

# Machine learning assisted unfolding for neutrino cross-section measurements with the OmniFold technique

R. G. Huang<sup>1</sup>, A. Cudd<sup>2</sup>, M. Kawaue<sup>3</sup>, T. Kikawa<sup>3</sup>, B. Nachman<sup>1</sup>, V. Mikuni<sup>1</sup> and C. Wilkinson<sup>1</sup>

<sup>1</sup>*Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

<sup>2</sup>*University of Colorado Boulder, Boulder, Colorado 80309, USA*

<sup>3</sup>*Kyoto University, Department of Physics, Kyoto, Japan*



(Received 14 April 2025; accepted 19 June 2025; published 9 July 2025)

The choice of unfolding method for a cross-section measurement is tightly coupled to the model dependence of the efficiency correction and the overall impact of cross-section modeling uncertainties in the analysis. A key issue is the dimensionality used in unfolding, as the kinematics of all outgoing particles in an event typically affect the reconstruction performance in a neutrino detector. OmniFold is an unfolding method that iteratively reweights a simulated dataset, using machine learning to utilize arbitrarily high-dimensional information, that has previously been applied to proton-proton and proton-electron datasets. This paper demonstrates OmniFold's application to a neutrino cross-section measurement for the first time using a public T2K near detector simulated dataset, comparing its performance with traditional approaches using a mock data study.

DOI: [10.1103/sp1f-n9k2](https://doi.org/10.1103/sp1f-n9k2)

## I. INTRODUCTION

Unfolding is a general term for removing the impact of a measuring device from a measurement [1,2]. In the context of cross-section measurements in high-energy physics, this can be thought of as the deconvolution of detector smearing and efficiency effects, necessary to present data in a way which does not require detailed information about the particle detector to interpret. Neutrino cross-section measurements typically use binned unfolding methods and a small number of dimensions to present results. Most current results are single-dimensional, but multidimensional results of up to three dimensions have been extracted [3]. A simple, but representative formula used to extract a binned cross-section  $d\sigma/dx_i$  as a function of the true kinematic variable  $x$  is [4]

$$\frac{d\sigma}{dx_i} = \frac{\sum_j \tilde{U}_{ij}^{-1} (N_j - B_j)}{\Phi_\nu T \Delta x_i \epsilon_i}, \quad (1)$$

where  $N_j$  ( $B_j$ ) is the number of selected (predicted background) events in reconstructed bin  $j$ ,  $\Phi_\nu$  is the total integrated flux,  $T$  is the number of target nuclei,  $\Delta x_i$  is the width of the true bin  $i$  (or area/volume for multidimensional extractions),  $\epsilon_i$  is the selection efficiency, and  $\tilde{U}_{ij}^{-1}$  is the

unfolding matrix. The unfolding matrix is a regularized inverse of the matrix  $U_{ij}$  that encapsulates the effects of detector smearing. The efficiency and the detector smearing are closely related, and in some formulations, the efficiency correction is absorbed into the smearing matrix. A number of unfolding methods have been explored in neutrino physics [5–7], between which the most important differences are how the result is regularized and how the potential for bias that regularization introduces is handled [4,8]. One of the most commonly used unfolding methods in neutrino physics, as in high-energy physics more generally, is D'Agostini unfolding [9,10], also known as Iterative Bayesian Unfolding (IBU).<sup>1</sup> This method has been independently described multiple times [11–13] and is known in other fields by other names, such as Lucy-Richardson deconvolution [14,15].

Experimental data from neutrino experiments have a number of unique challenges: they typically suffer from low statistics; detector volumes are large with nonuniform detector efficiency and smearing; they use relatively low-fidelity, high-threshold detectors; and systematic uncertainties are generally large. These issues couple to the unfolding problem in complex ways, which risks degrading the quality of extracted results. Because the detector smearing and efficiency depend on a large number of variables, expressing a smearing matrix with a small number of dimensions implicitly integrates over all of

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP<sup>3</sup>.*

<sup>1</sup>Although this is somewhat of a misnomer, as it is a fully frequentist method [11].

the other relevant variables. This integration relies not only on the detector model in the simulation but is convolved with a prediction for how signal events will be distributed in those variables, leading to model dependence and the potential for bias [4,8]. Additionally, by expressing the result in a small number of dimensions, the result loses much of the richness of the original data, particularly as the binning must be determined in advance using simulation studies. Experiments often extract results in a number of variables, but generally without correlations between the different variables. This has presented difficulties in the use of these measurements, although methods have been proposed to overcome this problem [7]. Additionally, the challenges of unfolding have led to discussions of whether to unfold data at all, both in neutrino physics [16] and more generally [17].

New unfolding methodologies are possible using machine learning, many of which naturally generalize to high-dimensional spaces, and a number of approaches have been explored [18–20]. In this work, we apply the OmniFold unfolding method [21,22] to a publicly available simulated dataset from the T2K experiment and compare its performance and relative advantages with respect to IBU.

This paper is organized as follows. Section II briefly introduces machine learning-based unbinned unfolding with OmniFold. The T2K dataset used for numerical results is described in Sec. III. Our methods for testing different approaches are documented in Sec. IV. Numerical results are presented in Sec. V, and the paper ends with conclusions and outlook in Sec. VI.

## II. UNBINNED UNFOLDING WITH OMNIFOLD

Unfolding methods using neural networks and other machine learning tools process continuous data and are thus naturally unbinned [18–20]. As these tools can also process high-dimensional feature spaces, unfolding many observables simultaneously is readily accommodated. Existing approaches are based on likelihood-ratio estimation with machine learning classifiers [21–23] or are based on direct likelihood-estimation with generative machine learning tools [24–36]. The last couple of years have seen the first unbinned cross-section studies in  $ep$  using H1 data [37–39] and in  $pp$  using data from LHCb [40], ATLAS [41,42], STAR [43,44], and CMS [45,46]. Since all of the existing experimental results use the OmniFold method, we focus on that technique here.

OmniFold is an iterative, two-step procedure that is illustrated in Fig. 1. It exploits the manner in which each event from a simulated dataset will have both “particle-level” and “detector-level” features. Particle-level features refer to the truth-level list of particles that resulted from an interaction along with their associated kinematics, provided by some kind of event generator. Detector-level features refer to the reconstructed quantities one obtains after pushing the particle-level features through the detector

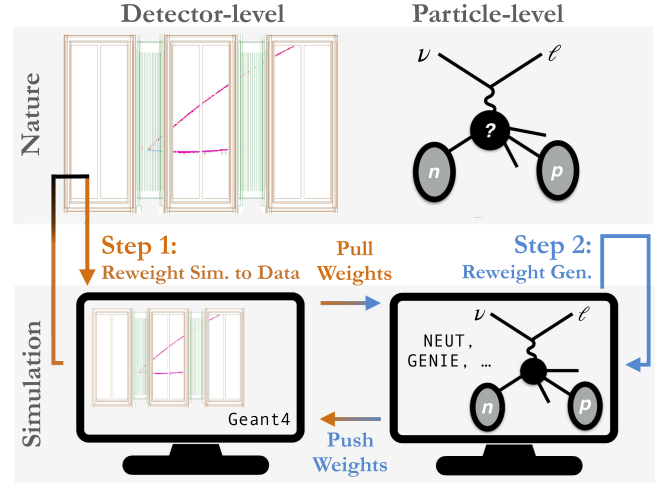


FIG. 1. A sketch of the OmniFold method applied to T2K data and simulation, with an example of how a typical charged current neutrino interaction (right) appears in the T2K near detector (left). Step 1 works with detector-reconstructed quantities, comparing simulated detector events against data and reweighting the former to fit the latter. These weights are pulled to step 2, which compares truth-level generator events against themselves with and without the pulled weights to provide reweighting of the true particle-level events.

simulation and reconstruction algorithms, and these are the quantities one has access to in an actual experiment. In the first step of OmniFold, the detector-level simulated data are reweighted to match the observed data. Then, these weights are “pulled back” to the particle level by assigning the weight obtained from the detector level in the first step. Events that do not pass the detector-level event selection are assigned the average weight for a given particle-level phase space region either at the end of the unfolding or at each iteration (comparisons described below). This pullback induces a new spectrum at particle level, but since the detector response is stochastic, the resulting reweighting is not a proper function of the particle-level phase space. The second step of OmniFold reweights the starting particle-level simulation to this induced simulation from the pulled-back weights. This weighting function is a proper function of the particle-level phase space by definition. One can then “push forward” these weights to the detector level and repeat the entire process. Iterating a finite number of times is a form of regularization, and there are a number of metrics that can be used to decide when to stop. The final result of the unfolding is a set of particle-level simulated events and a corresponding set of event weights.

While the reweighting can be performed with any method, the algorithm is unbinned when based on machine learning classifiers. Interpreting classifiers as likelihood ratio estimators, sometimes called the “likelihood ratio trick,” is well known in statistics [47,48] and has been frequently used in particle physics [49].

### III. T2K PUBLIC DATASET

This paper utilizes a public simulated T2K dataset [50] based on a previous analysis of T2K ND280 near detector data [51], which was a measurement of charged-current  $\nu_\mu$  events that produce zero pions in the final state (referred to as the  $CC0\pi$  topology) on a hydrocarbon target. The same dataset or subsets of it has been used in other related T2K publications [6,52,53]. The public simulated dataset represents signal and select background events in both truth and reconstructed space along with associated systematic uncertainties for the flux, detector, and neutrino interaction models. The simulated dataset covers T2K runs 2, 3, 4, and 8 in neutrino mode beam running and contains events corresponding to  $17.24 \times 10^{21}$  protons on target, roughly a factor of 20 more than the data statistics analyzed in Ref. [51]. The ND280 simulation uses the NEUT package [54–56] to generate neutrino interactions in the detector, GEANT4 [57] for particle propagation and energy deposition, and a custom package to simulate the detector electronics response.

ND280 comprises a number of subdetectors installed inside the refurbished UA1/NOMAD magnet, which provides a 0.2 T field used to measure the charge and momentum of particles. The inner tracker region of ND280 uses three time projection chambers (TPCs) interleaved with two plastic scintillator fine-grained detectors (FGDs) and is surrounded by an electromagnetic calorimeter. The FGDs each contain 1.1 tons of target mass for neutrino interactions and provide tracking of the charged particles coming from the interaction vertex, while the TPCs provide momentum and particle identification. The dataset used here is focused on neutrino interactions occurring on the hydrocarbon in the first FGD.

