Data-driven model validation for neutrino-nucleus cross section measurements

P. Abratenko,[38] O. Alterkait,[38] D. Andrade Aldana,[14] L. Arellano,[21] J. Asaadi,[37] A. Ashkenazi,[36] S. Balasubramanian,[12] B. Baller,[12] A. Barnard,[28] G. Barr,[28] D. Barrow,[28] J. Barrow,[25] V. Basque,[12] J. Bateman,[21] O. Benevides Rodrigues,[14] S. Berkman,[24] A. Bhanderi,[21] A. Bhat,[7] M. Bhattacharya,[12] M. Bishai,[3] A. Blake,[18] B. Bogart,[23] T. Bolton,[17] M. B. Brunetti,[40] L. Camilleri,[10] Y. Cao,[21] D. Caratelli,[4] F. Cavanna,[12] G. Cerati,[12] A. Chappell,[40] Y. Chen,[32] J. M. Conrad,[22] M. Convery,[32] L. Cooper-Troendle,[29] J. I. Crespo-Anadón,[6] R. Cross,[40] M. Del Tutto,[12] S. R. Dennis,[5] P. Detje,[5] R. Diurba,[2] Z. Djurcic,[1] K. Duffy,[28] S. Dytman,[29] B. Eberly,[34] P. Englezos,[31] A. Ereditato,[7,12] J. J. Evans,[21] C. Fang,[4] W. Foreman,[14,19] B. T. Fleming,[7] D. Franco,[7] A. P. Furmanski,[25] F. Gao,[4] D. Garcia-Gamez,[13] S. Gardiner,[12] G. Ge,[10] S. Gollapinni,[19] E. Gramellini,[21] P. Green,[28] H. Greenlee,[12] L. Gu,[18] W. Gu,[3] R. Guenette,[21] P. Guzowski,[21] L. Hagaman,[7] M. D. Handley,[5] O. Hen,[22] C. Hilgenberg,[25] G. A. Horton-Smith,[17] Z. Imani,[38] B. Irwin,[25] M. S. Ismail,[29] C. James,[12] X. Ji,[26] J. H. Jo,[3] R. A. Johnson,[8] Y.-J. Jwa,[10] D. Kalra,[10] G. Karagiorgi,[10] W. Ketchum,[12] M. Kirby,[3] T. Kobilarcik,[12] N. Lane,[21] J.-Y. Li,[11] Y. Li,[3] K. Lin,[31] B. R. Littlejohn,[14] L. Liu,[12] W. C. Louis,[19] X. Luo,[4] T. Mahmud,[18] C. Mariani,[39] D. Marsden,[21] J. Marshall,[40] N. Martinez,[17] D. A. Martinez Caicedo,[33] S. Martynenko,[3] A. Mastbaum,[31] I. Mawby,[18] N. McConkey,[30] V. Meddage,[17] L. Mellet,[24] J. Mendez,[20] J. Micallef,[22,38] K. Miller,[7] A. Mogan,[9] T. Mohayai,[16] M. Mooney,[9] A. F. Moor,[5] C. D. Moore,[12] L. Mora Lepin,[21] M. M. Moudgalya,[21] S. Mulleriababu,[2] D. Naples,[29] A. Navrer-Agasson,[15,21] N. Nayak,[3] M. Nebot-Guinot,[11] C. Nguyen,[31] J. Nowak,[18] N. Oza,[10] O. Palamara,[12] N. Pallat,[25] V. Paolone,[29] A. Papadopoulou,[1] V. Papavassiliou,[27] H. B. Parkinson,[11] S. F. Pate,[27] N. Patel,[18] Z. Pavlovic,[12] E. Piasetzky,[36] K. Pletcher,[24] I. Pophale,[18] X. Qian,[3] J. L. Raaf,[12] V. Radeka,[3] A. Rafique,[1] M. Reggiani-Guzzo,[11] L. Ren,[27] L. Rochester,[32] J. Rodriguez Rondon,[33] M. Rosenberg,[38] M. Ross-Lonergan,[19] I. Safa,[10] D. W. Schmitz,[7] A. Schukraft,[12] W. Seligman,[10] M. H. Shaevitz,[10] R. Sharankova,[12] J. Shi,[5] E. L. Snider,[12] M. Soderberg,[35] S. Söldner-Rembold,[15,21] J. Spitz,[23] M. Stancari,[12] J. St. John,[12] T. Strauss,[12] A. M. Szelc,[11] N. Taniuchi,[5] K. Terao,[32] C. Thorpe,[21] D. Torbunov,[3] D. Totani,[4] M. Toups,[12] A. Trettin,[21] Y.-T. Tsai,[32] J. Tyler,[17] M. A. Uchida,[5] T. Usher,[32] B. Viren,[3] J. Wang,[26] M. Weber,[2] H. Wei,[20] A. J. White,[7] S. Wolbers,[12] T. Wongjirad,[38] M. Wospakrik,[12] K. Wresilo,[5] W. Wu,[29] E. Yandel,[4,19] T. Yang,[12] L. E. Yates,[12] H. W. Yu,[3] G. P. Zeller,[12] J. Zennamo,[12] and C. Zhang[3]

(MicroBooNE Collaboration)[*]

[1]*Argonne National Laboratory (ANL), Lemont, Illinois 60439, USA*
[2]*Universität Bern, Bern CH-3012, Switzerland*
[3]*Brookhaven National Laboratory (BNL), Upton, New York 11973, USA*
[4]*University of California, Santa Barbara, California 93106, USA*
[5]*University of Cambridge, Cambridge CB3 0HE, United Kingdom*
[6]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid E-28040, Spain*
[7]*University of Chicago, Chicago, Illinois 60637, USA*
[8]*University of Cincinnati, Cincinnati, Ohio 45221, USA*
[9]*Colorado State University, Fort Collins, Colorado 80523, USA*
[10]*Columbia University, New York, New York 10027, USA*
[11]*University of Edinburgh, Edinburgh EH9 3FD, United Kingdom*
[12]*Fermi National Accelerator Laboratory (FNAL), Batavia, Illinois 60510, USA*
[13]*Universidad de Granada, Granada E-18071, Spain*
[14]*Illinois Institute of Technology (IIT), Chicago, Illinois 60616, USA*
[15]*Imperial College London, London SW7 2AZ, United Kingdom*
[16]*Indiana University, Bloomington, Indiana 47405, USA*
[17]*Kansas State University (KSU), Manhattan, Kansas 66506, USA*
[18]*Lancaster University, Lancaster LA1 4YW, United Kingdom*
[19]*Los Alamos National Laboratory (LANL), Los Alamos, New Mexico 87545, USA*
[20]*Louisiana State University, Baton Rouge, Louisiana 70803, USA*
[21]*The University of Manchester, Manchester M13 9PL, United Kingdom*
[22]*Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139, USA*
[23]*University of Michigan, Ann Arbor, Michigan 48109, USA*
[24]*Michigan State University, East Lansing, Michigan 48824, USA*
[25]*University of Minnesota, Minneapolis, Minnesota 55455, USA*

[26]*Nankai University, Nankai District, Tianjin 300071, China*
[27]*New Mexico State University (NMSU), Las Cruces, New Mexico 88003, USA*
[28]*University of Oxford, Oxford OX1 3RH, United Kingdom*
[29]*University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*
[30]*Queen Mary University of London, London E1 4NS, United Kingdom*
[31]*Rutgers University, Piscataway, New Jersey 08854, USA*
[32]*SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA*
[33]*South Dakota School of Mines and Technology (SDSMT), Rapid City, South Dakota 57701, USA*
[34]*University of Southern Maine, Portland, Maine 04104, USA*
[35]*Syracuse University, Syracuse, New York 13244, USA*
[36]*Tel Aviv University, Tel Aviv, Israel, 69978*
[37]*University of Texas, Arlington, Texas 76019, USA*
[38]*Tufts University, Medford, Massachusetts 02155, USA*
[39]*Center for Neutrino Physics, Virginia Tech, Blacksburg, Virginia 24061, USA*
[40]*University of Warwick, Coventry CV4 7AL, United Kingdom*

(Received 10 November 2024; accepted 29 April 2025; published 15 May 2025)

Neutrino-nucleus cross section measurements are needed to improve interaction modeling to meet the precision needs of neutrino experiments in efforts to measure oscillation parameters and search for physics beyond the Standard Model. We review the difficulties associated with modeling neutrino-nucleus interactions that lead to a dependence on event generators in oscillation analyses and cross section measurements alike. We then describe data-driven model validation techniques intended to address this model dependence. The method relies on utilizing various goodness-of-fit tests and the correlations between different observables and channels to probe the model for defects in the phase space relevant for the desired analysis. These techniques shed light on relevant mismodeling, allowing it to be detected before it begins to bias the cross section results. We compare more commonly used model validation methods which directly validate the model against alternative ones to these data-driven techniques and show their efficacy with fake data studies. These studies demonstrate that employing data-driven model validation in cross section measurements represents a reliable strategy to produce robust results that will stimulate the desired improvements to interaction modeling.

DOI: 10.1103/PhysRevD.111.092010

I. INTRODUCTION

The desire to measure neutrino-nucleus cross sections is motivated by the needs of modern neutrino experiments. Moving forward, precision measurements of muon-to-electron neutrino oscillations [1–4] will enable the characterization of charge-parity violation in the neutrino sector [5], the determination of the neutrino mass ordering [6], and searches for physics beyond the Standard Model. These oscillations are studied through measurements of neutrino-nucleus interactions, which represent the nucleus's response to a neutrino probe [7]. In order to interpret these measurements properly and to disentangle any new physics from background Standard Model processes [3,4], this data must be accompanied by precise modeling of neutrino-nucleus scattering in the ∼GeV energy region [8] benchmarked with rigorous cross section measurements.

Neutrino-nucleus interactions present a challenging theoretical problem. They involve both the electroweak force and the strong force, which is nonperturbative in the relevant energy regime [9], all within the complex multibody environment of the nucleus. This results in an incomplete theoretical description of neutrino-nucleus interactions in the ∼GeV regime. However, these challenges do not necessarily prevent the success of accelerator-based neutrino oscillation experiments [1,2]. As long as the nucleus's response to the neutrino probe can be described with sufficient detail, the desired precision can still be achieved in oscillation measurements. For this purpose, experiments utilize event generators, which simulate neutrino interactions through a collection of effective models constructed to explain different modes of neutrino-nucleon interactions [8] and are used to estimate event selection efficiencies and detector responses. In this light, neutrino-nucleus interaction cross section measurements are calibration points which help ensure that simulations provide a robust description of nature.

*Contact author: microboone_info@fnal.gov

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 SCOAP³.

However, this naturally raises the question of model dependence in cross section measurements, which likewise utilize event generators to correct for backgrounds, efficiencies, finite resolution, and biases in the reconstruction of kinematic quantities. The process of "extracting" or "unfolding" the cross section, which constitutes mapping the reconstructed distributions onto physics quantities, assumes that the model captures the true value of these corrections within its uncertainties. Though this is required in order to obtain a robust result, the validity of this assumption is not known *a priori*, and must be verified in any cross section measurement.

To address this issue of model dependence, we propose utilizing a data-driven model validation procedure to test whether the model, together with its uncertainties, can describe the data in a self-consistent manner. The validation is based upon constructing a variety of data-driven tests in order to identify mismodeling relevant to the desired cross section measurement. When the model passes validation, it suggests that the data is a suitable realization of the range of possibilities afforded by the model's uncertainties. We demonstrate that, in general, when this condition is met, any bias introduced in the cross section extraction will be within the quoted uncertainties of the measurement, thereby building confidence in a robust result. The MicroBooNE experiment has previously used these data-driven techniques in a variety of analyses [10–13]. These results include measurements of visible kinematic variables, such as the energy and angle of the outgoing muon in a charged current muon neutrino (ν_μCC) interaction, as well as measurements of derived quantities, such as the cross section as a function of the incoming neutrino energy or the energy transferred to the nucleus. We emphasize that these techniques can be employed equally to extract neutrino energy dependent cross sections and to extract cross sections as a function of visible variables.

This paper is organized as follows. In Sec. II, we motivate why model validation is critical in all cross section measurements and explore where model dependence may arise. In Sec. III, we describe various techniques which may be used to detect relevant mismodeling and general considerations for designing a sufficiently sensitive model validation procedure. In Sec. IV, we compare the usage of the fake datasets in two cases: one in evaluating the model uncertainties and the other one on the model validation procedure. We then present fake data studies (FDSs) as described in the latter case to demonstrate the efficacy of the data-driven model validation procedure. These points are then summarized in Sec. V.

II. IMPORTANCE OF MODEL VALIDATION

A. *A priori* information about event generators

The complex nature of neutrino-nucleus scattering in the ∼GeV energy regime necessitates the use of effective models to describe these interactions. Event generators are formed from a collection of these effective models and are used by neutrino experiments to simulate neutrino interactions and interpret experimental data. A variety of generator codes exist, including GENIE [14], NEUT [15], NuWro [16], and GiBUU [17]. Though these generators are built upon similar underlying theory and may even employ some of the same models, the details of the implementation and choice of model parameters can have a large impact on the generator's prediction and its ability to describe data. Despite their crucial role, each effective model generally suffers from not being able to describe the corresponding interaction modes across the complete phase space [18,19]. This poses issues for experiments because event generators are often required to be capable of describing the complete contents of the final state and to provide coverage over the entirety of the available phase space, but usually require substantial interaction uncertainties to do so.

This motivates the need for data-driven inputs to inform the allowed parameters for these models and reduce their uncertainties. Experiments often supplement generators with tailored "tunes" to better represent their data [20–25]. The success of this strategy in fulfilling the requirement of generating complete final-state particle kinematics in neutrino-nucleus interactions has been demonstrated with the consistent $|\Delta m^2_{\text{atm}}|$ values extracted from the accelerator neutrino oscillation experiments [1,2] and reactor antineutrino oscillation experiments [26,27], in which the inverse β decay process allows for a simpler and more precise reconstruction of neutrino energy. Nevertheless, because of their hybrid nature, event generators tend to better describe inclusive processes or phase spaces where sufficient data were accumulated. For other exclusive processes, which require a more detailed description of the hadronic final states, event generators are more likely to fall short.

This places experiments in the following situation. On one hand, it is unlikely the parameters available in current event generators are sufficient in describing interaction modes in the complete phase space. On the other hand, the conservative uncertainties assigned on these parameters partially mitigate this shortcoming but often lead to large systematic uncertainties. As such, cross section measurements that treat the reliance on event generators with care remain essential in stimulating improvements to simulation that will enable the desired level of precision in current and future neutrino experiments.

B. Sources of model dependence

Extracting cross section measurements from experimental data constitutes mapping reconstructed distributions to truth counterparts that can be more readily compared to external predictions. In this process, an overall model, generally consisting of flux, detector, and interaction models, is required to estimate the

mapping from reconstructed quantities to true quantities. This leaves such measurements susceptible to model dependence, hence the need for model validation.

To see more explicitly how such a dependence arises, consider the general equation describing the process of extracting a flux-averaged differential cross section, $\frac{d\sigma}{dx}$, as a function of a given truth variable x from a measured distribution $n_a = d_a - b_a$, where, for the ath bin, d_a is the number of selected events, b_a is the background prediction, and n_a is the estimated number of selected signal events. In generic terms, this equation takes the form

$$\left(\frac{d\sigma}{dx}\right)_\beta = \frac{\Sigma_a U_{\beta a}(d_a - b_a)}{\Phi \cdot T \cdot \varepsilon_\beta \cdot (\Delta x)_\beta}, \tag{1}$$

where $U_{\beta a}$ is the smearing matrix describing the predicted probability that an event in reconstructed bin a belongs to truth bin β, ε_β is the estimated selection efficiency for signal events in truth bin β, $(\Delta x)_\beta$ are the widths of the truth bins, and T is the number of target nuclei. The total integrated flux prediction Φ is the integral of the predicted neutrino flux $\phi(E_\nu)$ over the entire neutrino energy spectrum

$$\Phi = \int \phi(E_\nu) dE_\nu. \tag{2}$$

In the case of extracting the cross section as a function of the neutrino energy, $\sigma(E_\nu)$, the integrated flux prediction acquires a dependence on the truth bin, Φ_β, to account for the fact that any truth bin only corresponds to a specific subset of the neutrino flux spectrum.

From Eq. (1), it is clear that cross section extraction depends on a model prediction for the neutrino flux, the rate of signal and background events, the selection efficiency, and detector effects that result in the imperfect reconstruction of kinematic quantities accounted for in the smearing matrix. This dependence on the overall model means that mismodeling in the phase space relevant to the cross section extraction can introduce bias into the measurement. For measurements that must contend with low efficiencies, purities, or substantial detector effects, this phase space may extend beyond the measured one into background channels, where mismodeling could lead to an incorrect background correction, or into channels and observables adjacent to the measurement where mismodeling could suggest an incorrect efficiency estimation. The fundamental challenge here is that one does not know if the model used in the extraction can describe nature. Moreover, when the overall model is unable to describe nature, it is usually unclear if the discrepancy is due to detector, cross section, or flux effects and may even be due to a combination of a variety of sources of mismodeling. In these cases, there may be a need for additional uncertainties beyond those inside the model, and therefore a form of model validation is required to verify that the existing uncertainties provide sufficient coverage of the data.

Model validation is especially important to any analysis that extracts cross sections as a function of the neutrino energy or energy transferred to the nucleus. These quantities are not directly observable and must be estimated from the measurement of the visible leptonic and hadronic energy. The way the unfolding maps from the reconstructed hadronic energy to the true energy transfer depends on the overall model, particularly the cross section model, to correct for the missing hadronic energy going to particles that cannot be reconstructed by the detector. Care must be taken to avoid introducing model dependence that biases these measurements beyond stated uncertainties, and a rigorous examination of the model is essential.

