Improving neutrino-nuclei interaction models: Recommendations and case studies on Peelle's Pertinent Puzzle

S. Abe,[17] L. Aliaga-Soplin,[16] J. Barrow,[9] L. Bathe-Peters,[10] B. Bogart,[7] L. Cooper-Troendle,[11,*] R. Diurba,[1] S. Dytman,[11] S. Gardiner,[4] L. Hagaman,[3] M. S. Ismail,[11] J. Isaacson,[4,8] J. Kim,[12] L. Liu,[4] J. McKean,[5] N. Nayak,[2] A. Papadopoulou,[6] L. Pickering,[14] X. Qian,[2] K. Skwarczyński,[13] J. Tena Vidal,[15] and J. Wolfs[12]

[1]*University of Bern, CH-3012 Bern, Switzerland*
[2]*Brookhaven National Laboratory, Upton, New York 11973, USA*
[3]*Columbia University, New York, New York 10027, USA*
[4]*Fermi National Accelerator Laboratory, Batavia, Illinois 60510, USA*
[5]*Department of Physics, Imperial College London, London SW7 2BZ, United Kingdom*
[6]*Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[7]*University of Michigan, Ann Arbor, Michigan 48109, USA*
[8]*Michigan State University, East Lansing, Michigan 48824, USA*
[9]*University of Minnesota Twin Cities, Minneapolis, Minnesota 55455, USA*
[10]*University of Oxford, Oxford OX1 3RH, United Kingdom*
[11]*University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA*
[12]*University of Rochester, Rochester, New York 14627, USA*
[13]*Royal Holloway University of London, Egham TW20 0EX, United Kingdom*
[14]*STFC Rutherford Appleton Laboratory, Didcot OX11 0QX, United Kingdom*
[15]*Tel Aviv University, Tel Aviv-Yafo, Israel*
[16]*University of Texas at Austin, Austin, Texas 78712, USA*
[17]*Kamioka Observatory, Institute for Cosmic Ray Research, University of Tokyo, Kamioka, Gifu 506-1205, Japan*

(Received 24 September 2025; accepted 18 December 2025; published 20 January 2026)

Improving the modeling of neutrino-nuclei interactions using data-driven methods is crucial for high-precision neutrino oscillation experiments. This paper investigates Peelle's Pertinent Puzzle (PPP) in the context of neutrino measurements, a longstanding challenge to fitting theoretical models to experimental data. Inconsistencies in data-model comparisons hinder efforts to enhance the accuracy and reliability of model predictions. We analyze various sources contributing to these inconsistencies and propose strategies to address them, supported by practical case studies. We advocate for incorporating model fitting exercises as a standard practice in cross section publications to enhance the robustness of results. We use a common analysis framework to explore PPP-related challenges with `MicroBooNE` and `T2K` data in an unified manner. Our findings offer valuable insights for improving the accuracy and reliability of neutrino-nuclei interaction models, particularly by systematically tuning models using data.

DOI: 10.1103/615w-kjk2

I. INTRODUCTION

Current and next-generation accelerator neutrino oscillation experiments [1–4] in the GeV neutrino energy range hold the potential to address some of the most pressing questions in neutrino physics, such as the matter-antimatter asymmetry through charge-parity violation [5], the precise ordering of neutrino masses among the three generations [6], and the search for sterile neutrinos [7]. These experiments, particularly DUNE [3] and Hyper-Kamiokande [4], aim to provide unprecedented precision in measuring neutrino oscillation parameters. However, achieving such precision necessitates a comprehensive understanding and meticulous control of the systematic uncertainties affecting these measurements [8].

One of the critical sources of systematic uncertainties in these experiments is the modeling of interaction between neutrinos and nuclei [9–11]. Neutrino-nuclei interactions are inherently complex because, while the interaction itself is governed by the electroweak force, the nuclear structure of the target is shaped by the strong force and therefore requires an understanding of the nonperturbative nature of QCD in this energy regime. This complexity leads to

*Contact author: lcoopert@proton.me

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.

incomplete theoretical descriptions and necessitates the use of effective models to describe these interactions [12].

To mitigate these uncertainties, it is common practice to tune cross section models through the use of event generators. These event generators, such as GENIE [13], NEUT [14], NuWro [15], GiBUU [16], and ACHILLES [17], incorporate various interaction models to describe neutrino-nuclei interactions. These models are built under different assumptions of the nuclear ground state, its structure function as well as the dynamics of intranuclear absorption and scattering of various particles produced in the tree level interaction. Tuning these models based on experimental data allows researchers to align theoretical predictions, which often include necessary approximations, with real observations, thereby improving the reliability of predictions. Examples of such tuning efforts include the work done by MINOS [18], T2K [1], NOvA [19], MINERvA [20], and MicroBooNE [21]. This process not only enhances the precision of neutrino oscillation experiments but also advances our understanding of neutrino scattering and nuclear physics.

In this paper, we investigate Peelle's Pertinent Puzzle (PPP) [22], a phenomenon where unexpected normalization values, often smaller than anticipated, are obtained from fitting a model to experimental data with correlated systematic uncertainties. An example of this is the reactor antineutrino anomaly [23,24], where the original best-fit data-to-prediction ratio is lower than the intuitive mean, as highlighted in Ref. [25]. In this specific case, it is possible to reformulate the fit to avoid a PPP issue [26], but this procedure does not generalize well to all situations. PPP-related behavior has also been observed in the model fitting of neutrino-nuclei cross section data [27]. While traditionally the study of PPP has focused on erroneous deviations in the normalization of a model fit to data, we use a broader definition here to more comprehensively address inaccurate fits to data. In this work, we use PPP to refer to any fit degradation that results from an improper treatment in the data with its full covariance, model, or their comparison. This definition includes fit distortions outside of the overall normalization, which we refer to as "non-normalization PPP".

As discussed in Ref. [22], improper estimates of uncertainties, such as covariance matrices, can contribute to PPP. Generally, PPP arises from inconsistencies between experimental data and theoretical models, and can serve as an indicator of such inconsistencies. However, the absence of PPP does not necessarily indicate consistency between experimental data and theoretical models. Other metrics beyond the normalization of the fitted mean, such as the p-value, can also serve as indicators of an inconsistency.

In this work, we examine PPP through case studies of model fitting of neutrino-nuclei cross section data and propose recommendations to enhance the fidelity of tuning neutrino-nuclei interaction models with experimental data. In Sec. II, we discuss how inconsistencies between data and models can introduce PPP and provide recommendations on how to avoid such inconsistencies. In Sec. III, we describe the covariance matrix formalism approach to fitting models to cross section data, highlighting its flexibility. In Sec. IV we describe methods that can help detect cases of PPP. In Sec. V we describe various mitigation strategies that can enable successful model fitting in the presence of PPP issues. In Sec. VI, we detail case studies of fitting neutrino-nuclei interaction models to experimental data and discuss how to identify data-model mismatches beyond abnormal best-fit normalization. Finally, in Sec. VII, we present strategies to address PPP when it is detected, and we summarize our findings in Sec. VIII.

II. INCONSISTENCIES BETWEEN DATA AND MODELS

Accurate comparisons between experimental data and theoretical models are fundamental to improve our understanding of neutrino-nucleus interactions. However, mismatches often arise due to limitations in the models, biases in the data, or inconsistencies in how the two are defined, interpreted, or communicated. These discrepancies can obscure underlying physics, distort model tuning efforts, and ultimately hinder progress in understanding fundamental interactions. This section examines the origins and implications of data-model inconsistencies, beginning with an illustrative case of parameter pull pathology, before exploring model limitations, the validity of Gaussian approximations, and communication mismatches in the interpretation and reporting of experimental results. Together, these discussions highlight both conceptual and practical challenges that must be addressed to ensure robust and reliable comparisons between theory and data.

The PPP phenomenon, which is often visually apparent from an abnormally low best-fit normalization, may occur under the following conditions: (i) strong correlations within the measurements of a dataset, (ii) model limitations, and (iii) an inconsistency between model predictions and data preferences. Often, a data-model inconsistency, also referred to as a mismatch, tension, or incompatibility, will lead to a high χ^2 value in the goodness-of-fit (GoF) assessment for the best-fit model. This allows high-χ^2 GoF to be taken as a symptom indicative of a potential data-model inconsistency and a PPP issue, although causation cannot be inferred with certainty. However, even in the absence of any visible PPP issues, a high χ^2 value still reflects significant data-model disagreement, posing a major challenge for effective model tuning. In Fig. 1, we illustrate the relationship between PPP and inconsistencies between data and the model through fitting a model with only one degree of freedom, the normalization of $N_1 + N_2$, to a two-dimensional measurement. The right panel shows a case of a data-model inconsistency, where the model is

unable to describe the relation between N_1 and N_2 in the data, resulting in a poor overall normalization in the fitted model after minimizing the χ^2.

Several scenarios can lead to mismatches between the data and the model. These may stem from deficiencies in the model, inaccuracies or biases in the data, or inconsistencies in how the model predictions and data are defined and interpreted. In the following sections, we will examine each of these possibilities in more detail.