Each event in the dataset includes the energy-momentum 4-vector for the leading muon and, if it exists, the leading proton in both truth and reconstructed space. For events that have been reconstructed, a detector sample identification (ID) is also included based on the subdetectors used to measure the muon candidate (and proton candidate, if any), defined as one of

- (1)  $\mu TPC$ . Only a reconstructed muon in the TPC;
- (2)  $\mu TPC + p TPC$ . One muon and one proton in the TPC;
- (3)  $\mu TPC + p FGD$ . One muon in the TPC and one proton in the FGD;
- (4)  $\mu FGD + p TPC$ . One muon in the FGD and one proton in the TPC;
- (5)  $\mu FGD$ . One muon in the FGD;
- (6)  $CC1\pi$ . One muon and one  $\pi^+$  in the TPC;
- (7)  $CCOther$ . One muon and multiple pion tracks in the TPC; and
- (8)  $CCMichel$ . One muon and a michel tagged pion in the FGD.

The final three sample IDs here are intended as background-enhanced samples and are used to help constrain

backgrounds to the  $CC0\pi$  analysis. For events in the truth space, a topology ID is included, indicating what observable final-state particles result from the interaction, defined as one of:

- (1)  $CC0\pi0p$ . Only an outgoing muon from the charged current interaction;
- (2)  $CC0\pi1p$ . One muon and one proton;
- (3)  $CC0\pi Np$ . One muon and two or more protons
- (4)  $CC1\pi$ . One muon and one charged pion; and
- (5)  $CCOther$ . Other events with a muon (neutral pion or multiple pion production).

Events are designated as signal or background at truth level by these topology IDs, with 1–3 serving as signal and 4–5 serving as background. Because of limitations of the data release, this is not the full list of backgrounds typically included in a T2K analysis. In particular, it omits contributions from neutral current interactions,  $\nu_e$  interactions with erroneous particle identification, wrong-sign events from  $\bar{\nu}$  interactions, and out-of-fiducial-volume interactions.

The dataset also contains information about systematic uncertainties associated with the flux, detector, and neutrino interaction models for the analysis. This comes in the form of a vector of 500 weight variations for each individual event that represent coherent random throws of these uncertainties. However, the detector systematic uncertainties do not include dedicated proton reconstruction effects, since the proton kinematics were not part of the original analysis associated with this data release.

### IV. TEST SETUP

We create a set of “fake data” to serve as the unfolding target by reweighting a subset of the available events in the dataset with a modified charged-current quasielastic (CCQE) cross-section based on the random-phase approximation (RPA) calculation from Ref. [58]. This fake data represent a nontrivially different signal model from that used in the simulated dataset, which did not include this weak-charge screening effect caused by the nuclear medium. The modification is performed by applying an event weight to CCQE events as a function of the true  $Q^2$  based on a parametrization of the RPA calculation using Bernstein polynomials, which is described in detail in Ref. [59]. The original set of non-reweighted events and weights then serve as the simulation for the unfolding procedure. Truth-level information for the fake data is not used during the unfolding procedures, but it is brought in for evaluation of the performance of each method in the final stage. For convenience, in the rest of this article, we will refer to the fake data simply as “data,” as during unfolding it is treated the same way that real data would be. Code used in the following analysis is publicly available [60].

For the events designated as simulation, we generate 500 throws of the event weights that simultaneously account for detector, cross-section, and flux systematic uncertainties as

well as Monte Carlo statistical uncertainties from the number of generated events. The weight variations from systematic effects are taken from the T2K public simulated data release. Monte Carlo statistical uncertainties are obtained by bootstrap resampling the available events with replacement. We similarly create 500 throws of the statistical uncertainties in the observed data by applying to it a weighted bootstrap resampling, where each data event with weight  $w$  receives a resampled weight  $w_{\text{resampled}} = \text{Poisson}(\lambda = w)$ . This reduces the number of data events to the same statistics as Ref. [51] and simulates the corresponding statistical uncertainty of that result. The unfolded results are then obtained by a toy universe method, where each of the 500 throws is independently run through the unfolding procedure. The resulting ensemble of 500 unfolded truth distributions gives our systematic and statistical uncertainties on the result. To maintain comparability between the unfolding methods under consideration, we use the same set of 500 variations for each of the procedures.

We consider four observables for evaluating the performance of each method:

- (i)  $(p_\mu, \cos \theta_\mu)$ . The two-dimensional muon kinematics defined by its momentum and forward angle, which were used in the differential cross-section analysis associated with the original data release [51];
- (ii)  $\delta p_T \equiv |\vec{p}_T^\mu + \vec{p}_T^p|$ . The transverse momentum imbalance between the muon and leading proton;
- (iii)  $\delta \alpha_T \equiv \arccos \frac{-\vec{p}_T^\mu \cdot \delta \vec{p}_T}{p_T^\mu \delta p_T}$ . The boosting angle between the muon and leading proton in the transverse plane; and
- (iv)  $\delta \phi_T \equiv \arccos \frac{-\vec{p}_T^\mu \cdot \vec{p}_T^p}{p_T^\mu p_T^p}$ . The angular difference between the muon and leading proton in the transverse plane compared to the case where they are back to back.

In the definitions of the latter three variables,  $\vec{p}_T^\mu$  and  $\vec{p}_T^p$  are, respectively, the projections of the momenta of the muon and leading proton into the plane transverse to the incident neutrino direction. These variables are sometimes referred to as single transverse variables (STVs) and have been used to measure transverse kinematic imbalances in neutrino-nucleus interactions [6,61]. A visualization of the physical meaning of the STVs is provided in Fig. 2. The STVs are not calculated for events where there is no leading proton, which can happen either because the proton was not reconstructed or because no proton was kicked out at truth level.

While OmniFold is an unbinned procedure, we must choose some binning of the variables for final evaluation and comparison against classical binned techniques. The two-dimensional binning of the muon kinematics is summarized in Table I, and the binning of the STVs is summarized in Table II. These binnings were chosen in previous T2K analyses [6,51] based on the detector resolution and efficiency within the relevant kinematic phase spaces as well as the expected statistics. For each of the binnings, we extract a differential cross-section from

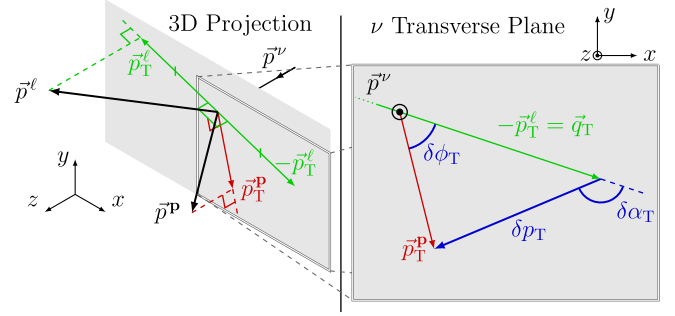


FIG. 2. Schematic view of the three single transverse variables for a quasilastic neutrino interaction producing a lepton with momentum  $\vec{p}^\ell$  and a proton with momentum  $\vec{p}^p$ . The neutrino approaches from the  $-z$  direction. On the right side, the variables are shown in the plane transverse to the initial neutrino momentum  $\vec{p}^\nu$ . Figure reproduced from Ref. [62].

the unfolded result and compare against the true cross sections used to generate the data.

In the OmniFold approach, we have the freedom to choose all of our inputs independently of the observables we wish to inspect in the end. We define three variations that we refer to as UniFold, MultiFold, and OmniFold in the same way as Ref. [21], distinguished by which observables are provided as inputs in each case. UniFold is effectively an unbinned version of IBU, with the only kinematic input being a single observable we are trying to unfold. Correspondingly, we run four separate instances of UniFold: one for each of the observables used in evaluation of the unfolding performance. MultiFold takes as input  $(p_\mu, \cos \theta_\mu, p_p, \delta p_T, \delta \alpha_T, \delta \phi_T)$ , which directly includes all of the observables that we are interested in and allows for the simultaneous unfolding of all of them. OmniFold has the most general inputs,<sup>2</sup> using just the kinematics of the muon and leading proton  $(p_\mu, \cos \theta_\mu, \phi_\mu, p_p, \cos \theta_p, \phi_p)$ , where, in the coordinates of Fig. 2,  $\phi \equiv \arctan(p_y/p_x)$  is the direction of the particle in the plane transverse to the neutrino beam direction. The STVs can be derived from these inputs, but we can expect they would be more difficult to learn compared to MultiFold. We refer to these three variations collectively as the OmniFold method or approach, since they all operate on the same principles.

During training and evaluation of the classifiers used in the OmniFold method, the momentum variables are log transformed to mitigate the effects of long tails and then standardized to a mean of 0 and standard deviation of 1. In events where one of the input variables is missing, either because it is missing at truth level or because of a failed reconstruction, a value of 0 is used for the missing quantity. For any particular setup, we use the same kinematic

<sup>2</sup>These are the most general inputs given what is provided in the available dataset, but a more detailed dataset would allow the use of an even wider phase space, potentially including all true/reconstructed particles.