Visible kinematic variables do not entirely avoid model dependence either. Whenever a measurement relies upon mapping from reconstructed quantities to true quantities this mapping must be estimated through a model, motivating the need for model validation in these cases. In particular, the mapping for quantities like the available energy E_{avail} [12,18,28–32], often defined as the sum of reconstructable energy deposited by visible particles, serves as an alternative to the energy transfer but has a strong dependence on the accurate simulation of particles that deposit energy in the detector. Since different modeling and reconstruction failures may be present in different final states, the mapping from reconstructed to true E_{avail} requires a robust description of the complete contents of the hadronic final state, thereby making it susceptible to model dependence.

Similar forms of mismodeling may impact measurements of visible kinematic variables that depend on the final state hadronic kinematics. Measurements of differential cross sections for more exclusive final states are also susceptible to mismodeling as they generally impose a detection threshold based on the reconstruction performance of the detector. Because selection efficiencies can be a complex function of the particle kinematics and contents of the final state, substantially different modeling of nuclear effects between generators can lead to drastically different predictions for the number of above threshold particles [18]. This can have a large impact on the estimated selection efficiencies, background predictions, and bin migration effects, thereby introducing model dependence into these results.

C. Real versus nominal flux

Measurements of differential cross sections inherently depend on the mapping between neutrino energy and the visible kinematic variable when unfolding or generating predictions. Due to the broad energy spectrum of neutrino beams, cross section measurements are typically averaged over the integral of the entire flux spectrum, Φ, as described in Eq. (1). This may reduce the dependence on this

mapping, but does not entirely remove it. For example, a 1 GeV neutrino cannot produce a muon with 2 GeV of kinetic energy. Therefore, despite being averaged over the entire flux, any given muon energy measurement bin naturally maps to a subrange of the neutrino energy spectrum. This introduces model dependence related to the neutrino flux prediction.

The extent of this model dependence depends on subtleties in the treatment of the flux and its related uncertainties. In the literature, this has been described as whether the measurement is averaged over an assumed well defined *nominal* neutrino flux spectrum, or averaged over the unknown *real* neutrino flux spectrum impingent on the detector [33]. When extracting cross sections in the real flux, unfolding amounts to correcting for detector effects with a presumed cross section and flux model. When extracting cross sections in a nominal flux, one corrects for detector effects but then also extrapolates from the (unknown) real flux to the nominal flux in the unfolding. This distinction is subtle but leads to important differences in the treatment of flux uncertainties. These uncertainties can be quite prominent in the context of accelerator-based neutrino cross section measurements, for which flux uncertainties are typically 5%–10%, though can be even larger in instances where the flux is poorly known.

The distinct advantage of cross section extraction in the real neutrino flux spectrum is that it minimizes the usage of the mapping between neutrino energy and visible kinematic variables in the extraction of cross sections. This makes these results less model-dependent. However, extractions in the real neutrino flux spectrum do not eliminate the dependence on the mapping from neutrino flux to observables but, rather, push it down the line to future analyzers of the data who are required to supply their own flux model and associated uncertainties when making comparisons. In other words, theorists must provide their own model of the mapping, which is usually just the prediction of the nominal flux and its uncertainties at the measurement location as reported by the experiment. However, in order to make a robust comparison between predictions and the data, these flux uncertainties should include correlations between the assumed neutrino energy spectrum and the extracted cross section result, which typically already include flux normalization uncertainties. These correlations are generally not reported by the experimental collaboration and may lead to incorrect conclusions about how well alternative models describe the data [33].

If an analysis extracts cross sections in the nominal neutrino flux spectrum, model-dependence arises when the mapping between neutrino energy and the visible kinematic variable is used to extrapolate from the real to the nominal neutrino spectrum. In this approach, flux uncertainties are estimated by varying the assumed real flux, but not the well-defined nominal flux, when determining the impact of flux effects. Referring to Eq. (1), this amounts to keeping Φ constant while reevaluating n_a based on the real flux in each flux systematic universe. This approach allows the covariance matrix to include the uncertainties of extrapolating the data from the unknown real neutrino flux to the nominal flux and corresponding Φ that the results are averaged over. In this case, comparisons to external theory or event generator predictions are straightforward because there is no need for future analyzers of the data to utilize additional flux uncertainties. Flux systematics are entirely accounted for by the experimentalist when estimating uncertainties for the cross section extraction. In this case, data-driven model validation can serve as a mechanism to quantify the bias introduced in the extrapolation from real to nominal flux and help avoid underestimating or overestimating flux uncertainties. This is similar to the utility of data-driven model validation when extracting derived quantities dependent on the neutrino flux spectrum, such as the cross section as a function of the neutrino energy or energy transferred to the nucleus, where it likewise serves to validate the mapping between reconstructed quantities and the incoming neutrino energy spectrum. The dependence on the model is greater for these derived quantities, hence a greater need for careful validation, but one should still be cautious in entirely ignoring this model dependence in the context of visible kinematic variables.

To reiterate, the primary difference between measurements in the real and nominal flux is the amount they depend on the mapping from neutrino energy to visible kinematics variables and their treatment of flux shape uncertainties, which are fully included in the covariance matrix extracted in a nominal flux measurements but not in a real flux measurement. The challenges of comparing predictions to measurements made in the real flux are described in Ref. [33]. Here, we illustrate those challenges with a toy example. We consider an ideal ν_μCC cross section measurement as a function of the muon energy with perfect efficiency, no background, and perfect muon energy reconstruction. In this case, the smearing matrix $U_{\beta a}$, background b_a, and the efficiency ε_β disappear from Eq. (1) leaving just the reconstructed signal event counts n_a, which are now exactly equal to the true signal event counts n_β, the integrated flux prediction Φ, the number of target nuclei T, and the bin widths $(\Delta E_\mu)_\beta$

$$\left(\frac{d\sigma}{dE_\mu}\right)_\beta = \frac{n_\beta}{\Phi \cdot T \cdot (\Delta E_\mu)_\beta}. \tag{3}$$

The only uncertainties on the extracted cross section from this equation are statistical uncertainties on n_β and an uncertainty on the integrated neutrino flux prediction Φ. For the latter, we use the MicroBooNE ν_μ flux with systematic uncertainties taken from Ref. [34]. We then generate fake data distributions that fluctuate n_β according to statistical and flux model uncertainties in the same 11

true muon energy bins used in Ref. [10]. This is performed by sampling the multivariate Gaussian distribution using a singular value decomposition, as well as Poisson statistical fluctuations.

In this toy study, we assume the true cross section is exactly the prediction from GENIE with CCQE and CC2p2h parameters set according to [20], which will be referred to as the "MicroBooNE tune." The real flux used to produce the observed n_β before additional statistical fluctuations is different in each fake dataset. The flux prediction used in the cross section extraction is kept constant across all universes and corresponds to the central value (CV) MicroBooNE ν_μ flux prediction. In this case, besides statistical fluctuations, any deviation in the results from the MicroBooNE tune CV is due to systematic fluctuations in the flux model.

Three different methods are utilized to compare the real flux-averaged cross section result to prediction; the "incorrect method," the "flawed method," and the "correct method." The "incorrect method" includes no additional flux uncertainties when comparing the extracted cross section with the prediction. This method is easy for analyzers to perform as it only requires them to supply a CV for the flux. The flawed method includes flux uncertainties on the prediction, but does not account for correlations between the extracted cross section results and the assumed neutrino energy spectrum. This is achieved by generating many alternative predictions, each of which uses a unique flux drawn according to the uncertainties on the flux model, and using them to build a covariance matrix. The flawed method is more difficult to perform, since it requires future analyzers of the data to have access to and utilize a published flux covariance matrix to derive an uncertainty for their prediction. The correct method is similar to the flawed method and also takes into account the flux uncertainty in the generator prediction by building a covariance matrix from predictions obtained with the flux model varied according to its uncertainties. The difference is that in the correct method one includes the correlations between the integrated flux normalization uncertainty in the extraction and the flux uncertainty in the prediction.

In Fig. 1, we illustrate the three methods by comparing the extracted real flux-averaged cross sections to a generator prediction that assumes the CV of the MicroBooNE flux model. The associated uncertainties on the real flux measurement and central value of the extracted cross section are the same in each method. The differences arise from the treatment of flux uncertainties for the prediction. As an additional point of reference, we verify that the correct method gives identical χ^2 values to those extracted with the nominal flux technique, within numerical errors. Figure 1(a) shows the distribution of χ^2 values obtained when the extracted cross section from each of 10000 fake datasets is compared to the MicroBooNE tune prediction,

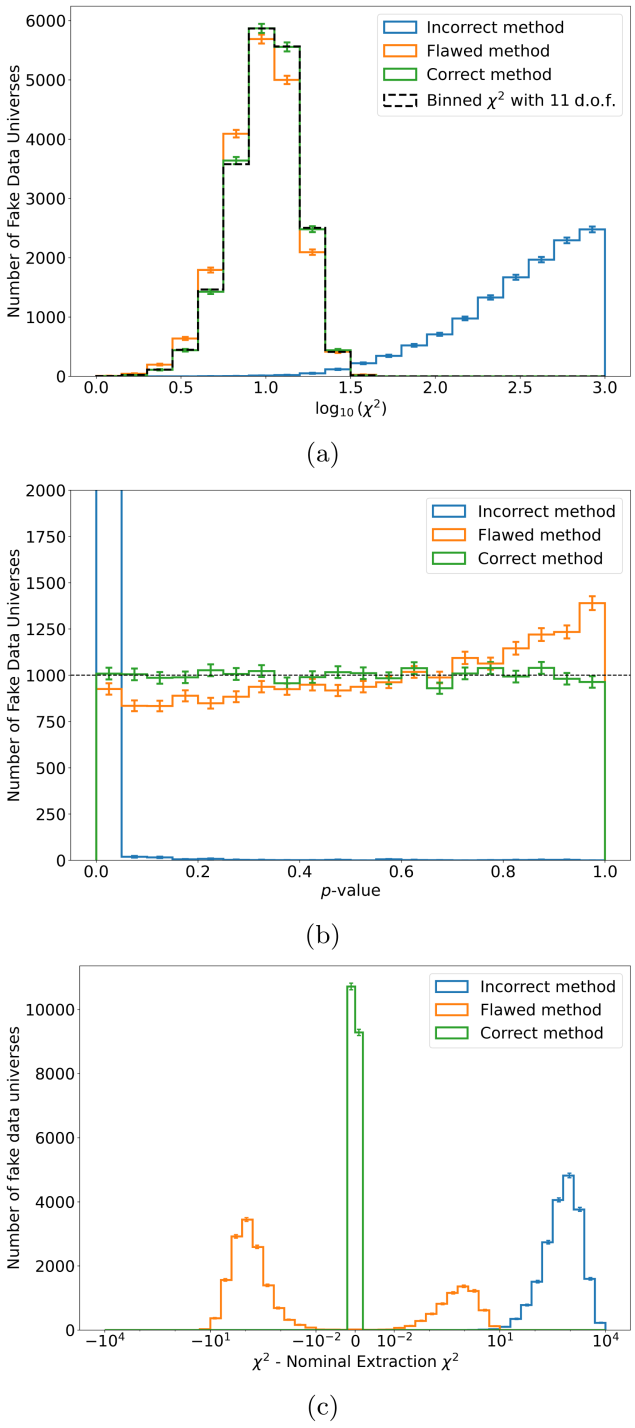


FIG. 1. Toy study illustrating three methods of comparing a real flux measurement with prediction. The incorrect method neglects flux uncertainties on the prediction. The flawed method includes flux uncertainties on the prediction, but neglects flux correlations between the prediction and extraction. The correct method includes flux uncertainties on the prediction and correlations with the extracted result. The distributions of χ^2 values obtained for each method is shown in (a) alongside a binned χ^2 distribution. The corresponding distribution of p-values is shown in (b). The distribution of the difference between χ^2 values and the nominal flux extraction χ^2 obtained for each method is shown in (c).

which corresponds to the truth for these fake datasets. Figure 1(b) shows the same, but with the χ^2 values converted into p-values. Figure 1(c) expands this comparison further by showing the difference between χ^2 values obtained with the various methods instead.

The incorrect method, though easiest to perform, generates a distribution of χ^2 values shifted toward larger values than expected from a χ^2 distribution with 11 degrees of freedom. This arises from the fact that, in a given fake dataset, there may be a sizable difference between the real flux and the CV flux, which produces a large χ^2 value. The possibility of this difference is not accounted for due to the fact that flux uncertainties were neglected on the supplied flux model and thus this mismodeling of the flux is wrongly attributed to defects in the MicroBooNE tune prediction.

For the flawed method, the distribution of χ^2 values is significantly closer to what is expected from a χ^2 distribution. However, it is still shifted to smaller values than the χ^2 distribution with 11 degrees of freedom. This is the result of double counting the flux uncertainty, which is done on both the extraction and prediction side.

The correct method generates a distribution of χ^2 values that agrees with a χ^2 distribution with 11 degrees of freedom. However, in most current real-flux-averaged cross section data releases, this method is impossible for analyzers of the data to perform, since it requires additional information about how the extracted cross section is correlated with flux spectrum variations. This information is rarely provided by experimentalists, rendering the correct method of comparing cross sections extracted in the real neutrino flux to external prediction likely impossible.

The deviation between the three methods is further illustrated in Fig. 1(c), which shows the distribution of the universe by universe differences between χ^2 values obtained for the measurement reported in the nominal flux and those obtained for the measurement reported in the real flux. From this figure, it is apparent that the incorrect method consistently produces a larger than accurate χ^2. Though the flawed method produces a distribution of χ^2 values shifted toward smaller values than a χ^2 distribution with 11 degrees of freedom, it can still produce a larger χ^2 value than obtained for the cross section extracted in the nominal flux. Since the nominal flux method is able to reproduce a χ^2 distribution with 11 degrees of freedom, this suggests that the flawed method does not strictly underestimate the χ^2. The aforementioned phenomena are also described in [33]. These observations further complicate the use of real flux-averaged measurements as they prevent one from using the flawed method and incorrect method as lower and upper bounds on the GoF between data and prediction and necessitates the use of the correct method for a rigorous comparison.

As demonstrated by this toy study, the requirements of reporting and using neutrino flux uncertainties and their correlations with cross section results make it very difficult, if not impossible, for theorists to properly compare their predictions with cross sections extracted using the real neutrino spectrum. Indeed, this may be related to some of the issues encountered in recent efforts to tune event generators to experimental data [20,24,35]. Because of this, we advocate for extracting cross sections in the nominal neutrino flux spectrum, which allows the results to be reported in a single well-defined flux. With this method, all flux uncertainties are included in the covariance matrix obtained in the unfolding. However, this method also introduces additional model dependence on the experimental side associated with the mapping between the neutrino energy spectrum and the measured kinematic variables. We thus advocate for data-driven model validation when extracting cross sections in any physics variable at the nominal neutrino flux spectrum. These methods are well suited to examine the entire uncertainty budget and thus help to ensure that the uncertainties associated with extrapolating from the true to the nominal neutrino flux spectrum are sufficient.

III. DATA-DRIVEN MODEL VALIDATION

Data calibration and data-driven model validation are closely related. Instead of using data to replace part of the model in a calibration procedure, the model validation procedure focuses on testing whether the model used for the extraction can describe the data in a self-consistent manner. When the validation indicates that the data falls within the allowed parameter space of the model, this builds confidence that the bias introduced in the cross section extraction will, in general, be within the quoted uncertainties of the measurement. This is demonstrated with several case studies in Sec. IV. The data-driven methods we propose are in contrast with other approaches to model validation, which usually examine the variation between multiple different model predictions for backgrounds, efficiencies, or biases in the reconstruction of kinematic quantities. This is commonly done through FDSs used to inform additional uncertainties to be added to the primary model used for extracting the data cross sections. Differences between the role of FDSs in model validation based on comparisons with alternative models and the data-driven model validation we propose is discussed in more detail in Sec. IV A.

A. Tools for model validation

1. Goodness of fit tests

The data-driven model validation procedure is based on comparing the model prediction to the data with goodness of fit (GoF) tests that quantify the ability of the overall model to describe the data. Any GoF tests performed over a reconstructed space distribution should be evaluated in such a way that correlations between bins are accounted for.

This can be achieved with a χ^2 test statistic constructed via the covariance matrix formalism given by

$$\chi^2 = (M-P)^T \cdot V^{-1} \cdot (M-P), \tag{4}$$

where M is the measurement vector, P is the prediction vector, and V is the total covariance matrix which includes the uncertainties on the reconstructed distribution and bin-to-bin correlations. These χ^2 values are interpreted by using the number of degrees of freedom, ndf, which corresponds to the number of bins, to obtain p-values.

To obtain sufficient stringency, we require that all tests which probe the model in the phase space relevant for the cross section extraction yield a p-value greater than 0.05, which indicates that the model is able to describe the data at the 2σ significance level. If all tests pass this level of stringency, then the model is considered to be validated and may be used for the cross section extraction. In this case, given a set of models that pass validation, the discrepancy between the results extracted using different models are expected to be smaller than the total uncertainties.

2. χ^2 decompositions

In an overall GoF test, it is possible that conservative uncertainties have hidden a significant discrepancy between data and model in some bins that may bias the cross section extraction in select regions of phase space. However, it is challenging to further evaluate the GoF in select regions of phase space because there are generally strong bin-to-bin correlations in the reconstructed distributions. To address this, the overall test statistic can be decomposed by diagonalizing the covariance matrix, thereby making all bins uncorrelated. In this transformed basis, the absence of correlations allows for a rigorous quantification of the GoF between the data and model prediction within a single bin given the uncertainties of the diagonalized covariance matrix. This allows one to test the local GoF of distributions in the decomposed space. Though it is challenging to draw physics conclusion from this decomposed space, the technique is a useful statistical tool in evaluating the compatibility between data and prediction when significant correlations in reconstructed distributions may render the overall χ^2 test statistic insensitive to local discrepancies.

