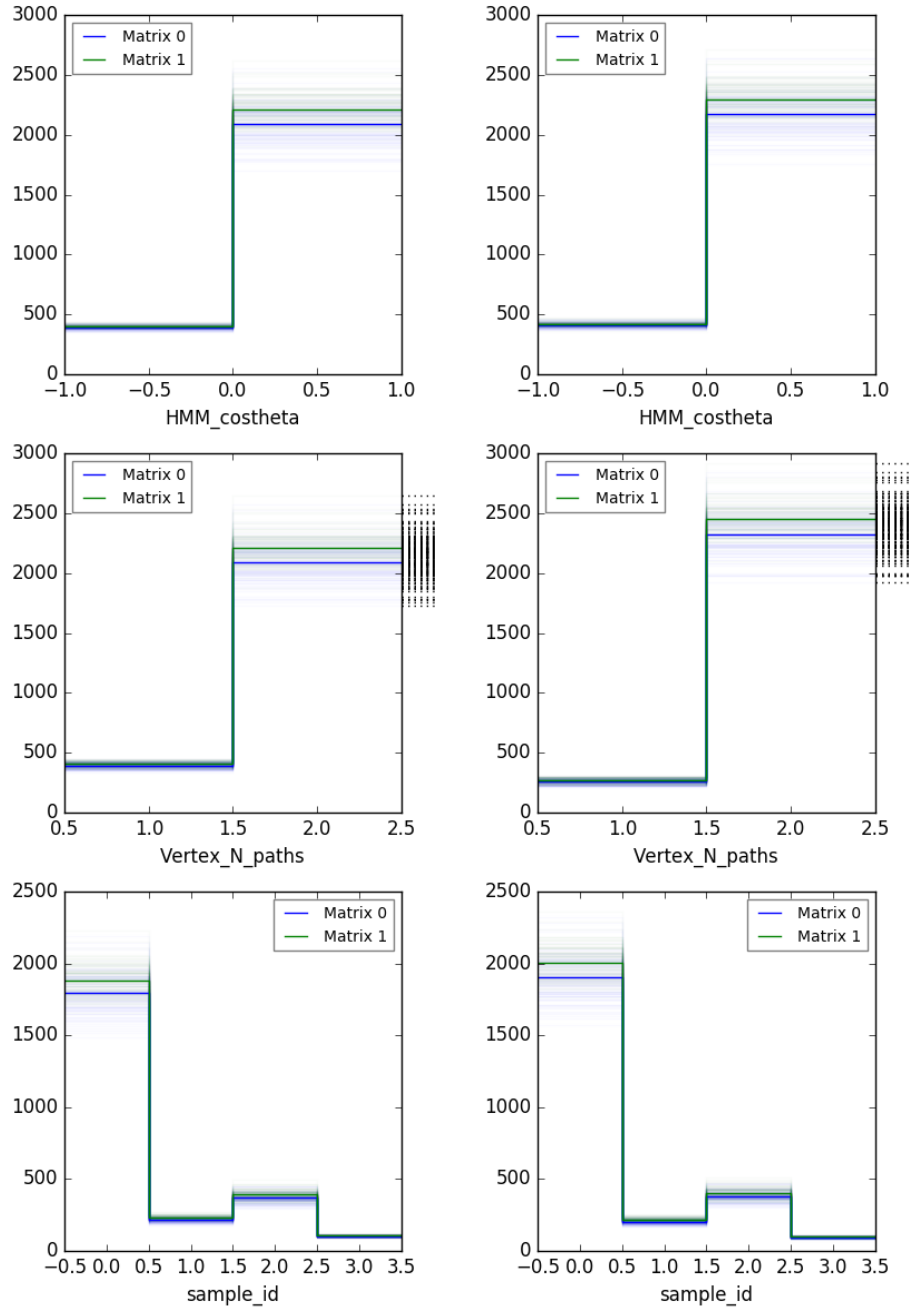


(a) Ignoring differences < 0.1



(b) Ignoring differences < 1.0

Figure B.1.: Matrix element comparison Nominal-Crazy. Shown are the difference of matrix element expectation values, normalised by the statistical uncertainty (see Section 3.2.1): $(\hat{R}_{ij}^{\text{Nominal}} - \hat{R}_{ij}^{\text{Crazy}})/\sqrt{\sigma^2(R_{ij}^{\text{Nominal}}) - \sigma^2(R_{ij}^{\text{Crazy}})}$. A normal and a student's t distribution are fitted to the histogram to estimate the variance of the distribution. (a) and (b) show the same data, but with different thresholds for the minimum difference.



(a) Genie truth

(b) Neut truth

Figure B.2.: Forward-folding test Nominal-Crazy. Neut and Genie truth data are folded through the response matrices (one generated with nominal and one with arbitrarily weighted Monte Carlo). Each semi-transparent line corresponds to one toy-matrix, and the opaque lines to the average matrix. The predictions differ, but are still compatible within the uncertainties of the matrices. The dotted lines outside the axes indicate that the neighbouring bin is an overflow bin.

```
from remu import binning
from remu import likelihood

# Load the response matrix
resp = np.load("response.npy")
gen_truth = np.load("generator-truth.npy")
with open("truth-binning.yml", 'rt') as f:
    truth_binning = binning.yaml.load(f)

# Load data and initialise likelihood machine
data = np.load("data.npy")
lm = likelihood.LikelihoodMachine(data, resp, truth_limits=gen_truth,
        eff_indices=np.where(gen_truth>0)[0], is_sparse=True)

# Load provided background
bg = np.load("background.npy")

# Load own models from event-by-event CSV files
# These contain *only* truth information
# No knowledge of the detector performance is needed to create them
truth_binning.reset()
truth_binning.fill_from_csv_file("my-model-0.csv")
model0 = truth_binning.get_values_as_ndarray()
truth_binning.reset()
truth_binning.fill_from_csv_file("my-model-1.csv")
model1 = truth_binning.get_values_as_ndarray()

# Create composite hypotheses, with model and bg weights as parameters
H0 = likelihood.TemplateHypothesis([model0, bg],
        parameter_limits=[(0,None), (0,None)])
H1 = likelihood.TemplateHypothesis([model1, bg],
        parameter_limits=[(0,None), (0,None)])

# Compare maximum likelihoods of models
ret0 = lm.max_log_likelihood(H0)
ret1 = lm.max_log_likelihood(H1)
print("Maximum likelihood model weights: %f, %f"%(ret0.x[0], ret1.x[0]))
if ret0.L > ret1.L:
    print("Model 0 is more likely than model 1.")
else:
    print("Model 1 is more likely than model 0.")
```

Listing 1: Simple example of how to use the `ReMU` package. Note that "response.npy" is actually a set of systematically and statistically varied response matrices. The likelihood results thus include the effects of the detector uncertainties.

B.1. The matrix

B.1.1. Used data samples

We use all available Monte Carlo samples to build the matrix. This includes the "signal only" samples and the sand muon samples.

B.1.2. Truth binning

The used truth binning is shown in listing 2. It has been tuned to make the signal efficiency in each bin sufficiently flat, while ensuring enough Monte Carlo statistics for each bin. The efficiencies of the Neut model in comparison with the bin edges are shown in Fig. B.3 through Fig. B.7. Note that the efficiency drop for backward going tracks in Fig. B.4 is caused by the prevalence of low-momentum tracks in that direction. This can be seen in Fig. B.5, where only tracks with $p > 400\,\mathrm{MeV/c}$ are considered. Thus it is not necessary to employ a finer binning in that area. The fine binning around $\cos(\theta) = 0$ does not just deal with the efficiency change there, but also with the fact that the reco binning has a bin edge at that point.

The total number of truth bins is 11760. Out of these, 5353 are filled with at least one MC event for the matrix generation. Ignoring the empty bins, the average truth bin filling is 46.3, slightly below the target of 50 events per bin. The median bin is filled with only 4 events, but 90% of the MC events are filled into bins with at least 29 events per bin. 80% of the events are filled into bins of at least 92 events per bin. Since the MC data used for the matrix construction should be close to the real distribution of events, this means that most real events are covered by bins with sufficiently high MC statistics to suppress the influence of the priors.

```
!RecBinning
binedges:
- - Truth_MIP_mom
  - [0.0, 150.0, 200.0, 300.0, 400.0, 500.0, 1200.0, .inf]
- - Truth_MIP_costheta
  - [-1.0, -0.2, -0.1, 0.0, 0.1, 0.2, 0.9, 1.0]
- - Truth_MIP_mom_sep
  - [-.inf, 0.0, 0.2, 0.4, 0.6, .inf]
- - Truth_MIP_mom_sep_back
  - [-.inf, 0.0, 0.2, 0.4, 0.6, 0.8, .inf]
- - MIPoofv
  - [0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, .inf]
include_upper: false
```

Listing 2: Truth binning yaml file.

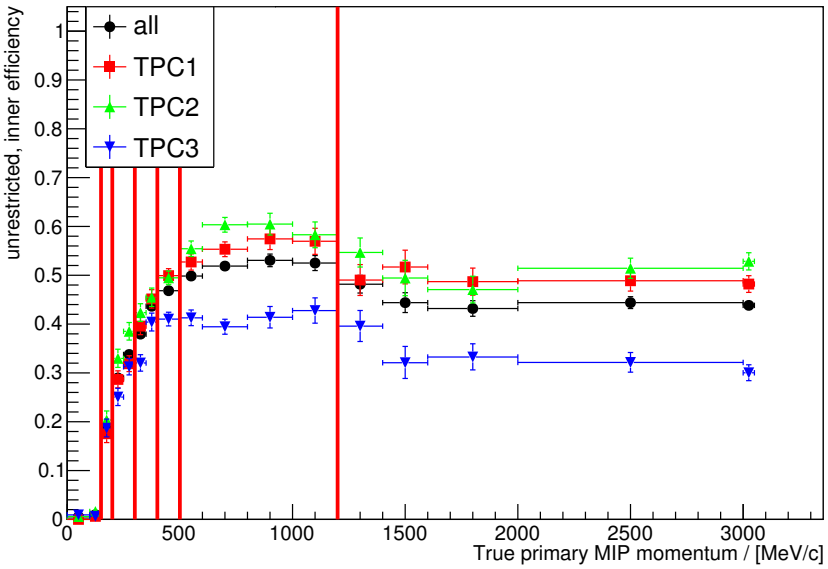


Figure B.3.: Efficiency and bin edges vs true momentum.

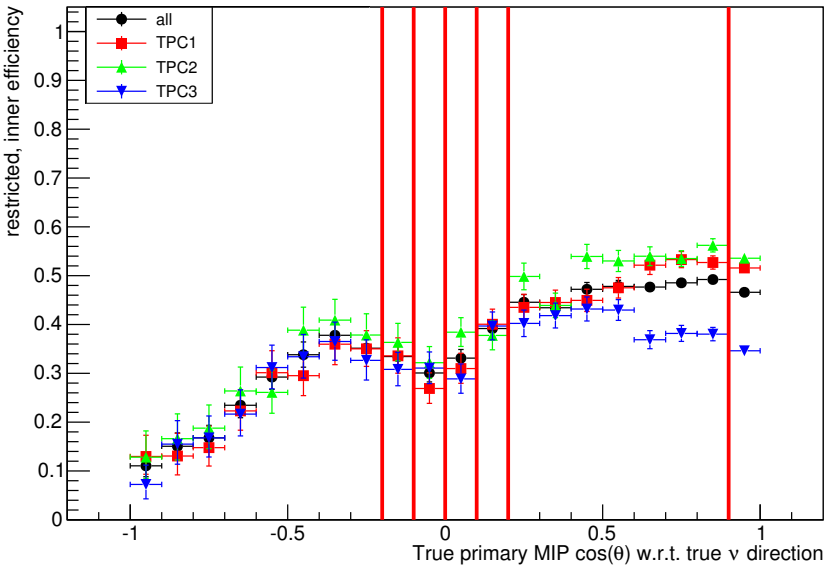


Figure B.4.: Efficiency and bin edges vs true angle.

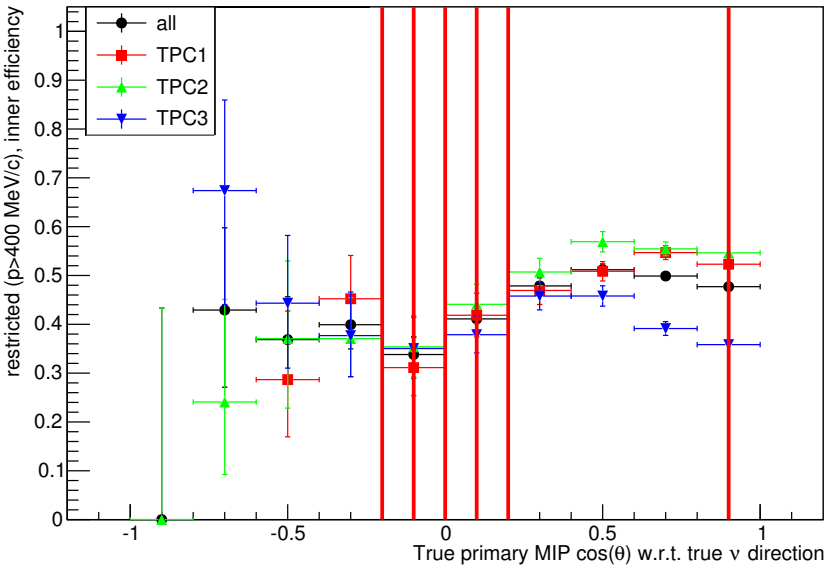


Figure B.5.: Efficiency and bin edges vs true angle at high momenta.

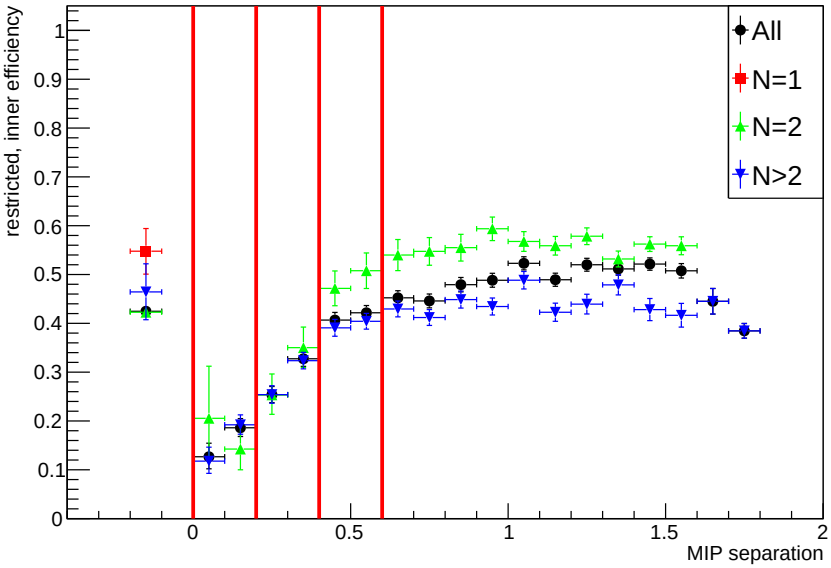


Figure B.6.: Efficiency and bin edges vs true forward separation for different charged-particle multiplicities.

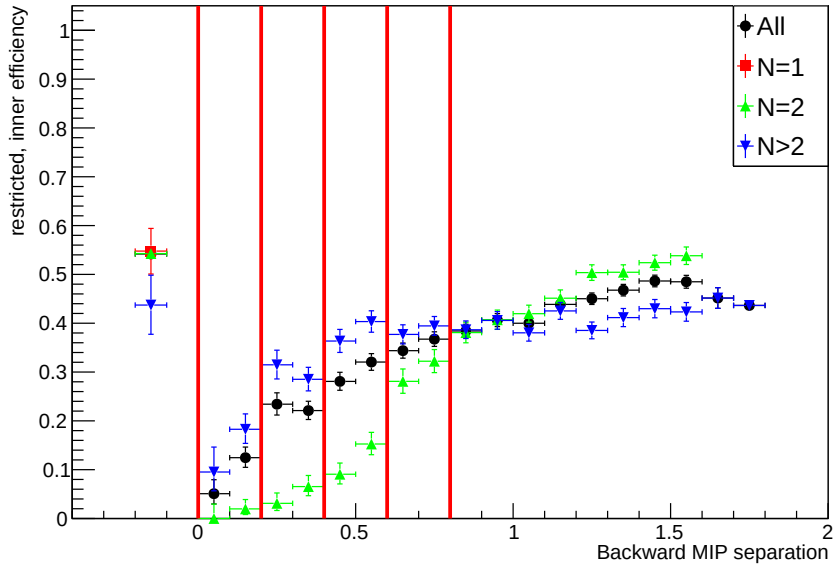


Figure B.7.: Efficiency and bin edges vs true backward separation for different charged-particle multiplicities.

B.1.3. Reco binning

The used reco binning is shown in listing 3. The sample ID is an integer that describes in which sample an event was reconstructed: the main selection (0) or one of the control samples (1-3). Since the expected number of recorded events is low, there are only four bins per sample. Most events of the signal models are recorded in forward going direction, while the OOFV background events are more evenly distributed. Thus differentiating between forward going and backward going events leads to an improved separation power. The same is true for the number of reconstructed paths (i.e. particle tracks) at the vertex. The OOFV is much more likely to produce events with a single track, while most signal events produce at least two charged particles leaving the vertex. Additionally, the biggest detector uncertainty at the moment is the "ad hoc" weight, which varies very strongly between 1-track and n-track events. Differentiating between the two allows the fitter to compensate this to a certain extent.

```
!RecBinning
binedges:
- - HMM_costheta
  - [-1.0,
    0.0,
    1.0]
- - Vertex_N_paths
  - [0.5,
    1.5,
    .inf]
- - sample_id
  - [-.5,
    0.5,
    1.5,
    2.5,
    3.5]
include_upper: false
```

Listing 3: Reco binning yaml file.

B.2. The templates

The templates used in the fit are shown in Fig. B.8 through Fig. B.43. Shown are the 1D projections onto each binned variable and 2D projections onto every possible combination of two variables.