TABLE I. Binning of the muon momentum  $p_\mu$  and forward angle  $\theta_\mu$  used in extracting a binned differential cross-section, identical to that used in the T2K analysis associated with the data release used in this work [51].

| Bin index | $\cos \theta_\mu$               | $p_\mu$ (GeV/c)      |
|-----------|---------------------------------|----------------------|
| 0         | $-1.0 < \cos \theta_\mu < 0.2$  | $0 < p_\mu < 30$     |
| 1         | $0.2 < \cos \theta_\mu < 0.6$   | $0 < p_\mu < 0.3$    |
| 2         |                                 | $0.3 < p_\mu < 0.4$  |
| 3         |                                 | $0.4 < p_\mu < 0.5$  |
| 4         |                                 | $0.5 < p_\mu < 0.6$  |
| 5         |                                 | $0.6 < p_\mu < 30$   |
| 6         | $0.6 < \cos \theta_\mu < 0.7$   | $0 < p_\mu < 0.3$    |
| 7         |                                 | $0.3 < p_\mu < 0.4$  |
| 8         |                                 | $0.4 < p_\mu < 0.5$  |
| 9         |                                 | $0.5 < p_\mu < 0.6$  |
| 10        |                                 | $0.6 < p_\mu < 0.8$  |
| 11        |                                 | $0.8 < p_\mu < 30$   |
| 12        | $0.7 < \cos \theta_\mu < 0.8$   | $0 < p_\mu < 0.3$    |
| 13        |                                 | $0.3 < p_\mu < 0.4$  |
| 14        |                                 | $0.4 < p_\mu < 0.5$  |
| 15        |                                 | $0.5 < p_\mu < 0.6$  |
| 16        |                                 | $0.6 < p_\mu < 0.8$  |
| 17        |                                 | $0.8 < p_\mu < 30$   |
| 18        | $0.8 < \cos \theta_\mu < 0.85$  | $0 < p_\mu < 0.3$    |
| 19        |                                 | $0.3 < p_\mu < 0.4$  |
| 20        |                                 | $0.4 < p_\mu < 0.5$  |
| 21        |                                 | $0.5 < p_\mu < 0.6$  |
| 22        |                                 | $0.6 < p_\mu < 0.8$  |
| 23        |                                 | $0.8 < p_\mu < 1.0$  |
| 24        |                                 | $1.0 < p_\mu < 30$   |
| 25        | $0.85 < \cos \theta_\mu < 0.9$  | $0 < p_\mu < 0.3$    |
| 26        |                                 | $0.3 < p_\mu < 0.4$  |
| 27        |                                 | $0.4 < p_\mu < 0.5$  |
| 28        |                                 | $0.5 < p_\mu < 0.6$  |
| 29        |                                 | $0.6 < p_\mu < 0.8$  |
| 30        |                                 | $0.8 < p_\mu < 1.0$  |
| 31        |                                 | $1.0 < p_\mu < 1.5$  |
| 32        |                                 | $1.5 < p_\mu < 30$   |
| 33        | $0.9 < \cos \theta_\mu < 0.94$  | $0 < p_\mu < 0.4$    |
| 34        |                                 | $0.4 < p_\mu < 0.5$  |
| 35        |                                 | $0.5 < p_\mu < 0.6$  |
| 36        |                                 | $0.6 < p_\mu < 0.8$  |
| 37        |                                 | $0.8 < p_\mu < 1.25$ |
| 38        |                                 | $1.25 < p_\mu < 2.0$ |
| 39        |                                 | $2.0 < p_\mu < 30$   |
| 40        | $0.94 < \cos \theta_\mu < 0.98$ | $0 < p_\mu < 0.4$    |
| 41        |                                 | $0.4 < p_\mu < 0.5$  |
| 42        |                                 | $0.5 < p_\mu < 0.6$  |
| 43        |                                 | $0.6 < p_\mu < 0.8$  |
| 44        |                                 | $0.8 < p_\mu < 1.0$  |
| 45        |                                 | $1.0 < p_\mu < 1.25$ |
| 46        |                                 | $1.25 < p_\mu < 1.5$ |

(Table continued)

TABLE I. (Continued)

| Bin index | $\cos \theta_\mu$               | $p_\mu$ (GeV/c)      |
|-----------|---------------------------------|----------------------|
| 47        |                                 | $1.5 < p_\mu < 2.0$  |
| 48        |                                 | $2.0 < p_\mu < 3.0$  |
| 49        |                                 | $3.0 < p_\mu < 30$   |
| 50        | $0.98 < \cos \theta_\mu < 1.00$ | $0 < p_\mu < 0.5$    |
| 51        |                                 | $0.5 < p_\mu < 0.7$  |
| 52        |                                 | $0.7 < p_\mu < 0.9$  |
| 53        |                                 | $0.9 < p_\mu < 1.25$ |
| 54        |                                 | $1.25 < p_\mu < 2.0$ |
| 55        |                                 | $2.0 < p_\mu < 3.0$  |
| 56        |                                 | $3.0 < p_\mu < 5.0$  |
| 57        |                                 | $5.0 < p_\mu < 30$   |

TABLE II. Binning of the STVs  $\delta p_T, \delta \alpha_T, \delta \phi_T$  for analysis, copied from a previous T2K analysis that extracted differential cross sections as functions of these observables [6].

| $\delta p_T$ (GeV) | $\delta \phi_T$ (radians) | $\delta \alpha_T$ (radians) |
|--------------------|---------------------------|-----------------------------|
| 0.0–0.08           | 0.0–0.067                 | 0.0–0.47                    |
| 0.08–0.12          | 0.067–0.14                | 0.47–1.02                   |
| 0.12–0.155         | 0.14–0.225                | 1.02–1.54                   |
| 0.155–0.2          | 0.225–0.34                | 1.54–1.98                   |
| 0.2–0.26           | 0.34–0.52                 | 1.98–2.34                   |
| 0.26–0.36          | 0.52–0.85                 | 2.34–2.64                   |
| 0.36–0.51          | 0.85–1.50                 | 2.64–2.89                   |
| 0.51–1.1           | 1.50– $\pi$               | 2.89– $\pi$                 |

variables in both step 1 (reweighting of detector-level events) and step 2 (reweighting of truth-level events) of the OmniFold procedure, although this is not necessary in general [38]. In addition, in step 1 for each case, we also include the detector sample information, and in step 2, we include the event’s truth-level interaction topology information, both of which are described in Sec. III. These labels are one-hot encoded for input.<sup>3</sup>

The OmniFold method is formally agnostic to the specific classifier architecture; for this work, we use a dense neural network with four hidden layers of 100 nodes each, using LeakyReLU activation functions for the hidden nodes and a sigmoid activation for the final output. The input features for each event are structured as a simple vector and received by an input layer with the corresponding number of nodes. The dataset that is the weighting target is given a label of 1, and the dataset that is to be reweighted is given a label of 0. In the case of step 1, these are, respectively, the data and the reconstructed simulation,

<sup>3</sup>One-hot encoding is a common way to deal with categorical information in neural networks. An input with N possible values is structured as N separate inputs  $x_1, \dots, x_N$ . An event that belongs to category  $i \in \{1..N\}$  is given a value of 1 for input  $x_i$ , while all other inputs  $x$  receive a value of 0 for that event.

and in the case of step 2, these are the generator events with pulled weights and the generator events with original weights.<sup>4</sup> We train with a weighted binary cross-entropy loss so that with an event from the 0-labeled dataset with a classifier result  $0 < q < 1$  will receive a reweighting factor of  $q/(1-q)$ . This reweighting factor can be capped to some maximum value for events where  $q$  is close to 1, but we did not find this to be necessary in our results. We note that obtaining extreme values of  $q$  from the classifier would imply some kind of severe disagreement between data and simulation, which would generally indicate the need to reconsider the choice of unfolding inputs or phase space.

The network is implemented with Keras [63] in TensorFlow2.9.0 [64] and trained with the Adam optimizer [65]. We use an initial learning rate of  $1 \times 10^{-4}$ , a batch size of 1024, and a 80/20 training/validation split of the available events. Training continues until there has been no improvement in the validation loss for 15 epochs, after which the model weights that gave the best validation loss are restored as the final result for that iteration. Training of a single network generally took less than 30 sec on a NVIDIA A100 GPU. On subsequent iterations past the first, the same total set of available events is used, but with a different random training/validation split each time. These later iterations also usually take less time to train, as they are warm started using the weights from the previous iterations.

We performed a coarse scan of the training parameters and did not find significant improvement from other values. However, we did find that an insufficiently large neural network size resulted in notable degradation of the performance, as one would expect when the network loses the capability to properly learn the likelihood ratio between the data and simulated distributions. A comparison of the performance with different network sizes is shown in Fig. 3. The final choice of four hidden layers with 100 nodes each was chosen as a point beyond which expanding the network size further had no obvious benefit.

To serve as a benchmark of a typical conventional unfolding approach, we perform an unfolding of each observable of interest with a binned version of the OmniFold method, which with perfect training is mathematically equivalent to IBU [21,66]. In this setup, which we call IBU-UniFold, we follow the OmniFold reweighting procedure, but for each event, we provide only the bin index for the observable in question instead of actual kinematic information. The bin index is provided as one-hot encoded input, in addition to one-hot encoded information about the sample ID in step 1 and topology ID in step 2 like what was provided in the other OmniFold methods. The use of IBU-UniFold allows us to have a

<sup>4</sup>Note that the choice of labels is arbitrary. Flipping the labels would simply mean flipping the expression for the reweighting factor that comes from the classifier output.

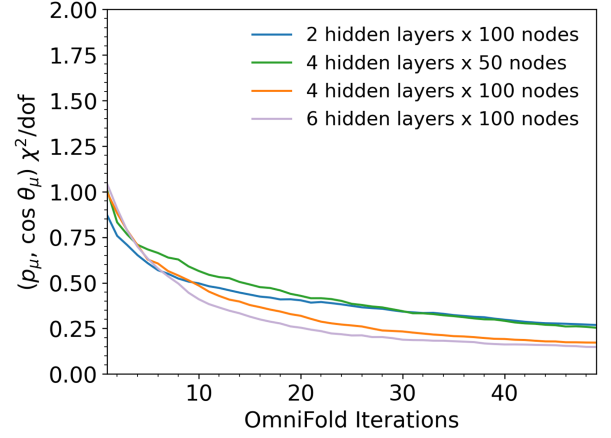


FIG. 3. Comparison of  $\chi^2$  performance for the unfolded differential cross-section as a function of  $(p_\mu, \cos \theta_\mu)$  with OmniFold using different sizes for the neural networks that serve as the classifiers in the method. Networks that are too small lack the ability to fully learn the high-dimensional likelihood ratio and result in worse unfolding performance, but performance also flattens out once additional network capacity becomes extraneous.

background treatment equivalent to what is used in the OmniFold methods, which makes direct comparison of the performance between methods more straightforward. This is in contrast with the background subtraction techniques typically used for IBU, which are different enough that the results and uncertainties are complicated to compare against those from OmniFold.

As an additional benchmark, we also run a binned variation that we call Binned UniFold. This differs from IBU-UniFold only in that we provide the bin information via the value of the bin center in the corresponding kinematic phase space instead of as one-hot encoded bin indices. In principle, there is no change in the amount of provided information between the Binned UniFold setup and the IBU-UniFold setup. However, providing bin centers results in the classifier needing to smoothly interpolate between adjacent bins in the phase space, which can make training more difficult compared to the one-hot encoded case.