In particular, the symmetric covariance matrix can be decomposed into $V = \tilde{Q} \cdot \Lambda \cdot \tilde{Q}^T$ where Λ contains the eigenvalues of V along its diagonal and $\tilde{Q}$ has the corresponding eigenvectors as its columns. Defining $Q = \tilde{Q}^{-1}$ and $\Delta = (M-P)$ allows Eq. (4) to be written as

$$\chi^2 = (Q \cdot \Delta)^T \cdot (Q \cdot V \cdot Q^T)^{-1} \cdot (Q \cdot \Delta). \tag{5}$$

By further defining $\epsilon_i = \Delta'_i / \sqrt{\Lambda_{ii}}$, where $\Delta' = Q \cdot \Delta$, the above expression can be written as

$$\chi^2 = \Delta'^T \cdot \Lambda^{-1} \cdot \Delta' = \sum_i \epsilon_i^2. \tag{6}$$

Because Λ is diagonal, the ϵ_i are all independent, and the χ^2 is now written in terms of independent components, also known as the χ^2 decomposition format. These ϵ_i are normally distributed and may be interpreted as the significance of the tension between data and prediction in the corresponding ith bin of the eigenvalue basis. Compared to the deviations on individual bins in the correlated reconstructed space distribution, the deviations between data and model in ϵ_i take into account the correlations and can be individually evaluated quantitatively in a consistent manner.

A local χ^2 and corresponding p-value, χ^2_{local} and p_{local} respectively, can be computed from any number of large ϵ_i which indicate the presence of a local discrepancy with

$$\chi^2_{\text{local}} = \sum_i^r \epsilon_i^2, \tag{7}$$

where r is the number of points summed over and the number of degrees of freedom in the χ^2_{local} distribution. When computing local p-values in this way, a large number of tests can be examined, increasing the odds of randomly producing a larger value. As such, one must correct for the look-elsewhere effect [36,37] by converting the local p-value into a global p-value which describes the probability of observing such a discrepancy in any combination of the ϵ_i. Several previous MicroBooNE analyses, such as [12], computed p_{local} from all ϵ_i above 2σ and then performed the p_{local} to p_{global} conversion according to

$$p_{\text{global}} = 1 - (1 - p_{\text{local}})^{\binom{n}{r}} = 1 - (1 - p_{\text{local}})^{\frac{n!}{(n-r)!r!}}, \tag{8}$$

where n is the total number of bins and r is the number of ϵ_i above the 2σ threshold, which we refer to as "extreme values." This accounts for the fact that for n independent ϵ_i there are $\binom{n}{r}$ ways to chose r with extreme values. However, calculating a p_{local} and converting it in a p_{global} in this manner does not produce an unbiased estimator, which is an undesirable attribute for a test statistic. Moreover, with this method, one must choose their definition of an extreme value and the choice of the threshold can impact the resulting p_{global}. As an example, consider a case in which a 2σ threshold is chosen and a distribution with 10 bins shows one ϵ_i at 3σ, another at 1.9σ and the rest all less than 1σ. This produces $p_{\text{global}} = 1-(1-0.0027)^{10} = 0.027$. However, if this distribution were to have its second most extreme ϵ_i at 2σ instead, this yeilds $p_{\text{global}} = 1-(1-0.0015)^{45} = 0.065$. This is a undesired result; the worse agreement in the second most extreme ϵ_i increases the p_{global} rather than decreasing it. This property arises from the fact that Eq. (8) assumes that all

observations are uncorrelated. Though each ϵ_i is uncorrelated, the observation of ϵ_1 above threshold is correlated with the observation of both ϵ_1 and ϵ_j above threshold, thereby violating the assumption behind Eq. (8).

As such, we suggest two alternatives. Rather than examining all ϵ_i above an arbitrary threshold, one can instead select only the largest ϵ_i. In this case $r = 1$ and Eq. (8), which is now valid, simplifies to

$$p_{\text{global}} = 1 - (1 - p_{\text{local}})^n. \tag{9}$$

If one still wished to examine multiple ϵ_i in accordance with Eq. (7), they could instead employ a frequentist method. This would entail simulating many pseudoexperiments with n bins, then, in each pseudoexperiment, finding the minimum possible local p-value, $p^{\text{min}}_{\text{local}}$, out of all combinations of bins. The same quantity would then also be calculated for the observed data distribution. By computing the fraction of pseudoexperiments with a $p^{\text{min}}_{\text{local}}$ below the $p^{\text{min}}_{\text{local}}$ of the data, one obtains a rigorous p_{global} for the data distribution.

Both of these alternative methods of calculating a p_{local} and converting it in a p_{global} produce an unbiased estimator that will not decrease the p_{global} if any of the p-values for individual ϵ_i increases. They also remove any dependence on an arbitrary threshold. Nevertheless, in most circumstances, the differences between these three methods will be small and we choose to utilize Eq. (9) in the fake data studies presented in Sec. IV.

An example of utilizing the χ^2 decomposition is illustrated in Fig. 2 on MicroBooNE data. The distribution of interest in this figure is the ν_μCC selection from [38] binned as a function of the reconstructed hadronic energy, which Fig. 2(b) shows in reconstructed space. Figure 2(c) then shows the χ^2 decomposition and includes both the resulting distribution of ϵ_i values in the decomposition space and the matrix used to transform from the reconstructed space to the decomposition space. This matrix, though generally challenging to interpret, can provide some insight into the mapping between the two spaces and the types of discrepancies that would result in significant tension in individual ϵ_i. Taking the first decomposition bin as an example, we see that this bin receives a negative contribution from the first three reconstructed bins and a positive contribution from all higher energy bins. As such, an effect that modifies the event rate differently for the first three low energy bins and the higher energy bins would likely cause a large discrepancy in this decomposition bin which would likewise result in a large ϵ_i indicative of significant mismodeling.

3. Conditional constraints

The conditional constraint procedure [39] can be used to increase the stringency of the validation by providing an additional means of probing for relevant mismodeling. In this procedure, a constraint from one set of data distributions is used to narrow the allowed model parameter space of a different set of distributions. This amounts to an updated central value and a reduced uncertainty band on the constrained prediction. More explicitly, consider two distributions, X and Y, with model predictions μ^X and μ^Y, and a covariance matrix containing these two channels (X, Y):

$$\Sigma = \begin{pmatrix} \Sigma^{XX} & \Sigma^{XY} \\ \Sigma^{YX} & \Sigma^{YY} \end{pmatrix}.$$

Here, the distributions are assumed to be jointly Gaussian with Σ^{XX} describing the uncertainties on channel X as well as the correlations between its bins, and Σ^{YY} analogously describing the uncertainties on channel Y as well as the correlations between its bins. The correlations between the bins of X and Y are described by Σ^{YX} and Σ^{XY}. If a data measurement of channel Y results in the distribution n^Y, one can derive the prediction for X given this observation in Y to be

$$\mu^{X,\text{const}} = \mu^X + \Sigma^{XY} \cdot (\Sigma^{YY})^{-1} \cdot (n^Y - \mu^Y), \tag{10}$$

$$\Sigma^{XX,\text{const}} = \Sigma^{XX} - \Sigma^{XY} \cdot (\Sigma^{YY})^{-1} \cdot \Sigma^{YX}. \tag{11}$$

In this case, Y is referred to as the *constraining* channel and X is referred to as the *constrained* channel with its posterior prediction and uncertainties given by $\mu^{X,\text{const}}$ and $\Sigma^{XX,\text{const}}$, respectively. This method can be understood as deriving a conditional probability distribution for X given Y. With this context, Eq. (10) describes the conditional mean for X and Eq. (11) describes the conditional covariance.

The amount by which the conditional mean differs from the prior central value prediction will be dictated by the three factors in the second term on the right of Eq. (10). Smaller uncertainties on the constraining distribution results in a more powerful constraint and a larger update to the prediction, this enters through $(\Sigma^{YY})^{-1}$. The correlations between the two channels enters through Σ^{XY}, with a higher level of correlation increasing the size of the update to the prediction. The amount the prediction in the constraining channels differs from the data observation enters through $(n^Y - \mu^Y)$, with a larger deviation from the prediction producing a larger update to the prediction in the constrained channel. The reduction in uncertainties on the constrained distribution is described by the second term on the right in Eq. (11). Likewise, this term has factors that depend on the size of the uncertainties on the constraining distribution and the size of the correlation between the two channels, namely $(\Sigma^{YY})^{-1}$ and Σ^{XY}, respectively. Smaller uncertainties on the constraining channel and larger correlations will increase the resulting reduction in uncertainties on the constrained channel. However, unlike the second term in Eq. (10), the second term in Eq. (11) does not depend on the data observation

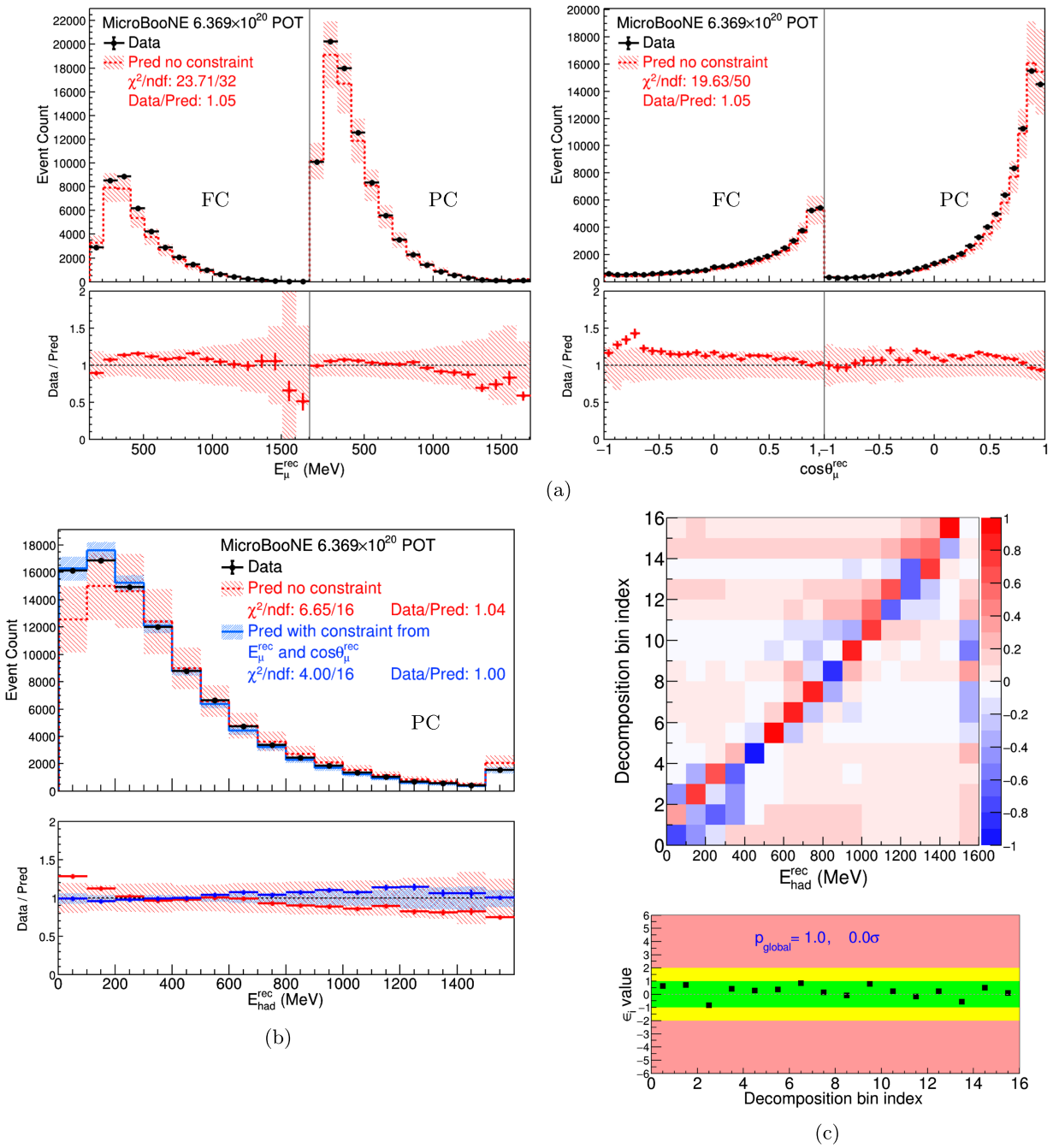


FIG. 2. Demonstration of the conditional constraint and χ^2 decomposition techniques using MicroBooNE data and the ν_μCC selection from [38]. (a) The reconstructed muon energy and muon scattering angular distributions. Both distributions utilize separate bins for FC and PC events. These are used to constrain the distribution shown in (b) and thus correspond to channel Y in Eq. (10). (b) The reconstructed hadronic energy distribution for PC events. This distribution is constrained by the distributions shown in (a) and thus corresponds to channel X in Eq. (10). The MC prediction before (after) constraint from the observed muon energy and angle distributions is shown in red (blue). In both (a) and (b), the top panels show the data and MC reconstructed distributions and the bottom panels show the data to MC ratio. The uncertainties of the prediction are shown in the bands and the data statistical uncertainties are shown on the data points. (c) Further examination of the GoF for (b) with the χ^2 decomposition. The top panel shows the matrix used to transform the constrained reconstructed hadronic energy distribution into the eigenvalue basis of the covariance matrix. The bottom panel shows the significance of the tension between data and MC, otherwise known as ϵ_i values, in the eigenvalue basis.

in the constraining channel; the amount of "additional information" provided by the constraint remains constant regardless of the data observation.

The procedure described in Eqs. (10) and (11) allows a GoF test to be performed on X after the constraints from Y. This allows the simultaneous examination of the correlated modeling of X and Y, which provides additional information about the compatibility between the model and data. Typically, the constrained and constraining channels have highly correlated model predictions due to detector and physics effects that impact both sets of distributions. The model's description of the correlations is dictated by the modeling of these effects. The reduction in uncertainties for the prediction on the constrained distribution provides additional sensitivity by providing a means of exploring the modeling of these correlations. This direct incorporation of the correlations into the updated central value prediction, uncertainties, and subsequent GoF tests make it straightforward to perform a rigorous examination of the consistency of the modeling across channels and observables, which is a tangible advantage of this method. The modeling of X, Y, and the relationship between X and Y must all be sufficient for the model to pass a constrained validation test.

An example of using the conditional constraint can be seen in Fig. 2. In this example, Y is defined to be the distributions of the reconstructed muon kinematics, namely, the reconstructed muon energy, E_μ^{rec}, and the muon angle, $\cos\theta_\mu^{\text{rec}}$, where θ_μ^{rec} is defined as the angle the muon scatters with respect to the incoming neutrino beam. These distributions are shown in Fig. 2(a). Events fully contained (FC) and partially contained (PC) within the detector are placed into separate bins to better separate these two classes of events, which may be sensitive to different forms of mismodeling. Distribution X is chosen to be that of reconstructed hadronic energy, $E_{\text{had}}^{\text{rec}}$, for events partially contained (PC) within the detector. This distribution is shown in Fig. 2(b). Note that in this case, X and Y contain the same events, and therefore statistical correlations are present in the covariance matrix. The MicroBooNE Monte Carlo (MC) prediction before constraint is shown in red and the prediction after constraint is shown in blue. The shift between the red and blue predictions is dictated by the observation in data for the muon kinematics and by the model's predicted relationship between the muon kinematics and the reconstructed hadronic energy. This difference between the red and blue predictions indicates how large of a correction to the central value is induced by the constraint through the rightmost term in Eq. (10). The bands surrounding the prediction, which show the uncertainties before and after constraint, illustrate the large reduction in uncertainties and enhanced sensitivity to mismodeling that is obtainable with the conditional constraint. In particular, this test is expected to provide sensitivity to the modeling of the missing hadronic energy as is described in more detail in Sec. III B. Other examples of similar constraint tests can be found in [10–12].

The conditional constraint procedure borrows principles from data calibration. It uses the data to reduce the uncertainty on the model prediction thereby allowing for more stringent examinations of the model. However, it should be emphasized that, in this context, the sole purpose of these constraints is model validation and they are not used in the cross section extraction. Since the constraining distributions are often using the same set of events as for the unfolding, this reduction of the systematic uncertainties would be superficial in the case of cross section extraction. As an extreme example, utilizing the exact same distribution for the constraining channel as the distribution to be unfolded results in a complete elimination of the uncertainties, which is obviously not realistic or useful for the cross section extraction. Nevertheless, as described in [40], one could also utilize the conditional constraint formalism directly in the unfolding to employ background constraints from side-band channels. Though the mathematics of this technique is identical to the data-driven validation we describe, its utility is distinct and we advocate for still employing a model validation when a background constraint is part of the analysis strategy.

B. Developing a model validation procedure

Data-driven model validation focuses on the total model uncertainties, which includes flux, detector, and cross section effects as well as any other modeling or systematic uncertainties utilized for the cross section extraction. When the overall model passes data-driven validation, it suggests that the data is a suitable realization of the range of possibilities afforded by the model. This does not mean there is no underestimation of individual components of uncertainty, but rather suggests that any underestimation of individual components is small compared to the overall uncertainty budget. This approach is consistent with the evaluation of significance in discrepancies [41,42]. As described in Sec. IV A, this differs from more commonly used approaches to model validation that tend to focus on the cross section modeling. A data-driven approach naturally evaluates the overall model, thereby making the assessment of the total uncertainty band more straightforward than many traditional approaches. This can be seen as a distinct advantage of data-driven approaches.