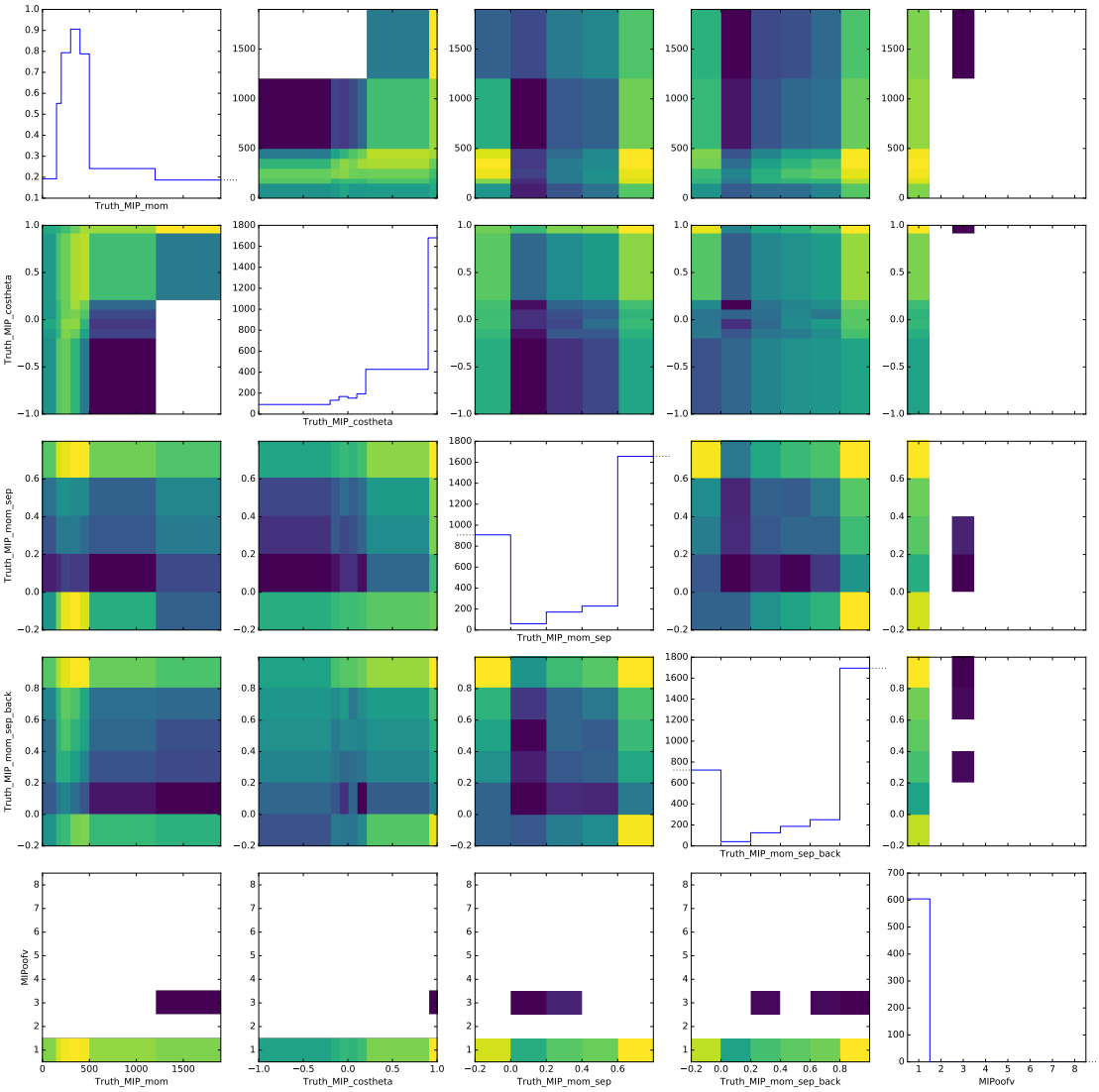


Figure B.8.: Neut ν_μ CC template

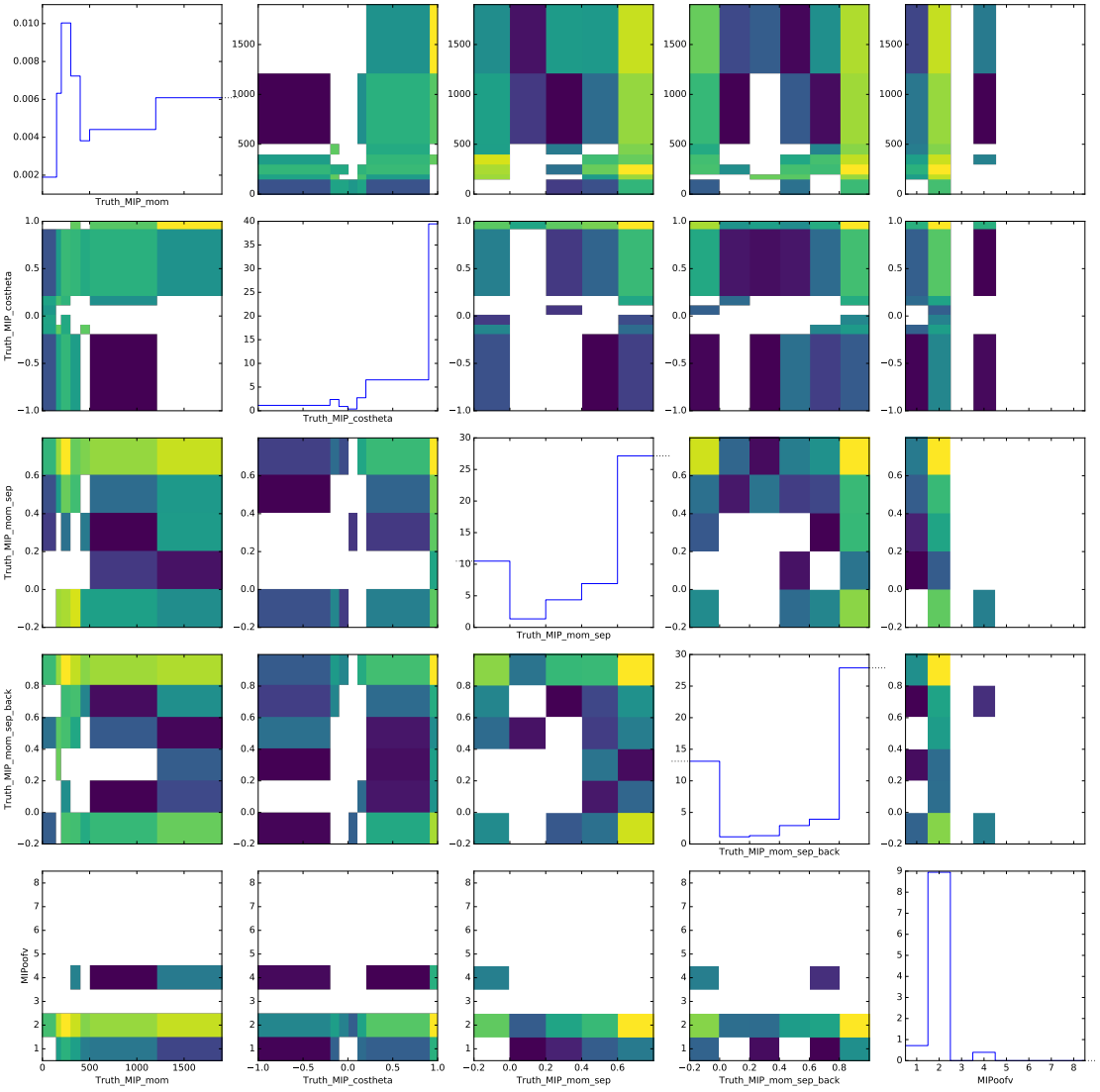


Figure B.9.: Neut $\bar{\nu}_\mu$ CC template

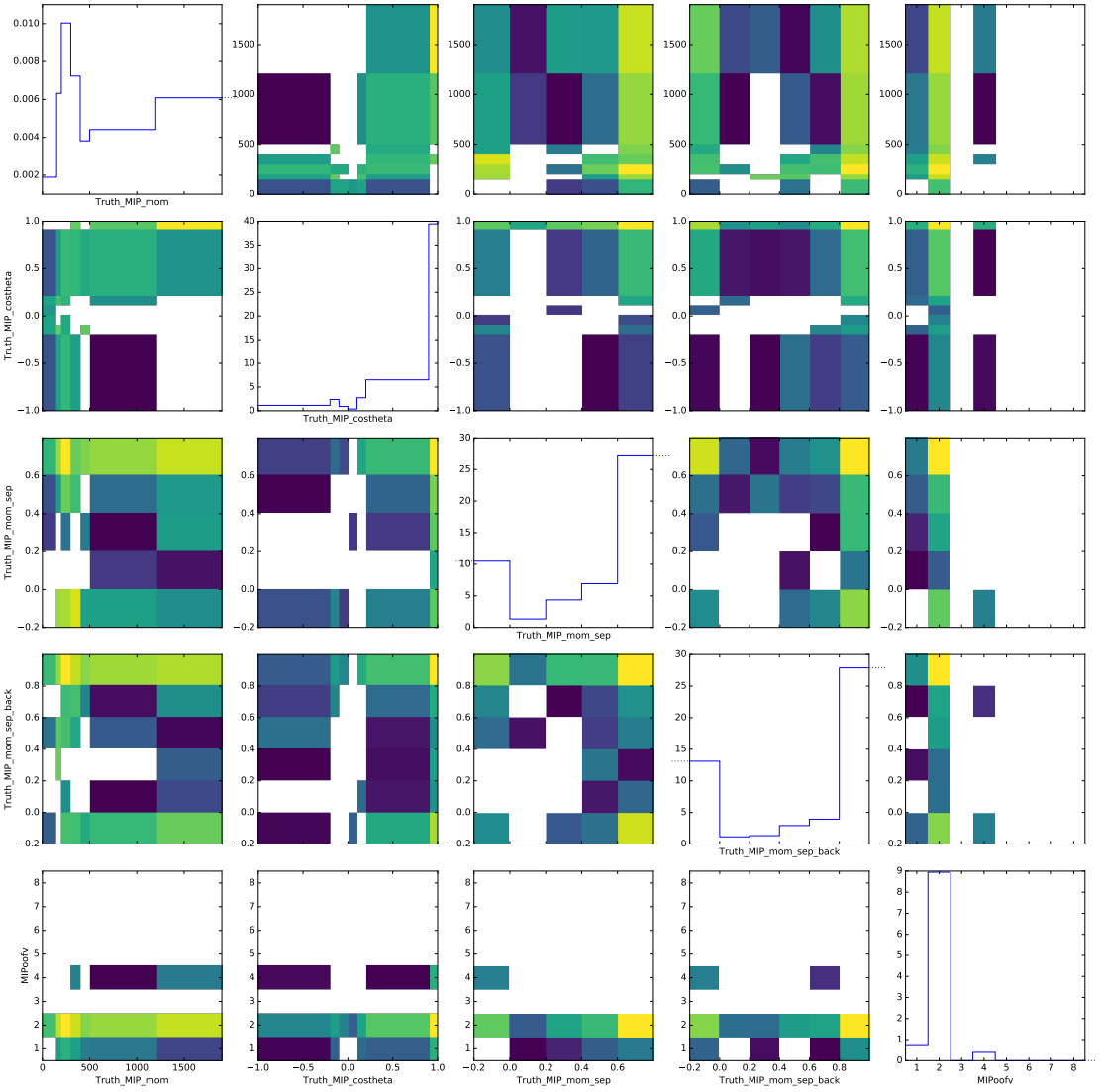


Figure B.10.: Neut $\nu_e/\bar{\nu}_e$ CC template

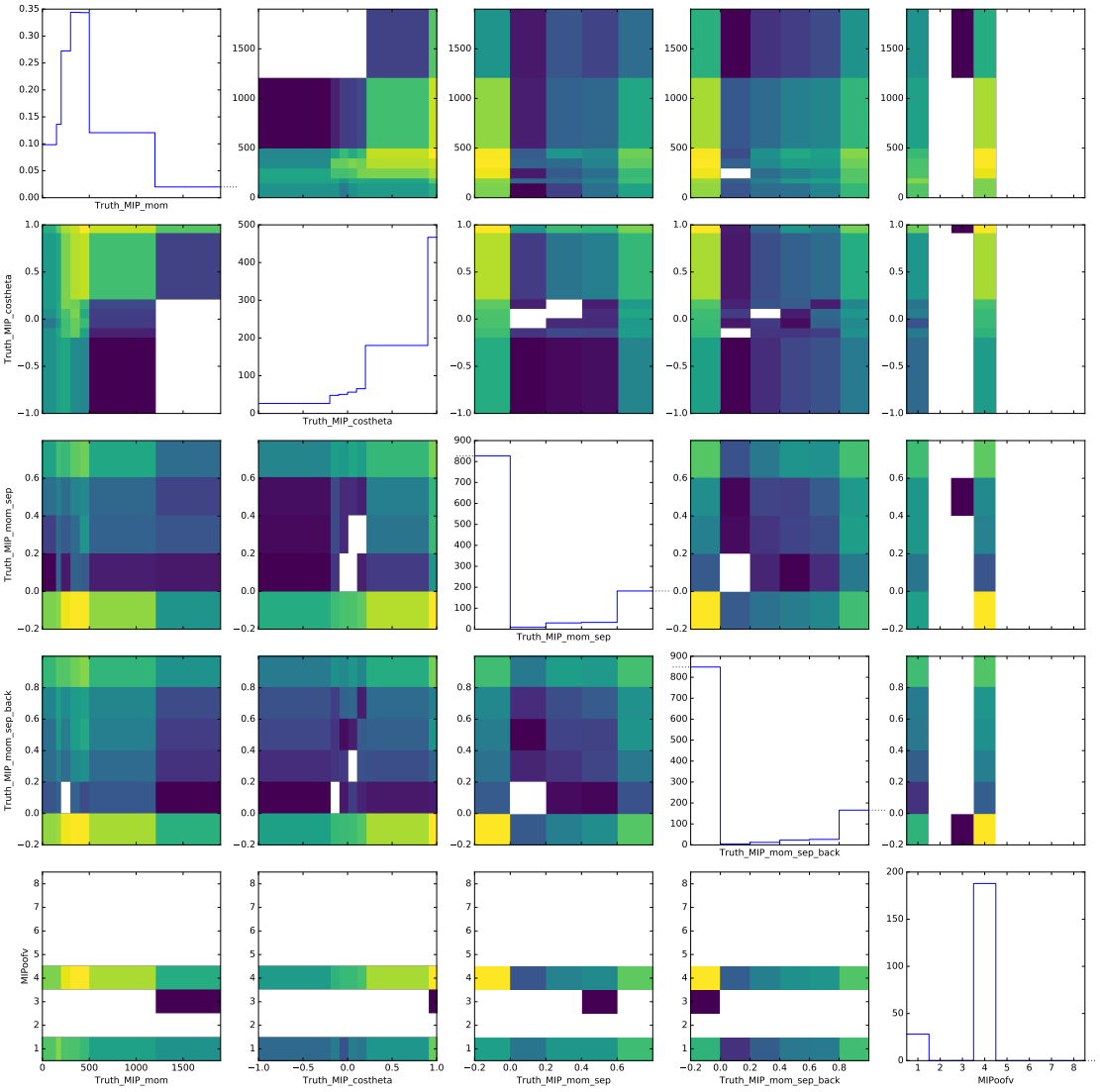


Figure B.11.: Neut NC template

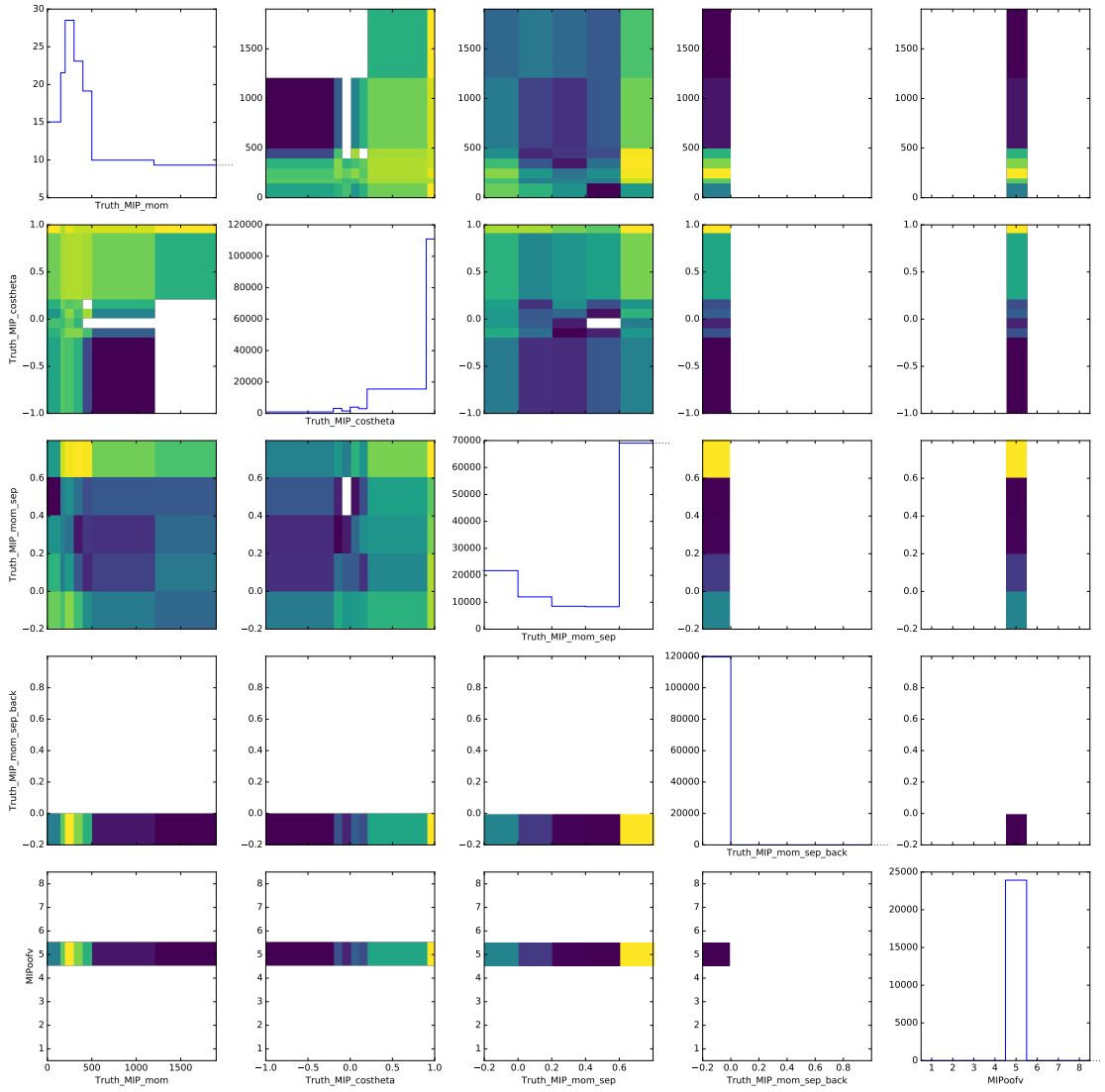


Figure B.12.: Neut 1-track OOFV delta-ray template

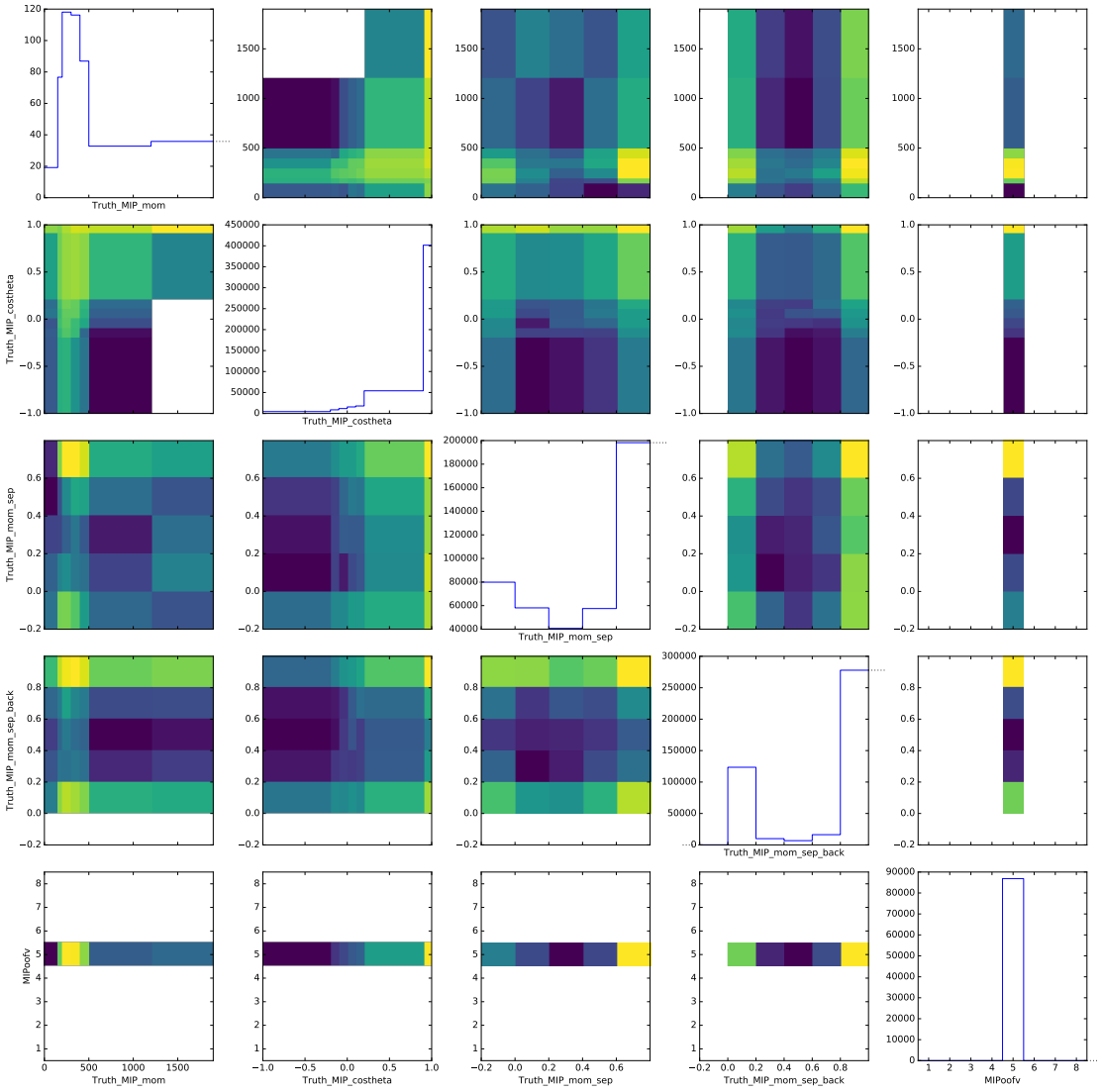