A. Model limitations

When addressing issues related to the model, we limit our discussion by excluding cases where the model is entirely unsuitable. Instead, we focus specifically on situations where the model has insufficient range of model parameters, or degrees of freedom, needed to fully capture the underlying physics. Modern event generators used to model neutrino-nuclei interactions typically include dozens of parameters that control the simulation of neutrino-nucleus interactions, but still rely on approximations in their modeling. Sometimes, parameter uncertainty estimates are provided by event generator experts, helping to compensate for model limitations. Additionally, it is possible for computational limitations to restrict the ability to fully explore the parameter space available in a model during fitting. When many parameters must be examined to find the minimum χ^2 test statistic for a best-fit search, the accumulated time required to generate new predictions for each set of parameters in a high-dimensional space can become impractical. One strategy in such cases is to focus on a selection of parameters deemed most relevant, while keeping all other parameters' central values fixed at their nominal values. This approach can often be effective, but it is not universally applicable. Fixing certain parameters' central values at their nominal values restricts the overall phase space and can lead to sizable mismatches between the data and the model, particularly when non-negligible parameters are fixed. In Sec. III, we present a covariance matrix-based fitting procedure that enables the simultaneous inclusion of numerous parameters without substantially increasing computational time.

B. Improper treatments of data

In addition to addressing issues related to the model, it is equally important to consider potential errors and approximations in the measurement of the data itself. These data frequently rely on Gaussian approximations when reporting their results. For example, neutrino-nuclei interaction cross section models are used to help unfold from reconstructed quantities to a measurement over truth quantities. These models are typically presented as central values accompanied by a multivariate Gaussian distribution defined by a covariance matrix. While the validity of this approximation remains a subject of active discussion, particularly for results driven by systematic uncertainties [27–29], model validation procedures such as those outlined in Refs. [30–33] provide effective means to mitigate these concerns for unfolding models used in producing data measurements. There are also other validation tools that can help form a comprehensive validation approach, such as the use of fake data studies to examine potential model dependence and side bands to examine the modeling of backgrounds.

Depending on the underlying Monte Carlo simulation model, the effects of systematic uncertainties may not always conform to a Gaussian approximation due to nonlinear and asymmetric relations between model parameters and the overall model prediction. These behaviors deviate from the Gaussian approximation, which can introduce discrepancies in evaluating a p-value when comparing results from the original base model to those derived using the covariance matrix. However, as demonstrated in Ref. [29], this Gaussian approximation tends to be overly aggressive at estimating low p-values, underestimating their true value. As a result, the Gaussian approach is more sensitive to model discrepancies in the low p-value regime, meaning that a test under this approximation with a p-value threshold such as 0.05 is conservative compared to one using the true p-value. This means it is possible to accurately perform model validation tests under the Gaussian approximation to check whether the model can describe the data within its uncertainties. In this context, while the fidelity of the covariance-matrix approximation to the underlying model and the validity of the Gaussian approximation remain important, greater emphasis should be placed on its ability to robustly capture the critical features relevant to the analysis. Data-driven model validation, such as the methodologies presented in [30], can be especially helpful in this task.

C. Improper data-model comparisons

So far, in the context of PPP, we've discussed data-model mismatches arising separately from either the model or the unfolded measurement, both from the Gaussian approximation of its uncertainties or during the unfolding procedure itself. Beyond these, data-model mismatches can also arise from an incorrect comparison between the model prediction and the unfolded data result, based on the formalism used to derive either of them. Two such examples include the regularization matrix, A_C, involved in the unfolding procedure [34,35] that accounts for the effects of regularization and the flux over which the cross-section result is averaged over, either the real unknown flux or the nominal central value flux [30,36].

In the extraction of cross sections from experimental data, properly accounting for detector resolution and efficiency effects is essential. One common approach is the use of an unfolding procedure that corrects for the smearing of measured quantities. In this case, to prevent nonintuitive results arising from the noninvertibility of this

smearing, regularization techniques, such as imposing smoothness or non-negativity constraints, are commonly applied. The influence of regularization on the final unfolding results is captured in the regularization matrix, originally defined in the context of Wiener singular value decomposition (SVD) unfolding and expanded to other methods in Ref. [35]. Publishing this matrix is critical, as it eliminates the need for introducing an *ad hoc* and often poorly defined "unfolding error". Providing the A_C matrix allows model tuning teams to apply it directly to their models, ensuring consistent and fair comparisons between data and model predictions. In contrast, the absence of the A_C matrix can create discrepancies between data and model results, even if an unfolding regularization uncertainty is provided in the absence of the exact A_C matrix. This can complicate tuning efforts and reduce the reliability and utility of the extracted cross sections. As an alternative, avoiding regularization altogether in the unfolding process requires using wider binning, which, while sidestepping regularization effects, often results in the loss of valuable information. Another approach is to report both regularized and unregularized measurements, such as in Ref. [37], although this essentially passes the problem of choosing the unfolding approach and corresponding tradeoffs onto anyone who intends to use the measurement.

During model tuning, the model prediction relies on a known neutrino spectrum input, typically the nominal neutrino flux reported by experiments. However, the measured event counts in a given kinematic bin are observed under an unknown real neutrino spectrum to which the experiment is exposed. The difference between the real and nominal neutrino spectrum is incorporated into the reported neutrino flux uncertainties. Since extracted cross sections are derived from these measured counts, their central values inherently depend on the unknown real neutrino spectrum. To avoid mismatches between the data and the model, it is crucial to reconcile these differences. This can be achieved by extrapolating the measured cross sections to the nominal neutrino spectrum or vice versa. As shown in Refs. [31–33], cross sections can be extracted directly at the nominal neutrino flux, provided that the mapping between the neutrino flux and the visible kinematic variables is supported by a rigorously validated extrapolation model. Alternatively, the correlation between the extracted cross sections and the neutrino flux model can be reported [30], allowing model tuning teams to account for uncertainties when extrapolating from cross sections at the nominal spectrum to those at the true spectrum, while taking into account the correlation with reported cross section results. Failure to address these differences and correlations can lead to significant mismatches [30], as comparing model predictions for cross sections at a nominal neutrino spectrum with measured cross sections based on an unknown real neutrino spectrum introduces inconsistencies and reduces the reliability of the analysis [36].

III. COVARIANCE MATRIX FORMALISM IN FITTING MODELS TO CROSS SECTION DATA

This section describes the conditional covariance matrix formalism, which can be used to fit models to data. Given two sets of random variables (X, Y) and the full covariance matrix Σ describing their correlations:

$$\Sigma = \begin{pmatrix} \Sigma^{XX} & \Sigma^{XY} \\ \Sigma^{YX} & \Sigma^{YY} \end{pmatrix}, \tag{1}$$

where n is used to describe the measurement vector and μ is used to describe the prediction vector. We can then derive the prediction for X given the constraints on Y provided by a measurement:

$$\begin{aligned} \mu^{X,\text{const}} &= \mu^X + \Sigma^{XY} \cdot (\Sigma^{YY})^{-1} \cdot (n^Y - \mu^Y) \\ \Sigma^{XX,\text{const}} &= \Sigma^{XX} - \Sigma^{XY} \cdot (\Sigma^{YY})^{-1} \cdot \Sigma^{YX}. \end{aligned} \tag{2}$$

This is a general result given that X and Y are jointly Gaussian distributed and can be obtained by conditioning one distribution by the other [38]. When using the covariance matrix formalism in fitting models to cross section data, we define X with dimension m_X to represent the model parameters under tuning, and Y with dimension m_Y to represent the actual cross section prediction in a given binning. The full covariance matrix Σ is then given by

$$\Sigma = \Sigma_n + \Sigma_\mu, \tag{3}$$

where Σ_n is obtained by expanding the original m_Y-dimensional measured cross section covariance matrix to $m_X + m_Y$ dimensions by filling zeros in the additional m_X dimensions. The Σ_μ represents the model covariance matrix and is obtained by simulating different universes of cross section predictions while varying the model parameters. With a sufficient number of universes simulated and their model parameters recorded, the model covariance matrix, including both model parameters (X) and model predictions of cross sections (Y), is constructed following the standard definition of a covariance matrix.

This covariance matrix formalism has both advantages and limitations. The primary advantage is computational efficiency. By generating different universes of model predictions in parallel and constructing the overall covariance matrix beforehand, one can directly calculate the best-fit solution, thus significantly reducing the time for the actual fitting process. Once the best-fit parameters are obtained, the conditional covariance matrix formalism can be reused to make a constrained model prediction based on these parameters. Since the computational effort is dominated by the number of universes rather than the number of model parameters, this formalism allows for easy marginalization and fitting of multiple model parameters with fewer concerns about hyperparameters and fit instability,

thereby enhancing the robustness and scalability of the model-fitting process. This procedure naturally addresses the concern of insufficient model coverage discussed in Sec. II.

However, this formalism also has a few limitations. In particular, the conditional mean and variance approach given by Eq. (2) does not naturally enforce physical boundaries on model parameters (e.g., constraints such as non-negativity or bounded ranges). One alternative is to construct a test statistic using the full joint covariance matrix Σ over both X and Y:

$$T(X) = (X - \mu^X, n^Y - \mu^Y) \cdot \Sigma^{-1} \cdot (X - \mu^X, n^Y - \mu^Y)^T, \quad (4)$$

where n^Y and μ^Y are the observed and expected values of Y, respectively; X is the model parameter being varied during the minimization, while μ^X is its expected value under the nominal model and remains fixed. In this formulation, μ^Y is treated as fixed, consistent with the expectation under the nominal model. The test statistic $T(X)$ can be minimized with software such as Minuit2 [39], allowing the inclusion of physical boundaries on X via the minimizer's configuration. Although the conditional approach is computationally efficient and suitable for many applications, using the full joint covariance matrix can be helpful in cases where enforcing such constraints is essential. Additionally, since the covariance matrix Σ is independent of the minimization and can be inverted ahead of time, the evaluation of $T(X)$ remains computationally efficient. We recommend that analysts enforce physical boundaries on parameter values in this manner so that meaningful fits are produced.