Background subtraction and efficiency corrections do not require explicit treatment with the OmniFold method. Instead, these are implicitly [22] applied when we obtain the reweighted generator events, which can be individually identified as being signal or background and passing selection cuts or not. Different options are available for technical details of how to handle selection cuts during the OmniFold procedure, the impact of which we discuss further in Sec. V. For evaluation of the performance, all results are converted into differential cross sections using the same bins.

We run 50 iterations of each of the unfolding procedures so that their convergence rates can be compared and a

smaller cutoff on the number of iterations can be imposed later to serve as the regularized final result. To mitigate uncertainties from the stochastic nature of the neural network training process, we run five trials of each OmniFold procedure using different random training/test splits of the events shown to the neural network each time. The reweighting factors from the resulting networks are then averaged to obtain the unfolded result.

## V. PERFORMANCE AND RESULTS

### A. Convergence criteria

To evaluate performance, we calculate the  $CC0\pi$  (signal) differential cross sections in the chosen kinematic bins. For the single transverse variables, we impose an additional requirement that the proton have a minimum momentum of 450 MeV at truth level, which is roughly the proton reconstruction threshold in the ND280 detector [6]. We calculate a metric  $\chi^2 = (\vec{\mu} - \vec{x})^T \Sigma^{-1} (\vec{\mu} - \vec{x})$  for each of the binned differential cross sections in truth space, using the unfolded result obtained from each of the methods. The vectors  $\vec{\mu}$  are the mean binned results for the observable in question over the 500 simulated pseudoexperiments with the previously described systematic and statistical variations, and the covariance matrix  $\Sigma$  is obtained from the spread of results over those same pseudoexperiments. The vectors  $\vec{x}$  are the truth-level differential cross-sections for the data that we have created as the unfolding target.

The  $\chi^2$  comparison of the OmniFold results against the truth-level data as a function of number of iterations of the unfolding procedure is shown in Fig. 4. For the purposes of

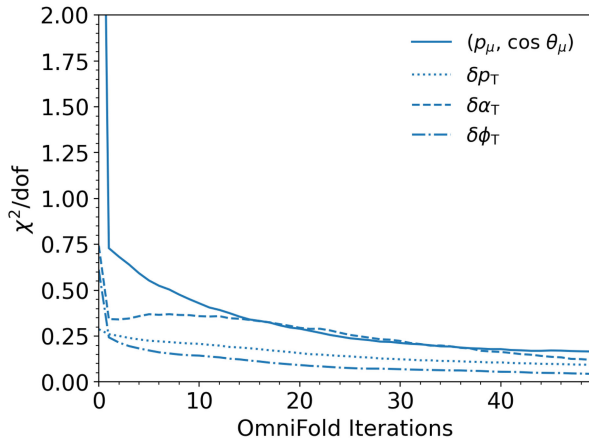


FIG. 4.  $\chi^2/\text{d.o.f.}$  (degrees of freedom) for the OmniFold result as a function of number of iterations of the procedure, for each of the differential cross-sections under consideration. The zeroth iteration is the prior, which begins outside the y-axis limits for  $(p_\mu, \cos \theta_\mu)$ . As with any iterative procedure, using a larger number of iterations reduces the dependence on the prior and reduces the amount of bias in the result. The point at which the  $\chi^2$  appears to converge depends on the observable and binning under inspection.

this study where we wish to compare the bias and uncertainties between methods, we choose an iteration number at which the  $\chi^2$  curve has flattened out to be considered the converged result, leaning toward a larger number of iterations to minimize bias. We choose a cutoff of 20 iterations for IBU-UniFold and 45 iterations for the OmniFold methods. However, this is not a feasible approach for a real data analysis, since it relies on knowing the truth-level target distribution. Furthermore, it relies on having a particular binning in mind already, which is potentially at odds with the general promise of the OmniFold method allowing for unbinned unfolding of multiple observables simultaneously.

For one possibility for an unbinned convergence criterion, we note that the OmniFold method functions by reweighting each individual event after each iteration. When the result is converged, running further iterations should only result in random jitter of the event weights instead of any significant changes. We propose considering the weight change per event averaged over the last  $N$  iterations, for some choice of  $N$ . As the result converges, the distribution of these weight changes should approach a Gaussian centered at 0 with some characteristic width determined by the problem. A sample of how these distributions look for  $N = 5$  in our dataset is shown in Fig. 5. While here we have not imposed a quantitative metric for convergence from these distributions, qualitatively, it is clear that the distributions at larger numbers of iterations are closer to the expected distribution of a converged result than the ones at smaller numbers of

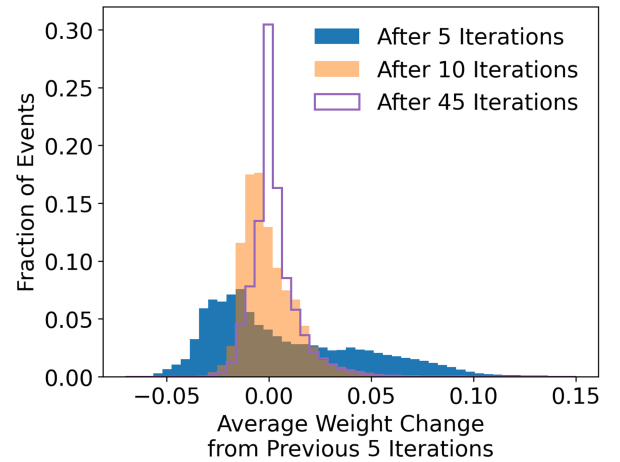


FIG. 5. The average weight change that each event has received over the previous five iterations of the OmniFold procedure, after 5, 10, and 45 total iterations. After 5 and 10 iterations, one can see the distributions of weight changes are still obviously skewed, indicating that there are significant reweightings still occurring across the whole dataset. After 45 iterations, we see the distribution is centered around 0 with not much of a tail. This suggests the unfolding procedure is mostly done at this point, which is also consistent with the behavior of the  $\chi^2$  curves shown in Fig. 4.

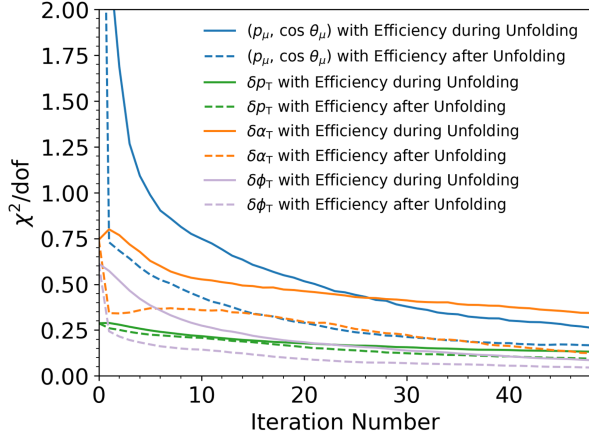


FIG. 6. OmniFold  $\chi^2$  performance for each differential cross-section of interest as a function of number of iterations, using different efficiency treatments. Including efficiency effects during unfolding (solid lines) means including nonreconstructed events in step 2 of the OmniFold procedure, where truth-level event reweighting is done. Applying efficiency corrections after unfolding (dashed lines) means excluding nonreconstructed events from step 2 but applying reweighting factors to them when they are brought in for the final unfolded result, which we see yields faster and better convergence.

iterations. Future work could explore this further as an avenue toward a principled way to choose the degree of regularization in an unbinned manner.

One more issue that impacts the convergence rate is the question of efficiency treatment with the OmniFold method, which is worth explicit discussion. Broadly speaking, there are two options: including efficiency effects during unfolding or applying them afterward. In the first case, we would include truth events that have no reconstructed counterpart during the step 2 reweighting of OmniFold, simply applying either a pull weight of 1 or some other average weight for those events. In the second case, we would exclude these truth events with failed reconstructions from the iterative procedure completely but reintroduce them when evaluating the final result. In this case, they would receive a reweighting factor that is

TABLE III.  $\chi^2$  values from comparing unfolded differential cross sections using each method to the truth-level cross sections underlying the data. Results are divided by observable of interest.

| Method         | $\chi^2$                                  |                            |                                 |                               |
|----------------|-------------------------------------------|----------------------------|---------------------------------|-------------------------------|
|                | $(p_\mu, \cos \theta_\mu)$<br>d.o.f. = 58 | $\delta p_T$<br>d.o.f. = 8 | $\delta \alpha_T$<br>d.o.f. = 8 | $\delta \phi_T$<br>d.o.f. = 8 |
| Prior          | 298.2                                     | 2.3                        | 5.9                             | 4.9                           |
| IBU-UniFold    | 2.1                                       | 0.2                        | 0.4                             | 0.1                           |
| Binned UniFold | 21.4                                      | 1.4                        | 0.9                             | 0.5                           |
| UniFold        | 27.1                                      | 1.1                        | 0.6                             | 1.1                           |
| MultiFold      | 3.1                                       | 0.3                        | 0.2                             | 0.3                           |
| OmniFold       | 10.0                                      | 0.8                        | 1.1                             | 0.4                           |

TABLE IV. Triangular discriminator values from comparing unfolded differential cross sections using each method to the truth-level cross sections underlying the data. Results are divided by observable of interest.

| Method         | Triangular discriminator   |              |                   |                 |
|----------------|----------------------------|--------------|-------------------|-----------------|
|                | $(p_\mu, \cos \theta_\mu)$ | $\delta p_T$ | $\delta \alpha_T$ | $\delta \phi_T$ |
| Prior          | 545.6                      | 27.5         | 31.2              | 26.7            |
| IBU-UniFold    | 17.1                       | 1.9          | 3.4               | 0.8             |
| Binned UniFold | 29.9                       | 2.8          | 6.0               | 1.9             |
| UniFold        | 17.3                       | 5.7          | 5.8               | 1.7             |
| MultiFold      | 2.7                        | 0.7          | 0.6               | 0.6             |
| OmniFold       | 9.4                        | 1.7          | 3.0               | 2.1             |

an extrapolation by the classifier trained from the OmniFold procedure, and this is equivalent to applying an unbinned efficiency correction after the unfolding procedure. A comparison of the performance of these approaches for our setup is shown in Fig. 6. We expect little difference in results between the two approaches when efficiencies are generally high for all bins, but this is not the case for efficiencies in our analysis. We observe that including efficiency effects during unfolding considerably slows the rate of  $\chi^2$  convergence, since this increases the bias toward the prior in regions of low efficiency. Correspondingly, the rest of the results in this paper are all obtained using the approach of an efficiency correction after unfolding.