However, this also brings about its own set of downsides. A combination of flux, detector and cross section effects could conspire in such a way to cancel out, thereby making reconstructed space tests less sensitive to mismodeling that would still bias the extracted cross sections. The potential for such a situation is investigated in Sec. IV B 1 and should be kept in mind when deciding which tests to include in the model validation procedure. In addition, it is important to note that the procedures we outline are more applicable when the background and signal models are both well

defined and relatively well understood up to their uncertainties. If a signal significantly deviates from existing models or is completely unknown then these techniques cannot be properly applied as validation. In that case, other techniques should be employed to validate the background modeling and the signal efficiency within uncertainties before extracting a cross section.

When developing a model validation procedure for a desired cross section extraction, one must make sure to design a set of tests that are sufficiently sensitive to mismodeling in the phase space relevant for the cross section extraction. For example, validating only the overall event rate will not be sufficient for a differential cross section measurement. Such a test averages over the entire reconstructed phase space and will thus be insensitive to any mismodeling related to the shape of the distribution that could bias the extraction of a differential cross section. Selecting an appropriately sensitive set of tests represents the key to a well-designed model validation procedure. To achieve this, one should consider what forms of mismodeling may be capable of introducing bias into their cross section result and choose tests that address these possibilities. mismodeling in distributions irrelevant to the extraction do not need to be examined. It may be useful to use FDSs to evaluate the ability of the selected set of tests to detect relevant mismodeling before it begins to bias the extracted cross section results beyond stated uncertainties. A case study containing this type of FDS is presented in Sec. IV. This study mirrors the analysis done in [10] and thus also provides an example of what a full suite of tests used in a data-driven validation may look like. However, such studies are not mathematical proofs that the intended validation procedure constructed for the given analysis is guaranteed to detect relevant mismodeling. As such, it is pertinent to probe the model from a variety of different angles in order to maximize the probability of detecting problematic forms of mismodeling that would bias the cross section results beyond their uncertainties.

The more challenging a variable or channel is to reconstruct or model, the more stringently it should be examined in the validation. For visible kinematic variables in more inclusive cross section measurements, a direct data to prediction comparison over the reconstructed distribution used in the unfolding is likely sufficient. However, more challenging distributions to model or reconstruct require additional validation. Examples include distributions containing events that are not fully contained within the detector volume, distributions used to map to quantities that cannot be fully reconstructed, or a more exclusive channel with kinematics cuts that are near detector thresholds. When selecting the constraining and constrained distributions, it is useful to keep in mind Eqs. (10) and (11). The second term on the right of both these equations contain factors which dictate that smaller uncertainties on the constraining channel and larger correlations between the constraining and constrained channel will increase the power of the constraint.

Conditional constraints are particularly useful when validating aspects of the model that are impacted by substantial modeling or reconstruction challenges. One way to use this technique is to use a better understood and more easily reconstructed distribution to constrain the less understood and harder to reconstruct distribution that is more likely to show mismodeling. For example, in the case of partially contained events, a constraint from the analogous distribution of fully contained events, which are correlated with the partially contained events through common physics and detector modeling, helps provides a more stringent test. In the case of measuring more exclusive channels, where the modeling of events near the detection threshold may become especially important, additional examination of the variables relevant to the detection threshold is warranted. In this situation, the constraint provides a means of examining multiple variables simultaneously, the relationship between which may contain important information on the sufficiency of the model.

One particularly prominent example of this is the reconstruction of low energy protons. Low energy nucleons are known to be modeled significantly differently among event generators [43,44] thus leading to very different predicted selection efficiencies for protons near the detection threshold [18]. This can have a large impact on measurements related to the proton's energy, such as the kinetic energy of the leading proton or transverse kinematic imbalance variables [45–47], which are projections of the lepton and proton momentum onto the plane perpendicular to the neutrino direction. In such a case, a conditional constraint from a related visible kinematic variable less susceptible to the threshold may be useful. Examples of this include the energy and angle of the outgoing lepton. These constraints will reduce the allowed parameter space through correlations due to shared detector, flux, and cross section modeling thereby allowing for a more thorough investigation of the modeling of the proton kinematics near the detection threshold. Including an additional constraint from a closely related channel, such as one without protons, can provide additional information on the modeling of backgrounds or signal events that did not pass selection cuts and may also be useful in probing for relevant mismodeling. This can be seen as similar to studying side-bands, which are often used to evaluate the modeling of backgrounds, but intends to evaluate the modeling used in the cross section extraction more directly. The central role of the correlations in this procedure makes it straightforward to evaluate if the model can consistently treat multiple distributions which provides a means of identifying when an observation in a side-band distribution is inconsistent with the observation in the signal distribution. This is a distinct advantage over evaluating side-bands in

isolation, which does not provide such sensitivity. Though one could also obtain much of the same information through other approaches, such as tuning a model to a side-band channel or alternative kinematic variable, the ease with which one can rigorously account for and evaluate correlations between distributions makes the conditional constraint an attractive option.

When attempting to extract variables that cannot be fully reconstructed, such as the neutrino energy or energy transferred to the nucleus, a direct comparison between data and MC is likely insensitive to relevant mismodeling. These quantities include both visible and missing portions, the latter of which cannot be examined with sufficient scrutiny in a direct comparison. In this case, the conditional constraint becomes a crucial tool in evaluating the mapping from reconstructed to true quantities. The choice of these constraints can be based upon physics arguments to help provide the required level of stringency.

For the case of energy transfer ν, one can use conservation of energy to design a test that is sensitive to the modeling of missing hadronic energy, $E_{\text{had}}^{\text{missing}}$, through the examination of the correlations between the leptonic and visible hadronic energy. The visible portion of the energy transfer, $E_{\text{had}}^{\text{vis}}$, can be measured through the reconstructed hadronic energy, $E_{\text{had}}^{\text{rec}}$, without needing the model to correct for contributions that are not directly measurable. The energy of the outgoing lepton, E_ℓ, can likewise be measured through the reconstructed lepton energy, E_ℓ^{rec}, without such corrections. Together, these quantities account for the total energy of the incoming neutrino,

$$E_\nu = E_\ell + \nu = E_\ell + E_{\text{had}}^{\text{vis}} + E_{\text{had}}^{\text{missing}}. \tag{12}$$

In the rightmost equality any contributions from the initial state nucleus have been absorbed into $E_{\text{had}}^{\text{missing}}$, that is, $E_{\text{had}}^{\text{missing}}$ is defined as the difference between the initial state neutrino energy and the observable leptonic and hadronic energy present in the final state. Through the use of E_ℓ^{rec} as a constraint, the simultaneous measurement of $E_{\text{had}}^{\text{rec}}$ and E_ℓ^{rec} is able to validate the predicted relationship between these distributions. This is achieved by using E_ℓ^{rec} as channel Y and $E_{\text{had}}^{\text{rec}}$ as channel X in Eq. (10)

$$\mu^{E_{\text{had}}^{\text{rec}},\text{const}} = \mu^{E_{\text{had}}^{\text{rec}}} + \Sigma^{E_{\text{had}}^{\text{rec}} E_\ell^{\text{rec}}} \cdot (\Sigma^{E_\ell^{\text{rec}} E_\ell^{\text{rec}}})^{-1} \cdot (n^{E_\ell^{\text{rec}}} - \mu^{E_\ell^{\text{rec}}}), \tag{13}$$

where $\Sigma^{E_\ell^{\text{rec}} E_\ell^{\text{rec}}}$ describes the uncertainties on E_ℓ^{rec} and $\Sigma^{E_\ell^{\text{rec}} E_{\text{had}}^{\text{rec}}}$ describes the correlations between E_ℓ^{rec} and $E_{\text{had}}^{\text{rec}}$. The same is done for Eq. (11) to obtain the reduced uncertainty band, $\Sigma^{E_{\text{had}}^{\text{rec}} E_{\text{had}}^{\text{rec}},\text{const}}$, for the $E_{\text{had}}^{\text{rec}}$ prediction.

From Eq. (13), we see that the modeling of E_ℓ^{rec}, $E_{\text{had}}^{\text{rec}}$, and the correlations between them all play a role in obtaining the posterior prediction. Following Eq. (4), these can then be simultaneously examined with a GoF test on the constrained $E_{\text{had}}^{\text{rec}}$ prediction

$$\chi^2 = (n^{E_{\text{had}}^{\text{rec}}} - \mu^{E_{\text{had}}^{\text{rec}},\text{const}})^T \cdot (\Sigma^{E_{\text{had}}^{\text{rec}} E_{\text{had}}^{\text{rec}},\text{const}})^{-1} \cdot (n^{E_{\text{had}}^{\text{rec}}} - \mu^{E_{\text{had}}^{\text{rec}},\text{const}}). \tag{14}$$

Given a flux prediction with its associated uncertainties, the correlations that yield $\mu^{E_{\text{had}}^{\text{rec}},\text{const}}$ will be dictated by the modeling of $E_{\text{had}}^{\text{missing}}$. Thus, if E_ℓ^{rec} and $E_{\text{had}}^{\text{rec}}$ are measured, and a constraint from E_ℓ^{rec} is applied to $E_{\text{had}}^{\text{rec}}$ to reduce uncertainties on the overall E_ν prediction through the high correlation in the flux prediction shared between the channels, a precise description of $E_{\text{had}}^{\text{missing}}$ becomes necessary for Eq. (12) to be satisfied. It is this narrowed parameter space for the correlated E_ℓ, $E_{\text{had}}^{\text{vis}}$, and $E_{\text{had}}^{\text{missing}}$ predictions that provide sensitivity to mismodeling in the missing energy through conservation of energy. Nevertheless, it should be noted that Eq. (12) still allows for a conspiracy of flux mismodeling (also known as mismodeling of E_ν) and mismodeling of $E_{\text{had}}^{\text{missing}}$ to cancel out and pass the test described by Eq. (14). However, the use of the constraint brings the examination of $E_{\text{had}}^{\text{missing}}$ through the correlations between E_ℓ^{rec} and $E_{\text{had}}^{\text{rec}}$ onto more equal footing with the examination of a directly observable variable, like E_ℓ^{rec}, in an unknown neutrino flux spectrum. As in the case of $E_{\text{had}}^{\text{missing}}$, an E_μ prediction is always going to depend on the assumed E_ν and it is challenging to fully disentangle the two.

The test described in Eq. (13) is demonstrated on MicroBooNE data in Fig. 2. This technique of using the lepton energy to constrain reconstructed hadronic energy has been employed by MicroBooNE to enable the extraction of energy-dependent inclusive ν_μCC cross sections on argon [10–12]. Similar constraints could be applied to other variables potentially sensitive to the modeling of missing hadronic energy. Similarly, when extracting nominal flux-averaged cross sections, which requires extrapolating from the real to the nominal neutrino flux spectrum, examining the mapping between the true and reconstructed neutrino energy is a route to evaluate if the overall uncertainty budget is sufficient for propagating the reported result from the real to the nominal flux.

Constraints could also be used to explore other physics that is challenging to probe with a direct data to MC comparison. Examples include using the lepton energy distribution to constrain the lepton angular distribution. Since quasielastic (QE), resonance (RES), meson exchange current (MEC), and deep-inelastic scattering (DIS) processes have distinct predictions for lepton kinematics, a mismodeling of the relative contribution of these events could be detected with such a constraint. This could also be explored by using the observed muon kinematics to constrain another variable that already

shows some separation between interaction modes, such as the opening angle between the lepton and the leading proton. This is demonstrated in the fake data studies presented in Sec. IV B 2.

When the model passes a well designed validation procedure, it suggests that the difference between the data and simulation are within the quoted total model uncertainty. When the model fails the validation procedure, it reveals that the difference between the data and simulation is beyond the model uncertainty and should trigger actions to improve the model for use in the analysis. This may include an expansion of the uncertainties or an updated central value prediction. These may be derived in a data-driven way [12] similar to the overall model validation procedure or derived from alternative models or event generators. As long as the expanded model is able to pass all relevant model validation tests, it may be used in place of the original model to extract the desired cross section results.

IV. FAKE DATA STUDIES

A. Usage of fake data studies

It is common practice in neutrino-nucleus cross section measurements to perform model validation on the primary interaction model used for cross section extraction through comparisons to alternative model or event generator predictions. Often, this comes in the form of FDSs designed to determine if the unfolding is able to recover the underlying true distribution when the "data" are produced by an alternative model rather than by nature. This allows one to assess how the variations between model predictions for backgrounds, efficiencies, or biases in the reconstruction of kinematic quantities can impact the results. In the case of poor closure on the underlying true distribution, these studies can be used to inform additional uncertainties to be applied to the primary model prediction used for extracting the data cross sections. Using fake data studies in this way is thus akin to using the data-driven model validation that we advocate for, which likewise aims to verify that the model contains sufficient uncertainties for the intended measurement.

In these FDSs, it is especially useful to consider alternative models that are expected to provide a particularly good description of the data or contain significantly different physics than the nominal model. Ideally, these FDSs should be conducted at statistics equal to, or higher than, that of the actual data. In many cases, the fake data is generated using the same detector and flux model as the nominal model. This makes these uncertainties superfluous and FDSs are therefore conducted removing these sources of uncertainty. This allows for a more direct test of the interaction model and the extent to which it may bias the unfolding.

The benefit of directly validating the interaction model used for unfolding with fake data is that this approach reduces the dependence on any given model. It helps ensure that the discrepancy in results extracted using the examined set of event generators is insignificant compared to the uncertainty on the results. While this method is a viable strategy for validating the interaction model used for unfolding, it has the following shortcomings:

(1) There is no guarantee that the combined phase space covered by the set of tested models and event generators is able to completely describe nature. Therefore, this approach lacks a guarantee of sufficiency in determining the primary model's underestimation of uncertainties.
(2) It is also possible that the phase space covered by these event generators leads to a significant overestimation of uncertainties. While overestimating systematic uncertainties is better than underestimating them, overestimated uncertainties will reduce the power of the data.
(3) Since one can always invent new effective models or event generators, it is not clear when one should stop in evaluating the model uncertainty under this approach.

It is even possible for both issues 1 and 2 to exist simultaneously in different regions of phase space and it is challenging to know where the spread of event generators lies relative to these two extremes. The fundamental issues of these shortcomings are related to the earlier discussion in Sec. II A on the nature of event generators.

We propose utilizing a data-driven model validation procedure to reduce the reliance on alternative models to validate the model used for unfolding and address issues 1 and 2. As described in Sec. III, when a model passes a carefully designed data-driven validation procedure, it suggests that the difference between the data and simulation are within the quoted total model uncertainty in the phase space relevant for the cross section extraction. This approach also addresses issue 3 because it is based upon real data rather than alternative models, and is thus better suited to determine the adequacy of the model used for the extraction and whether an expansion of its uncertainties is appropriate. Nevertheless, the data-driven validation does not totally eliminate these issues. One can still run into issue 1 if one selects tests that do not probe the full phase space of the measurement, or issue 2 if one's suite of tests is overly sensitive to mismodeling irrelevant to the desired measurement, and the choice of which tests to utilize is analogous to issue 3. However, having the validation centered around the data rather than alternative models minimizes the potential downsides of these issues.

When evaluating the data-driven model validation methods presented here, we use FDSs in a very different way. Instead of testing the robustness of the model, these FDSs aim to test the robustness of the model validation itself.

In particular, FDSs in the case of data-driven model validation should aim to demonstrate the following points:

(i) The model validation procedure is indeed sensitive to mismodeling in the phase space relevant for the cross section extraction.
(ii) When the model passes validation, the potential bias in the extracted cross sections is small compared to the total uncertainties.

These FDSs may additionally demonstrate that even in certain cases where the model fails validation, the biases in the extracted cross sections are still small compared to the total uncertainties. Though an analysis should not proceed to unfolding with a model that has not passed validation, observing such cases in these studies provides additional evidence that the validation is able to identify mismodeling before bias is introduced. In short, FDSs employed in the context of data-driven model validation aim to examine the *procedure* used to validate the model, whereas more typical FDSs aim to directly examine the *model* used for the extraction. As was done in the Supplemental Material of Refs. [10–13], we illustrate how to test the sufficiency of a model validation procedure through FDSs presented in IV B.

When evaluating a data-driven model validation procedure with FDSs, it is perhaps more useful to utilize all systematic uncertainties rather than only the cross section uncertainties, even if some uncertainties are superficial in the context of fake data. Such studies provide a more realistic test of the stringency of the model validation as it is performed on the real data. Through the conditional constraint, the validation cancels shared systematics from all sources including neutrino flux, cross section, and detector effects. As seen in Eq. (11), the way the constraint reduces these uncertainties on the model prediction is independent of the data observation and only depends on the uncertainties on and correlations between the model predictions for the different distributions. Because of this, the way the constraint reduces the uncertainties is exactly the same in these studies as in real data. This is not true when only the cross section uncertainties are included, in which case the constraint may behave very differently due to the significantly different treatment of the systematic uncertainties. Thus, a test with the complete set of systematics is required to demonstrate that the constraint is able to reduce uncertainties on the reconstructed distributions enough to detect relevant mismodeling that will bias the extraction beyond stated uncertainties. Furthermore, many FDSs could be interpreted as a detector effect rather than a cross section effect, thereby testing the ability of the model validation to detect model discrepancies not attributable to the cross section modeling.