Furthermore, as discussed in Sec. II, the use of a covariance matrix naturally assumes that the underlying distribution follows a multivariate Gaussian distribution, which may deviate from the original model. This implies that the minimum found by minimizing $T(X)$, or obtained through the conditional covariance-matrix formalism, may differ from the true minimum defined by comparing the original model prediction with the data observation using only the data covariance matrix. Generally, this discrepancy should not pose a problem as long as the two minima are reasonably close. Therefore, we advocate checking the difference in parameter values between these two minima. In practice, one can initially fix the less relevant parameters at their best-fit values and allow only the important parameters to vary during the original model fit. Once the minimum is found, the comparison between the two minima can be performed with respect to the derived uncertainties on the best-fit parameters from the conditional covariance-matrix model fit. An example of the covariance matrix model fitting approach was created to support this work. It uses one of the models and measurements considered in this paper and is provided on Github [40].

IV. DETECTING TENSIONS BETWEEN DATA AND MODELS

There are many reasons why a model prediction may be in tension with a measurement. If the tension arises from inaccuracies in the neutrino interaction modeling, such comparisons can provide valuable insights for refining these models. Neutrino interaction models, when constructed with sufficient uncertainties and flexibility, allow for meaningful tuning. Even if the *a priori* central value prediction significantly differs from the measurement, the fitting process can yield a reasonable χ^2/ndf comparison and provide useful updates to the model parameters. However, tensions may also result from improper treatments, such as those discussed in Sec. II. In these cases, the fitting process may fail, resulting in inaccurate or unphysical parameter values.

In practice, it is difficult to distinguish between cases where data-model tension arises from genuine inaccuracies in the neutrino interaction modeling and cases where it stems from improper treatments. While this distinction cannot be made with certainty without full knowledge of the methodologies used in the measurement and the model, a systematic approach, similar to the data-driven model validation procedure [30], can help identify likely inconsistencies in cases where unexpected large discrepancies are encountered. This approach attempts to assess whether a data measurement can be accurately unfolded through the use of a supporting model which includes the Gaussian approximation of its uncertainties through a covariance matrix. This validation is achieved by testing whether the unfolding model can accurately describe the data within its stated uncertainties. The procedure also utilizes the conditional constraint formalism, similar to the one employed in Sec. III, which incorporates correlations between observables to refine model predictions and reduce uncertainties based on observed data in reconstructed space. This approach uses Bayes' theorem to update the model prediction on one observable given a constraint from a measurement on a different observable. The result is a more sensitive test of the model used to unfold the data into truth space. Note that this updated model prediction is only used in the validation test and not the unfolding procedure.

Similar to how data unfolding can benefit from an examination of data-model consistency, attempts to fit models to data can benefit from a detailed examination of data-model comparisons to assess their plausibility. We recommend employing a goodness-of-fit metric to detect potential mismatches due to improper treatments between a model and a measurement after fitting the model to the data. A straightforward method involves constructing a (global) χ^2 test statistic using the model prediction, including its uncertainties, measurement, and corresponding covariance matrices, and calculating the associated p-value. This p-value quantifies the probability of observing a level of tension at least as extreme as the measured one if the model

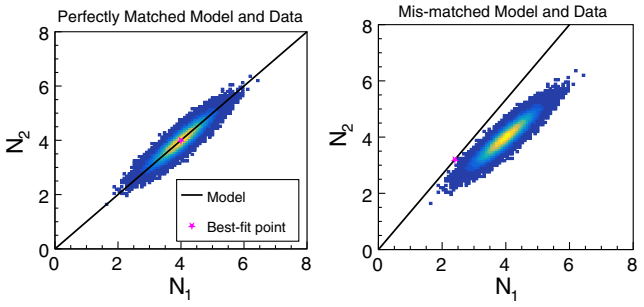


FIG. 1. Illustration of PPP with a simple two-dimensional measurement. Data preference is represented by the color contour, model preference is represented by the black line, and the model best-fit point is represented by the magenta star. Scenarios of data-model agreement and data-model mismatch are shown on the left and right, respectively.

is correct. If the p-value is sufficiently low, such as below 0.05 (corresponding to a significance above 2σ), it may indicate that the data-model comparison is invalid due to an inconsistency, which can be caused by an improper treatment in the analysis.

This global χ^2 approach can sometimes detect improper treatments but may also be overly conservative in other cases. As illustrated in Fig. 1, a PPP-driven mismatch in a measurement-prediction comparison can result in a best fit with a significant difference in normalization. In such cases, the global χ^2/ndf may fail to detect an issue, as poor agreement in a single degree of freedom (e.g., normalization) might not significantly impact the overall χ^2. To address this, the combined error covariance matrix Σ_C can be diagonalized through a basis change:

$$\Sigma_C = \Sigma_n + \Sigma_\mu, \qquad \Sigma_D = R^{-1}\Sigma_C R, \tag{5}$$

where Σ_n and Σ_μ are the measurement and model covariance matrices, and R is the matrix of eigenvectors of Σ_C. In this diagonalized basis, the difference between measurement and prediction becomes

$$D = R^{-1}(n-\mu). \tag{6}$$

Here, correlations between bins are removed, allowing the direct evaluation of individual bin contributions χ_i^2 to the total χ^2:

$$\chi_i^2 = (D^T\Sigma_D^{-1}D)_i = D_i^2/\Sigma_{D,ii}, \tag{7}$$

where D_i is the ith element of the vector D, and $\Sigma_{D,ii}$ is the ith diagonal component of the matrix Σ_D. Note that it is possible for degrees of freedom (DoF) other than normalization to exhibit tension between data and model, which may indicate a "non-normalization PPP" issue. If only one or a few DoF contribute disproportionately large χ_i^2 terms, the total χ^2/ndf may remain below threshold, making the global comparison insensitive to specific mismatches.

Building on this observation, it is possible to identify mismatches at a local level by examining the χ_i^2 values and corresponding p-values of individual bins. It is important to control the familywise error rate by correcting local p-values when comparing them to a significance threshold to account for the look-elsewhere effect [41] and avoid over-reporting extreme values. There are many useful correction procedures; here we apply the Šidák correction [42] to each local p-value:

$$p_{\mathrm{corrected}} = 1 - (1 - p_{\mathrm{local}})^N, \tag{8}$$

where N represents the total number of local p-values (bins). If any corrected p-values fall below the stated threshold of 0.05, the measurement-prediction comparison can be rejected on the grounds of suspected improper treatment. This approach is particularly effective when mismatches are confined to a subset of the degrees of freedom in the measurement space. By focusing on individual bins in the diagonalized space, this additional examination is more likely to detect localized discrepancies that may be missed in a global comparison.

V. MITIGATING INCONSISTENCIES BETWEEN DATA AND MODELS

Once serious tension between data and model is detected such that an improper treatment is considered likely, the next step is to determine the appropriate course of action. One option is to abandon attempts to fit the model to the measurement, acknowledging that the inconsistencies render the process fundamentally flawed. Alternatively, if one believes that the model contains inaccuracies that are not sufficiently covered by parameter uncertainties, one can enlarge the model uncertainties. However, typically models are already assigned conservatively large uncertainties by default, so it may be difficult to sufficiently enlarge model uncertainties in a plausible manner. Finally, one can adopt a mitigation strategy aimed at addressing the presence of the tension, thereby enabling a meaningful and constructive comparison between the measurement and the model prediction. In this section we discuss potential mitigation strategies under the assumption that there is a PPP issue that must be addressed in a fit attempt as best as possible.

A. Mitigating PPP through covariance matrix simplification

Since strong correlations within the data measurements are a necessary condition for PPP, one approach to mitigate this issue is to eliminate the nondiagonal terms in the data covariance matrix. While this explicitly avoids the PPP problem, it effectively invalidates the uncertainties reported by the experiments and creates challenges when results

from multiple experiments are included in the model tuning exercise. However, in cases where only a single experimental result is included and there are alternative ways to estimate the uncertainties of the model parameters, this approach can yield decent results. For example, in Ref. [21] a series of iterative fits were performed to data with only diagonal uncertainties used, and postfit model uncertainties were chosen to cover the range of best-fit model parameter values encountered throughout the iterative fitting process. It is worth noting that the referenced work also investigated fits using the covariance matrix transform described in the next section and found good agreement with the diagonal-uncertainties-only fits.