## B. Results

The resulting  $\chi^2$  values comparing the unfolded truth-level differential cross sections against those underlying the

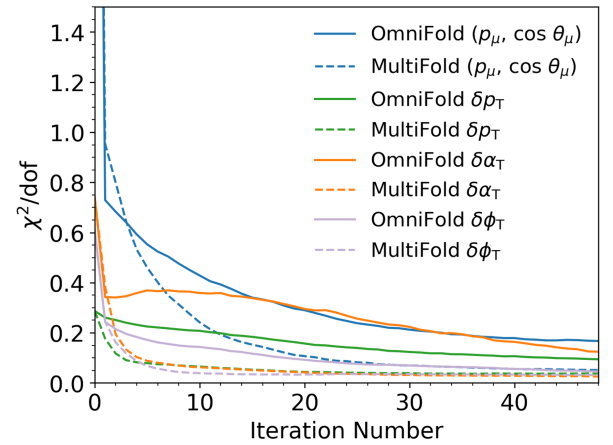


FIG. 7. Comparison of  $\chi^2$  performance for different inputs to the neural network in the OmniFold approach. For OmniFold (solid lines), the inputs are the muon and leading proton kinematics  $(p_\mu, \cos \theta_\mu, \phi_\mu, p_p, \cos \theta_p, \phi_p)$ , from which our observables of interest can be derived. With MultiFold (dashed lines), the kinematic variables used as input are  $(p_\mu, \cos \theta_\mu, p_p, \delta p_T, \delta \alpha_T, \delta \phi_T)$ , directly including the observables of interest.

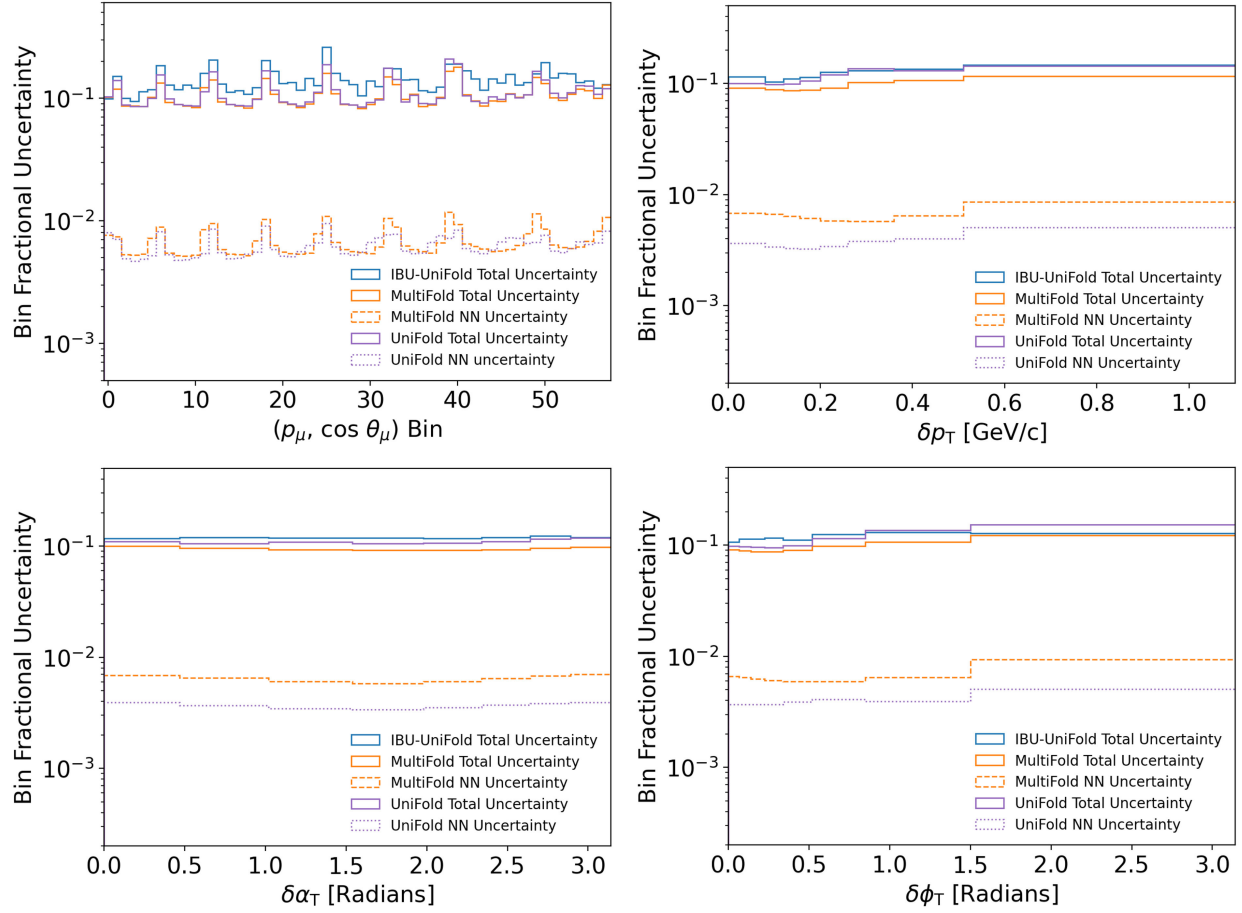


FIG. 8. Bin-by-bin uncertainty in the unfolded results from IBU-UniFold, UniFold, and MultiFold, excluding effects of correlations among bins. The neural network (NN) uncertainty indicates the standard error due to just the neural network’s stochastic training process, obtained by running the procedure five separate times. The MultiFold result contains generally lower uncertainties than IBU-UniFold. We also note that the NN uncertainty from UniFold is lower than that of MultiFold in the transverse variables, which is expected as a result of the simpler classifier problem that UniFold faces.

data are shown in Table III for each of the observables of interest and for each of the multiple unfolding approaches described in Sec. IV. Since the  $\chi^2$  metric does not clearly distinguish between results with large bias and large uncertainty and results with small bias and small uncertainty, we also consider the triangular discriminator  $\Delta(P, Q) = \sum_i |p_i - q_i|^2 / (p_i + q_i)$  [67], where  $p_i$  is the unfolded content of bin  $i$ ,  $q_i$  is the true content of bin  $i$ , and the summation runs over the available bins. The results from evaluating the triangular discriminator between each unfolded result and the true data distribution are shown in Table IV, serving as an estimate of solely the bias. We see that MultiFold and OmniFold are able to achieve  $\chi^2$  values comparable to those of IBU-UniFold across all observables, while also achieving less bias. We emphasize that the IBU-UniFold and UniFold results are obtained by running a separate unfolding procedure for each different observable, whereas in the MultiFold and OmniFold results we are simultaneously unfolding all of them in one pass.

IBU-UniFold performs better than Binned UniFold, even though in principle these methods have access to the same amount of information. The discrepancy in performance between these methods is attributed to difficulties in training the neural networks, as learning the bin-to-bin fluctuations is easier with discrete inputs.<sup>5</sup> These training difficulties are likely exacerbated by the generally low statistics in a neutrino dataset compared to previous applications of OmniFold, with only around 20,000 events in the data for our setup. Given this limitation, the similarity in performance between Binned UniFold and unbinned UniFold is unsurprising, too, as our analysis bins are already quite fine relative to the available statistics.

The clear improvement in performance seen when comparing MultiFold and OmniFold against UniFold demonstrates that the additional information provided in those cases is indeed helpful in unfolding. The observation

<sup>5</sup>Neural networks are known to be more effective in higher dimensions, which is the case for the one-hot-encoded entries.