If an analysis is employing data-driven validation, fake data studies that directly validate the model through comparisons to alternative event generators may still be useful to identify situations in which the extracted results are not biased beyond stated uncertainties but nevertheless show better agreement for the model used for unfolding over the truth. An example of this type of situation would be one in which the extracted cross section shows a mild discrepancy with the truth at 1.2σ significance, but shows tension at only 0.5σ significance with the model used for extraction. Such a situation cannot be identified with data-driven model validation, as it only examines if the bias introduced is within stated uncertainties. This can be seen as an advantage of more typical FDSs. When employing data-driven validation, one should be cautious if this type of bias is identified in a FDS, but such bias does not necessarily invalidate the model used for unfolding if it still passes the data-driven validation. Unless the alternative generator used to produce the fake data is expected to be a particularly good description of the real data, this bias may not carry over into the extraction of the real data, and, even if it does, the extracted results are still expected to fall within the uncertainties of the true value in nature if the model passes validation by a well-constructed set of tests.

B. Illustration of data-driven model validation using fake data studies

Using the model validation procedure and unfolding analogous to the methodology employed in [10], we conduct two sets of FDSs that serve as case studies for data-driven model validation. The first set utilizes a fake dataset produced from the nominal GENIE-based MicroBooNE model [14,20] with a shift in the reconstructed proton energy intended to mimic a mismodeling of the relative contribution of the missing hadronic energy to the total energy transfer. The second set of FDSs utilizes fake datasets produced from an alternative event generator; namely NuWro 19.02.2 [16]. Studies are performed with the full systematic uncertainties to probe the ability of the model validation to detect discrepancies under a treatment of systematic uncertainties akin to that of real data. A FDS utilizing the NuWro fake data is also performed with only the cross section systematics in order to more directly probe the cross section model and its associated uncertainties. As will be shown throughout this section, the results of these two sets of FDSs are consistent with the expectation that the data-driven model validation is able to detect relevant mismodeling before it begins to bias the extraction of cross sections.

These fake datasets represent significant deviations from the mapping between true and reconstructed energy transfer predicted by the nominal GENIE-based MicroBooNE MC. This is illustrated in Fig. 3, which shows the energy transfer resolution for the central value prediction of the MicroBooNE MC, the MicroBooNE MC prediction with a 15% reduction in reconstructed proton energy, and the NuWro prediction. The MicroBooNE MC prediction's cross section uncertainties are included on the nominal prediction without the proton energy scaling, but these uncertainties

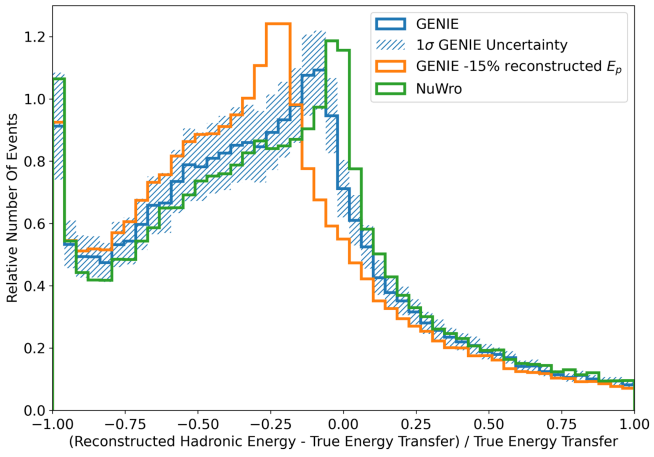


FIG. 3. The energy transfer resolution predicted by the GENIE-based MicroBooNE tune simulation and cross section uncertainties, the GENIE-based MicroBooNE tune simulation with 15% reduction in reconstructed proton energy, and the NuWro simulation. We show only the region between −1 and 1 for visual clarity.

are not large enough to cover the NuWro prediction. Furthermore, a 15% reduction in the visible hadronic energy is a more significant deviation than changing from the nominal MicroBooNE MC prediction to NuWro prediction.

For both sets of FDSs, we employ the same ν_μCC event selection and systematic uncertainties as in [10], which are described in detail in [12,38]. We then perform model validation designed to enable the extraction of the cross section as a function of E_ν, the differential cross section with respect to the energy transfer ν, and the differential cross section with respect to the muon energy E_μ. The validation is described in more detail in the following paragraphs with a complete list of tests presented in the Appendix. Following the validation, cross section results are extracted using the Wiener-SVD unfolding technique [48]. Results for E_μ are not extracted for the FDSs with the reconstructed proton energy scaling because these fake data samples do not change the underlying muon kinematics resulting in perfect closure in this variable for each FDS.

The model validation utilizes a multitude of tests in order to increase the chances of detecting problematic forms of mismodeling that would bias the cross section results beyond their uncertainties. One does not necessarily know *a priori* which test will be most sensitive, hence the numerous tests. In this case study, each test is performed on events fully contained (FC) and partially contained (PC) within the detector as well as jointly on all events (FC&PC) with separate bins for FC and PC. The only exception is in tests where the FC distribution is used to constrain the PC one, in which case only the PC distribution is examined. In each test, the GoF is quantified using Eq. (4) to determine how well the fake data distributions are described by the nominal MicroBooNE model used for the unfolding. The GoF is then further evaluated by decomposing the covariance matrix into linearly independent components via eigenvalue decomposition. This transformation to an uncorrelated basis allows a local p-value to be calculated from the decomposition bin that shows the largest discrepancy. The look-elsewhere effect is then corrected for by converting the local p-value into a global p-value via Eq. (9). This procedure is described in more detail in Sec. III. These global p-values and the p-values obtained from the χ^2 GoF test described in Eq. (4) are the metrics used to evaluate the sufficiency of the model, and must be greater than 0.05 in all tests for the model to pass validation.

The first set of tests examines the modeling of the muon kinematics in detail. These are not performed in the FDSs with the reconstructed proton energy scaling because these fake data samples do not modify the muon kinematics. The tests begin with evaluating the GoF on the muon energy distributions, E_μ^{rec}, and the muon angular distributions, $\cos\theta_\mu^{\text{rec}}$. Next, the FC distributions are used to constrain the PC distributions in the same variable. These PC distribution are, in general, more susceptible to mismodeling due to poorer reconstruction in events that escape the active volume of the detector, hence the additional tests. Then, to examine the muon kinematics more holistically, the E_μ^{rec} distributions are used to constrain the $\cos\theta_\mu$ distributions. These tests provide a significantly reduced posterior uncertainty. This allows the muon kinematics to be examined in detail, potentially exposing mismodeling of a variety of physics effects, such as the relative contribution of different interaction modes, which could bias the measurement of the muon kinematics or derived quantities such as E_ν or ν. If the model passes this suite of tests, it builds confidence that it is sufficient to extract cross sections as a function of the muon kinematics.

With the muon kinematics validated, the focus shifts to the hadronic energy distributions in order to enable the extraction of cross sections as a function of E_ν and ν. These tests begin the same as the ones on the muon kinematics. Events are binned as a function of the reconstructed hadronic energy, $E_{\text{had}}^{\text{rec}}$, or the reconstructed neutrino energy, $E_\nu^{\text{rec}} = E_{\text{had}}^{\text{rec}} + E_\mu^{\text{rec}}$, and GoF tests are performed both on the unconstrained distributions and on the PC distributions after constraints from the FC ones in the same variable. From here, the muon kinematics are used to constrain the E_ν^{rec} and $E_{\text{had}}^{\text{rec}}$ distributions. These tests, which are described in Eq. (14), examine the correlated prediction between the hadronic and leptonic energy thereby providing sensitivity to the missing hadronic energy. If the model passes this second suite of tests, it builds confidence that the model can describe the leptonic and hadronic energy as well as the correlations between them and is sufficient to extract cross sections as a function of E_ν and ν.

To evaluate how well the extraction has reproduced the underlying true distribution, a χ^2 test statistic is computed

between the underlying truth and the unfolded fake data. In the case of the FDSs conducted with the scaling of the reconstructed proton energy, the true distribution corresponds to the MicroBooNE MC. This is because neither the incoming neutrino energy, the energy transfer, nor the muon kinematics are modified by this scaling. These truth level distributions remain identical and only the relative contribution of missing and visible energy is modified. In the case of the FDSs conducted with the NuWro 19.02.2 fake data, the true distribution corresponds to a prediction from an independently produced high-statistics sample from the event generator. The χ^2 test statistic is constructed using the covariance matrix obtained in the cross section extraction and is converted to a p-value assuming a χ^2 distribution with degrees of freedom equal to the number of bins in the extracted result. To compare the stringency of the model validation to the amount of bias induced in the cross section extraction, this p-value is compared to the p-values used to evaluate the model validation. For ease of comparison, we convert all p-values to significance levels corresponding to multiples of the standard deviation σ of a normal distribution. The significance level corresponding to a model validation test indicates the amount of mismodeling detected by the validation and the significance level corresponding to a cross section extraction indicates the amount of bias induced in the unfolding. When a model validation test yields a larger significance level than the corresponding cross section extraction, this indicates that the model validation is more stringent than the cross section extraction. In this case, one is able to use the results of the model validation to identify situations in which the model is insufficient before such deficiencies become relevant to the unfolding. This allows for subsequent efforts to mitigate deficiencies in the overall model by using an updated CV prediction or expanded set of uncertainties before proceeding to the extraction.

1. Proton energy scaling fake data studies

The results of these FDSs with different scalings of the visible proton energy are shown in Fig. 4 and Table I. In Fig. 4, the gray band corresponds to the range of the significance values obtained across the validation tests, and the blue and orange points indicate the significance of the bias between the cross sections extracted in E_ν and ν and MC truth. Theses FDSs are conducted with all systematic uncertainties at 6.4×10^{20} POT, which corresponds to the exposure of the first three runs of MicroBooNE data taking that have been used for many recent MicroBooNE cross section measurements [11–13]. In these studies, we do not account for correlations in the data and MC statistical uncertainties arising from the fact that the fake data and MC utilize the same set of events. However, this is treated identically between the validation and the cross section extraction making this a fair comparison between the sensitivity of the model validation and the bias induced in the cross section extraction.

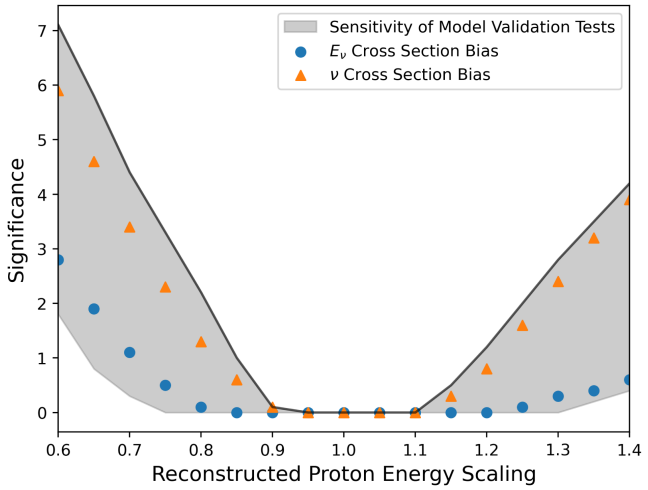


FIG. 4. Sensitivity of the model validation to discrepancies compared to the bias induced in the cross-section extraction in FDSs utilizing full-systematics. The fake datasets used for these FDSs each have their reconstructed proton energy scaled by different amounts to mimic mismodeling of the missing hadronic energy. The x-axis corresponds to this scaling. The y-axis indicates the agreement between the fake data and nominal MicroBooNE MC for the cross-section extractions and model validation tests in terms of significance level. For the model validation, a large significance indicates high sensitivity to mismodeling. For the extracted cross section, high significance indicates more biased results and worse closure on the underlying true distribution.

For each fake dataset, the significance of the bias between the extraction of the cross section as a function of E_ν is well below the significance of the discrepancy identified in the corresponding model validation tests. The significance of the bias for the differential cross section extracted in ν is always higher than it is for E_ν but is similarly always below the significance of the discrepancy identified in the most sensitive model validation test. Furthermore, even at large proton energy scalings past the point at which the validation indicates that there is relevant mismodeling, the agreement between the underlying truth and extracted results remains quite good for the cross section as a function of E_ν. These observations indicate that the stringency of the model validation is greater than the bias in the cross section extraction induced by mismodeling. This allows mismodeling of the missing hadronic energy to be detected before it becomes problematic to the extraction.

As demonstrative examples, the extracted fake data cross sections and several of the more sensitive model validation tests for the fake data studies with visible proton energy scalings of 0.85 and 0.75 are shown in Figs. 5 and 6, respectively. Specifically, we show the model validation test which consists of evaluating the GoF of the $E_{\text{had}}^{\text{rec}}$

TABLE I. Results of the model validation and cross-section extraction for each fake dataset that scales the reconstructed proton energy. All sources of uncertainty are included at data statistics equivalent to 6.4×10^{20} protons on target (POT) of exposure. The significance values in the "Bias in Extraction" columns indicate the level of closure the extracted fake data cross sections have with the underlying truth. The significance values in the "Sensitivity of the Model Validation to Discrepancies" columns indicate the level at which the fake data and the MC used for the cross section extraction disagree. The different columns correspond to different validation tests. All significance levels were obtained by converting the χ^2 test statistic to a p-value assuming a χ^2 distribution with degrees of freedom equal to the number of bins. These p-values are then interpreted as significance levels corresponding to multiples of the standard deviation σ of a normal distribution.

			Sensitivity of the model validation to discrepancies											
Visible K_p scaling	Bias in extraction		$E_{\text{had}}^{\text{rec}}$ GoF			$E_{\text{had}}^{\text{rec}}$ decomposition			E_{ν}^{rec} GoF			E_{ν}^{rec} decomposition		
	$\sigma(E_\nu)$	$d\sigma/d\nu$	FC&PC	FC	PC	FC&PC	FC	PC	FC&PC	FC	PC	FC&PC	FC	PC
0.6	2.8	5.9	5.9	2.0	6.3	6.0	3.6	6.6	5.3	1.8	6.4	6.6	3.6	7.1
0.65	1.9	4.6	4.2	1.3	4.9	4.9	2.8	5.5	3.3	0.8	4.6	5.3	2.9	5.8
0.7	1.1	3.4	2.7	0.6	3.7	3.8	2.0	4.4	1.5	0.3	2.9	3.9	2.2	4.4
0.75	0.5	2.3	1.3	0.2	2.4	2.7	1.1	3.3	0.2	0.0	1.3	2.5	1.4	3.1
0.8	0.1	1.3	0.2	0.0	1.2	1.6	0.4	2.2	0.0	0.0	0.2	1.1	0.6	1.8
0.85	0.0	0.6	0.0	0.0	0.3	0.4	0.0	1.0	0.0	0.0	0.0	0.0	0.1	0.4
0.9	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
0.95	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.15	0.0	0.3	0.0	0.0	0.1	0.1	0.0	0.5	0.0	0.0	0.0	0.1	0.1	0.3
1.2	0.0	0.8	0.0	0.0	0.5	0.5	0.2	1.2	0.0	0.0	0.0	0.7	0.6	1.1
1.25	0.1	1.6	0.3	0.1	1.3	1.1	0.8	1.9	0.0	0.0	0.2	1.6	1.3	2.0
1.3	0.3	2.4	1.1	0.4	2.2	1.7	1.4	2.5	0.0	0.0	0.7	2.5	2.0	2.8
1.35	0.4	3.2	2.0	0.9	3.0	2.2	2.1	3.0	0.2	0.2	1.3	3.3	2.6	3.5
1.4	0.6	3.9	2.8	1.5	3.8	2.7	2.7	3.4	0.6	0.4	2.1	3.9	3.2	4.2

distribution for PC events after the constraint from the muon kinematics. The corresponding χ^2 decomposition is also shown. The combination of these two FDSs demonstrate the points outlined in Sec. IV A. For the 0.85 scaling, the model passes the validation test depicted in Fig. 5(a) as well as all other validation tests. This indicates that the uncertainties on the model are sufficient for this 15% shift in the visible proton energy. The extracted cross sections likewise show little bias. This can be seen in Figs. 5(b) and 5(c). The success of the cross section extraction is confirmed by examining the differences between the extracted results and the underlying truth in the eigenvalue basis of the covariance matrix, which eliminates correlations between bins. In this basis, the tension with the underlying truth is less than 1σ significance for the majority of the bins, with only a single bin of the extracted $d\sigma/d\nu$ decomposition falling slightly outside 2σ significance. This is illustrated in the bottom subpanels of Figs. 5(b) and 5(c), which show the significance of tension in this basis for each bin.

The fake dataset with the proton energy scaling of 0.85, shown in Fig. 5, can be contrasted with what is seen for the proton energy scaling of 0.75, which is shown in Fig. 6. Here, the model validation indicates the presence of discrepancies not covered by the systematics. In particular, the decomposition test on the PC distributions after constraint from muon kinematics shown in Fig. 6(a) reveals disagreement at the 3.3σ significance level. If such a discrepancy were observed for real data, additional uncertainty or an otherwise expanded model would have been implemented to cover this difference and mitigate the potential for bias in the cross section extraction. Such a strategy was employed in Refs. [12] and [13] when a modeling discrepancy related to the leading proton kinetic energy distribution relevant to the desired cross sections was identified.