B. Mitigating PPP through covariance matrix transformation

As illustrated in Ref. [27], another approach to addressing the PPP issue is to perform a nonlinear norm-shape conversion on the data results and their covariance matrix. The new set of variables, $\mathcal{C} = \{C_1, \ldots, C_m\}$, are defined as follows:

$$C_i = f(n_i) = \begin{cases} \alpha \frac{n_i}{\sum_k n_k}, & 1 \le i \le m-1 \\ N_T = \sum_k n_k, & i = m, \end{cases} \tag{9}$$

where α is a scale parameter that can be chosen, and m is the dimension of the data measurement. Despite this transformation being nonlinear, the covariance matrix of $\{C_1, \ldots, C_m\}$ can be estimated using the error propagation rule:

$$\mathrm{Cov}[C] = J(f) \cdot \Sigma_{\mathrm{data}} \cdot J(f)^T, \tag{10}$$

where $J(f)$ is the Jacobian of the nonlinear transformation f. The new covariance matrix is expressed as follows:

$$\mathrm{Cov}[C] = \left(\begin{array}{c|c} \frac{\alpha^2}{N_T^2}\left(\sigma_{i,j} - \frac{n_i}{N_T}\sum_l \sigma_{j,l} - \frac{n_j}{N_T}\sum_k \sigma_{k,i} + \frac{n_i n_j}{N_T^2}\sum_{kl}\sigma_{k,l}\right) & \begin{array}{c} \frac{\alpha}{N_T}\left(\sum_l \sigma_{1,l} - \frac{n_1}{N_T}\sum_{kl}\sigma_{k,l}\right) \\ \vdots \\ \frac{\alpha}{N_T}\left(\sum_l \sigma_{m-1,l} - \frac{n_{m-1}}{N_T}\sum_{kl}\sigma_{k,l}\right) \end{array} \\ \hline \frac{\alpha}{N_T}\left(\sum_k \sigma_{k,1} - \frac{n_1}{N_T}\sum_{kl}\sigma_{k,l}\right) \cdots \frac{\alpha}{N_T}\left(\sum_k \sigma_{k,m-1} - \frac{n_{m-1}}{N_T}\sum_{kl}\sigma_{k,l}\right) & \sum_{kl}\sigma_{k,l} \end{array} \right), \tag{11}$$

where σ_{kl} is an element of the data covariance matrix Σ_{data}. We reproduce the above equation to correct typos in Ref. [27]. Figure 2 illustrates the transformation of the allowed measurement distributions in a simple two-bin measurement. While there is no information loss in this transformation, the nonlinear nature of the transformation alters the correlations between the data points. While this approach also avoids the PPP issue, it is not ideal because it applies a nonlinear transformation that alters the correlations reported by the experimental collaboration. These correlations that should be preserved to maintain the integrity of the original analysis.

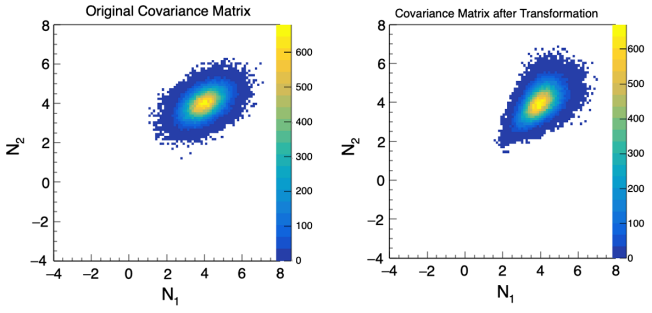


FIG. 2. Comparison of data correlations for the original two-bin data measurement (left) and those after the nonlinear norm-shape transformation (right).

Furthermore, these two approaches discussed above primarily address the normalization issue identified in the original PPP phenomenon. As discussed in Sec. IV, data-model mismatches can also occur in higher-dimensional shape analyses involving non-normalization PPP-like issues. Ideally, a comprehensive approach would address both normalization and shape dimension mismatches inclusively.

C. Mitigating inconsistencies between data and model through quantile mapping

We propose a strategy to mitigate inconsistencies between data and model predictions, addressing not only PPP but also broader issues that can arise in such comparisons. Central to this approach is the investigation of bin-by-bin disagreements in the diagonalized space discussed earlier. This basis allows for the identification and resolution of inconsistencies that extend beyond overall normalization effects, as demonstrated in Sec. VI. By adopting this comprehensive framework, we aim to provide a more robust solution for resolving data-model tensions.

A key challenge in fitting a model to data in the presence of improper treatments is managing the exaggerated

discrepancies between measurement and prediction. Such discrepancies often manifest as excessively large local χ_i^2 terms, which distort the overall χ^2 minimization process and can lead to extreme or nonphysical best-fit parameters. This issue arises because the covariance matrix Σ_D is insufficient to describe the observed differences D under the influence of a distorting improper treatment.

To address this, we propose a procedure inspired by the process of quantile mapping (QM) [43], which adjusts the uncertainties in the diagonalized data-minus-measurement covariance matrix Σ_D. By enlarging the uncertainties for specific elements, the adjusted covariance matrix $\tilde{\Sigma}_D$ becomes capable of accounting for the observed disagreements between measurement and prediction. This adjustment reduces the excessive contributions of individual bins to the overall χ^2, mitigating the impact of improper treatments and enabling more reliable and meaningful fit results.

We provide a simple description of QM here, and leave a more rigorous derivation of the approach to the Appendix. The central idea of the procedure is that when a model accurately describes a measurement, referred to here as the null hypothesis, the individual χ_i^2 terms will be sampled from a χ^2 distribution with one degree of freedom. These n χ_i^2 terms can then be arranged in increasing order and described by the cumulative distribution function (CDF) $g(x)$:

$$g(x) = K(x)/n, \tag{12}$$

where $K(x)$ counts the number of bins with a χ_i^2 value less than or equal to x over the domain $\mathbb{R} \geq 0$.

For a large number of bins, if the null hypothesis is true then the observed CDF will approximate the CDF of the χ^2 distribution with one degree of freedom. If the observed χ_i^2 terms are larger than expected, this would indicate that the corresponding uncertainties in Σ_D are too small. Therefore, if the observed CDF does not match the expected distribution in this manner, the data uncertainty of individual bins can be enlarged to map each χ_i^2 value onto the corresponding value for the CDF of a χ^2 distribution with one degree of freedom. We recommend that uncertainties are only maintained or enlarged, but not shrunk in this manner, so that the QM treatment remains conservative in terms of its uncertainty estimation. This approach will yield an adjusted data-model comparison where the newly determined uncertainties are able to describe the observed differences.

Before concluding this section, it is important to emphasize that the proposed procedures are designed to mitigate, rather than fully eliminate, the negative impacts of improper treatments in data-model comparisons. While the underlying causes of these improper treatments remain unaddressed, these strategies focus on minimizing their adverse effects to enable meaningful model fitting. Notably, the quantile mapping approach provides a flexible framework for incorporating multiple measurements, even when some are affected by issues like PPP. By enlarging uncertainties for comparisons suspected of improper treatments, this method allows such measurements to be included in a joint fit in a conservative manner. This approach naturally underemphasizes these problematic comparisons while prioritizing those with no suspected improper treatment, enhancing the robustness and reliability of the overall analysis. However, if QM is applied in cases where there is no improper treatment and only a moderate model inaccuracy, QM will decrease the capability of the measurement to inform legitimate model refinement. Therefore, it is important to only apply QM in cases where one is reasonably confident that an improper comparison or otherwise severe data-model mismatch is present and hampering successful fit attempts.

VI. CASE STUDIES OF MODELING FITTING

We present multiple case studies that illustrate types of potential mismatches and demonstrate the capabilities of the proposed methods in detecting and mitigating improper treatments. We compare published cross sections to model predictions from commonly used cross section event generators such as `GENIE` and `NEUT`, with a selection of postfit parameter values shown in the Appendix. In each case, the comparison to model predictions both before and after fitting, including calculated χ^2 values, is made using the nominal model uncertainties that are described for each model prediction. In some cases, part or all of the model used to predict a cross section is the same as the model used to unfold the data measurement. This means there are potential correlations between the data and the predictions they are compared against that are not well captured in this analysis. However, these correlations are likely small, given that neutrino interaction uncertainties are subdominant in the cross section measurements in these case studies, in part because they are naturally suppressed in the unfolding process. Furthermore, many case studies highlight particular improper treatments by comparing model fits with and without artificially introducing the improper treatment of interest, and only find evidence of PPP when the desired improper treatment is artificially added. This demonstrates that the artificially added improper treatment is responsible for the PPP effect observed. Therefore, overall we do not believe that the potential correlations in neutrino interaction modeling between a data measurement and a model prediction play a significant role in these studies.

A. Inconsistencies between data and models from regularizations in data unfolding

The first case study examines the impact of regularization in unfolding on the quality of model fits, specifically when the effects of regularization are not properly

accounted for in data-model comparisons. Unfolding methods are widely used to reconstruct event count distributions and produce cross section measurements in terms of true kinematics. Many of these methods, such as Wiener SVD [34] and D'Agostini unfolding [44], rely on regularization to stabilize the unfolding process and suppress statistical fluctuations. However, regularization can introduce nonphysical distortions into the unfolded data compared to an unregularized counterpart.

Methods like Wiener SVD address this issue by calculating a regularization matrix, A_C, which accounts for the effects of regularization. This process can be generalized to many other kinds of unfolding [35] as well. Applying the regularization matrix to model predictions ensures that both the data and model predictions are treated consistently, avoiding the introduction of bias in their comparison. By contrast, failing to account for regularization effects creates mismatches between the unfolded data and model predictions, leading to unreliable fit results and potentially improper interpretations.