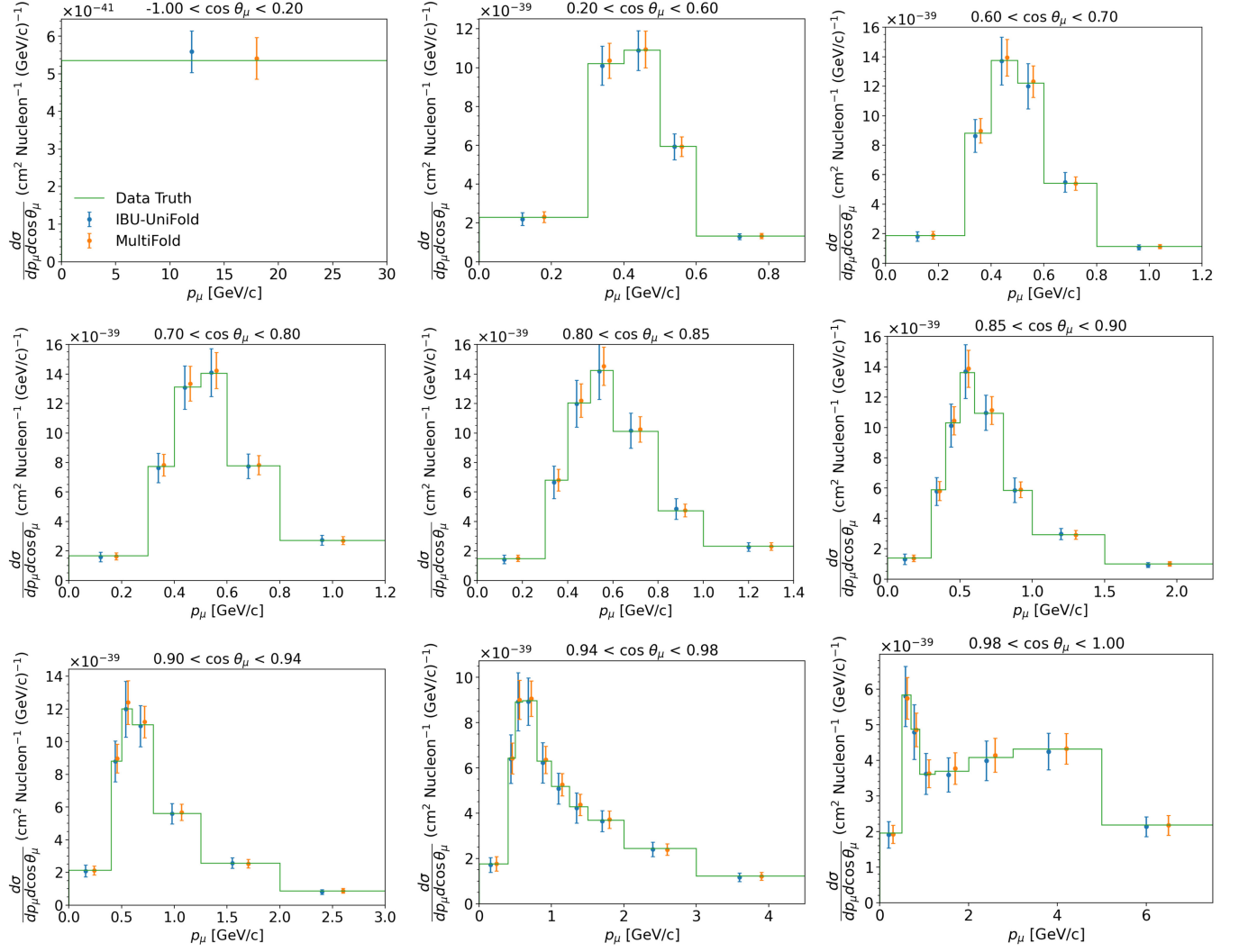


FIG. 9. Unfolded CC0 $\pi$  differential cross-sections as a function of the muon kinematics  $(p_\mu, \cos \theta_\mu)$ , compared against the truth-level cross sections underlying the data. Each plot contains the differential cross sections from a specified range of the forward angle  $\cos \theta_\mu$ . Error bars are the spread in results from 500 pseudoexperiments varying systematic and statistical uncertainties, while the nominal result in each bin is the mean from those 500 unfolded throws. Note that the highest momentum bin in each case extends up to 30 GeV but is truncated in the plots for readability.

that MultiFold performs better than OmniFold across all metrics also serves as another indication that the classifier training is imperfect. A direct comparison of the  $\chi^2$  convergence of these two methods is shown in Fig. 7, where we see MultiFold converges both faster and to better values. While in principle OmniFold has access to all the information that MultiFold did in our setup, it is a nontrivial function to derive the STVs from the muon and proton kinematics. With the limited statistics in our dataset, it is unsurprising that the OmniFold networks appear to be unable to perform as well as the MultiFold networks, given that the latter directly receive all of the variables of interest. From a practical perspective, an analyzer who knows the observables they will be interested in will usually want to include this information as direct input to the procedure

anyway, sidestepping this problem, but this calls for caution when attempting to unfold arbitrarily complicated derived quantities from simple inputs.

The fact that MultiFold achieves similar  $\chi^2$  performance to IBU-UniFold for any particular observable is noteworthy, given that it has the additional advantage that it is unfolding them all simultaneously. On top of this, the triangular discriminator performances indicate that MultiFold and OmniFold achieve less bias than IBU-UniFold. The discrepancy between the  $\chi^2$  and triangular discriminator comparisons among methods can be attributed to the generally lower uncertainties obtained by the OmniFold methods. A comparison of the uncertainty budgets between IBU-UniFold, UniFold, and MultiFold is provided in Fig. 8. We see that for the  $(p_\mu, \cos \theta_\mu)$

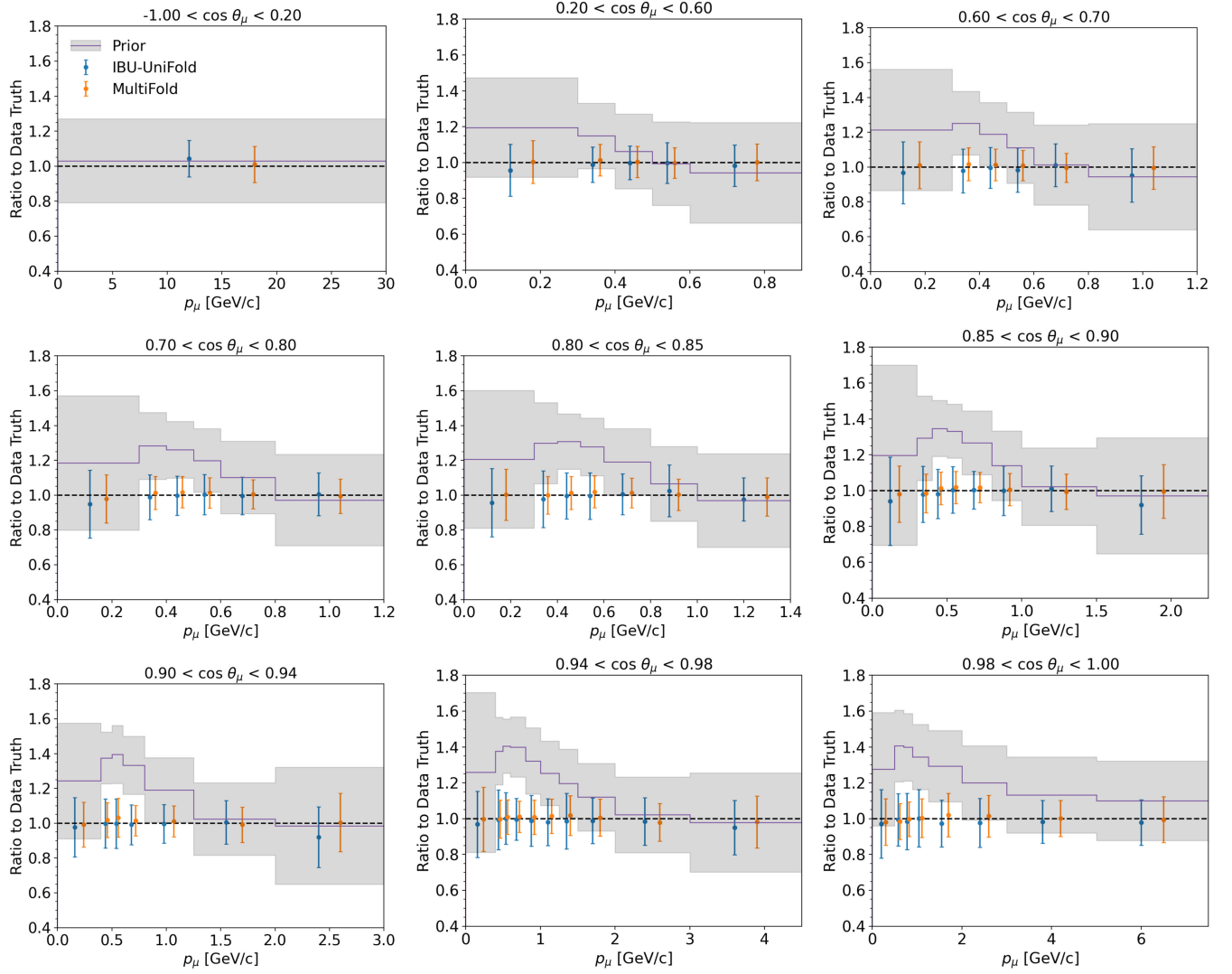


FIG. 10. Ratios of the unfolded  $\text{CC}0\pi$  differential cross-sections as a function of the muon kinematics ( $p_\mu, \cos \theta_\mu$ ), compared against the true cross sections used to generate the data. Each plot contains the ratios from a specified range of the forward angle  $\cos \theta_\mu$ . Error bars are the spread in results from 500 pseudoexperiments varying systematic and statistical uncertainties, while the nominal result for each bin is the mean of those 500 unfolded throws. Note that the highest momentum bin in each case extends up to 30 GeV but is truncated in the plots for readability.

distribution in particular with its relatively fine bins IBU-UniFold yields notably higher uncertainties. The unbinned nature of the OmniFold method's reweighting in combination with the penalization of overfitting by the training/validation split of the data results in a more regularized result compared to IBU-UniFold.

The uncertainty budget shown in Fig. 8 also includes the contribution to the uncertainty from neural network training randomness, evaluated by calculating the variation in results using identical data but different training initializations for the neural networks. However, this standard error can be reduced by simply running the procedure additional times and averaging all of the results. The unfolded results we have presented are the result of

ensembling five trials of neural networks, which reduces the standard error to a level that is negligible relative to the systematic and statistical uncertainties involved with the data. As expected given this small contribution to the uncertainty, we also find that the neural network ensembling has negligible impact on the overall unfolding performance. However, we note that the effect of ensembling will vary with the specific unfolding problem and classifier choice.

For direct visualization of the results, the unfolded differential cross sections from both IBU-UniFold and MultiFold compared against the true values underlying the data are shown for the muon kinematics in Figs. 9 and 10 and for the STVs in Fig. 11.