The level of tension identified in the model validation of the fake dataset with the proton energy scaling of 0.75 indicates that this analysis should not proceed to the cross section extraction with the model and associated uncertainties in their present form. Nevertheless, for the purposes of this study, we proceed to cross section extraction in order to compare the sensitivity of the model validation to the bias induced in the cross section extraction. When the fake data cross section as a function of E_ν is extracted with the nominal model, a reasonable result is obtained. The extracted cross section and underlying truth only disagree at 0.5σ and 1.1σ significance in regularized truth space and in the eigenvalue basis, respectively. This is seen in the top and bottom panel of Fig. 6(b), respectively. While the extracted $d\sigma/d\nu$ [Fig. 6(c)] is in tension with the truth at 2.3σ significance, the discrepancy is less than is seen in the most sensitive model validation test, which shows tension at the 3.3σ level [Fig. 6(a)]. Furthermore, examination of the tension between the extracted fake data cross section and truth in the eigenvalue basis shows that all of the bins

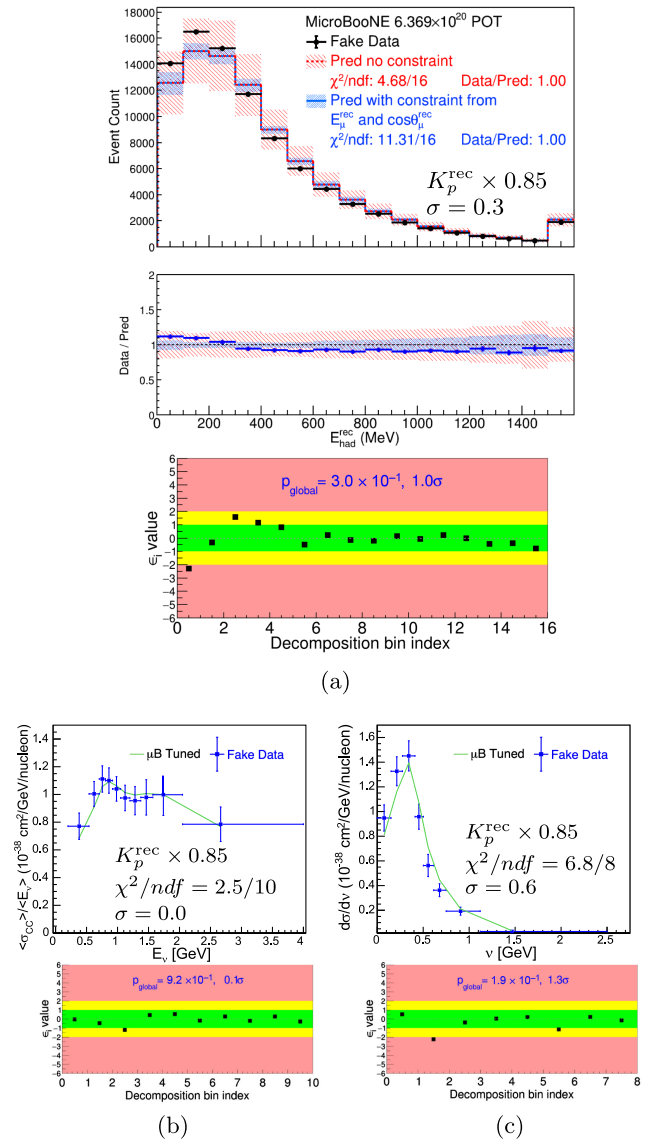


FIG. 5. The FDSs with the reconstructed proton energy scaled by 0.85. Select model validation tests are shown in (a). The extracted fake data cross section as a function of E_ν is shown in (b) and the extracted differential cross section as a function of ν is shown in (c). The χ^2 displayed on these panels is calculated between the true distribution indicated by the green line and the extracted result. The top plots of (b) and (c) show the extracted result and the bottom panels show the significance of the tension in each bin after the distribution has been transformed to the eigenvalue basis of the covariance matrix. (a) Model validation tests comparing the reconstructed hadronic energy for fake data and MC prediction for PC events. The top two pannels shows reconstructed space, with the red (blue) lines and bands showing the prediction without (with) the constraint from the muon kinematics. The uncertainties on the MC are shown in the bands and the statistical uncertainties on the data are shown on the data points. The bottom pannel shows the significance of the tension in each bin after the distribution has been constrained and transformed to the eigenvalue basis of the covariance matrix.

fall within 2σ significance, except for one, indicating a moderately successful unfolding despite the use of a model shown to be insufficient by the model validation.

In these FDSs, an interesting feature is illustrated by comparing the local p-values in the eigenvector basis of the covariance matrix obtained from the model validation to those of the cross section extraction. As can be seen by comparing Figs. 5(a)–5(c) and 6(a)–6(c), the greatest tension seen in an individual bin is at a similar level in the extraction and validation. This trend is seen for all scalings and is perhaps unsurprising, as it is likely that the tension in these bins is originating from the same source of mismodeling, which in the case of these FDSs, is the

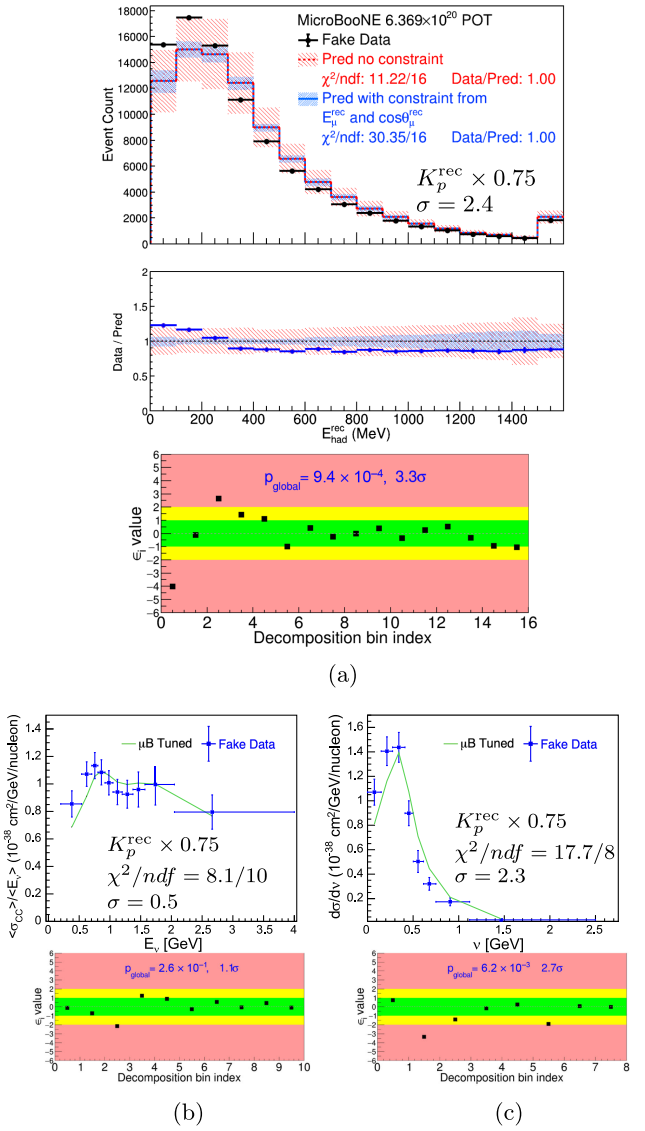


FIG. 6. Same as Fig. 5, but for the fake dataset with the reconstructed proton energy scaled by 0.75. (a) Model validation tests comparing the reconstructed hadronic energy for fake data and MC prediction for PC events. (b) Extracted $\sigma(E_\nu)$. (c) Extracted $d\sigma/d\nu$.

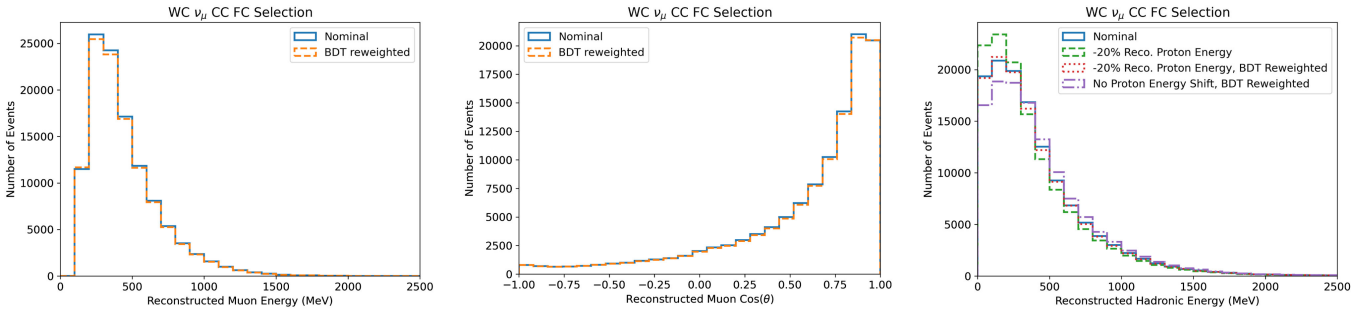


FIG. 7. [(a),(b)] Reconstructed muon kinematic distributions and (c) hadronic energy distribution for events passing the ν_μCC selection from [38] before and after the missing energy shift and multivariate event reweighting. The reweighting recovers good agreement in these variables despite the incorrect true-to-reconstructed energy mapping. Note that a proton energy shift has no impact on muon kinematic distributions.

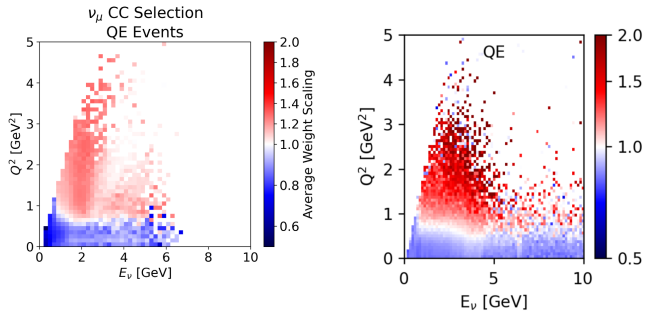
induced mismodeling of the fraction of the energy transfer which is visible. This observation could potentially be utilized when, instead of performing model validation with data, FDSs and alternative event generator predictions are used to evaluate the sufficiency of the model. Rather than informing the need for additional uncertainties by evaluating the bias introduced in the extraction bin-by-bin in truth space, it may be useful to evaluate bias bin-by-bin in the eigenvalue basis instead. Evaluating the bias in truth space with highly correlated bins may be unable to account for discrepancies that are amplified or suppressed by bin-to-bin correlations. Such a scenario could be avoided via a transformation to the eigenvector basis where bins become uncorrelated and the totality of systematics can be taken into account while evaluating bias.

The proton energy scaling FDSs can be viewed as an extension to the ones shown in the Supplemental Material of [10–12] which were inspired by a similar DUNE FDS performed in Ref. [49]. The DUNE study utilized a fake dataset with a 20% reduction in the reconstructed proton energy for the purposes of studying the impact of mismodeling on extracting oscillation parameters with a near and far detector. This fake dataset was then reweighted using a multivariate event reweighter to match the nominal MC in all the reconstructed distributions considered. However, as a result, there was still a significantly different mapping between the reconstructed and true neutrino energy in the reweighted fake data and the MC. Because of this, when a simultaneous near-far detector fit was performed on the reweighted fake data, the incorrect oscillation parameters were obtained despite the good agreement in the near detector.

Similar to the DUNE FDS, the FDSs presented in this section demonstrate that the total MicroBooNE MC uncertainties, of which the interaction model uncertainties are but a subset, cannot explain anything beyond a ∼20%–25% shift in the reconstructed proton energy. However, when comparing the conclusions of the DUNE FDS to the ones presented in this section, it is important to acknowledge the different set of assumptions in each. Though the multivariate event reweighter does not change reconstructed distributions and would indeed pass the model validation, the reweighting creates large shifts in the contributions from QE, RES, and DIS processes. As we will show, these modifications are likely beyond any reasonable event generator prediction. Since different interaction modes have distinct predictions for the muon kinematics, the multivariate event reweighter's freedom to make significant modifications to individual interaction modes would likely result in large tension when comparing the true Q^2 prediction from the event reweighter to any reasonable cross section model or event generator prediction and its associated uncertainties.

To demonstrate this point, we emulate the DUNE FDS by training a boosted decision tree (BDT) multivariate event reweighter [50] to restore agreement between the fake dataset with the 20% reduction in the proton energy and the MicroBooNE MC prediction. The reweighting is done in the true muon energy, true muon angle, and true transfer energy, which recovers good agreement between the MC and the fake dataset with the 20% reduction in the proton energy in the analogous reconstructed distributions, as shown in Fig. 7. This allows the reweighted model to pass model validation despite a significantly different mapping between true and reconstructed distributions, which would create bias in the cross section extractions. For this particular example, good agreement is seen in all model validation tests consistent with tension at the 0σ level, but the extracted cross section as a function of the energy transfer shows bias at the 0.7σ level when considering all systematic uncertainties. Though this is a relatively small amount of bias, it indicates that our model validation has not detected relevant mismodeling.

A comparison between the weights obtained in this study and those obtained in the DUNE FDS can be seen in Fig. 8. Similar behavior is observed with QE events at high Q^2 scaled up by as much as a factor of 2 and QE events at low Q^2 scaled down by as much as a factor of 0.6. As such, when the true Q^2 distribution for QE events is compared to the nominal model, a significant discrepancy is seen. This



(a) (b)

FIG. 8. The (a) MicroBooNE BDT reweighted model and (b) DUNE BDT reweighted model from [49].

is demonstrated in Fig. 9, where the χ^2/ndf calculated between the reweighted distribution and the nominal model and its associated cross section uncertainties is 135.04/25, indicating significant discrepancy. The basic lepton kinematics distributions for QE events are agnostic to final state interactions in the limit of complete factorization of the initial and final interactions, which is a well-realized approximation employed by state-of-the-art neutrino event generators [18,51]. As such, being supported by many neutrino and electron scattering experiments, these distributions are one of the better understood portions of cross section modeling, making such a large discrepancy with the nominal model hard to justify.

To explore this discrepancy further, the BDT reweighted model was compared to QE-like MicroBooNE ν_μCC1p data from [52,53]. This dataset is dominated by QE events, and, in select regions of phase space, it achieves an estimated ~95% pure QE sample. The impact of the BDT-reweighted model's significantly smaller QE prediction is illustrated in

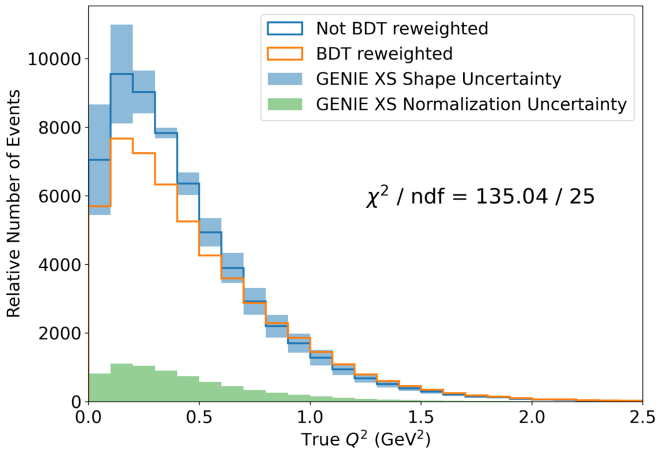


FIG. 9. Comparison between the BDT reweighted model and the GENIE MicroBooNE Tune prediction as a function of Q^2 for QE events. The normalization and shape components of the uncertainty bands are shown separately. The χ^2 is calculated using all cross section uncertainties of the MicroBooNE tune.

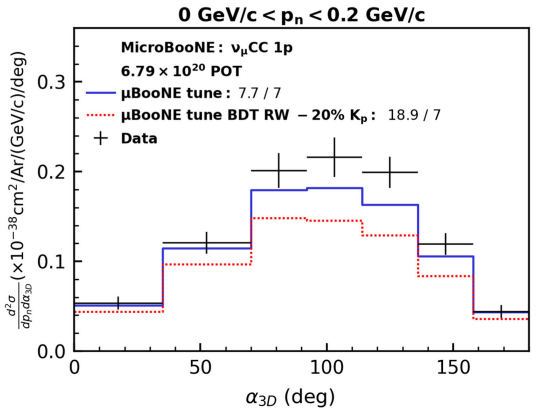


FIG. 10. Comparison of the BDT reweighted model prediction and various event generator predictions to MicroBooNE ν_μCC1p data from [53]. The χ^2/ndf values in the legend are calculated using the reported covariance matrix and each prediction is smeared by the reported A_C matrix.

Fig. 10, where the MicroBooNE tune prediction with and without the BDT weights is compared to the extracted double-differential cross section as a function of α_{3D} in the $p_n < 0.2$ GeV range. These variables describe the magnitude of the missing momentum (p_n) and the angle between the missing momentum vector and the momentum transfer vector (α_{3D}). The low phase space p_n region consists almost exclusively of QE events where the final state proton has not experienced significant final state interactions. The BDT reweighted model underestimates the data in this region and is in tension with the data at 2.6σ significance. This is noticeably worse than the nominal model without the BDT weights, which shows tension at only 0.9σ significance. Worsened agreement that approaches this level of increased tension is seen across the other distributions as well, indicating that the BDT model is not preferred by this data.

Encountering a situation in real data analogous to the multivariate event reweighter, which would pass the model validation but has an entirely different mapping between true and reconstructed neutrino energy, cannot be completely ruled out. However, such a scenario is unlikely as lepton kinematic distributions are well constrained by electron scattering data, particularly for QE events. The BDT reweighting we explore here shows noticeable tension with the QE prediction from the unreweighted model and performs worse when compared to the QE-like ν_μCC1p dataset from [52,53]. The DUNE study demonstrates that, through the use of PRISM, one can still obtain a robust measurement of oscillation parameters even in such a scenario with large deviations from reasonable models, which would be required if one wished to perform an oscillation analysis with minimal assumptions on the cross section model. However, techniques such as PRISM are not applicable to many experiments, such as MicroBooNE, and other strategies like data-driven model validation are

TABLE II. Results of the model validation and cross section extraction for the NuWro FDSs. Two treatments of systematic uncertainties are considered: one with full systematic uncertainties ("All Uncertainties") and the other with only cross section induced systematic uncertainties ("XS Syst + Stat Uncertainties"). Statistical uncertainties are always included. The Bias in Extraction columns indicate the significance of the tension between the extracted result and NuWro prediction. The entries in the Model Validation columns indicate the significance at which the fake data and the MC used for the cross section extraction disagree in the test that shows the most tension. The significance levels were obtained using p-values calculated from the χ^2 test statistic interpreted assuming a χ^2 distribution with degrees of freedom equal to the number of bins.