Figure B.13.: Neut n-track OOFV delta-ray template

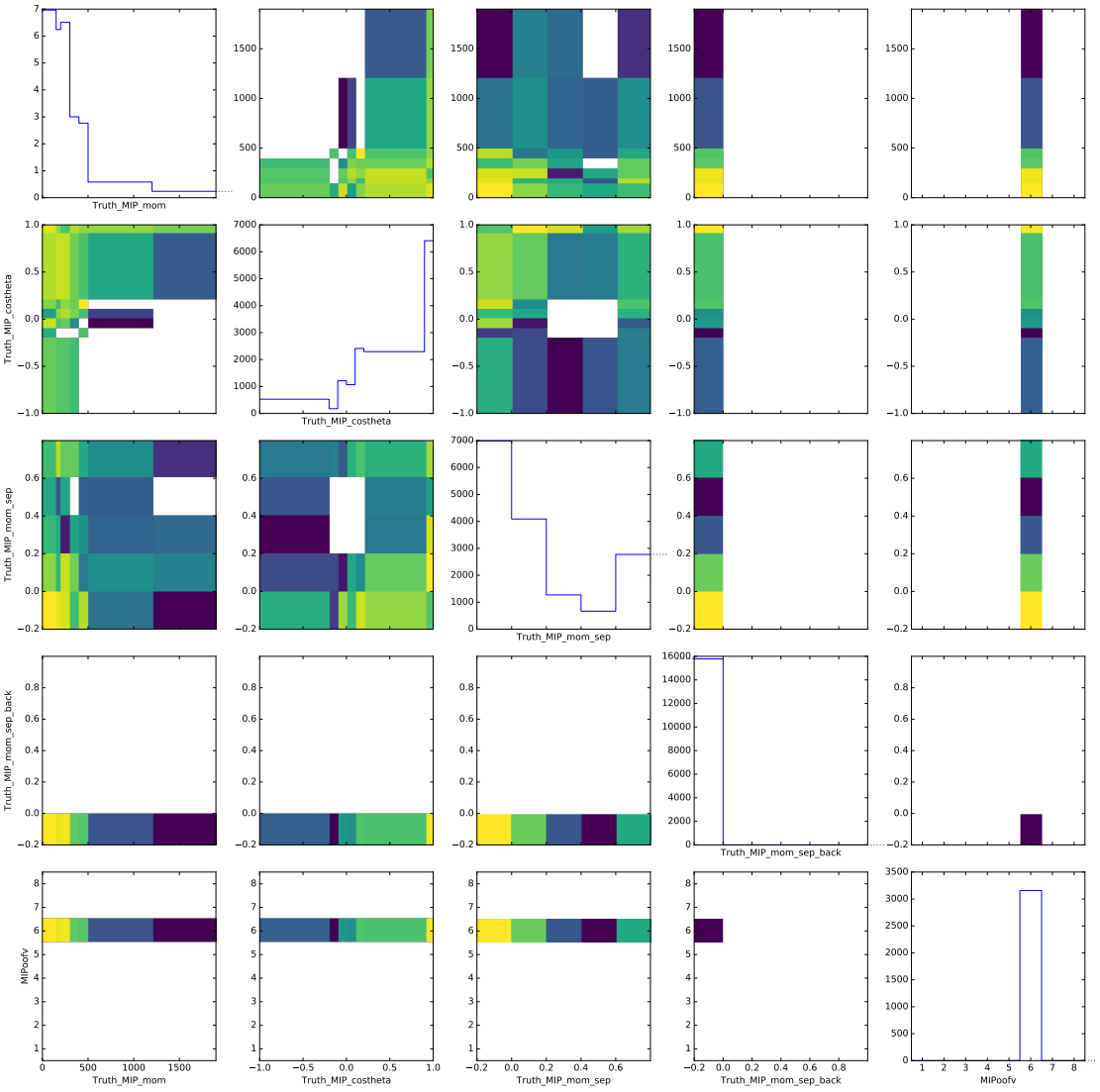


Figure B.14.: Neut 1-track OOFV decay template

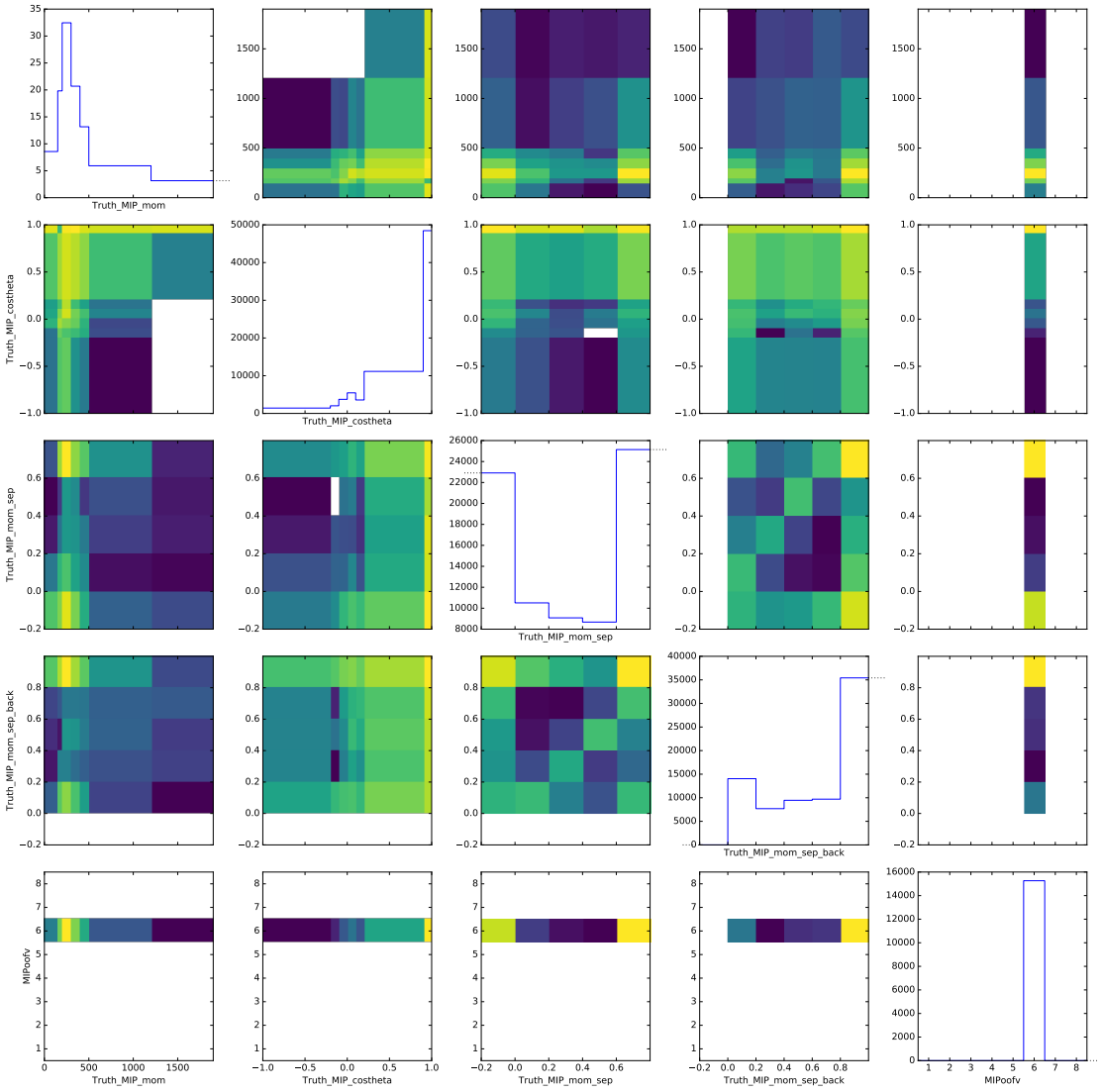


Figure B.15.: Neut n-track OOFV decay template

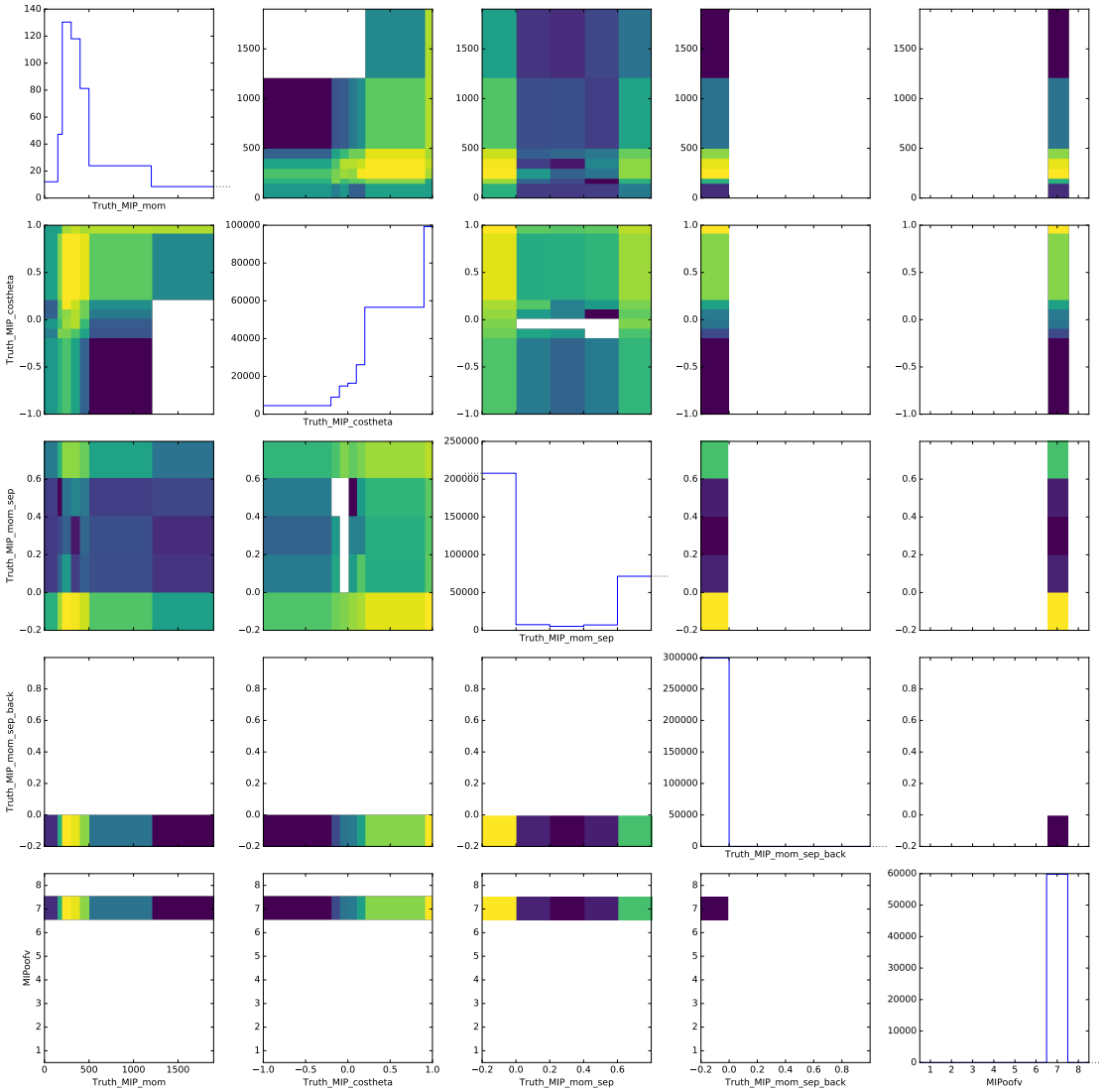


Figure B.16.: Neut 1-track timing-related OOFV template

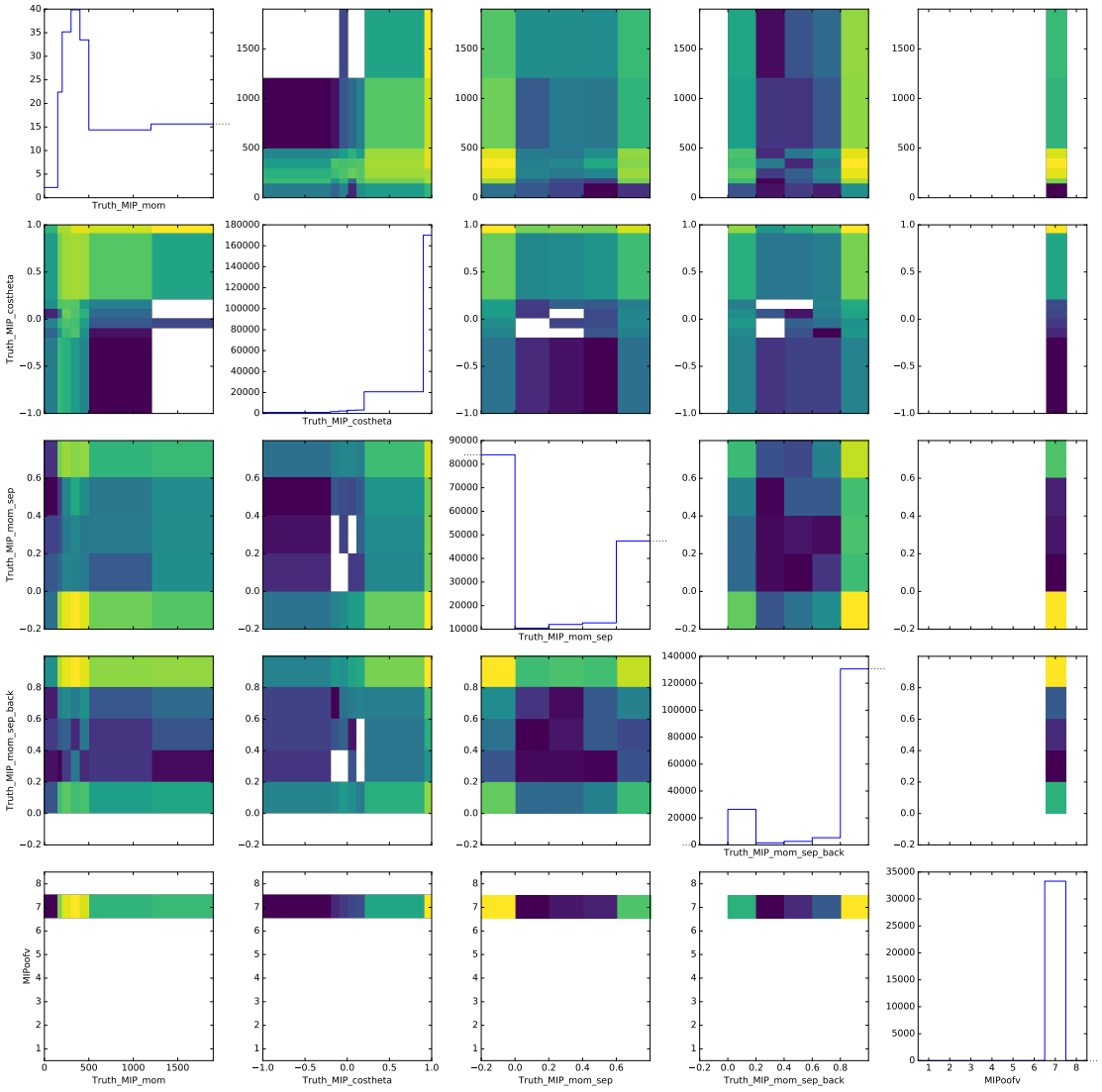


Figure B.17.: Neut n-track timing-related OOFV template

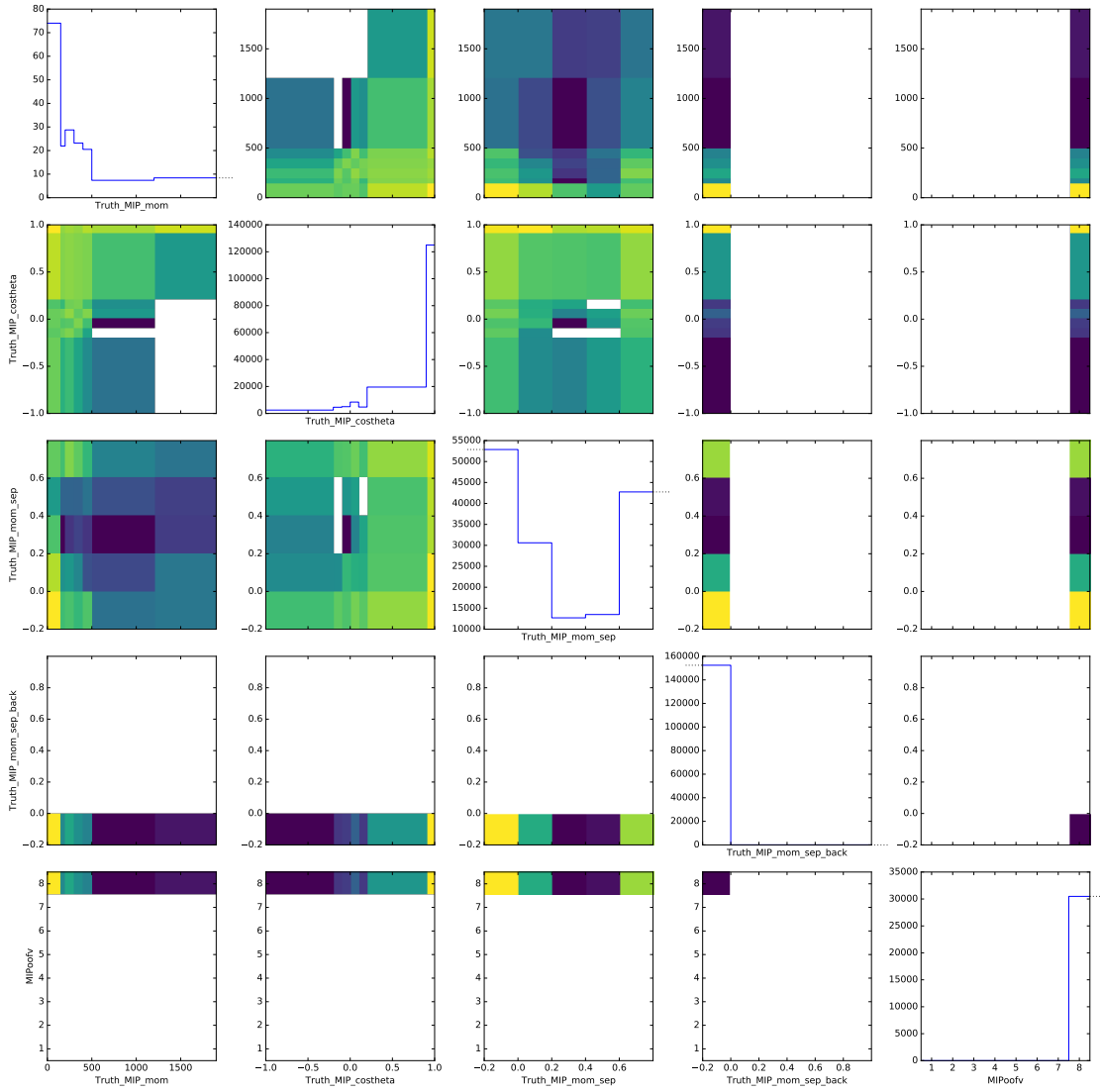