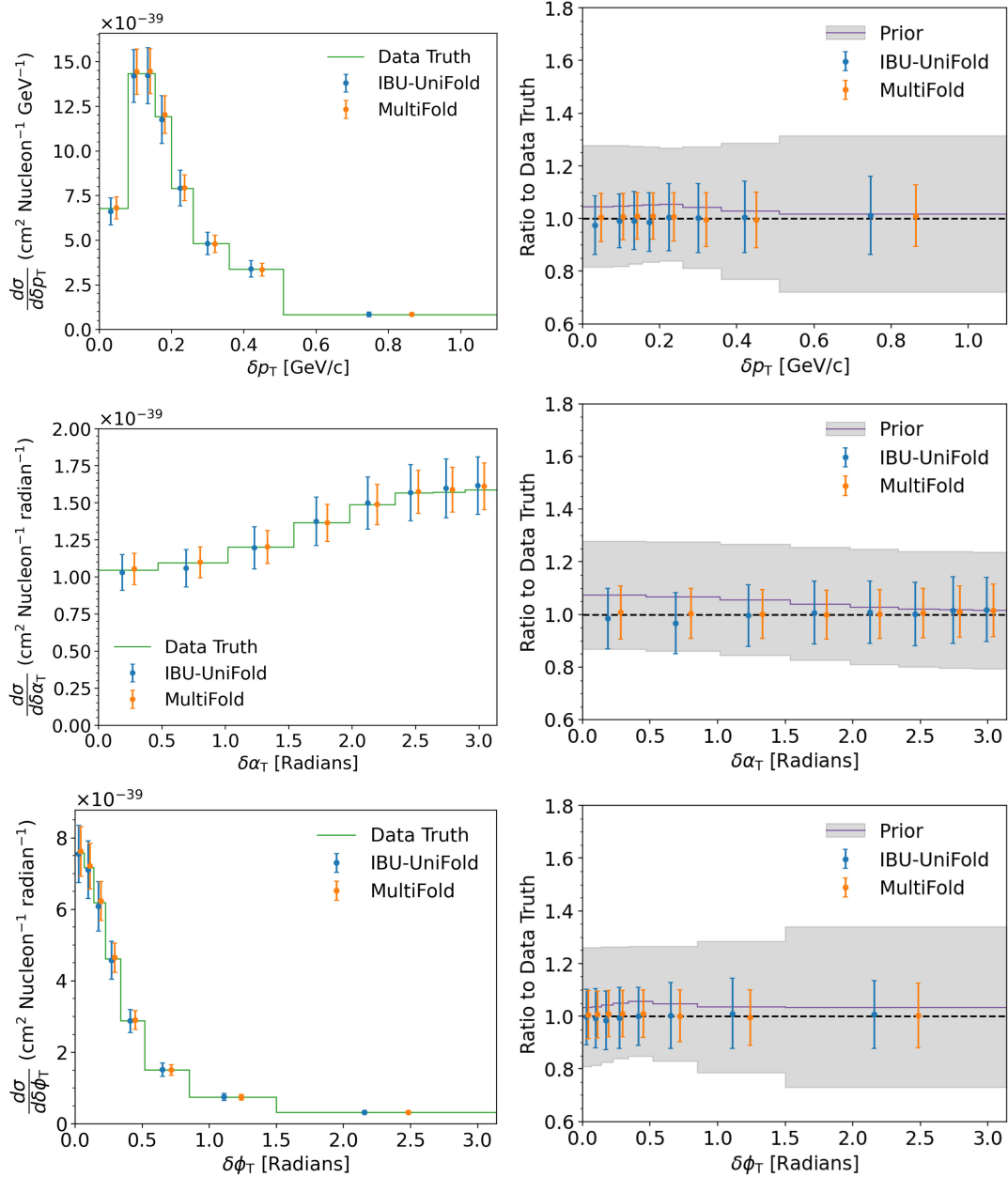


FIG. 11. Unfolded CC0 $\pi$  differential cross-sections as functions of the three STVs, compared against the true cross-sections underlying the data. Only events with a true proton momentum  $> 450$  MeV are included for these comparisons. Error bars are the spread in results from 500 pseudoexperiments varying systematic and statistical uncertainties, and the nominal result in each bin is the mean from those 500 unfolded throws. Left: unfolded cross sections from IBU-UniFold and MultiFold compared against the true cross sections used to generate the data. Right: per-bin ratios of the unfolded cross sections from IBU-UniFold and MultiFold compared against the true cross-sections for the data. The prior shows the mean and standard deviation of each bin from the 500 pseudoexperiments before applying any unfolding procedure.

## VI. CONCLUSIONS

With its approach of event-by-event reweighting with the assistance of machine learning, OmniFold offers the ability to perform unbinned unfolding of multiple observables simultaneously while accounting for background and efficiency corrections with high-dimensional dependencies. We have demonstrated for the first time its application to a

neutrino cross-section measurement, using public simulated data from the T2K ND280 detector. By creating fake data with nontrivial reweightings of the underlying interaction models, we were able to test the closure properties of the OmniFold approach, finding that with optimized inputs it was able to achieve similar  $\chi^2$  goodness of fit with less bias and smaller uncertainties compared to a conventional IBU approach across several observables of interest.

Our results did not show strict improvement in terms of  $\chi^2$  from using OmniFold-based approaches compared to a conventional unfolding method. Moving from a binned to unbinned approach without the addition of other information yielded generally worse results, which seems to stem from imperfect neural network training. We note that, while we used a dense neural network for all setups in this study for better comparability, in practice a simpler classifier architecture such as a boosted decision tree would be more appropriate if one actually wanted to do a UniFold-like analysis with only one or two input quantities. We do see improvement as we move to higher-dimensional inputs, and comparing MultiFold against OmniFold, we see that directly providing the most relevant observables yields better performance than requiring the procedure to derive this information from other more general inputs, telling us that thoughtful choices of input variables are important. For any future application of OmniFold to real data, we thus emphasize the importance of using fake data studies to optimize the classifier architecture, training settings, and inputs. Some of the issues we observed are partially caused by the generally low statistics associated with a neutrino dataset, which will continue to be a problem for real neutrino data. Since OmniFold requires training with the actual data, improving the statistics is not a trivial matter of running more simulations, and this problem will require particular attention. Other methods for data augmentation in unfolding, including pretraining [68,69], would be useful to explore in future work.

That being said, the performance of the OmniFold method was likely also limited by the nature of the public dataset that we used. For instance, the dataset only contained information about the primary muon and leading proton in each event, but we would expect OmniFold to do a better job of dealing with background contributions if we could provide pion kinematic information where it existed. For events where more than one proton is reconstructed, providing information about the subleading protons should be helpful, too. This is supported by our general observation that our MultiFold and OmniFold implementations outperformed UniFold by making use of additional information besides just the unfolding variable. The public dataset that we used lacked the full stack of potentially useful information because it was not used in the conventional analysis, but for a future application of OmniFold to actual data, we would want to maximize the amount of useful information that we provide as input. Our choice of a relatively simple neural network was also informed by the limited available information. More sophisticated classifier architectures could be used for datasets dealing with larger and variable-length particle stacks [70], although the available statistics will have to be considered when making this choice.

In conclusion, this work shows that OmniFold is a useful unfolding method to apply to neutrino measurements.

Even with a limited dataset, it is able to deliver on its theoretical advantages, using high-dimensional information to improve the quality of the unfolding and do this for multiple observables simultaneously. While we used a T2K ND280 dataset, the advantages should be similar for any detector. We have identified the difficulties encountered during this work, some of which may be particular to neutrino datasets and will require further study, but OmniFold shows promise as an unfolding technique for all neutrino experiments to improve the quality of their measurements.

## ACKNOWLEDGMENTS

We would like to thank T2K for making the simulated dataset extensively used in this work publicly available. We would also like to particularly thank *Ciro Riccio* for his invaluable help in understanding the T2K public dataset and in his efforts to produce the dataset initially. Additionally, we would like to thank *Ryan Milton* for validating the implementation of OmniFold used in this work with his RooUnfold implementation. The work of R. H., B. N., and C. W. was supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Contract No. DE-AC02-05CH11231. The work of M. K. and T. K. was supported by JSPS KAKENHI Grant No. JP23H04504. The work of A. C. was supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under DOE Award No. DE-SC0010005. Additional support was provided through the U.S.-Japan Science and Technology Cooperative Research Program in High Energy Physics. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC Grants No. ERCAP0023703, No. ERCAP0028625, and No. ERCAP0031852.

## DATA AVAILABILITY

The data that support the findings of this article are openly available [50,60].

## APPENDIX: CORRELATION MATRICES

Correlation matrices for the MultiFold result are provided in Fig. 12, for each of the observables of interest. The mapping of the bin indices to actual bin values can be found in Tables I and II. These correlation matrices are calculated using the unfolded results from 500 throws of the simulation that account for systematic and statistical uncertainties, but without accounting for flux normalization. Flux uncertainties would otherwise add a notable positive correlation to all bins in the result.

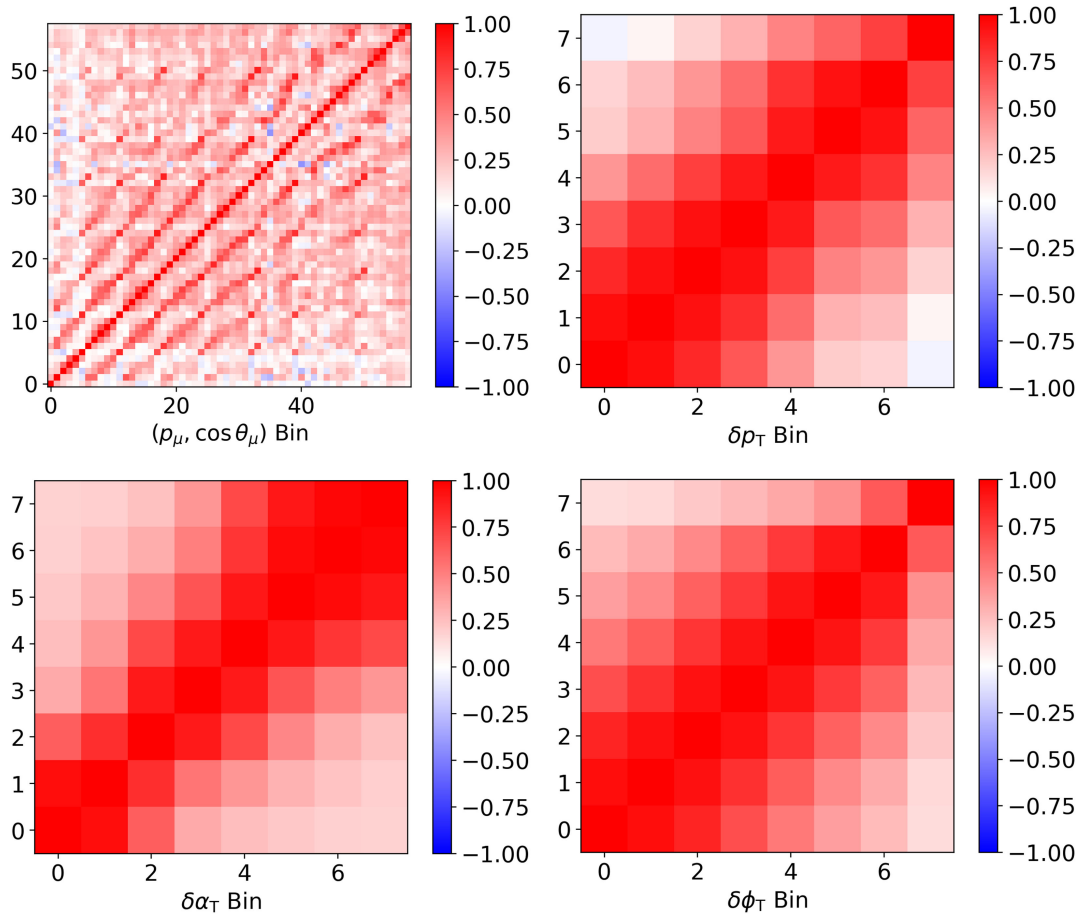


FIG. 12. Correlation matrices from the MultiFold method for the number of events per bin in the unfolded distributions of each observable of interest. The correlations in the results are roughly as expected given the nature of the systematic uncertainties applied to the simulation.