All uncertainties					XS Syst + Stat Uncertainties				
Bias in extraction			Model validation $\cos\theta_\mu^{\rm rec}\|E_\mu^{\rm rec}$		Bias in extraction			Model validation $\cos\theta_\mu^{\rm rec}\|E_\mu^{\rm rec}$	
E_ν	E_μ	ν	GoF	Decomposition	E_ν	E_μ	ν	GoF	Decomposition
0.1	0.0	0.1	0.0	0.4	2.6	0.1	1.5	3.1	3.9

required. Likewise, we caution that the general conclusions about the model validation drawn from the FDSs present in this work may not extrapolate as well onto scenarios far outside our current understanding of neutrino-nucleus scattering. The same holds true if more "tradition" FDSs are used to validate a model for a cross section measurement; one should be cautious extrapolating the results of successful closure on alternative models well beyond the bounds of reasonable models if these alternative models do not also extend well beyond these bounds. Indeed, this is directly related to the three issues raised earlier on "tradition" FDSs. There is no guarantee that a model spread will cover nature (issue 1), but including too many extreme scenarios may result in overestimated uncertainties (issue 2), and it is not clear when one should stop examining closure on these alternative models, particularly when extreme deviations are included (issue 3). Given the extremity of the example, this DUNE FDS is not inconsistent with the notion that one is able to use data-driven validation to detect mismodeling of the mapping between true and reconstructed neutrino energy to enable the extraction of energy dependent cross sections.

2. *Alternative event generator fake data studies*

The results of the FDSs utilizing fake data produced by NuWro 19.02.2 are summarized in Table II. This set of fake data was generated at 6.11×10^{20} POT, which is approximately equal to the exposure of the first three runs of data used for recent MicroBooNE cross section measurements [11–13]. The fake data contains statistical fluctuations in addition to the difference in event generators. As in Sec. IV B 1, the amount of discrepancy between the fake data and MicroBooNE MC prediction identified in the model validation is compared with the deviations of the extracted cross sections from the truth. This is done by converting the p-values obtained in each test or extraction into significance levels corresponding multiples of the standard deviation of a normal distribution. Both the full systematic uncertainty and cross section systematic plus statistical uncertainty situations are considered, which are labeled "all uncertainties" and "XS

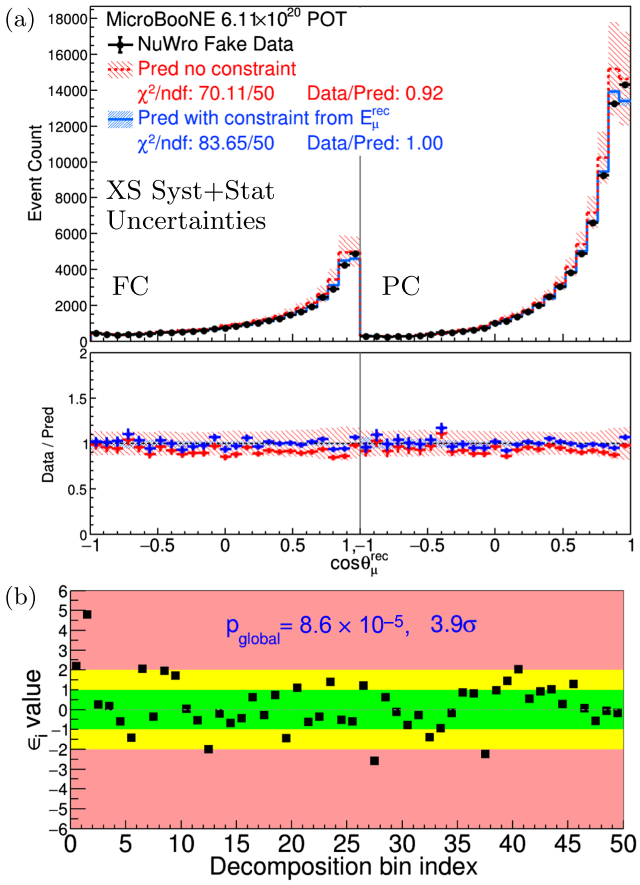


FIG. 11. Model validation tests comparing the NuWro fake data sample to the nominal MC prediction for the reconstructed muon scattering angular distribution for FC&PC events. In (a), the red (blue) lines and bands show the prediction without (with) the constraint from the observed muon energy distribution for FC&PC events. The uncertainties of the prediction only include the cross section and statistical terms and are shown in the bands. The data statistical uncertainties are shown on the data points. In (b), the significance of the tension in each bin after the distribution has been constrained and transformed to the eigenvalue basis of the covariance matrix is shown.

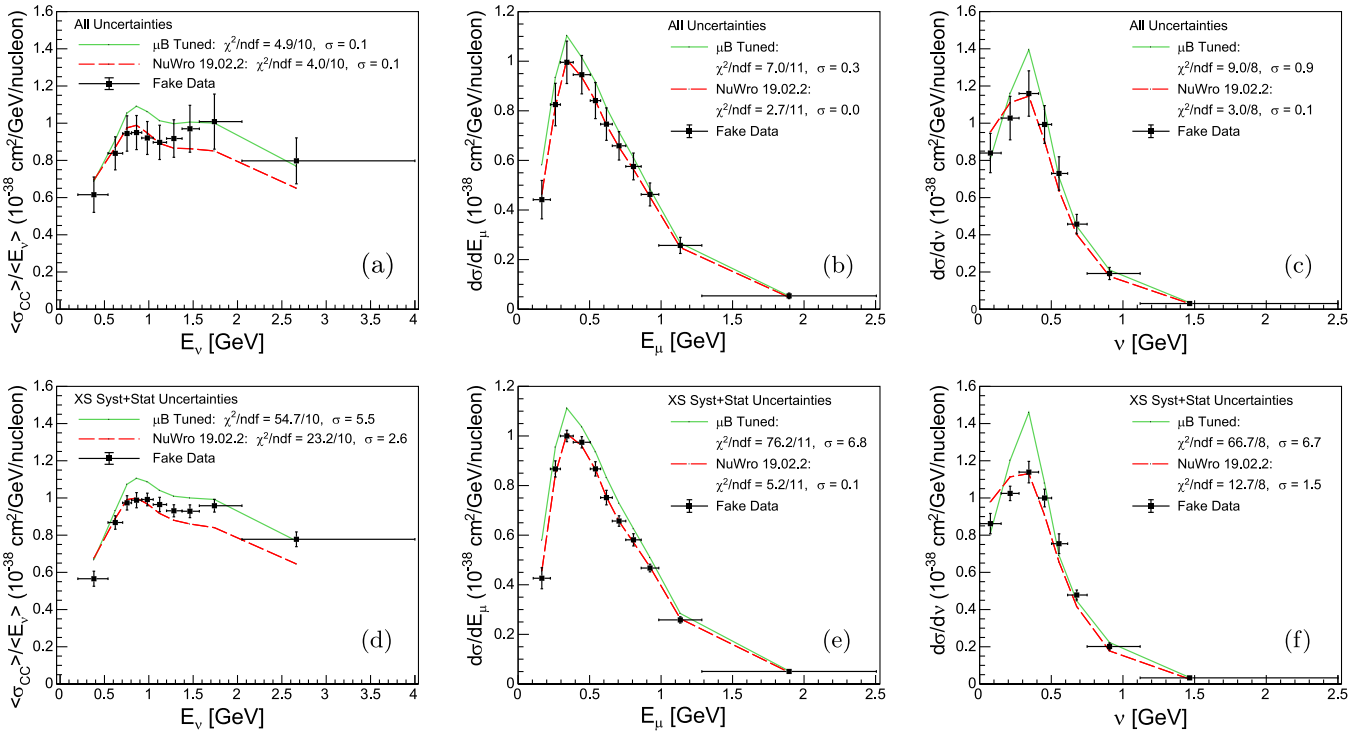


FIG. 12. The extracted ν_μCC cross section from the NuWro fake data [(a),(d)] as a function of neutrino energy, [(b),(e)] as a function of muon energy, and [(c),(f)] as a function of energy transfer. The statically independent NuWro prediction is compared to the measurements using full systematic and statistical uncertainties in (a)–(c), and only cross section related systematic and statistical uncertainties in (d)–(f). The nominal MicroBooNE MC prediction is also shown.

Syst + Stat uncertainties," respectively. The results for each scenario are described throughout this section.

Though the entire suite of tests described in Sec. IV B were performed, only the test that identified the most significant mismodeling is shown in Table II. In all cases, this test identifies tension that is equal to or greater than the tension between the extracted cross sections and NuWro truth. This is illustrated in Fig. 11, which shows the aforementioned most sensitive model validation test for the "XS Syst + Stat Uncertainties" study, and Fig. 12, which shows the extracted cross section for both sets of studies. These findings are consistent with the results shown in Sec. IV B 1 and are consistent with the point that, in general, one may design a data-driven model validation procedure that is more sensitive to the mismodeling than the extracted cross sections.

For the NuWro study with the full systematic uncertainties, the nominal MicroBooNE model passes validation. Similarly, the comparison between the extracted cross sections and the NuWro predictions yields χ^2/ndf values of $4.0/10$, $2.7/11$, and $3.0/8$ for E_ν, E_μ and ν, respectively, indicating that minimal bias was induced by the unfolding. This can be seen in Figs. 12(a)–12(c). In these figures, the extracted cross sections are compared to predictions from the NuWro truth and the nominal MicroBooNE MC. Though both predictions show good agreement with the extracted results, NuWro is slightly favored in all cases.

For the study with only the cross section systematic uncertainties, the nominal MicroBooNE model does not pass validation. The aforementioned model validation test that demostrates the most sensitivity to mismodeling shows this explicitly in Fig. 11. In this test, the FC&PC $\cos\theta_\mu^{\mathrm{rec}}$ distribution is constrained by the FC&PC E_μ^{rec} distribution. Shape differences in the forward scattering bins are present even after constraint, and the decomposition of the χ^2 reveals tension at the 3.9σ significance level. Given the choice of 2σ significance stringency for the model validation, this level of tension would prompt further investigation if it were encountered in real data and an alternative central value or expanded uncertainty budget would be implemented before unfolding. For the purposes of this study, the cross sections were extracted with the nominal model despite the tension identified in the validation. These results can be seen in Figs. 12(d), 12(e), and 12(f). The comparison between these extracted cross sections and NuWro predictions yields χ^2/ndf values of $23.2/10$, $5.2/11$, and $12.7/8$ for E_ν, E_μ, and ν, respectively. The corresponding significance of this bias is 2.6, 0.1, and 1.5, indicating that a moderate amount of bias was induced in the extraction of $\sigma(E_\nu)$ and, to a lesser extent $d\sigma/d\nu$. However, the significance of this bias is still less than the tension identified in the model validation procedure, which reaches the 3.9 significance level. This observation is again consistent with the previous result that the model

validation is, in general, more sensitive than the extraction of cross sections in any of these variables. Furthermore, we note that the NuWro truth is noticeably favored by the data for all three results and the large χ^2 values obtained for the MicroBooNE MC prediction indicate the stringency of this FDS.

3. Further exploration of the conditional constraint

In what follows, we explore the potential utility of conditional constraints beyond their role in model validation for cross section extraction. To do so, we further investigate the behavior of the constraint from the muon kinematics by examining its impact on the distribution of the reconstructed muon-proton opening angle, $\theta^{\text{rec}}_{\mu,p}$, for a semi-inclusive selection of ν_μCC events with protons (ν_μCC Np). This selection is identical to the ν_μCC Np selection described in Ref. [12] and consists of all events passing the inclusive ν_μCC selection that also have a primary reconstructed proton with greater than 35 MeV of kinetic energy. The $\theta^{\text{rec}}_{\mu,p}$ distribution is particularity interesting because it is sensitive to the breakdown of different interaction modes, with QE interactions dominating at negative $\cos\theta^{\text{rec}}_{\mu,p}$ and MEC and RES + DIS events more prominent at positive $\cos\theta^{\text{rec}}_{\mu,p}$. Figure 13, which displays the NuWro and MicroBooNE MC $\cos\theta^{\text{rec}}_{\mu,p}$ predictions broken down by interaction mode, illustrates this separation.

Figure 13 illustrates that there is a noticeable difference in the strength of the MEC predictions from NuWro and the MicroBooNE MC. In addition, the QE prediction from the MC is somewhat larger than NuWro, while the RES + DIS contributions are similar. This is quantified in Table. III, which contains the fraction contribute to the total selection for ν_μCC events from each interaction mode. These differences are also apparent when the NuWro fake data $\cos\theta_{\mu,p}$ distribution is compared to the MicroBooNE MC prediction with all its associated uncertainties. This is shown in Fig. 14. Even with the full uncertainty band, the more MEC and RES + DIS rich region ($\cos\theta^{\text{rec}}_{\mu,p} > 0$) is overpredicted to the point that the NuWro fake data is not completely covered by the uncertainties on prediction. The fake data is also overpredicted in the QE rich region

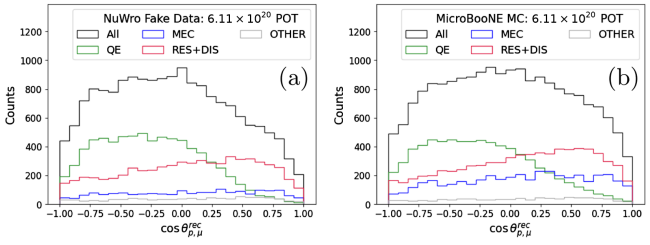


FIG. 13. The (a) NuWro and (b) MicroBooNE MC $\cos\theta^{\text{rec}}_{\mu,p}$ predictions for events passing the semi-inclusive $\nu_m u$CCNp selection broken down by interaction mode.

TABLE III. Fractional contribution of ν_μCC events from different interaction modes to the total ν_μCC Np selection. The remaining ∼5% of events are mostly background non-ν_μCC events that pass the selection.

	NuWro Fake Data	MicroBooNE MC	MicroBooNE MC with constraint
QE	0.449	0.366 ± 0.058	0.456 ± 0.025
MEC	0.116	0.218 ± 0.104	0.145 ± 0.032
RES + DIS	0.375	0.367 ± 0.093	0.352 ± 0.035

($\cos\theta^{\text{rec}}_{\mu,p} < 0$), but to a lesser extent and still falls within the uncertainties of the MC.

To further examine the data to MC differences in this observable, we expand the typical procedure of applying the constraint to the total prediction by also applying it to the prediction for each individual interaction mode. The constraint utilizes the muon kinematics distributions of the FC&PC ν_μCCNp selection and the FC&PC ν_μCC0p selection, which consists of the subset of events in the inclusive ν_μCC selection without a final state proton or with a final state proton below the 35 MeV threshold. The latter selection is likewise equivalent to the ν_μCC0p selection described in Ref. [12]. This application of the constraint to individual interaction modes rather than the prediction as a whole is only done in this study for the purposes of interpreting the action of the constraint, this procedure is not used in any of the model validation described throughout the rest of this work. As expected based on the

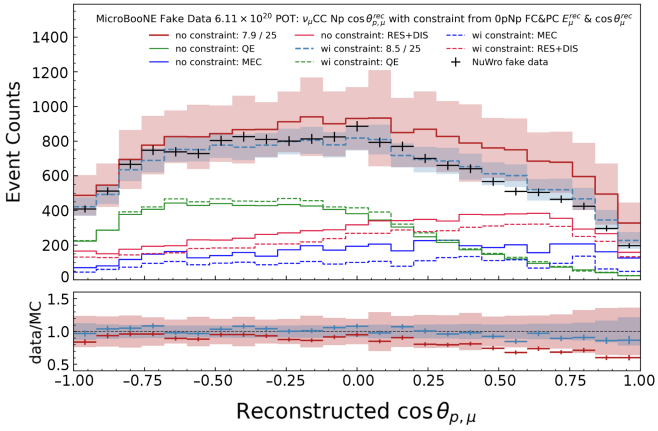


FIG. 14. Model validation test comparing the NuWro fake data sample to the nominal MC prediction for the reconstructed muon-proton opening angle distribution formed from events passing the semi-inclusive ν_μCCNp selection. The red (blue) lines and bands show the prediction without (with) the constraint from muon kinematics for the distribution formed from events passing the ν_μCCNp selection and the distribution formed from events passing the ν_μCC0p selection. The uncertainties of the prediction include all systematic and statistical terms. These are shown in the bands on the prediction. The statistical uncertainties on the fake data are shown on the data points.

difference in size of the MEC contribution between the two generators, the constraint produces a noticeable decrease in the MicroBooNE MC MEC prediction. A decrease in the RES + DIS prediction is also observed. Nevertheless, the fake data still falls at the lower edge of the resulting total posterior prediction's uncertainties for $\cos\theta^{\mathrm{rec}}_{\mu,p} > 0.5$ where the contribution from QE interaction modes is small and MEC and RES + DIS interaction modes dictate the strength of the prediction. For $\cos\theta^{\mathrm{rec}}_{\mu,p} < 0.5$, a small increase in the QE prediction offsets the decrease in the other interaction modes. This causes the total prediction to decrease slightly such that it agrees with the fake data within uncertainties across the QE dominated region of phase space.