Figure B.18.: Neut 1-track other OOFV template

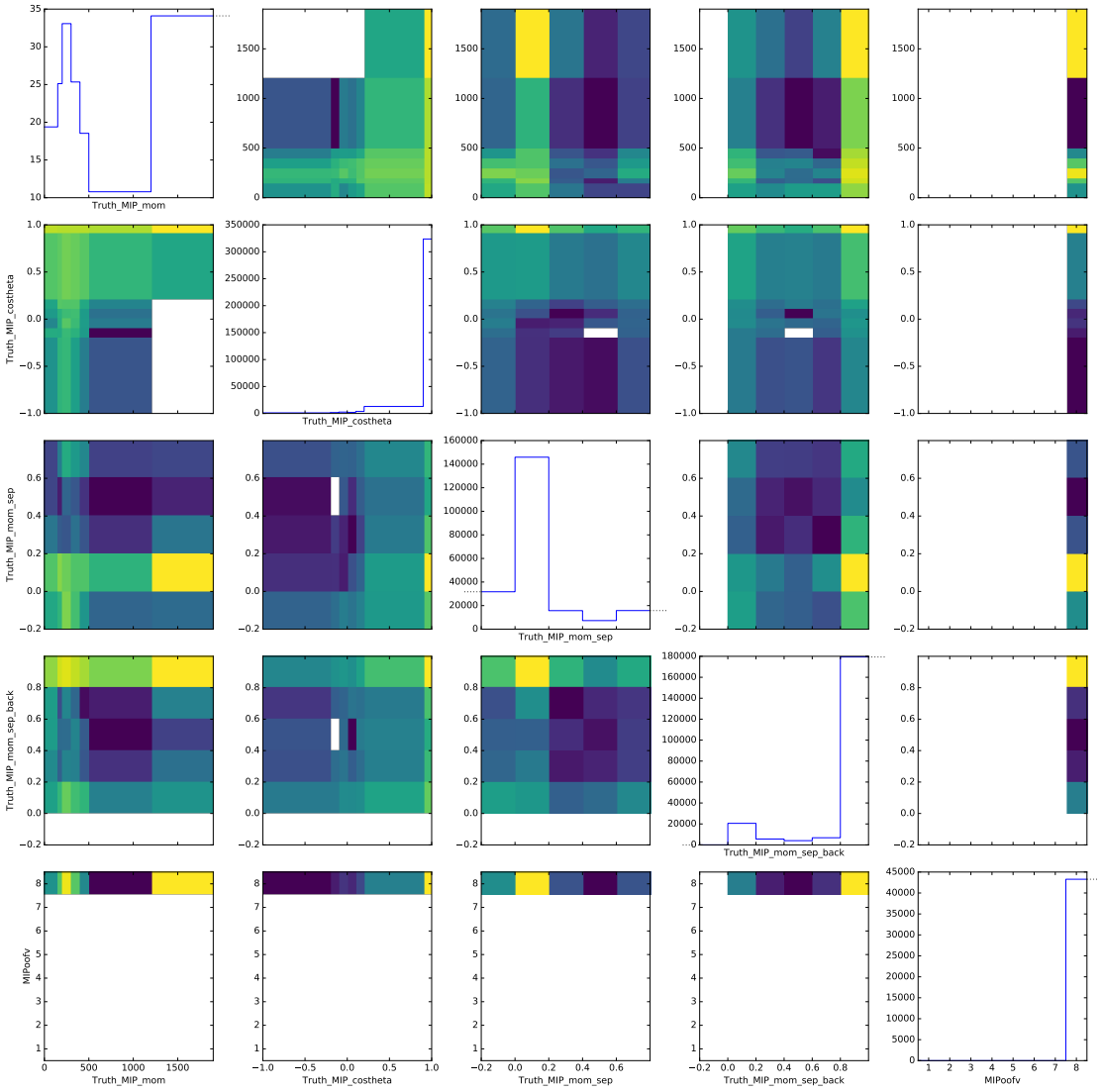


Figure B.19.: Neut n-track other OOFV template

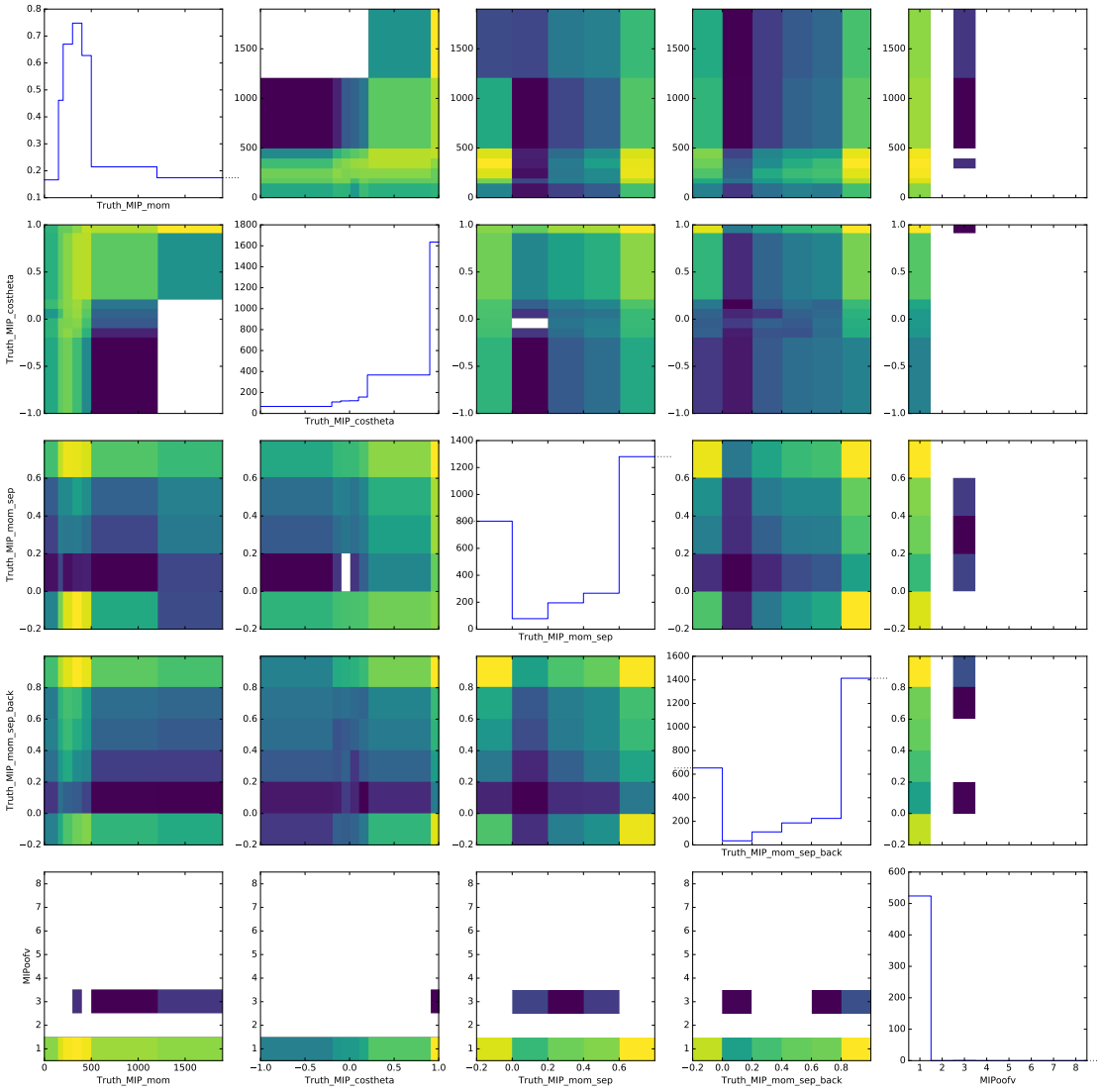


Figure B.20.: Genie ν_μ CC template

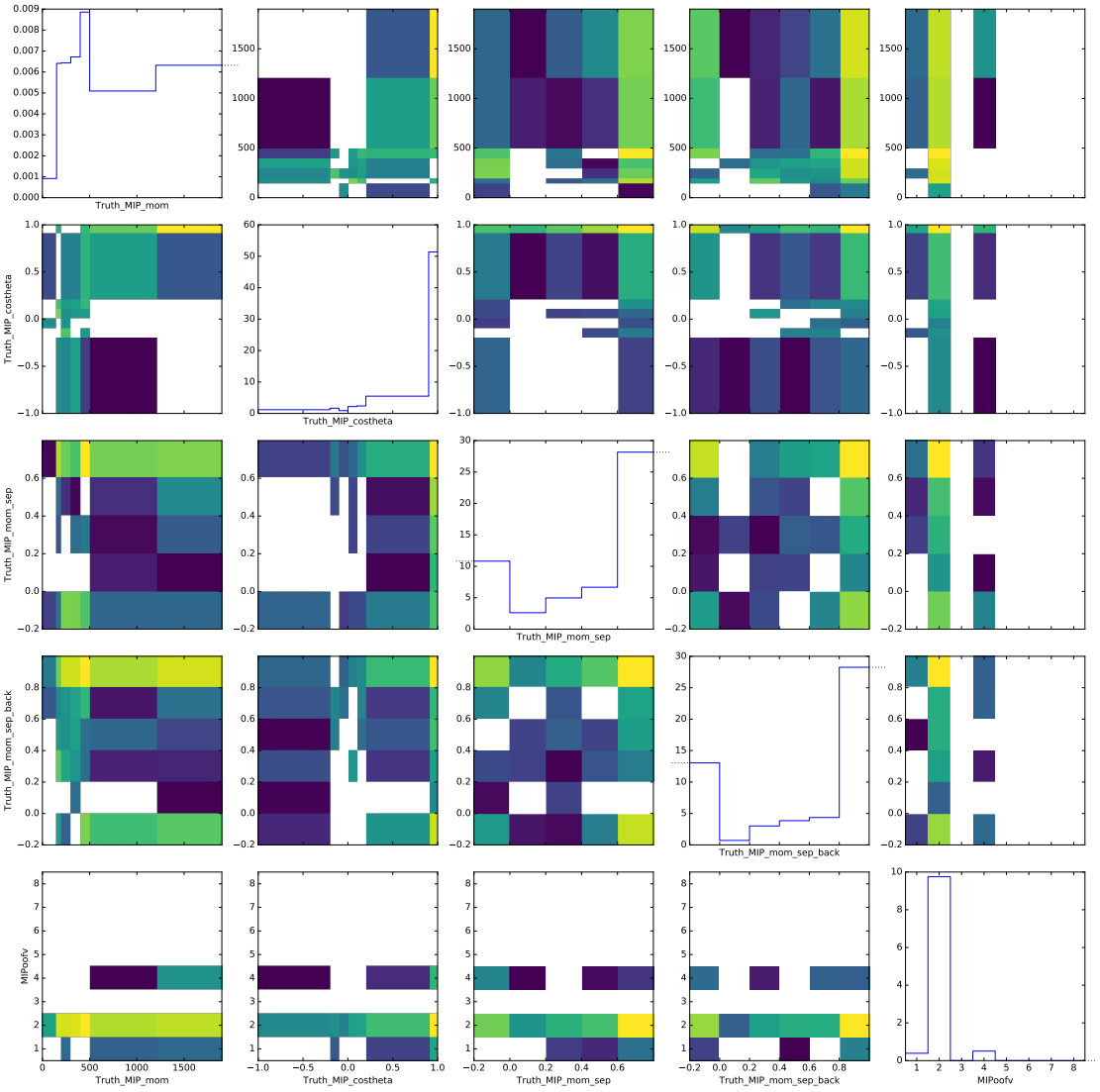


Figure B.21.: Genie $\bar{\nu}_\mu$ CC template

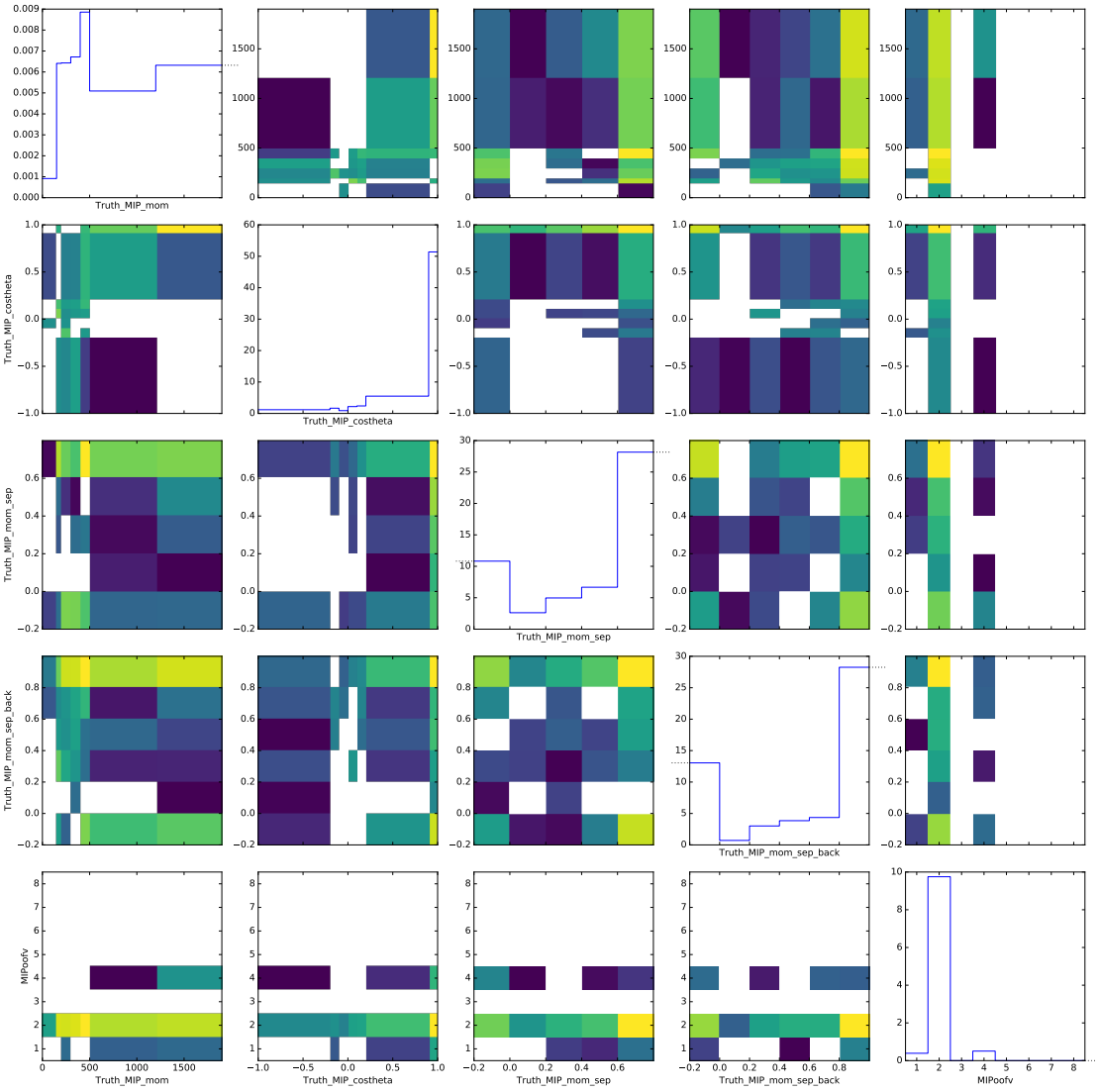


Figure B.22.: Genie $\nu_e/\bar{\nu}_e$ CC template

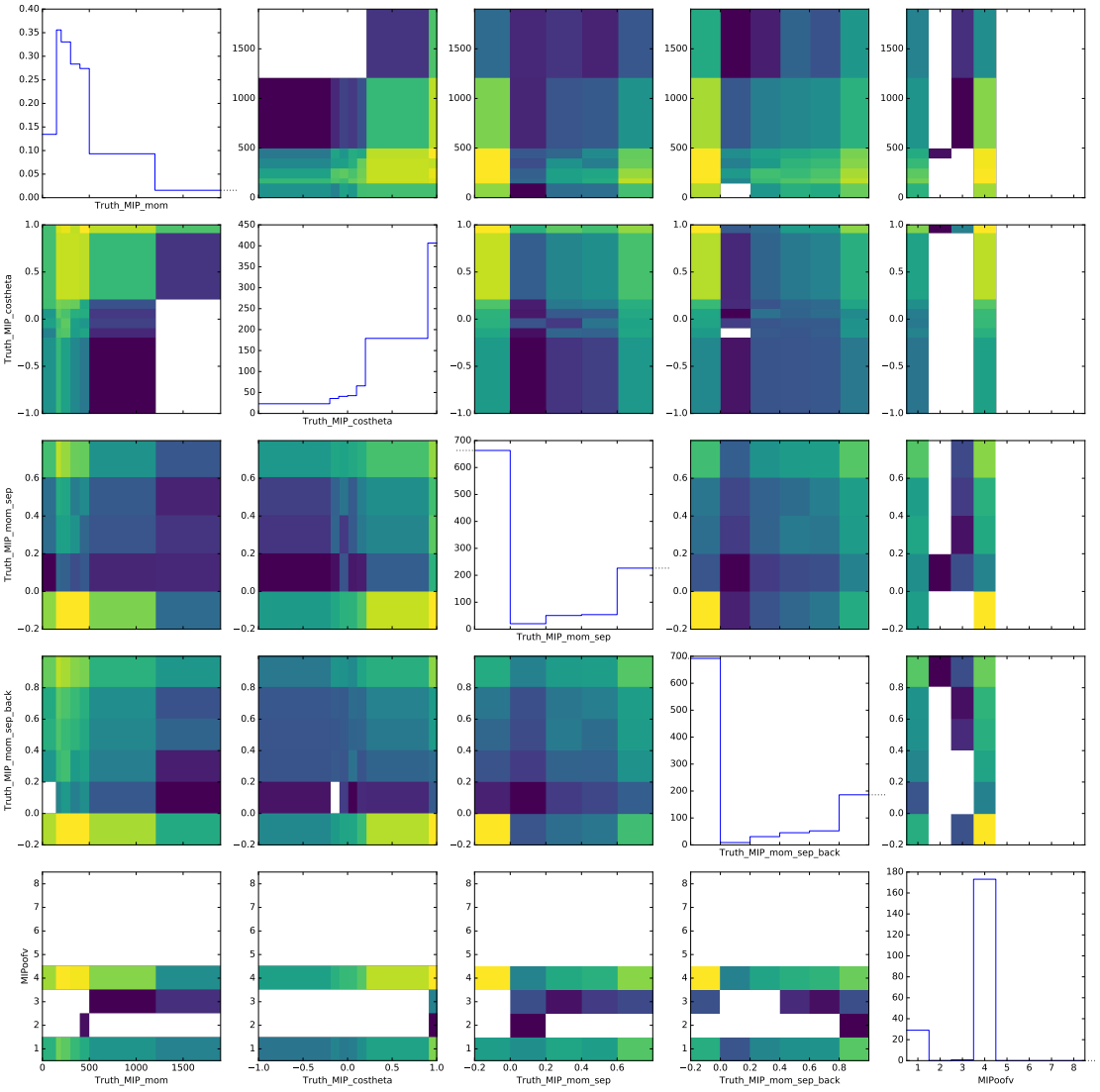