- 
- [1] G. Cowan, Conf. Proc. C **0203181**, 248 (2002), <https://www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings/cowan.pdf>.
  - [2] *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN, Geneva, Switzerland 17-20 January 2011*, edited by H.B. Prosper and L. Lyons (CERN, Geneva, 2011).
  - [3] S. Navas *et al.* (Particle Data Group), *Phys. Rev. D* **110**, 030001 (2024).
  - [4] K. Mahn, C. Marshall, and C. Wilkinson, *Annu. Rev. Nucl. Part. Sci.* **68**, 105 (2018).
  - [5] W. Tang, X. Li, X. Qian, H. Wei, and C. Zhang, *J. Instrum.* **12**, P10002 (2017).
  - [6] K. Abe *et al.* (The T2K Collaboration), *Phys. Rev. D* **98**, 032003 (2018).
  - [7] S. Gardiner, [arXiv:2401.04065](https://arxiv.org/abs/2401.04065).
  - [8] M. B. Avanzini *et al.*, *Phys. Rev. D* **105**, 092004 (2022).
  - [9] G. D'Agostini, *Nucl. Instrum. Methods Phys. Res., Sect. A* **362**, 487 (1995).
  - [10] G. D'Agostini, in *Proceedings of the Alliance Workshop on Unfolding and Data Correction* (2010), [arXiv:1010.0632](https://arxiv.org/abs/1010.0632).
  - [11] M. Kuusela, Statistical issues in unfolding methods for high energy physics, Master's thesis, Aalto University, 2012.
  - [12] Y. Vardi, L. A. Shepp, and L. Kaufman, *J. Am. Stat. Assoc.* **80**, 8 (1985).
  - [13] H. N. Multhei and B. Schorr, *Nucl. Instrum. Methods Phys. Res., Sect. A* **257**, 371 (1987).
  - [14] W. H. Richardson, *J. Opt. Soc. Am.* **62**, 55 (1972).
  - [15] L. B. Lucy, *Astron. J.* **79**, 745 (1974).
  - [16] L. Koch, *J. Instrum.* **14**, P09013 (2019).
  - [17] R. D. Cousins, S. J. May, and Y. Sun, [arXiv:1607.07038](https://arxiv.org/abs/1607.07038).
  - [18] M. Arratia *et al.*, *J. Instrum.* **17**, P01024 (2022).
  - [19] S. Badger *et al.*, *SciPost Phys.* **14**, 079 (2023).

- [20] N. Huetsch *et al.*, *SciPost Phys.* **18**, 070 (2025).
- [21] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, and J. Thaler, *Phys. Rev. Lett.* **124**, 182001 (2020).
- [22] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman, A. Suresh, and J. Thaler, in *Proceedings of the 9th International Conference on Learning Representations* (2021), [arXiv:2105.04448](https://arxiv.org/abs/2105.04448).
- [23] C.-C. Pan, X. Dong, Y.-C. Sun, A.-Y. Cheng, A.-B. Wang, Y.-X. Hu, and H. Cai, [arXiv:2406.01635](https://arxiv.org/abs/2406.01635).
- [24] K. Datta, D. Kar, and D. Roy, [arXiv:1806.00433](https://arxiv.org/abs/1806.00433).
- [25] J. N. Howard, S. Mandt, D. Whiteson, and Y. Yang, *Sci. Rep.* **12**, 7567 (2022).
- [26] S. Diefenbacher, G.-H. Liu, V. Mikuni, B. Nachman, and W. Nie, *Phys. Rev. D* **109**, 076011 (2024).
- [27] A. Butter, T. Jezo, M. Klasen, M. Kuschick, S. Palacios Schweitzer, and T. Plehn, *SciPost Phys. Core* **7**, 064 (2024).
- [28] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, and R. Winterhalder, *SciPost Phys.* **8**, 070 (2020).
- [29] Y. Alanazi *et al.*, *Phys. Rev. D* **106**, 096002 (2022).
- [30] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone, and U. Köthe, *SciPost Phys.* **9**, 074 (2020).
- [31] M. Vandegar, M. Kagan, A. Wehenkel, and G. Louppe, [arXiv:2011.05836](https://arxiv.org/abs/2011.05836).
- [32] M. Backes, A. Butter, M. Dunford, and B. Malaescu, *SciPost Phys. Core* **7**, 007 (2024).
- [33] M. Leigh, J. A. Raine, K. Zoch, and T. Golling, *SciPost Phys.* **14**, 159 (2023).
- [34] J. Ackerschott, R. K. Barman, D. Gonçalves, T. Heimel, and T. Plehn, *SciPost Phys.* **17**, 001 (2024).
- [35] A. Shmakov, K. Greif, M. Fenton, A. Ghosh, P. Baldi, and D. Whiteson, [arXiv:2305.10399](https://arxiv.org/abs/2305.10399).
- [36] A. Shmakov, K. Greif, M. J. Fenton, A. Ghosh, P. Baldi, and D. Whiteson, *SciPost Phys.* **18**, 117 (2025).
- [37] V. Andreev *et al.* (H1 Collaboration), *Phys. Rev. Lett.* **128**, 132002 (2022); H1 Collaboration, Report No. H1prelim-22-031, 2022.
- [38] V. Andreev *et al.* (H1 Collaboration), *Phys. Lett. B* **844**, 138101 (2023).
- [39] V. Andreev *et al.* (H1 Collaboration), [arXiv:2412.14092](https://arxiv.org/abs/2412.14092).
- [40] R. Aaij *et al.* (LHCb Collaboration), *Phys. Rev. D* **108**, L031103 (2023).
- [41] G. Aad *et al.* (ATLAS Collaboration), *Phys. Rev. Lett.* **133**, 261803 (2024).
- [42] G. Aad *et al.* (ATLAS Collaboration), [arXiv:2502.02062](https://arxiv.org/abs/2502.02062).
- [43] Y. Song (STAR Collaboration), [arXiv:2307.07718](https://arxiv.org/abs/2307.07718).
- [44] T. Pani (STAR Collaboration), *EPJ Web Conf.* **296**, 11003 (2024).
- [45] P. T. Komiske, S. Kryhin, and J. Thaler, *Phys. Rev. D* **106**, 094021 (2022).
- [46] CMS Collaboration, Measurement of event shapes in minimum bias events from pp collisions at 13 TeV, Report No. CMS-PAS-SMP-23-008, CERN, Geneva, 2024, <https://cds.cern.ch/record/2899591>.
- [47] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer New York Inc., New York, NY, USA, 2001).
- [48] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning* (Cambridge University Press, Cambridge, England, 2012).
- [49] K. Cranmer, J. Pavez, and G. Louppe, Approximating likelihood ratios with calibrated discriminative classifiers, [arXiv:1506.02169](https://arxiv.org/abs/1506.02169).
- [50] T2K Collaboration, Data release for simulated muon neutrino charged-current interactions on hydrocarbon without pions in the final state using multiple detectors with correlated energy spectra at T2K, [10.5281/zenodo.15183090](https://doi.org/10.5281/zenodo.15183090) (2025).
- [51] K. Abe *et al.* (T2K Collaboration), *Phys. Rev. D* **108**, 112009 (2023).
- [52] K. Abe *et al.* (T2K Collaboration), *Phys. Rev. D* **101**, 112001 (2020).
- [53] K. Abe *et al.* (T2K Collaboration), *Phys. Rev. D* **101**, 112004 (2020).
- [54] Y. Hayato, *Nucl. Phys. B, Proc. Suppl.* **112**, 171 (2002).
- [55] Y. Hayato, *Acta Phys. Pol. B* **40**, 2477 (2009), <https://www.actaphys.uj.edu.pl/R/40/9/2477>.
- [56] Y. Hayato and L. Pickering, *Eur. Phys. J. Spec. Top.* **230**, 4469 (2021).
- [57] S. Agostinelli *et al.*, *Nucl. Instrum. Methods Phys. Res., Sect. A* **506**, 250 (2003).
- [58] J. Nieves, R. Gran, I. Ruiz Simo, F. Sánchez, and M. Vicente Vacas, *Nucl. Part. Phys. Proc.* **273–275**, 1830 (2016).
- [59] K. Abe *et al.* (T2K Collaboration), *Phys. Rev. D* **103**, 112008 (2021).
- [60] <https://github.com/rhuang1/OmnifoldT2K>
- [61] X.-G. Lu, L. Pickering, S. Dolan, G. Barr, D. Coplowe, Y. Uchida, D. Wark, M. O. Wascko, A. Weber, and T. Yuan, *Phys. Rev. C* **94**, 015503 (2016).
- [62] L. Pickering, *J. Phys. Soc. Jpn. Conf. Proc.* **12**, 010032 (2016).
- [63] F. Chollet *et al.*, Keras, <https://keras.io> (2015).
- [64] M. Abadi *et al.*, [arXiv:1605.08695](https://arxiv.org/abs/1605.08695).
- [65] D. Kingma and J. Ba, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [66] R. Milton, V. Mikuni, T. Lee, T. Wamorkar, and B. Nachman, *J. Instrum.* **20**, P05034 (2025).
- [67] F. Topsoe, *IEEE Trans. Inf. Theory* **46**, 1602 (2000).
- [68] V. Mikuni and B. Nachman, *Phys. Rev. D* **111**, 054015 (2025).
- [69] V. Mikuni and B. Nachman, *Phys. Rev. D* **111**, L051504 (2025).
- [70] P. Komiske, E. Metodiev, and J. Thaler, *J. High Energy Phys.* **01** (2019) 121.