Quantitatively, the posterior QE prediction corresponds to 0.456 ± 0.025 of the selection. This is in good agreement with the NuWro fake data, for which 0.449 of the selection is from QE ν_μCC interactions. This indicates that the constraint has utilized the information provided by the muon kinematics, with the only hadronic information provided by splitting the distribution based on the presence of a > 35 MeV kinetic energy proton, to increase the QE such that it falls in line with the NuWro fake data. Similarly, the prediction for the fraction of MEC events decreases from 0.218 ± 0.104 to 0.145 ± 0.032 and it also agrees with the fake data for which the fraction of MEC events is 0.116. However, despite the initially similar RES + DIS agreement, the constraint reduces this component such that the posterior prediction of 0.352 ± 0.035 is further from the NuWro value of 0.375, but is nevertheless still within uncertainties. These metrics are summarized in Table III.

This study demonstrates how the model validation and the conditional constraint can be used to probe the modeling of different interaction modes and their composition. The constraint greatly reduces the uncertainties on the rates ($> 50\%$) for each mode and tends to brings their CVs closer to the truth from the fake data. More generally, this study demonstrates how these methods allow one to further examine where the data lies within the allowed model space with significantly more sensitivity. The versatility of these methods means they can be utilized to provide additional insight about the models in such a way to be complementary to the actual cross section extraction. This opens up additional avenues of investigation for future analyses. For example, one could use the conditional constraint to fit various free parameters within a given model more directly. This can be easily achieved by denoting the free parameter as the constrained variable and using the distribution of a given observable as the constraining channel. From then, the rest of the procedure follows from Eqs. (10) and (11), which updates the free parameter to its posterior prediction based on an assumed prior and the given model. However, we note that interpreting these results in this way is by construction, model dependent and care must be taken to do so across different models, given that there is likely a complex interplay between implementation of these free parameters and other parameters of that model. Nevertheless, this can be quite enlightening and also complementary to the extracted cross section which, though potentially not providing as direct of insight, describes a well-defined physics quantity and will be interpretable in the context of any other model prediction.

V. SUMMARY

Neutrino-nucleus cross section measurements are motivated by the desire to improve interaction modeling to meet the precision needs of modern neutrino experiments. More robust interaction models are needed for these experiments to reach their desired level of sensitivity in measurements of oscillation parameters and searches for physics beyond the Standard Model. The difficulties associated with modeling neutrino-nucleus interactions at this level of precision result in the reliance on event generators using effective models in both neutrino oscillation experiments and cross section measurements.

The heavy reliance on event generators makes model validation important when extracting neutrino-nucleus interaction cross sections. To this end, we have presented a set of data-driven validation procedures and demonstrated that they can be powerful for detecting deficiencies in the models used for cross-section extraction. This validation utilizes techniques based on goodness-of-fit tests and the conditional constraint procedure to determine if the overall model can describe the data within uncertainties in the phase space relevant for the unfolding. Since this approach is based upon real data, it is well-grounded to appropriately validate the unfolding model and to evaluate the need for additional uncertainties. If a carefully selected set of data-driven tests are passed by the model, this builds confidence that any bias introduced in the cross section extraction will be within the quoted uncertainties of the measurement. This applies to both measurements of visible kinematic variables, such as the outgoing muon kinematics, and to derived quantities, such as the energy transferred to the nucleus. We demonstrate the efficacy of these data-driven methods with fake data studies aimed at comparing the sensitivity of the validation to the amount of bias introduced in the cross section extraction, in which we find that the validation is able to detect mismodeling before it impacts the cross section extraction. These data-driven methods based upon the conditional constraint are highly versatility and can be utilized to provide additional insight that complements the actual cross section extraction.

These validations are particularly important in the case of extracting nominal flux-averaged cross sections, which we advocate for due to the additional challenges associated with flux uncertainties created for the future analyzers of the data when extracting cross sections in the real flux. Producing robust cross section measurements with proper

treatment of uncertainties is essential to tuning efforts and event generator improvements, which will enable the desired precision in neutrino experiments in the near future. Utilizing data-driven model validation to extract nominal flux-averaged cross sections represents a reliable strategy to achieve this goal.

ACKNOWLEDGMENTS

This document was prepared by the MicroBooNE collaboration using there sources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359. MicroBooNE is supported by the following: the U.S. Department of Energy, Office of Science, Offices of High Energy Physics and Nuclear Physics; the U.S. National Science Foundation; the Swiss National Science Foundation; the Science and Technology Facilities Council (STFC), part of the United Kingdom Research and Innovation; the Royal Society (United Kingdom); the UK Research and Innovation (UKRI) Future Leaders Fellowship; and the NSF AI Institute for Artificial Intelligence and Fundamental Interactions. Additional support for the laser calibration system and cosmic ray tagger was provided by the Albert Einstein Center for Fundamental Physics, Bern, Switzerland. We also acknowledge the contributions of technical and scientific staff to the design, construction, and operation of the MicroBooNE detector as well as the contributions of past collaborators to the development of MicroBooNE analyses, without whom this work would not have been possible.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX: FAKE DATA MODEL VALIDATION TEST

In this appendix, we present the complete list of model validation tests used in the fake data studies presented in Sec. IV B.

(i) Evaluation of the E_μ^{rec} FC, PC, and FC&PC distributions through overall χ^2 GoF tests and χ^2 decompositions (6 total tests).
(ii) Evaluation of the E_μ^{rec} PC distribution after constraint from the analogous FC distribution. The overall χ^2 GoF test and χ^2 decomposition are examined (2 total tests).
(iii) Evaluation of the $\cos\theta_\mu^{\mathrm{rec}}$ FC, PC, and FC&PC distributions through overall χ^2 GoF tests and χ^2 decompositions (6 total tests).
(iv) Evaluation of the $\cos\theta_\mu^{\mathrm{rec}}$ PC distribution after constraint from the analogous FC distribution. The overall χ^2 GoF test and χ^2 decomposition are examined (2 total tests).
(v) Evaluation of the $\cos\theta_\mu^{\mathrm{rec}}$ FC, PC, and FC&PC distributions after constraint from the FC&PC E_μ^{rec} distribution. The overall χ^2 GoF test and χ^2 decomposition is examined for each distribution (6 total tests).
(vi) Evaluation of the E_ν^{rec} FC, PC and FC&PC distributions through overall χ^2 GoF tests and χ^2 decompositions (6 total tests).
(vii) Evaluation of the E_ν^{rec} PC distribution after constraint from the analogous FC distribution. The overall χ^2 GoF test and χ^2 decomposition are examined (2 total tests).
(viii) Evaluation of the E_ν^{rec} FC, PC, and FC&PC distributions after constraint from the FC&PC E_μ^{rec} and $\cos\theta_\mu^{\mathrm{rec}}$ distributions. The overall χ^2 GoF test and χ^2 decomposition is examined for each distribution (6 total tests).
(ix) Evaluation of the $E_{\mathrm{had}}^{\mathrm{rec}}$ FC, PC, and FC&PC distributions through overall χ^2 GoF tests and χ^2 decompositions (6 total tests).
(x) Evaluation of the $E_{\mathrm{had}}^{\mathrm{rec}}$ PC distribution after constraint from the analogous FC distribution. The overall χ^2 GoF test and χ^2 decomposition are examined (2 total tests).
(xi) Evaluation of the $E_{\mathrm{had}}^{\mathrm{rec}}$ FC, PC, and FC&PC distributions after constraint from the FC&PC E_μ^{rec} and $\cos\theta_\mu^{\mathrm{rec}}$ distributions. The overall χ^2 GoF test and χ^2 decomposition is examined for each distribution (6 total tests).

[1] K. Abe *et al.* (T2K Collaboration), Measurements of neutrino oscillation parameters from the T2K experiment using 3.6×10^{21} protons on target, Eur. Phys. J. C **83**, 782 (2023).
[2] M. A. Acero *et al.* (NOvA Collaboration), Improved measurement of neutrino oscillation parameters by the NOvA experiment, Phys. Rev. D **106**, 032004 (2022).
[3] B. Abi *et al.* (DUNE Collaboration), Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume II: DUNE Physics, arXiv: 2002.03005.
[4] K. Abe *et al.* (Hyper-Kamiokande Collaboration), Physics potential of a long-baseline neutrino oscillation experiment

using a J-PARC neutrino beam and Hyper-Kamiokande, Prog. Theor. Exp. Phys. **2015**, 53C02 (2015).
[5] H. Nunokawa, S. J. Parke, and J. W. F. Valle, *CP* violation and neutrino oscillations, Prog. Part. Nucl. Phys. **60**, 338 (2008).
[6] X. Qian and P. Vogel, Neutrino mass hierarchy, Prog. Part. Nucl. Phys. **83**, 1 (2015).
[7] M. V. Diwan, V. Galymov, X. Qian, and A. Rubbia, Long-Baseline Neutrino Experiments, Annu. Rev. Nucl. Part. Sci. **66**, 47 (2016).
[8] J. A. Formaggio and G. P. Zeller, From eV to EeV: Neutrino cross sections across energy scales, Rev. Mod. Phys. **84**, 1307 (2012).
[9] F. Gross *et al.*, 50 years of quantum chromodynamics, Eur. Phys. J. C **83**, 1125 (2023).
[10] P. Abratenko *et al.* (MicroBooNE Collaboration), First measurement of energy-dependent inclusive muon neutrino charged-current cross sections on argon with the MicroBooNE detector, Phys. Rev. Lett. **128**, 151801 (2022).
[11] P. Abratenko *et al.* (MicroBooNE Collaboration), Measurement of triple-differential inclusive muon-neutrino charged-current cross section on argon with the MicroBooNE detector, arXiv:2307.06413.
[12] P. Abratenko *et al.* (MicroBooNE Collaboration), Inclusive cross section measurements in final states with and without protons for charged-current ν_μ-Ar scattering in MicroBooNE, Phys. Rev. D **110**, 013006 (2024).
[13] P. Abratenko *et al.* (MicroBooNE Collaboration), First simultaneous measurement of differential muon-neutrino charged-current cross sections on argon for final states with and without protons using MicroBooNE data, Phys. Rev. Lett. **133**, 041801 (2024).
[14] L. Alvarez-Ruso *et al.*, Recent highlights from GENIE v3, Eur. Phys. J. Spec. Top. **230**, 4449 (2021).
[15] Y. Hayato and L. Pickering, The NEUT neutrino interaction simulation program library, Eur. Phys. J. Spec. Top. **230**, 4469 (2021).
[16] T. Golan, J. Sobczyk, and J. Żmuda, NuWro: The Wrocław Monte Carlo generator of neutrino interactions, Nucl. Phys. B, Proc. Suppl. **229–232**, 499 (2012).
[17] O. Buss, T. Gaitanos, K. Gallmeister, H. van Hees, M. Kaskulov, O. Lalakulich, A. B. Larionov, T. Leitner, J. Weil, and U. Mosel, Transport-theoretical description of nuclear reactions, Phys. Rep. **512**, 1 (2012).
[18] M. B. Avanzini *et al.*, Comparisons and challenges of modern neutrino-scattering experiments (TENSIONS 2019 report), Phys. Rev. D **105**, 092004 (2022).
[19] W. Filali, L. Munteanu, and S. Dolan, Benchmarking neutrino interaction models via a comparative analysis of kinematic imbalance measurements from the T2K, MicroBooNE and MINERvA experiments, Phys. Rev. D **111**, 032009 (2025).
[20] P. Abratenko *et al.* (MicroBooNE Collaboration), New CC0π GENIE model tune for MicroBooNE, Phys. Rev. D **105**, 072001 (2022).
[21] P. Stowell *et al.* (MINERvA Collaboration), Tuning the GENIE pion production model with MINERvA data, Phys. Rev. D **100**, 072005 (2019).
[22] M. A. Acero *et al.* (NOvA Collaboration), Adjusting neutrino interaction models and evaluating uncertainties using NOvA near detector data, Eur. Phys. J. C **80**, 1119 (2020).
[23] E. S. P. Guerra *et al.*, Using world $\pi^\pm$-nucleus scattering data to constrain an intranuclear cascade model, Phys. Rev. D **99**, 052007 (2019).
[24] J. Tena-Vidal *et al.* (GENIE Collaboration), Neutrino-nucleus CC0π cross-section tuning in GENIE v3, Phys. Rev. D **106**, 112001 (2022).
[25] J. Tena-Vidal *et al.* (GENIE Collaboration), Neutrino-nucleon cross-section model tuning in GENIE v3, Phys. Rev. D **104**, 072009 (2021).
[26] F. P. An *et al.* (Daya Bay Collaboration), Precision measurement of reactor antineutrino oscillation at kilometer-scale baselines by Daya Bay, Phys. Rev. Lett. **130**, 161802 (2023).
[27] G. Bak *et al.* (RENO Collaboration), Measurement of reactor antineutrino oscillation amplitude and frequency at RENO, Phys. Rev. Lett. **121**, 201801 (2018).
[28] R. Gran *et al.* (MINERνA Collaboration), Antineutrino charged-current reactions on hydrocarbon with low momentum transfer, Phys. Rev. Lett. **120**, 221805 (2018).
[29] P. A. Rodrigues *et al.* (MINERνA Collaboration), Identification of nuclear effects in neutrino-carbon interactions at low three-momentum transfer, Phys. Rev. Lett. **116**, 071802 (2016).
[30] M. V. Ascencio *et al.* (MINERνA Collaboration), Measurement of inclusive charged-current ν_μ scattering on hydrocarbon at $E_\nu \sim 6$ GeV with low three-momentum transfer, Phys. Rev. D **106**, 032001 (2022).
[31] S. Henry *et al.*, Measurement of electron neutrino and antineutrino cross sections at low momentum transfer, Phys. Rev. D **109**, 092008 (2024).
[32] C. Wilkinson, S. Dolan, L. Pickering, and C. Wret, A substandard candle: The low-ν method at few-GeV neutrino energies, Eur. Phys. J. C **82**, 808 (2022).
[33] L. Koch and S. Dolan, Treatment of flux shape uncertainties in unfolded, flux-averaged neutrino cross-section measurements, Phys. Rev. D **102**, 113012 (2020).
[34] A. A. Aguilar-Arevalo *et al.* (MiniBooNE Collaboration), Neutrino flux prediction at MiniBooNE, Phys. Rev. D **79**, 072002 (2009).
[35] T. Bonus, J. T. Sobczyk, M. Siemaszko, and C. Juszczak, Data-based two-body current contribution to the neutrino-nucleus cross section, Phys. Rev. C **102**, 015502 (2020).
[36] E. Gross and O. Vitells, Trial factors for the look elsewhere effect in high energy physics, Eur. Phys. J. C **70**, 525 (2010).
[37] J. Conrad, Statistical issues in astrophysical searches for particle dark matter, Astropart. Phys. **62**, 165 (2015).
[38] P. Abratenko *et al.* (MicroBooNE Collaboration), Search for an anomalous excess of inclusive charged-current ν_e interactions in the MicroBooNE experiment using Wire-Cell reconstruction, Phys. Rev. D **105**, 112005 (2022).
[39] M. L. Eaton, *Multivariate Statistics: A Vector Space Approach* (John Wiley and Sons, New York, 1983), pp. 116–117.
[40] S. Gardiner, Mathematical methods for neutrino cross-section extraction, arXiv:2401.04065.

[41] J. R. Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 2nd ed. (University Science Books, Sausalito, CA, 1996).
[42] S. J. Kline, The purposes of uncertainty analysis, J. Fluids Eng. **107**, 153 (1985).
[43] S. Dytman, Y. Hayato, R. Raboanary, J. T. Sobczyk, J. Tena-Vidal, and N. Vololoniaina, Comparison of validation methods of simulations for final state interactions in hadron production experiments, Phys. Rev. D **104**, 053006 (2021).
[44] A. Ershova *et al.*, Study of final-state interactions of protons in neutrino-nucleus scattering with INCL and NuWro cascade models, Phys. Rev. D **106**, 032009 (2022).
[45] S. Dolan, Probing nuclear effects in neutrino-nucleus scattering at the T2K off-axis near detector using transverse kinematic imbalances, Ph.D. thesis, University of Oxford, 2017.
[46] L. Bathe-Peters, S. Gardiner, and R. Guenette, Comparing generator predictions of transverse kinematic imbalance in neutrino-argon scattering, arXiv:2201.04664.
[47] X.-G. Lu, L. Pickering, S. Dolan, G. Barr, D. Coplowe, Y. Uchida, D. Wark, M. O. Wascko, A. Weber, and T. Yuan, Measurement of nuclear effects in neutrino interactions with minimal dependence on neutrino energy, Phys. Rev. C **94**, 015503 (2016).
[48] W. Tang, X. Li, X. Qian, H. Wei, and C. Zhang, Data unfolding with Wiener-SVD method, J. Instrum. **12**, P10002 (2017).
[49] A. Abud *et al.* (DUNE Collaboration), Deep Underground Neutrino Experiment (DUNE) near detector conceptual design report, Instruments **5**, 31 (2021).
[50] A. Rogozhnikov, Reweighting with boosted decision trees, J. Phys. Conf. Ser. **762**, 012036 (2016).
[51] L. Alvarez-Ruso *et al.*, NuSTEC White Paper: Status and challenges of neutrino–nucleus scattering, Prog. Part. Nucl. Phys. **100**, 1 (2018).
[52] P. Abratenko *et al.* (MicroBooNE Collaboration), Multi-differential cross section measurements of ν_μ-argon quasielasticlike reactions with the MicroBooNE detector, Phys. Rev. D **108**, 053002 (2023).
[53] P. Abratenko *et al.* (MicroBooNE Collaboration), Measurement of nuclear effects in neutrino-argon interactions using generalized kinematic imbalance variables with the MicroBooNE detector, Phys. Rev. D **109**, 092007 (2024).