To illustrate the impact of regularization in data-model comparisons, we perform a series of model fittings to measurements reported by MicroBooNE in one, two, and three dimensions. These measurements all investigate inclusive muon neutrino charged-current (ν_μCC) interactions on argon using very similar signal definitions, event selections, and unfolding methodologies. They include single- and double-differential cross sections $d\sigma/dE_\mu$ and $d^2\sigma/dE_\mu d\cos\theta_\mu$ [45] as well as the three-dimensional inclusive cross section $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ [32]. For each measurement, model fits are attempted both with and without applying the regularization matrix, which represents realistic scenarios where the regularization matrix was not reported. In cases where the regularization matrix was omitted from the fit, it is also omitted from the χ^2 calculation.

The NUISANCE package [46] is used to compute model predictions for comparison with the reported cross sections. In this section we use a model based on GENIE v3.0.6 G18_10a_02_11a, tuned to T2K data [21,47], and referred to as the GENIE nominal model. Uncertainties for each of the 55 model parameters follow the treatment outlined in the referenced tune [21], with the exception of FrPiProd_N and FrPiProd_pi, which were omitted for technical limitations in the implementation. Using NUISANCE and the provided code framework [40], for each measurement we construct both the central value prediction of the model and a covariance matrix that describes the overall uncertainty in the model prediction, including correlations between the measurement space and model parameters. In cases where we apply the regularization matrix, this is done before constructing the covariance matrix for the model uncertainty, meaning that the covariance is computed in the regularized space. Leveraging the covariance matrix formalism described in Sec. III, we perform a fit of the model to the data measurement, enabling a detailed investigation of the role of regularization in the unfolding process.

First we investigate the quality of fits performed to the single-differential cross section measurement $d\sigma/dE_\mu$, shown in Figs. 3 and 4. When the regularization matrix is applied, the postfit model prediction achieves a χ^2/ndf of $8.1/11$ with an associated p-value of 0.70, indicating a good agreement between the model and data without any clear evidence of a normalization issue indicative of PPP. When the regularization matrix is omitted, the fit quality is degraded with a χ^2/ndf of $11.1/11$ and an associated p-value of 0.43. Although the quality of the fit has worsened, there is no significant sign of a PPP issue. This indicates that omission of the regularization matrix will harm the quality of model fit attempts, but may not necessarily present a detectable PPP issue.

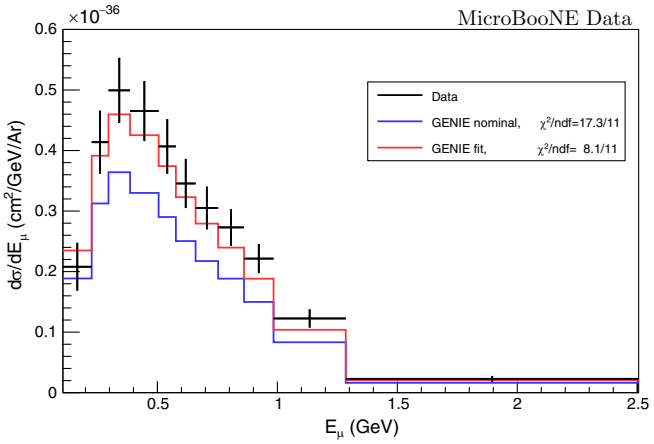


FIG. 3. Comparison of the measured MicroBooNE $d\sigma/dE_\mu$ cross section data (black) to the nominal GENIE model prediction with the regularization matrix applied (blue), and the fit result with the regularization matrix applied (red).

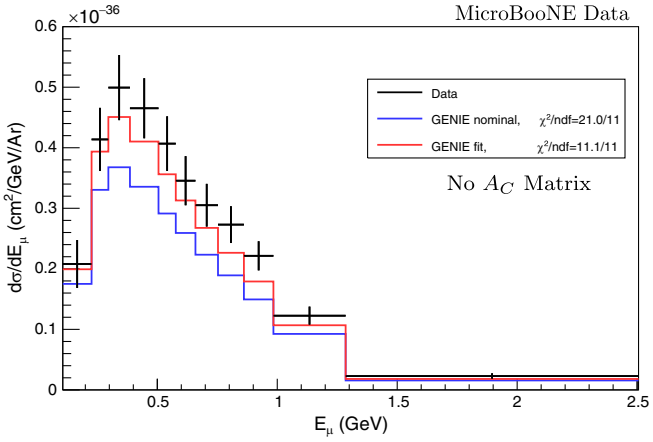


FIG. 4. Comparison of the measured MicroBooNE $d\sigma/dE_\mu$ cross section data (black) to the nominal GENIE model prediction without applying the regularization matrix (blue), and the fit result without applying the regularization matrix (red).

Next we investigate the quality of fits performed to the double-differential cross section measurement $d^2\sigma/dE_\mu d\cos\theta_\mu$, shown in Figs. 5 and 6. When the regularization matrix is applied, the postfit model prediction achieves a χ^2/ndf of 55.8/69 with an associated p-value of 0.87. Furthermore, an analysis of individual χ^2_i terms in the diagonalized covariance matrix basis reveals a largest individual χ^2_i value of 6.2, corresponding to a p-value of 0.59 after applying the Šidák correction [42] to control the familywise error rate. These comparisons all indicate a good agreement between the model and data without any clear evidence of a normalization issue indicative of PPP.

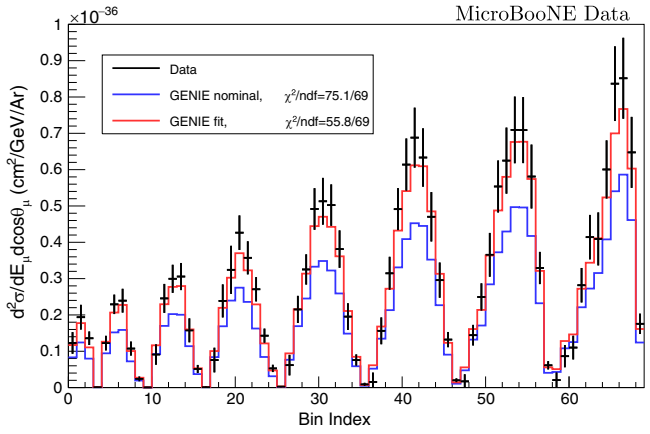


FIG. 5. Comparison of the measured MicroBooNE $d^2\sigma/dE_\mu d\cos\theta_\mu$ cross section data (black) to the nominal GENIE model prediction with the regularization matrix applied (blue), and the fit result with the regularization matrix applied (red). The x-axis represents the bin index.

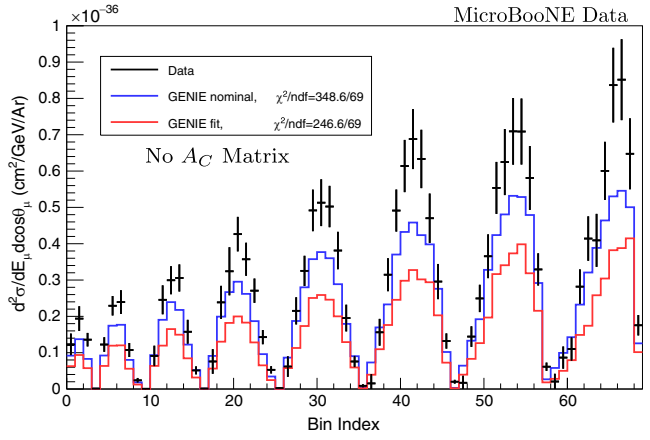


FIG. 6. Comparison of the measured MicroBooNE $d^2\sigma/dE_\mu d\cos\theta_\mu$ cross section data (black) to the nominal GENIE model prediction without applying the regularization matrix (blue), and the fit result without applying the regularization matrix (red). The x-axis represents the bin index. The absence of the regularization matrix leads to significant normalization discrepancies.

When the regularization matrix is omitted, shown in Fig. 6, the fit quality is significantly degraded with a clear normalization disagreement. The poor agreement between model and data is also apparent in the global χ^2/ndf, which shrinks from 348.6/69 from under the nominal model to 246.6/69 for the postfit model, corresponding to an extremely low p-value of 9×10^{-22}. Additionally, the most extreme individual χ^2_i term reaches 48, with a Šidák-corrected p-value of 4×10^{-12}. A detailed examination of individual χ^2_i terms suggests the presence of non-normalization PPP issues in the fit. Specifically, many DoF exhibit noticeably larger data-model tension through enlarged χ^2_i after fitting. This issue disappears when the regularization matrix is applied to the model prediction before fitting, indicating that it is directly related to the improper treatment and a symptom of non-normalization PPP. Since the fit tensions extend beyond the normalization DoF, PPP-mitigation strategies limited to separately addressing normalization and shape are insufficient to resolve complex non-normalization PPP issues.

Finally, we investigate the quality of fits performed to the three-dimensional cross section measurement $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$, shown in Figs. 7 and 8. When the regularization matrix is applied, the postfit model prediction achieves a χ^2/ndf of 144.96/138 with an associated p-value of 0.325. An analysis of individual χ^2_i terms in the diagonalized covariance matrix basis reveals a largest individual χ^2_i value of 8.95, corresponding to a p-value of 0.319 after applying the Šidák correction. As with double-differential measurement, these comparisons all indicate a good agreement between the model and data without any clear evidence of a normalization issue indicative of PPP.