Figure B.23.: Genie NC template

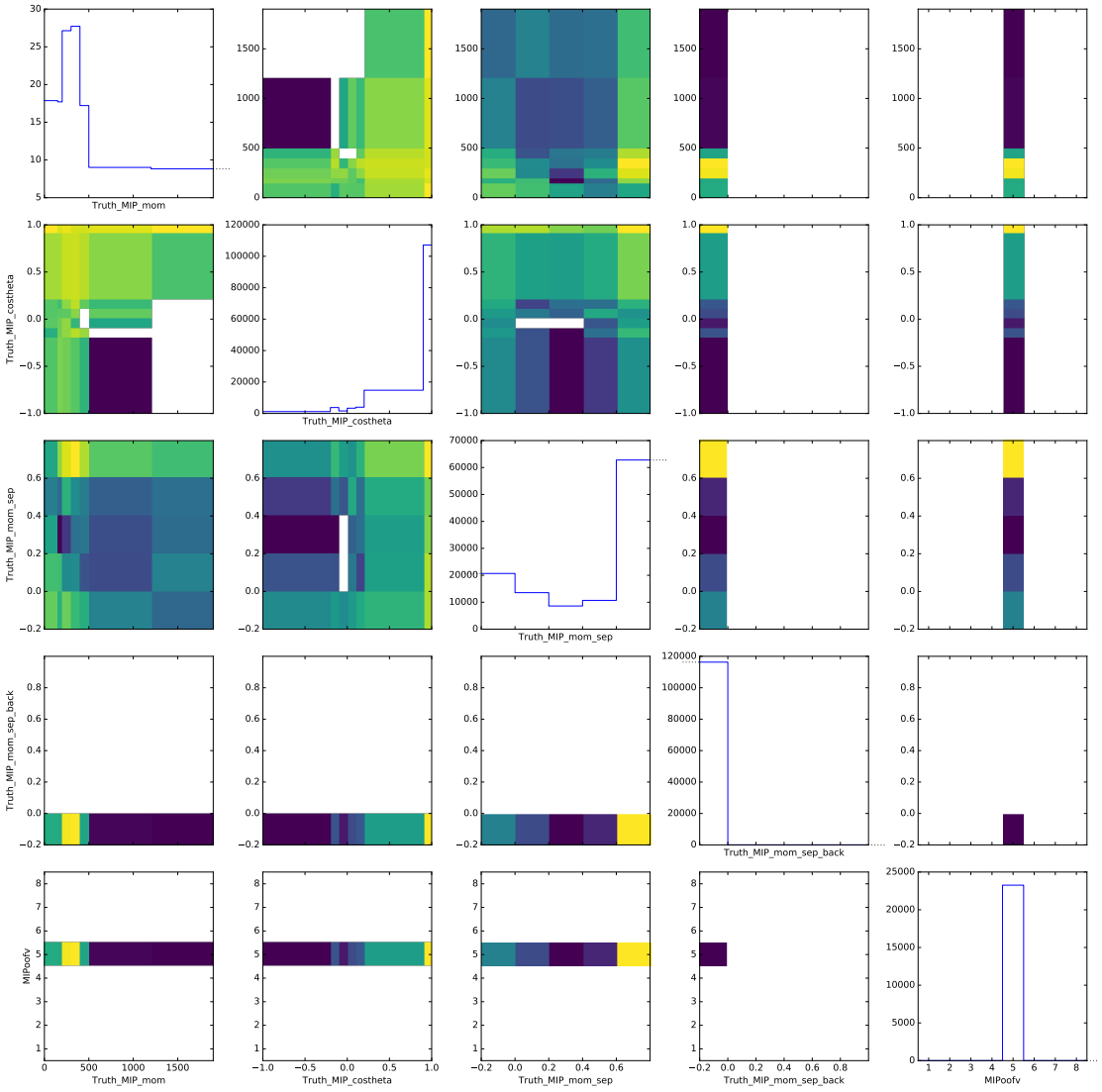


Figure B.24.: Genie 1-track OOFV delta-ray template

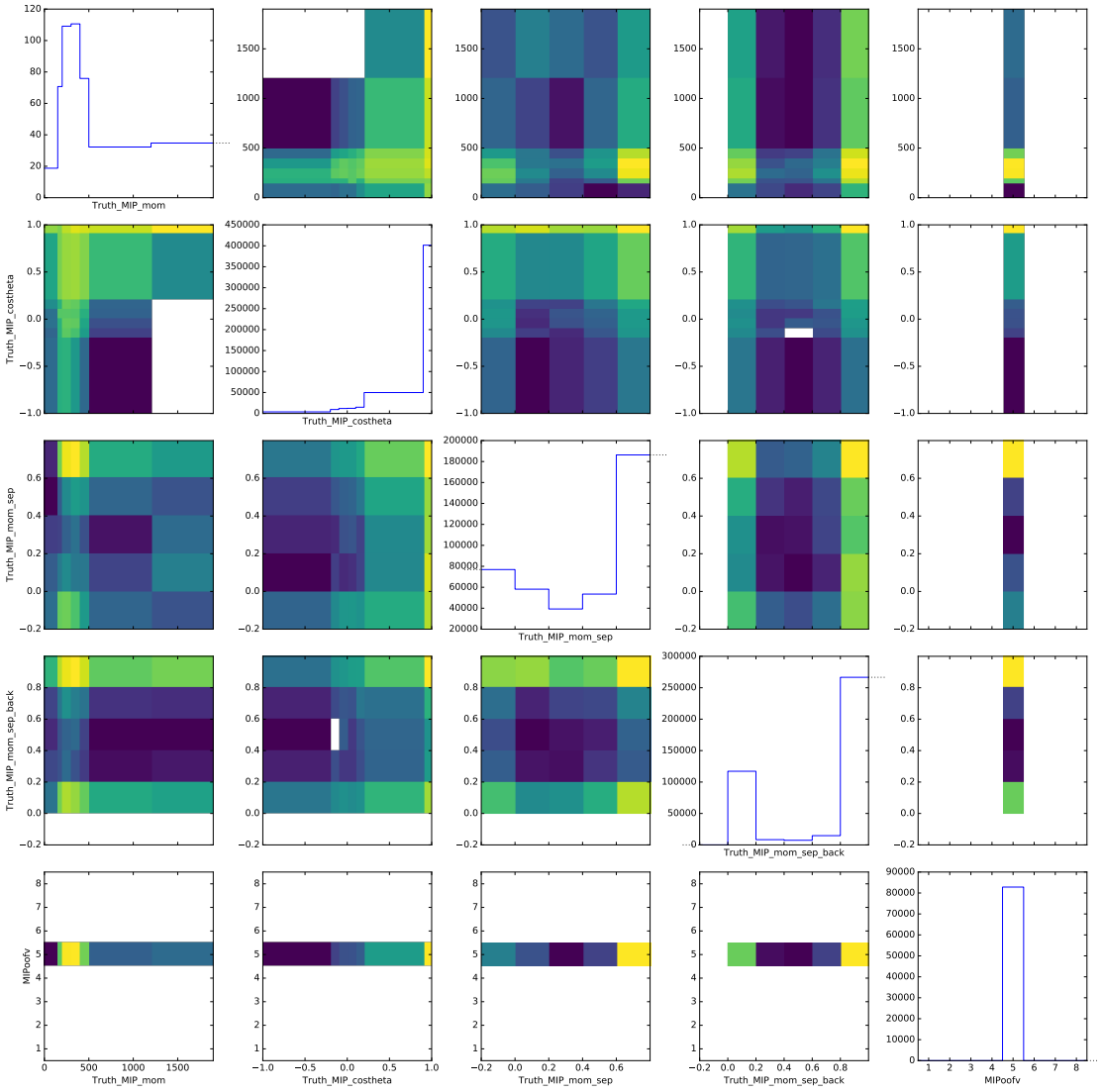


Figure B.25.: Genie n-track OOFV delta-ray template

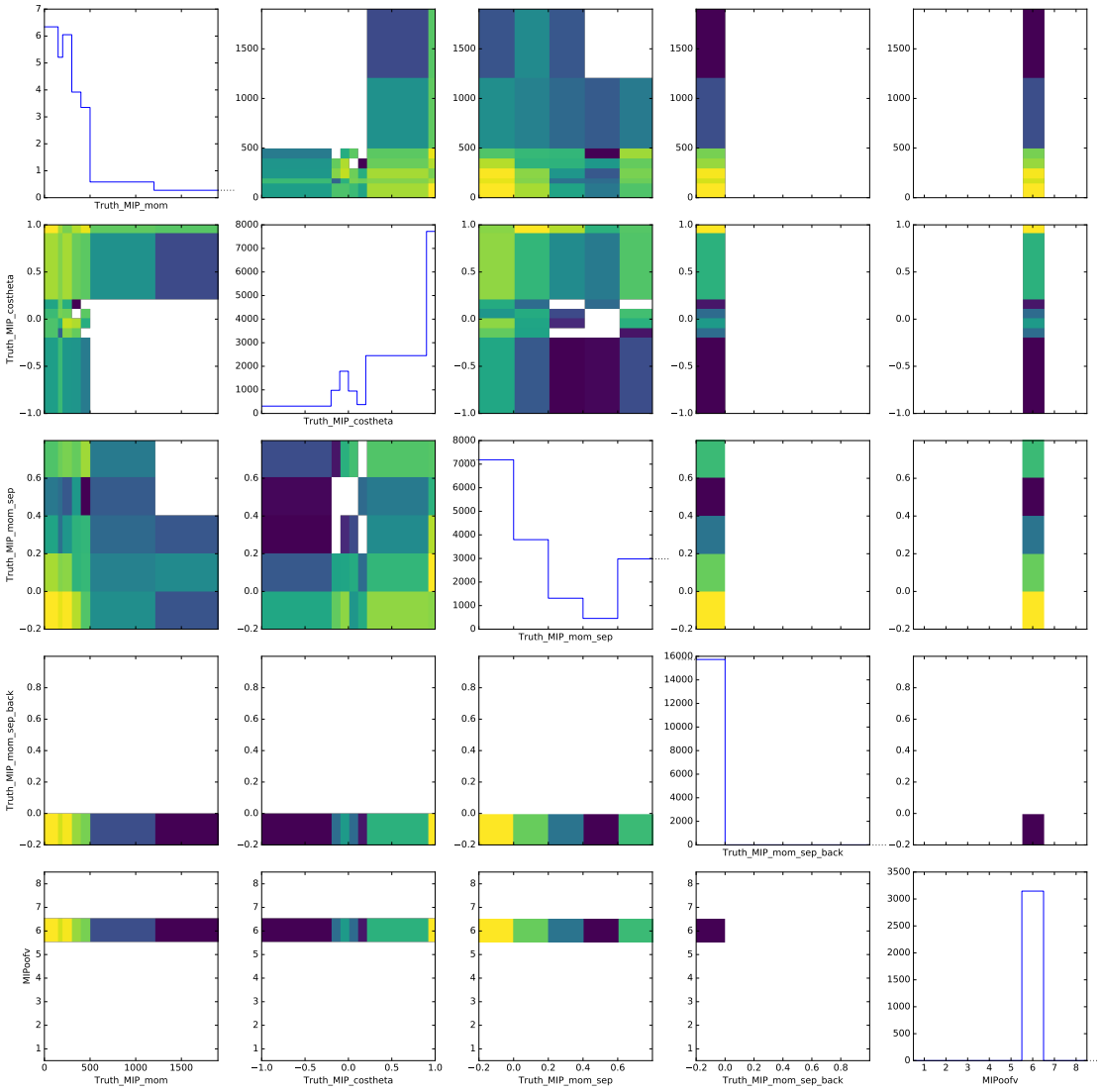


Figure B.26.: Genie 1-track OOFV decay template

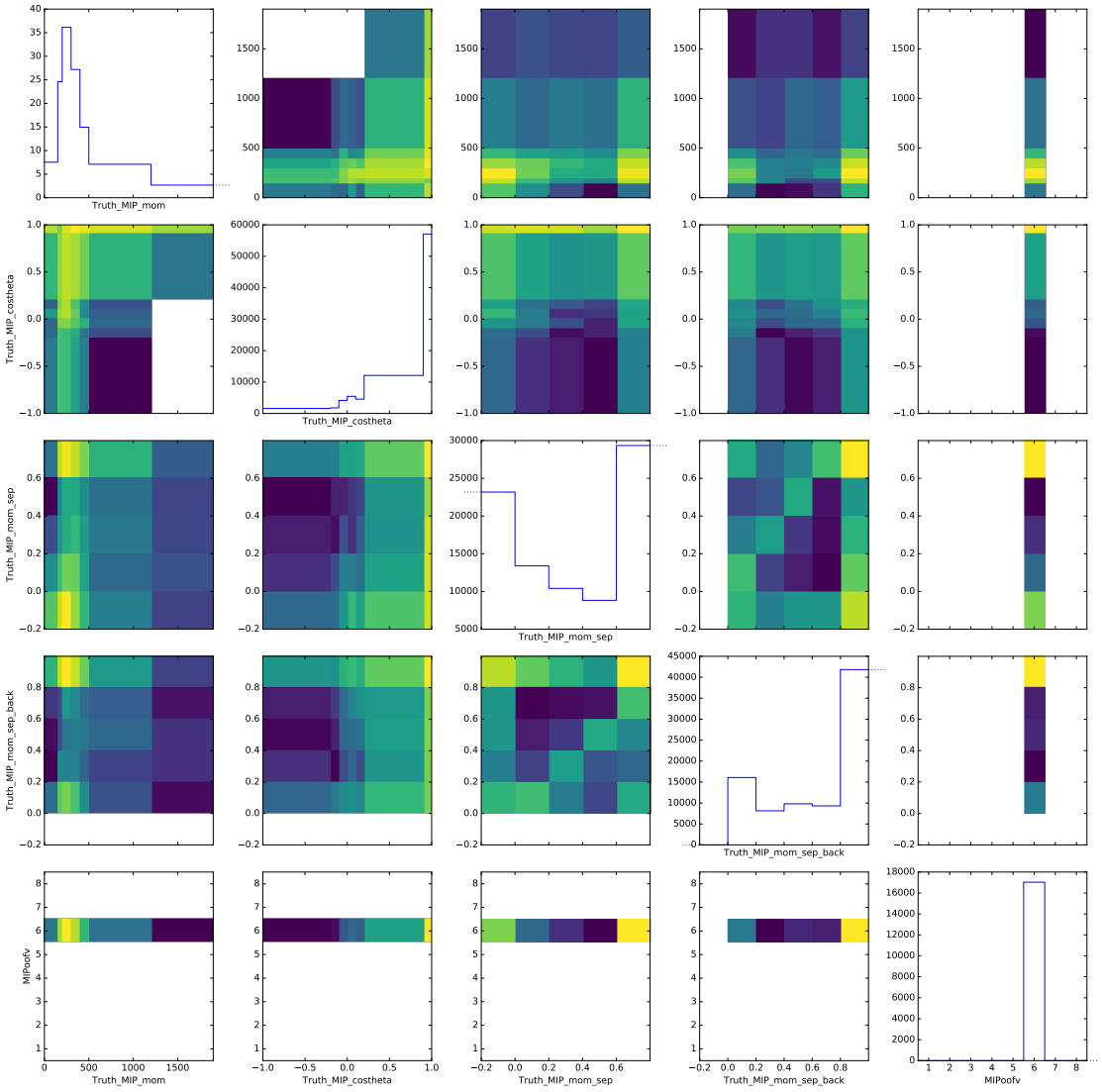


Figure B.27.: Genie n-track OOFV decay template

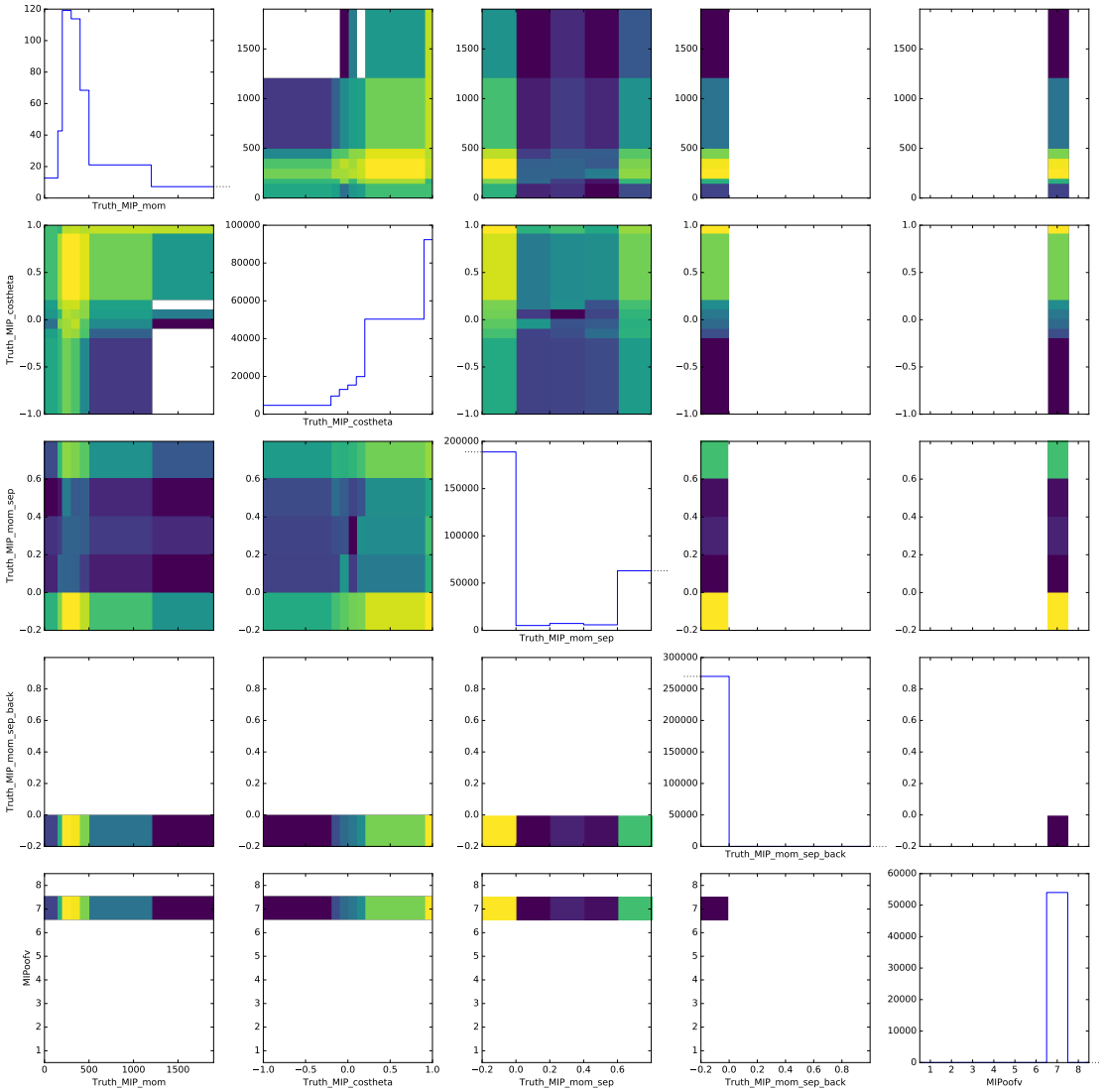


Figure B.28.: Genie 1-track timing-related OOFV template

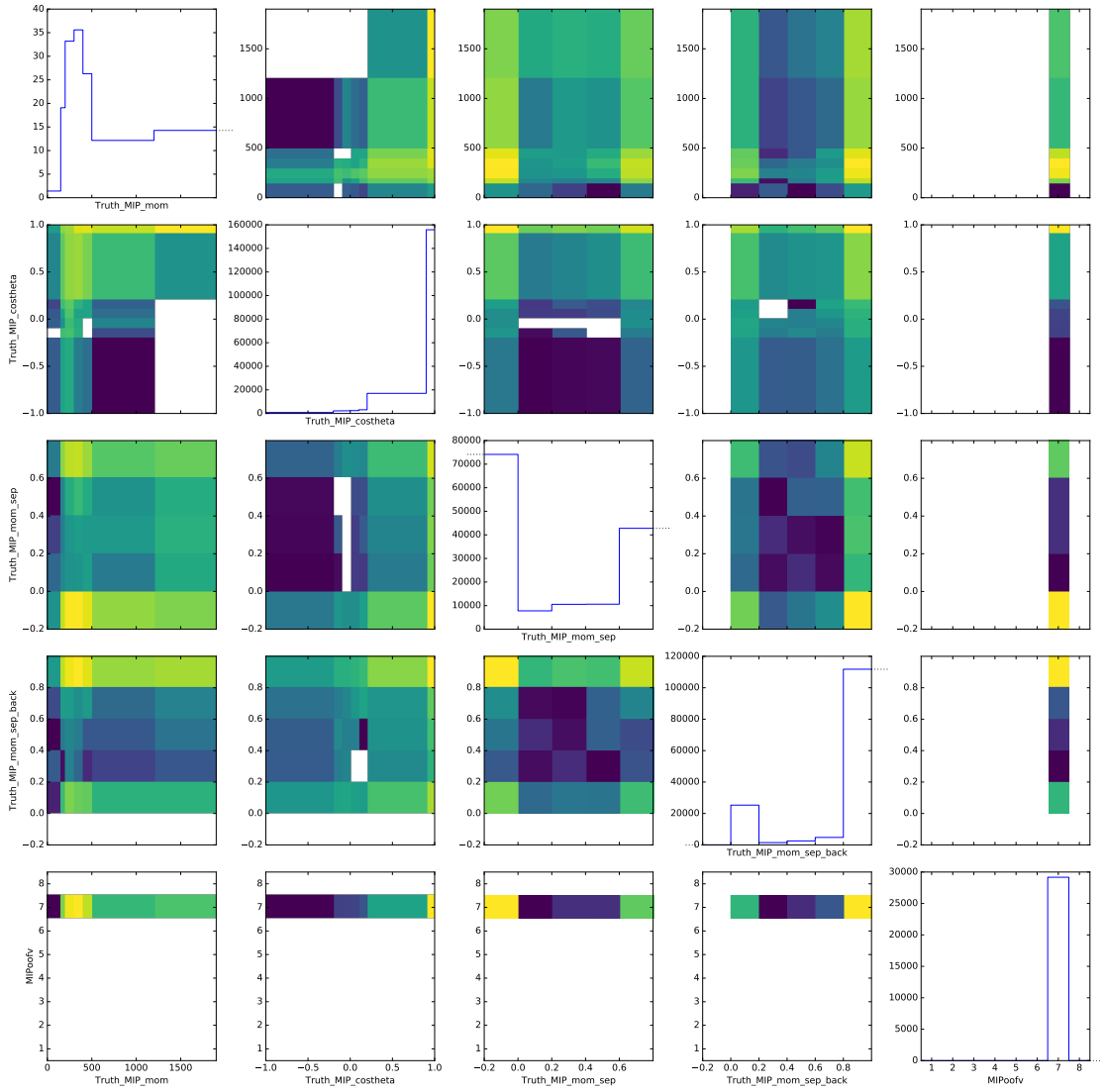


Figure B.29.: Genie n-track timing-related OOFV template

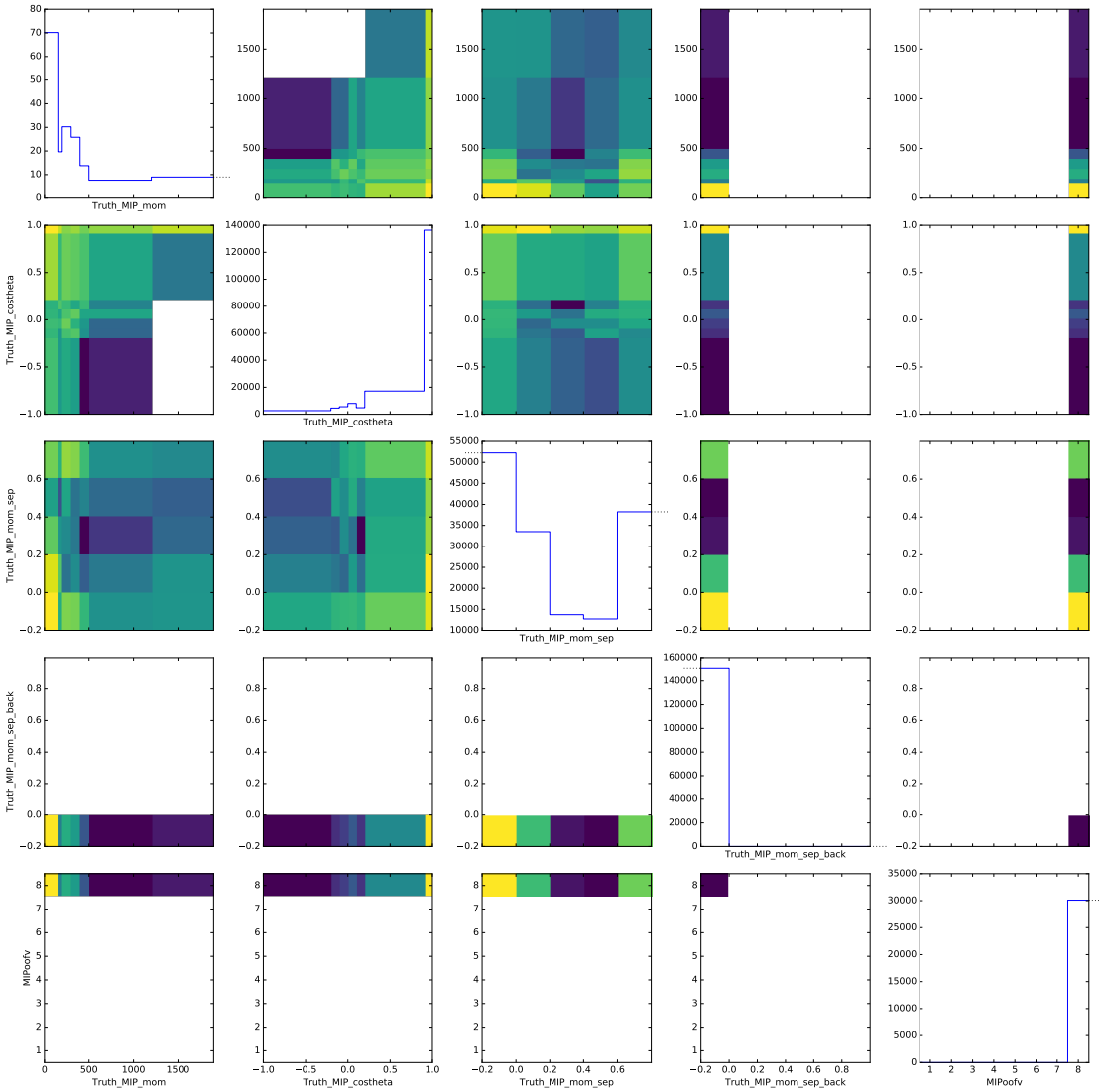


Figure B.30.: Genie 1-track other OOFV template

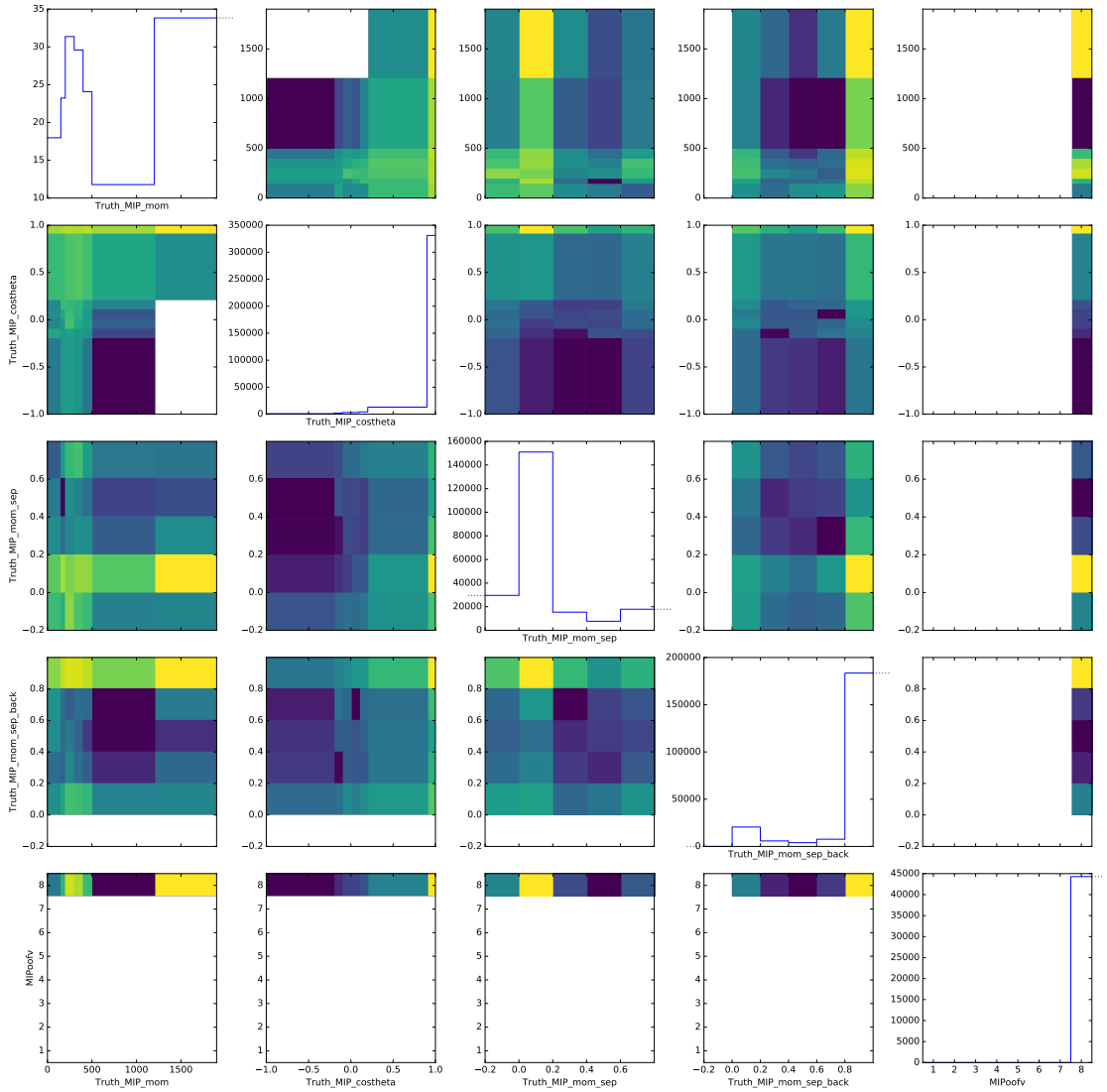


Figure B.31.: Genie n-track other OOFV template

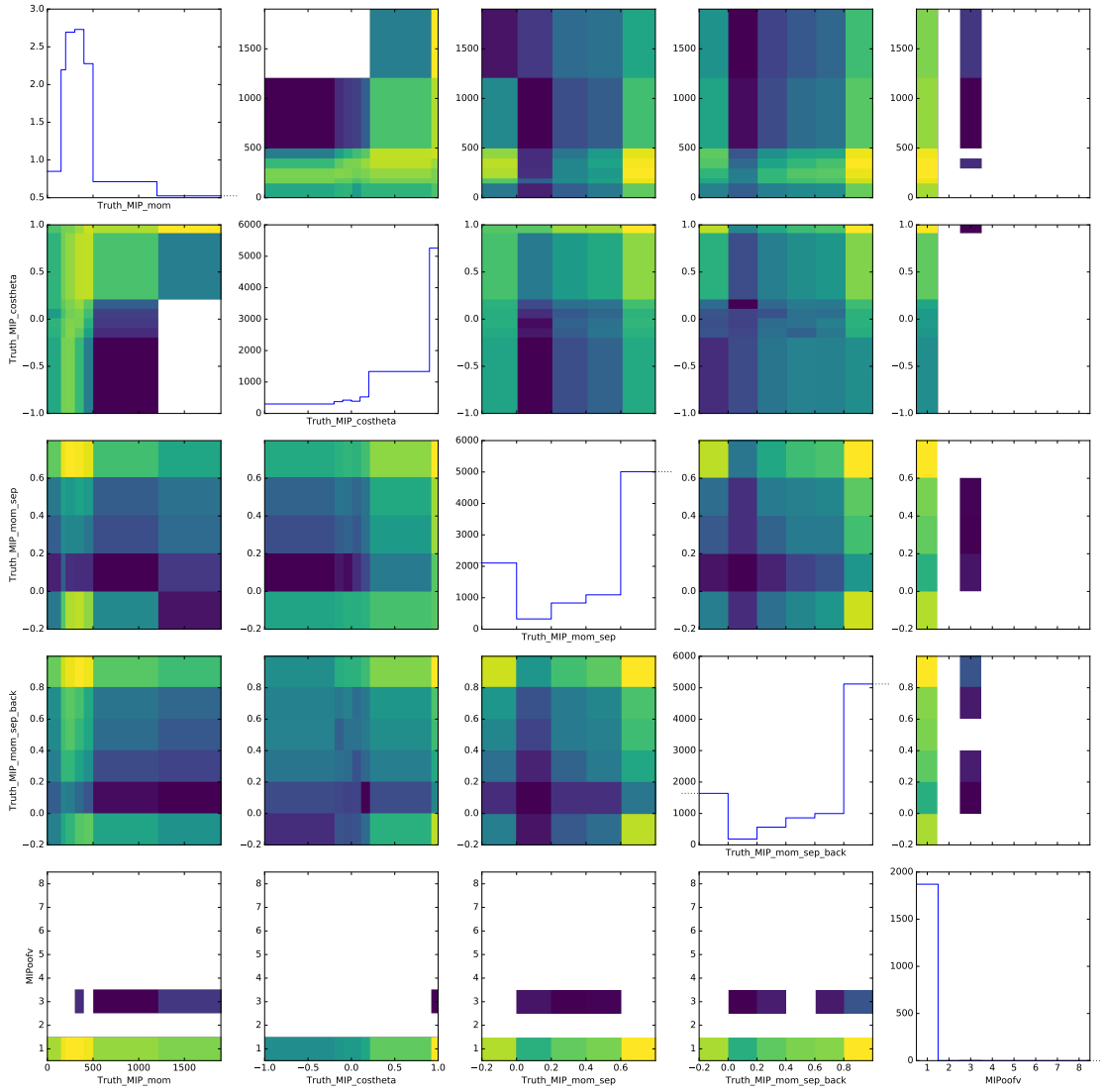


Figure B.32.: Crazy ν_μ CC template

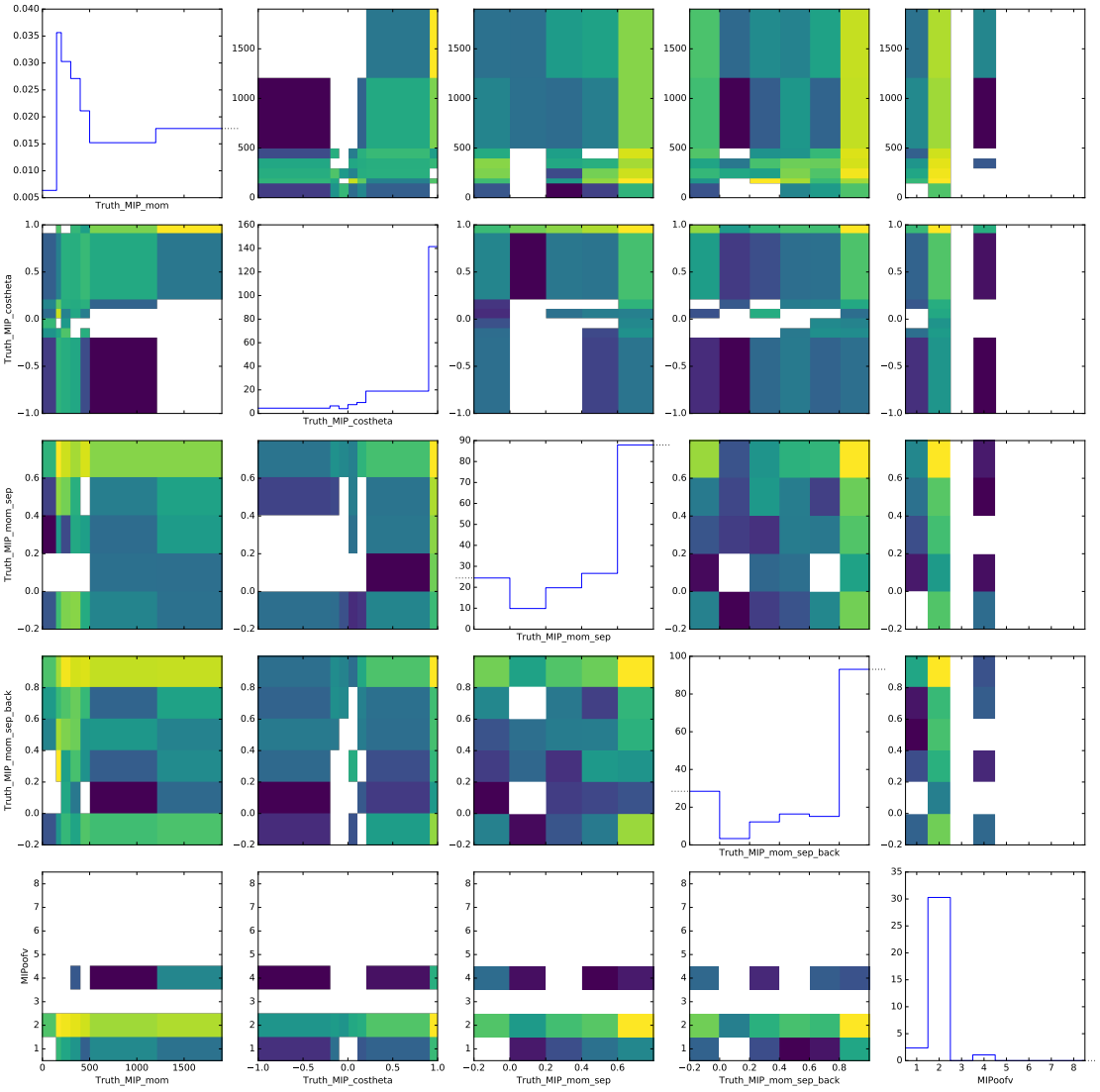


Figure B.33.: Crazy $\bar{\nu}_\mu$ CC template