Again, when the regularization matrix is omitted, shown in Fig. 8, the fit quality is significantly degraded with a clear normalization disagreement. The poor agreement between model and data is also apparent in the global χ^2/ndf value of 616.77/138, corresponding to an extremely low p-value of 4×10^{-19}. Additionally, the most extreme individual χ^2_i term reaches 52.76, with a Šidák-corrected p-value of 10^{-11}. Examination of individual χ^2_i terms shows non-normalization PPP issues through the existence of numerous DoF with significantly increased data-model tensions after fitting. As seen previously, the outliers disappear when the regularization matrix is applied to the model prediction before fit and comparison, demonstrating that they are driven by the improper treatment and a symptom of non-normalization PPP.

This case study highlights that omitting the regularization matrix, which arises from the data unfolding process, implicitly treats the unfolded results as truth-level

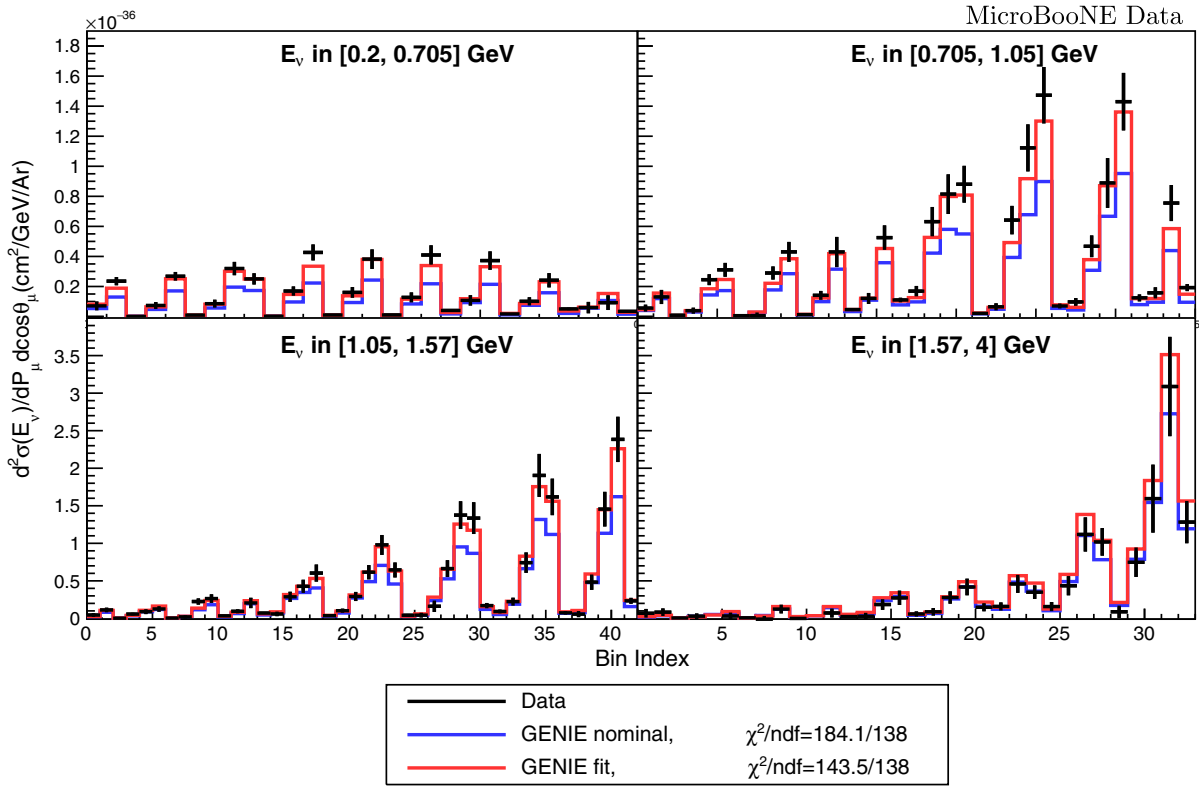


FIG. 7. Comparison of the measured MicroBooNE $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ cross section data (black) to the nominal GENIE model prediction (blue), and the fit result (red). Each subfigure corresponds to a different slice of neutrino energy, and the x-axis represents the bin index for the 2D differential cross section within each E_ν slice.

quantities. This assumption introduces significant challenges in data-model comparisons, including PPP-related issues, particularly evident in normalization discrepancies, as well as broader inconsistencies between the data and model predictions. Without the regularization matrix, the regularization-induced distortions are not properly accounted for, which can lead to exaggerated penalties in the χ^2 metric. Consequently, the fit may prioritize addressing artificial mismatches over resolving physically meaningful disagreements, resulting in unreliable and nonphysical fit parameters, with the normalization issue being a clear manifestation of PPP. The effect of the data-model mismatch varies across the measurements studied, suggesting that some measurements, such as those in higher dimensions, are more susceptible to a severe PPP issue situation than others. Additionally, the absence of the regularization matrix compromises the integrity of the comparison, as it prevents equal treatment of the data and model predictions, thereby undermining the reliability of the extracted results. Including the regularization matrix is therefore critical for achieving meaningful and consistent comparisons in unfolding-based analyses. In Sec. VI D, we will further build on this case study to demonstrate the application of the QM mitigation strategy.

B. Inconsistencies between data and models from real vs nominal neutrino flux

The second case study investigates the mismatch arising from assuming a certain flux to average the unfolded cross section over. They take the form of the "real flux", where mainly the detector effects are unfolded and the cross section is averaged over the actual unknown neutrino flux, or the "nominal flux", where a further translation is made to report a nominal flux-averaged cross section directly. While both are entirely self-consistent approaches, there are important trade-offs to consider in either one. As highlighted in [36], the cross section measurements that are extracted at the real flux are less model dependent because the observed events in the detector are directly related to the convolution of the real flux and cross section. However, comparisons to this cross section using various model predictions have to then assume a flux, which is typically the nominal neutrino flux. The differences between the two fluxes will therefore lead to inconsistencies in the comparison, in particular by not fully accounting for the uncertainties in the flux shape in the reported measurement and their correlation with the flux uncertainties in the theoretical prediction. The latter approach—extracting the cross section at the nominal flux—circumvents the issue,

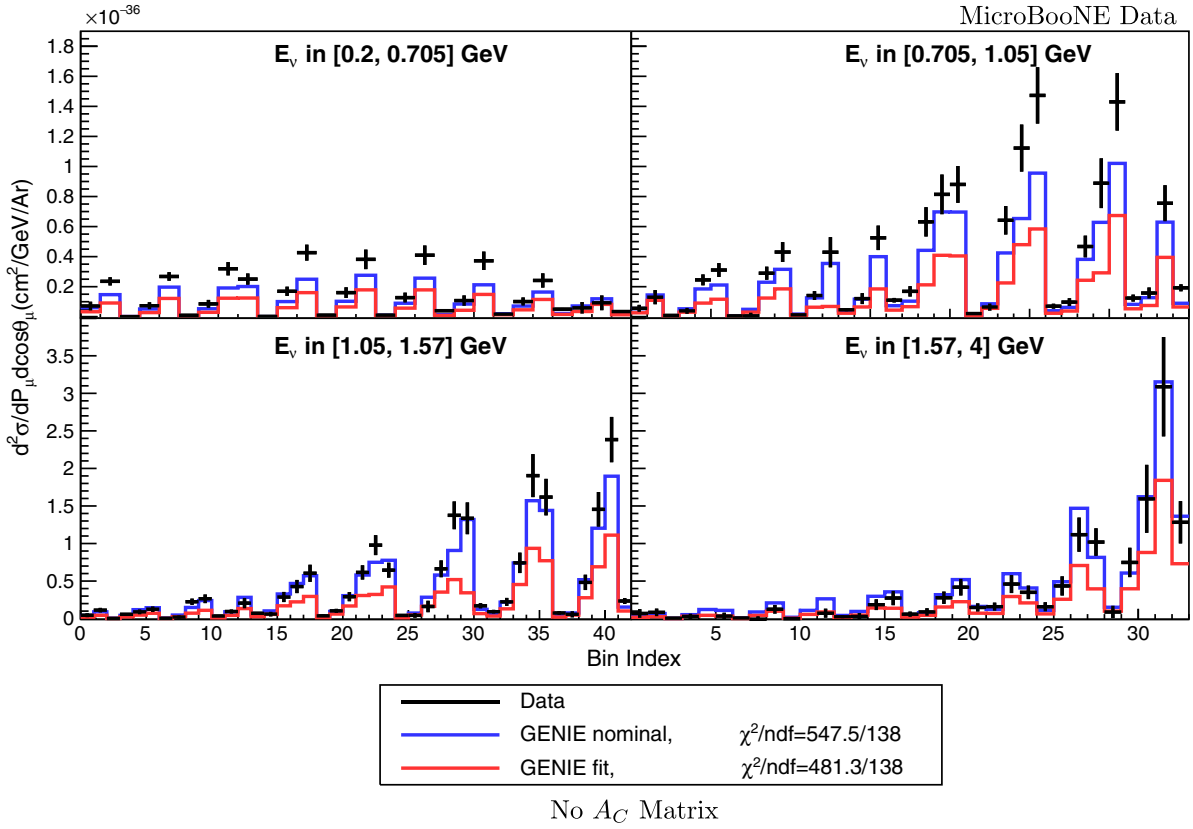


FIG. 8. Comparison of the measured MicroBooNE $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ cross section data (black) to the nominal GENIE model prediction without applying the regularization matrix (blue), and the fit result without applying the regularization matrix (red). Each subfigure corresponds to a different slice of neutrino energy, and the x-axis represents the bin index for the 2D differential cross section within each E_ν slice. The absence of the regularization matrix leads to significant normalization discrepancies.

but introduces additional model dependence. This potential bias, however, can be assessed and constrained prior to unfolding through the data-driven validation methodology introduced in Ref. [30].

In the current context, we are mainly interested in the impact of these assumptions to model fits. We construct a toy study where we use the MicroBooNE three-dimensional ν_μCC cross section results [32] but introduce artificial mismatches related to the flux when fitting to the GENIE nominal model. Since these three-dimensional cross sections were originally extracted at the nominal flux, we know exactly the input flux used in the measurement and are therefore able to induce two mismatches when comparing to the model:

(i) We first construct the model prediction by averaging over an alternative neutrino flux and not the nominal. This alternative flux is derived to be a plausible realization of the MicroBooNE flux covariance matrix. In particular, the first principal component of the flux covariance matrix is used to construct the deviation from the nominal flux at a level of 2σ, while the other principal components are not varied. This represents the situation where there is a systematic issue in the nominal flux prediction but is nevertheless plausible relative to its uncertainties.

(ii) Secondly, to mimic the situation in the "real flux" approach more closely, we remove the contribution of the flux-shape uncertainties in the measurement by subtracting the total flux covariance matrix and replacing it with a 10% normalization uncertainty denoting the overall error in the integrated flux.

We then compare the model prediction based on this alternative flux and re-factored uncertainties to the reported MicroBooNE measurement in Fig. 9 below. As can be seen, there is a clear PPP-like normalization disagreement in the best-fit prediction to the data, even if the overall χ^2 is much improved compared to the nominal case.

This case study highlights that mismatches between the real and nominal neutrino flux also introduce challenges in data-model comparisons, including PPP-related issues and broader inconsistencies. In practice, when cross sections are extracted at the real flux but model predictions are based on the nominal neutrino flux, systematic biases arise through the unmodeled correlations between measurement and prediction, leading to exaggerated penalties in the χ^2 metric. This forces the fit to absorb flux-driven mismatches,

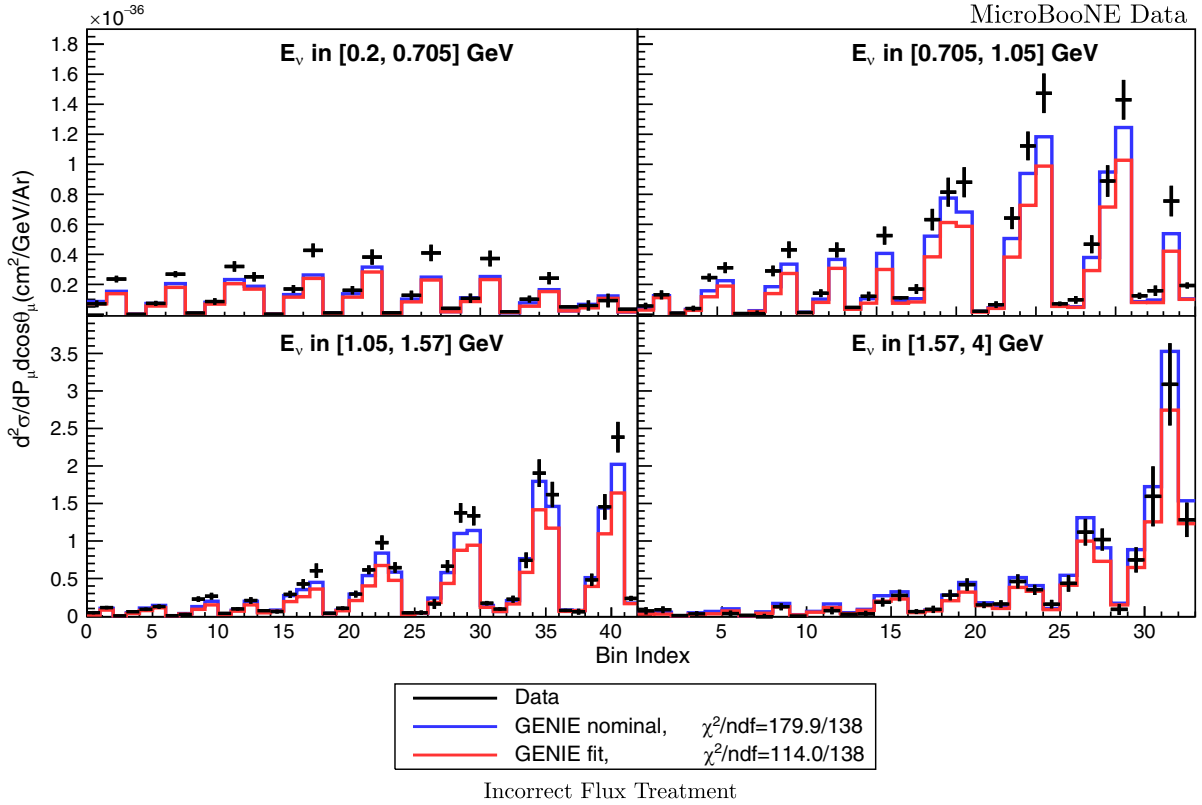


FIG. 9. Comparison of the measured MicroBooNE $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ cross section data (black) to the GENIE model prediction averaged over an alternative MicroBooNE flux (blue), and the corresponding fit result (red) to the data. The comparison as well as the fit is based on an alternative covariance matrix described above, which removes the effects of the flux-shape uncertainties.

potentially overshadowing meaningful physical discrepancies. As a result, the fit may yield unstable or nonphysical parameters, with normalization mismatches serving as a clear manifestation of the PPP effect.

C. Inconsistencies between data and models from model limitations

The third case study investigates issues related to insufficient model coverage in data-model comparisons. To illustrate this point, we introduce the NEUT 5.8.0 model [14] and the T2K 2020 ν_μCC0π measurement [48] in addition to aforementioned GENIE model and MicroBooNE three-dimensional ν_μCC cross section results. We use the unregularized version of the results reported by T2K, preventing the introduction of a PPP issue from any improper treatment related to regularization in the measurement.

The NEUT model incorporates the spectral function (SF) approach [49] based on plane wave impulse approximation [50], which neglects final state interactions via wave distortion, resulting in a significant overestimation of events at low Q^2. In order to illustrate the impact of insufficient models, we restrict the model fitting analysis to six impactful parameters from the full set available for NEUT. Among these parameters is the optical potential [51], which plays a critical role in describing nuclear effects. Other parameters are M_A^{QE}, one *ad hoc* high Q^2 normalizations for quasi-elastic (QE), a 2p2h normalization, and M_A^{RES} and C_5^A proposed by Graczyk and Sobczyk [52] for resonant production. By limiting the number of fitted parameters, the study focuses on how constrained model flexibility can exacerbate issues related to insufficient model coverage, particularly in regions of the phase space where the underlying physics is not fully captured.

As shown in Fig. 10(a), the NEUT model introduced above overpredicts the data while the best-fit model, despite achieving a significantly lower χ^2, under predicts the data, which is a behavior indicative of a PPP issue. To further investigate, the number of fit parameters are increased by introducing two additional *ad hoc* normalization adjustments at high Q^2 and an additional freedom to account for Pauli blocking at low Q^2. This expanded model fit allows for slightly better postfit data-model agreement, both in the χ^2 and the normalization, however a significant normalization discrepancy indicative of a PPP issue still remains. This suggests that both model fits lack the freedom to explain the observed data-model disagreement.

To evaluate whether the observed misfit is due to insufficient model coverage or issues with the underlying base model, the NEUT model is modified through the

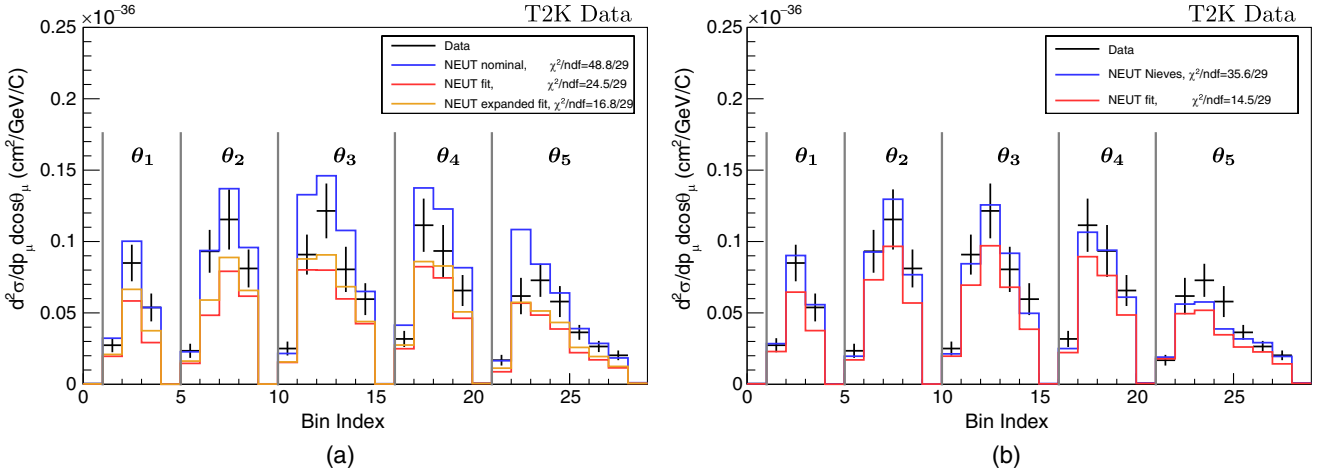


FIG. 10. Comparison of the measured T2K CC0π double differential cross section to (a) the prior prediction of NEUT as well as two fits featuring six and nine free parameters labeled "NEUT fit" and "NEUT expanded fit", respectively, and (b) the prior prediction of NEUT with Nieves model as well as a fit with six parameters. Each angle slice is separated with a vertical gray line and labeled, with a variable number of muon momentum bins spanning [0, 30] GeV/c spanning each slice.