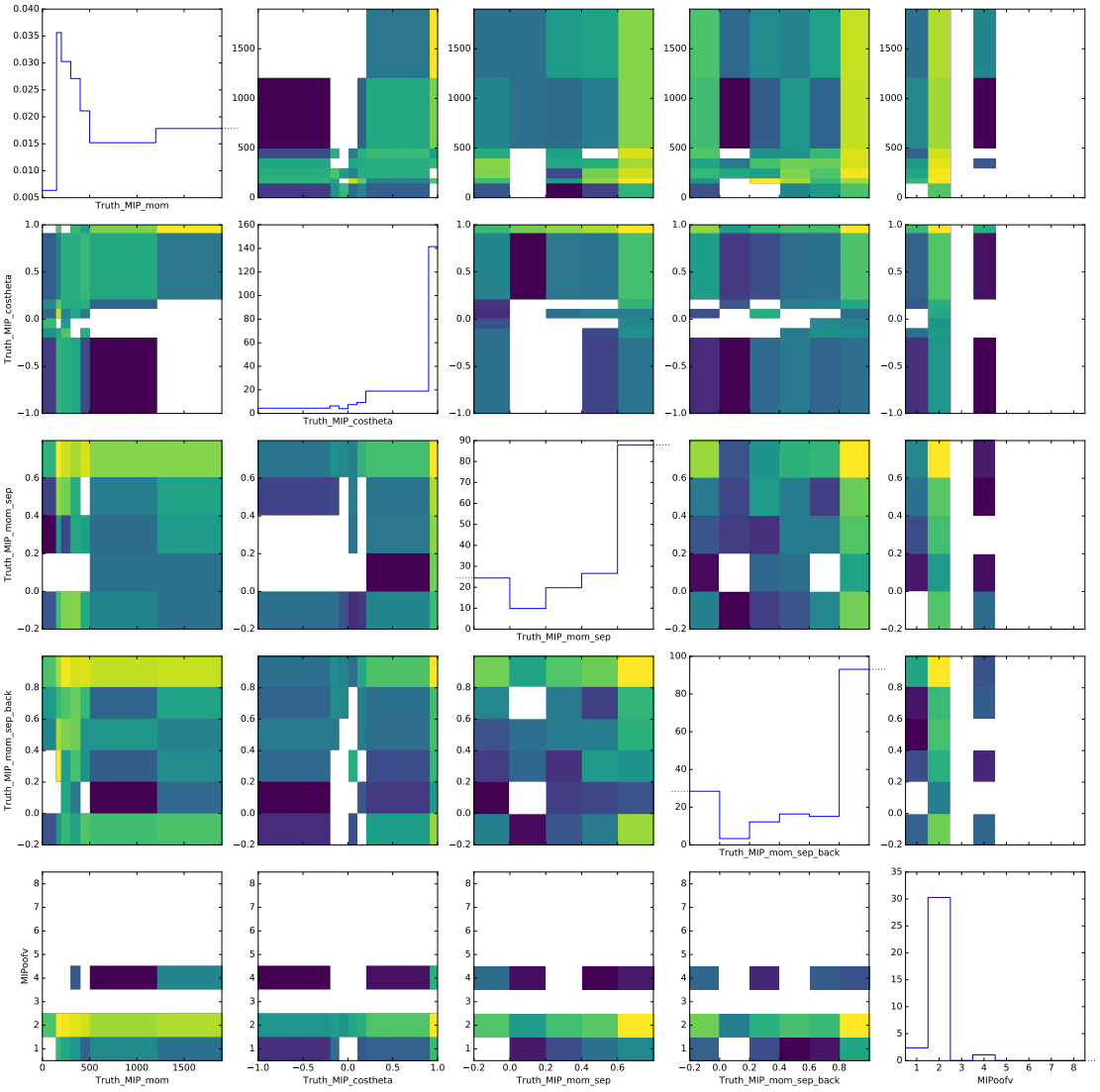


Figure B.34.: Crazy $\nu_e/\bar{\nu}_e$ CC template

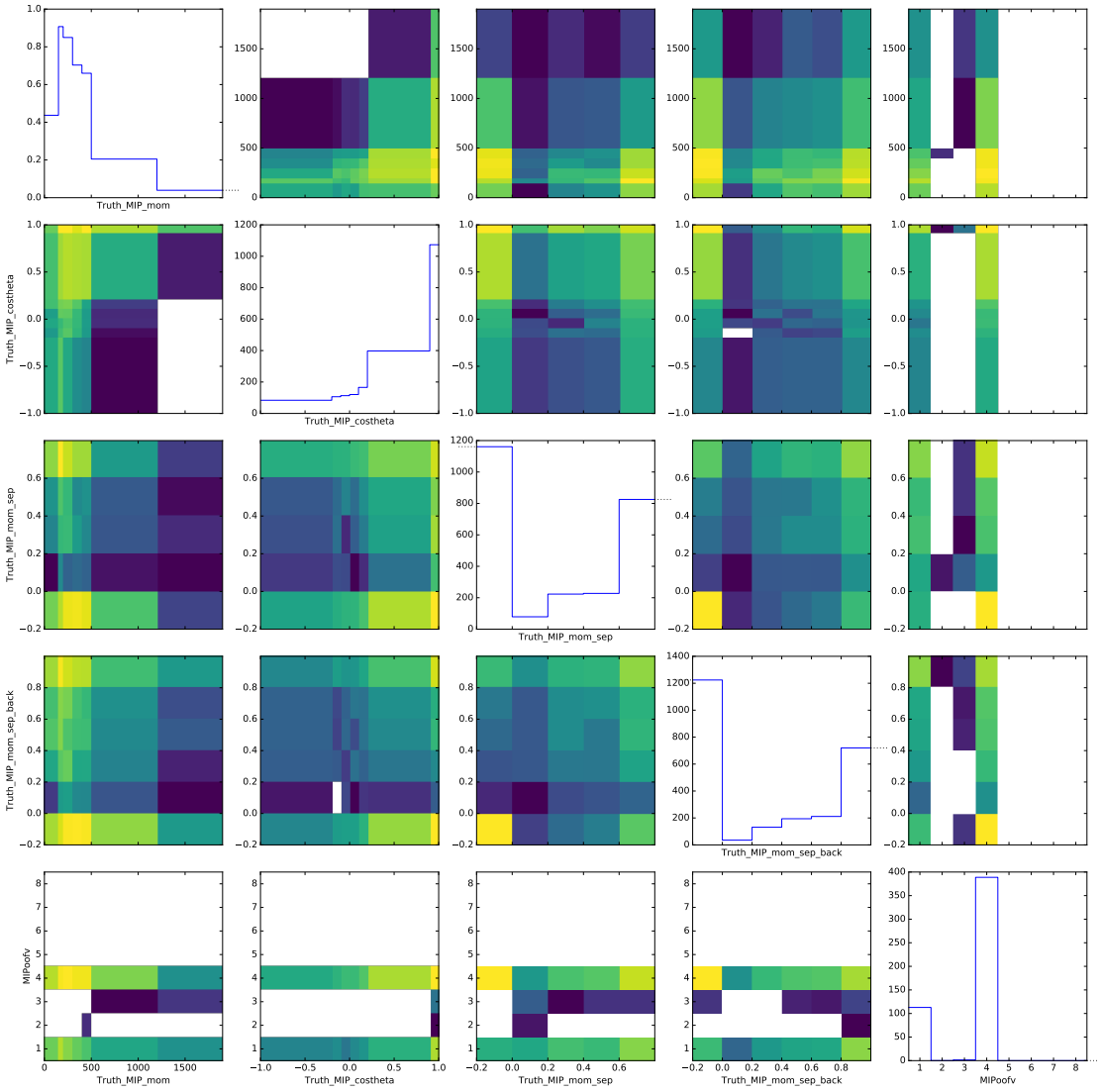


Figure B.35.: Crazy NC template

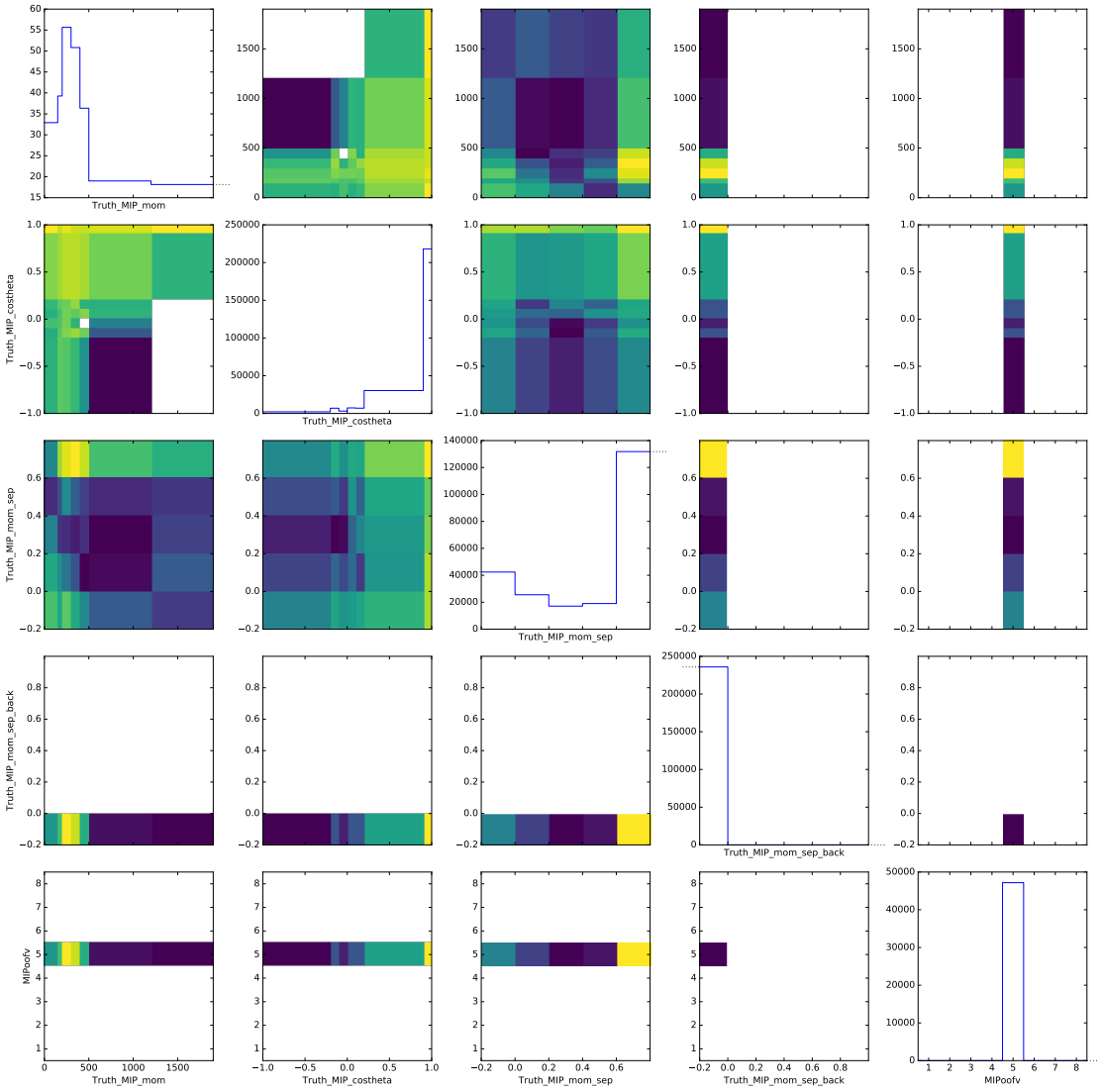


Figure B.36.: Crazy 1-track OOFV delta-ray template

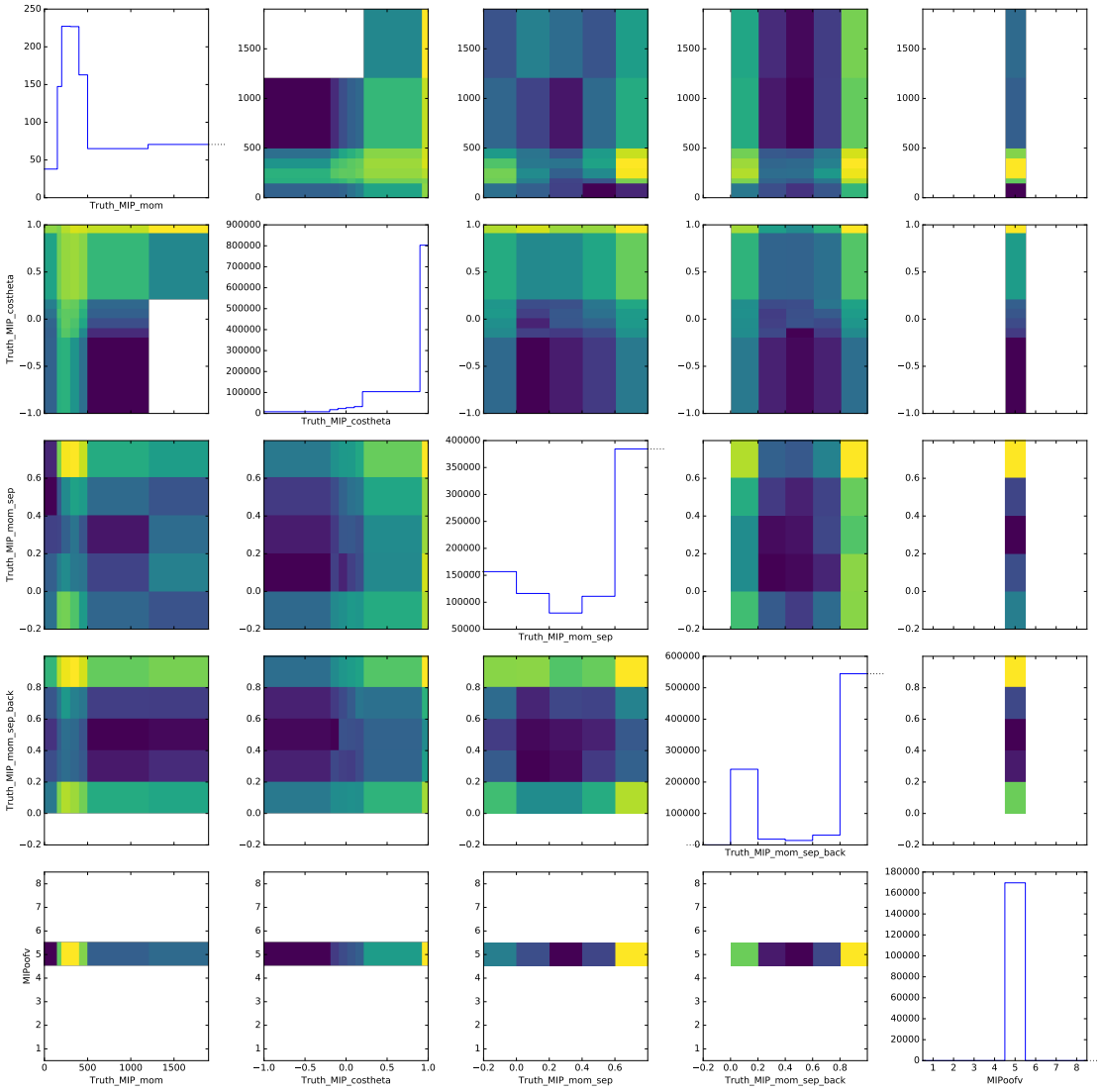


Figure B.37.: Crazy n-track OOFV delta-ray template

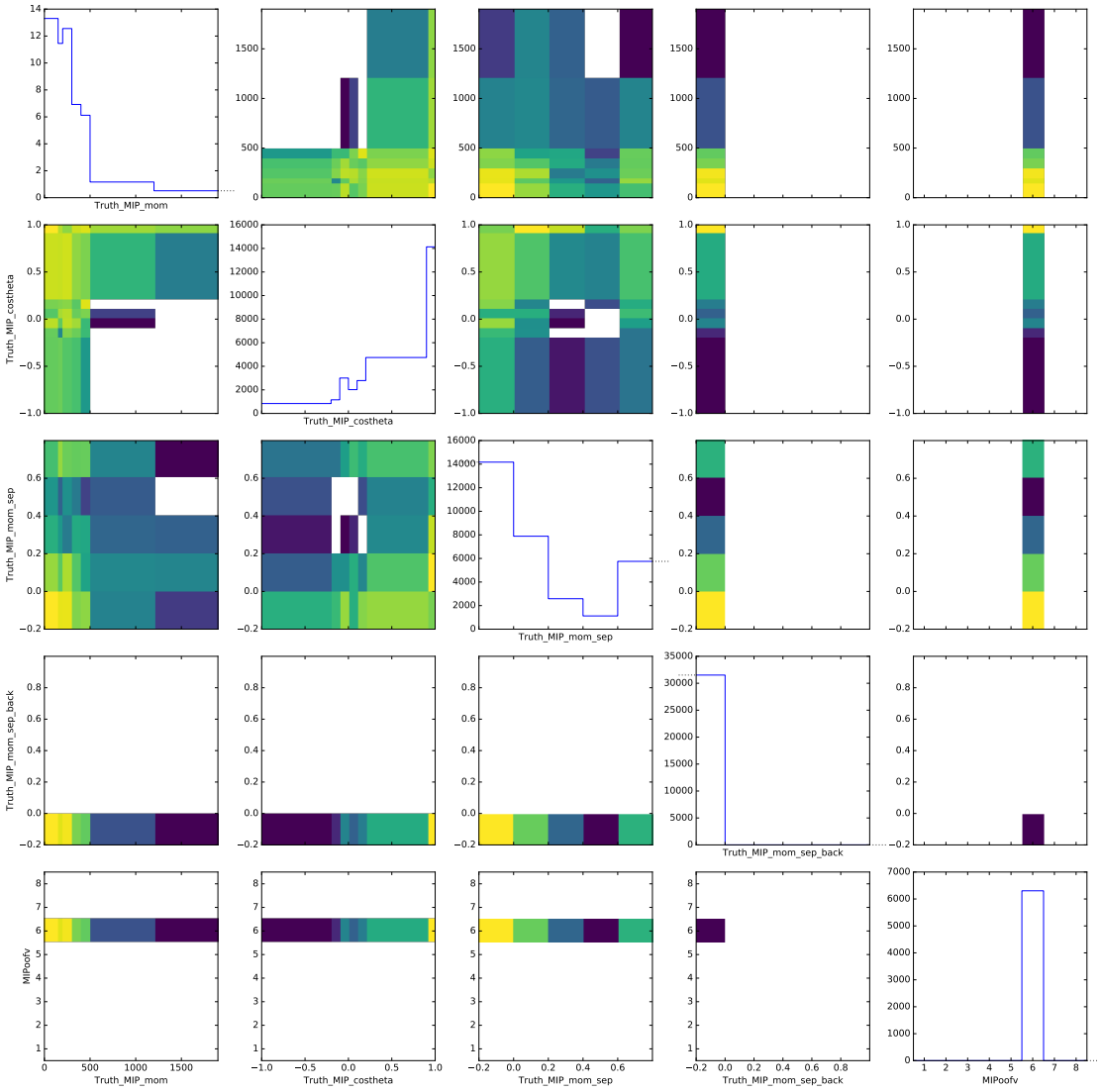


Figure B.38.: Crazy 1-track OOFV decay template

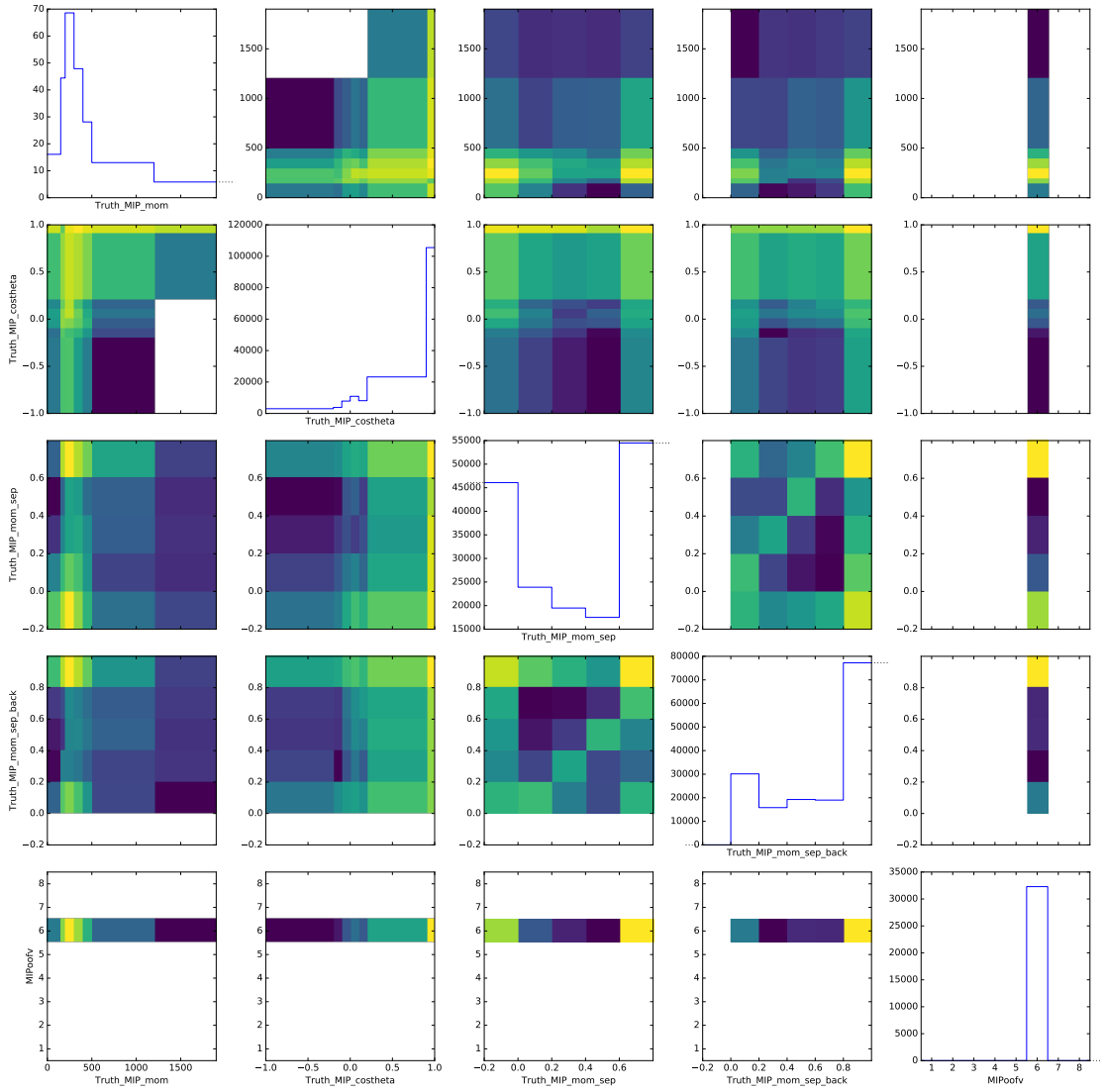


Figure B.39.: Crazy n-track OOFV decay template

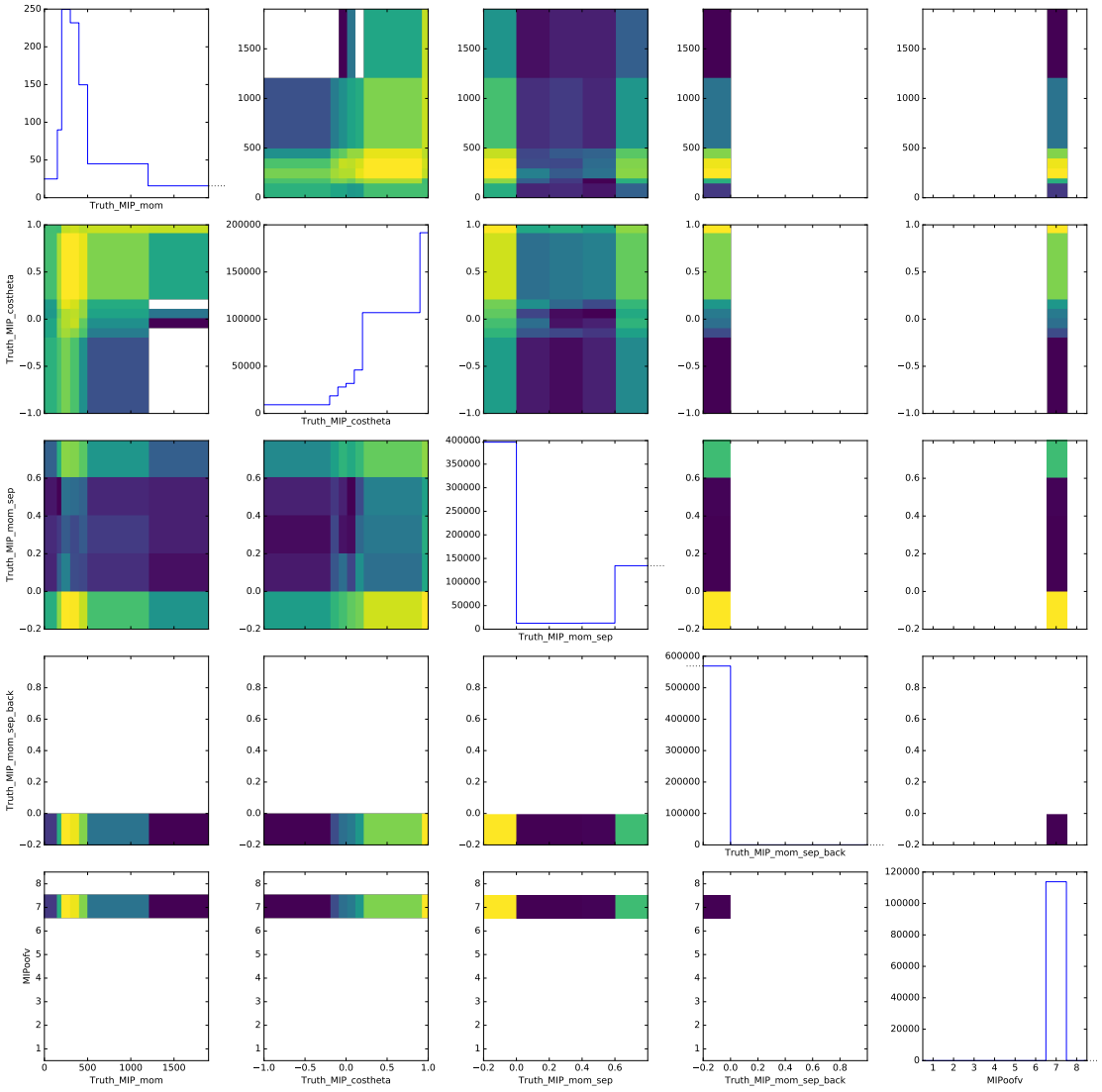


Figure B.40.: Crazy 1-track timing-related OOFV template

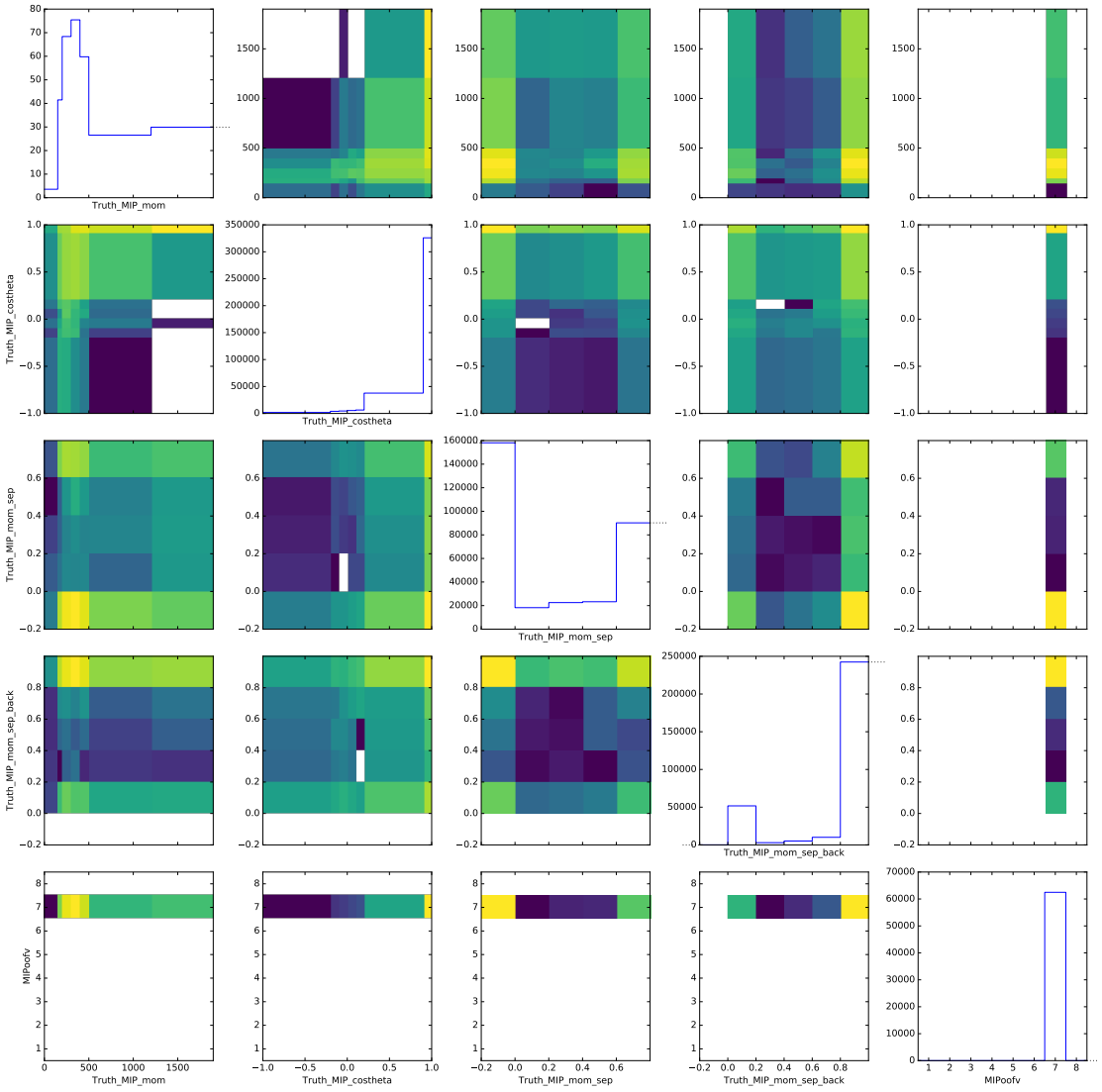


Figure B.41.: Crazy n-track timing-related OOFV template

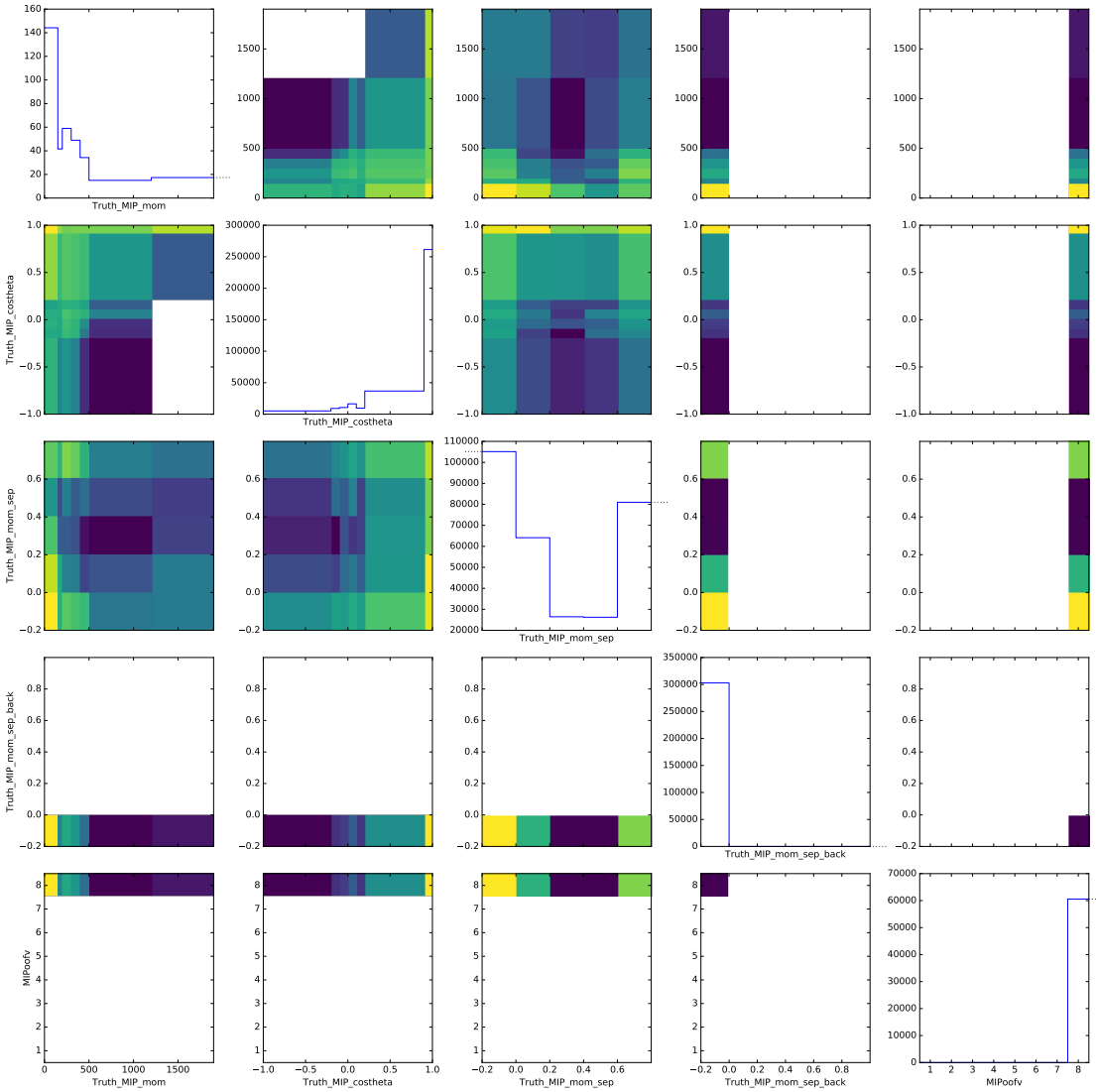


Figure B.42.: Crazy 1-track other OOFV template

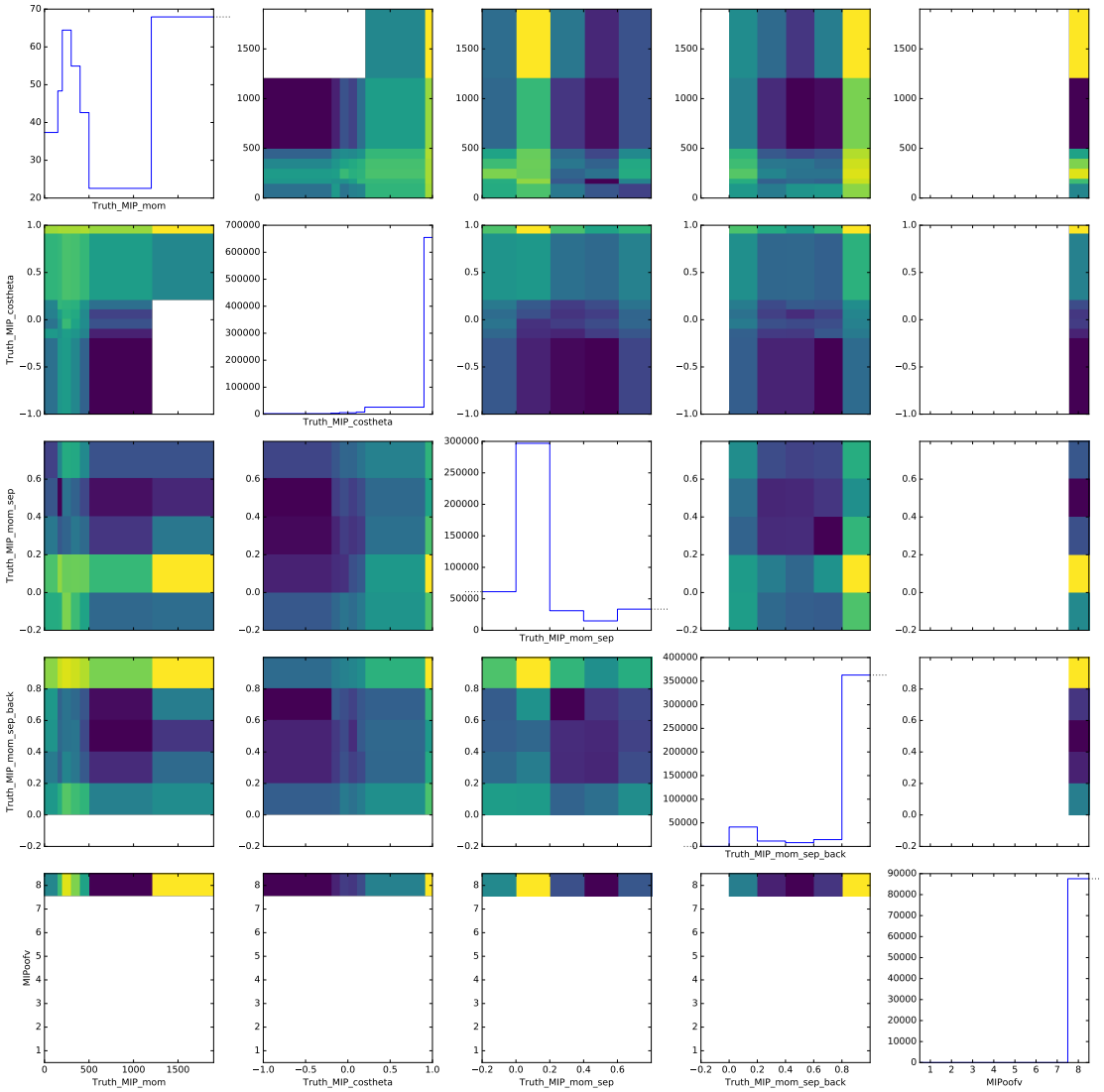


Figure B.43.: Crazy n-track other OOFV template