inclusion of the Nieves 1p1h model [53] as a replacement for the SF model. As shown in Fig. 10(b), the NEUT model using Nieves 1p1h achieves a better data-model agreement than the previous NEUT model using a SF model, demonstrating an improved baseline agreement. This model is then fit using the same limited set of six parameters, not including the expanded fit parameters. The resulting postfit χ^2 decreases significantly, although there is still a notable normalization disagreement that arises from the fit attempt, indicating a PPP issue. However, the disagreement is less severe than in the NEUT model fits using the SF model, suggesting that the improved baseline comparison under the Nieves model helps reduce the PPP issue. Still, the limited number of fit parameters appears to be a contributing factor to the normalization discrepancy in all fit attempts.

It is worth noting the difficulty in real-world scenarios in attributing any observed PPP symptom to a specific cause. So far we have attempted to demonstrate the potential for model limitations to induce a PPP issue. However, it is possible that the observed PPP symptoms may be primarily driven by another issue than model limitations. For example, the T2K measurement used in these fit exercises is reported as a "real-flux" measurement, while model predictions are averaged over a nominal flux prediction. As examined in detail in Sec. VI B, this data-model mismatch can cause a PPP issue, meaning that it is possible that the observed PPP symptoms seen here may be driven more by this source or even another unknown source than by the model limitations that have been posited. Although perfect information is not available, it is possible to gain insight on the most likely reality from combining information across multiple studies. Relevant to this particular dataset, Ref. [36] demonstrated that the extra flux uncertainties generated from transforming an older T2K measurement [47] with a similar phase space to a reference flux prediction [54] are small compared to the other uncertainties in the measurement. This suggests that in this case the impact of the improper treatment of comparing a "real-flux" measurement to a "nominal-flux" prediction may be small, reducing the likelihood that it is the primary cause of the observed PPP issues. Another way to test this hypothesis is to investigate additional model comparisons and fit attempts, and see whether they demonstrate similar PPP issues or not. If different outcomes are observed for different models, the flux mismatch hypothesis becomes disfavored.

Following this line of reasoning, we examine the comparison between the T2K data and the GENIE nominal and tuned predictions to the dataset, shown in Fig. 11. The latter prediction was tuned using four of the model's parameters; MaCCQE, NormCCMEC, RPA_CCQE, and

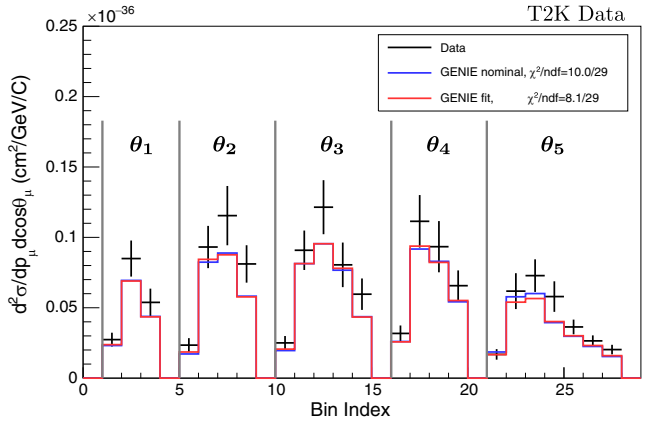


FIG. 11. Comparison of the measured T2K CC0π double differential cross section to GENIE nominal and tuned predictions before and after fitting with four free parameters.

XSecShape_CCMEC. The MaCCQE parameter adjusts the value of the axial mass in the dipole parametrization of the CCQE axial-vector form factor. The NormCCMEC parameter adjusts the normalization of the cross section of meson exchange current (MEC) interactions, which are the dominant type of charged-current 2-particle, 2-hole (CC2p2h) interactions in G18. The RPA_CCQE parameter adjusts the strength of the suppression of the CCQE cross section at low Q^2 due to nucleon-nucleon long range correlations. The XSecShape_CCMEC parameter adjusts the shape of the CC2p2h cross section between the Valencia model and the Empirical model, as described in Ref. [21].

Comparing the nominal GENIE and NEUT predictions in Figs. 11 and 10, respectively, the nominal GENIE model prediction achieves much better agreement with the data, with a much lower χ^2/ndf. Additionally, the nominal GENIE model underpredicts the data, in contrast with the nominal NEUT model prediction which overpredicts the data. After fitting the GENIE model to the data, the χ^2/ndf only improves slightly, with a minimal visual shift in the central value prediction. This may reflect a limitation in the fitted parameters in addressing the data-model discrepancy or may result from the fact that the nominal GENIE model prediction already achieved good agreement with the data, with a χ^2/ndf well below unity. Either way, the overall normalization of the model prediction does not worsen after fitting, meaning the visible sign of a PPP issue that was present in each NEUT model fit is not reproduced when fitting the GENIE model to the same data set. This further supports the view that model limitations, and not flux issues, are the primary cause of the PPP issue observed in fit attempts with the NEUT model. Integrated cross sections for each model comparison to the T2K measurement are shown in Table I. Comparing integrated cross section predictions between model versions demonstrates the presence of PPP normalization issues that were visually identified in Fig. 10.

To directly illustrate how mismatches between data and model can trigger a PPP effect, we further consider the artificial model deficiency that arises when comparing a model prediction on argon to experimental cross section data from T2K, which is measured on a carbon target. The comparison on T2K carbon data is made in Fig. 12 using the GENIE model but erroneously computed under the assumption of an argon target. In contrast with the good model agreement found in Fig. 11, now there is significant disagreement seen between nominal model prediction and the T2K measurement. Given this improper treatment, the fit is no longer successful and shows clear signs of PPP through a significant normalization deficit.

While the six-parameter NEUT model exhibits the PPP effect when fitting the T2K data, it is also valuable to explore its performance on the three-dimensional ν_μCC MicroBooNE data introduced in the previous section. Since NEUT does not include an SF for argon, the Nieves model has been used. Figure 13 presents the results of this fit. Interestingly, while the χ^2 value is not particularly low, there is no evident normalization discrepancy and no clear manifestation of a PPP effect. This outcome underscores an important nuance; although insufficient models can give rise to PPP, its emergence depends on the specific interplay between the data and the model. PPP is therefore not an inevitable consequence of model limitations, but rather arises under particular conditions where mismatches align in a way that accentuates the effect.

TABLE I. Integrated cross sections per carbon atom for the T2K measurement [48] and model predictions from NEUT and GENIE.

Model	$\sigma(10^{-38}\ \mathrm{cm}^2/C)$
T2K measurement	5.64 ± 0.72
NEUT nominal	6.27
NEUT fit	3.54
NEUT expanded fit	4.13
NEUT Nieves	5.73
NEUT Nieves fit	4.21
GENIE nominal	3.00
GENIE fit	3.02

D. Illustrating the QM mitigation strategy

In this section, we illustrate the QM mitigation strategy using the case study where a fit is attempted without applying the regularization matrix A_C to model predictions for the double-differential cross section $d^2\sigma/dE_\mu d\cos\theta_\mu$. This scenario, which can occur if the regularization matrix is not reported, results in significantly degraded fit performance, as demonstrated in Sec. VI A. These examples highlight the challenges posed by regularization-induced

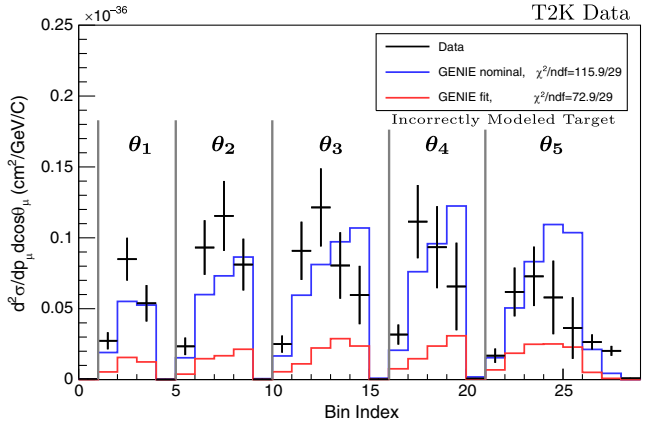


FIG. 12. Comparison of the measured T2K CC0π double differential cross section to the GENIE G18_10a_02_11a model prediction erroneously calculated assuming an argon target. Both data and model cross sections are plotted per twelve nucleons (carbon).

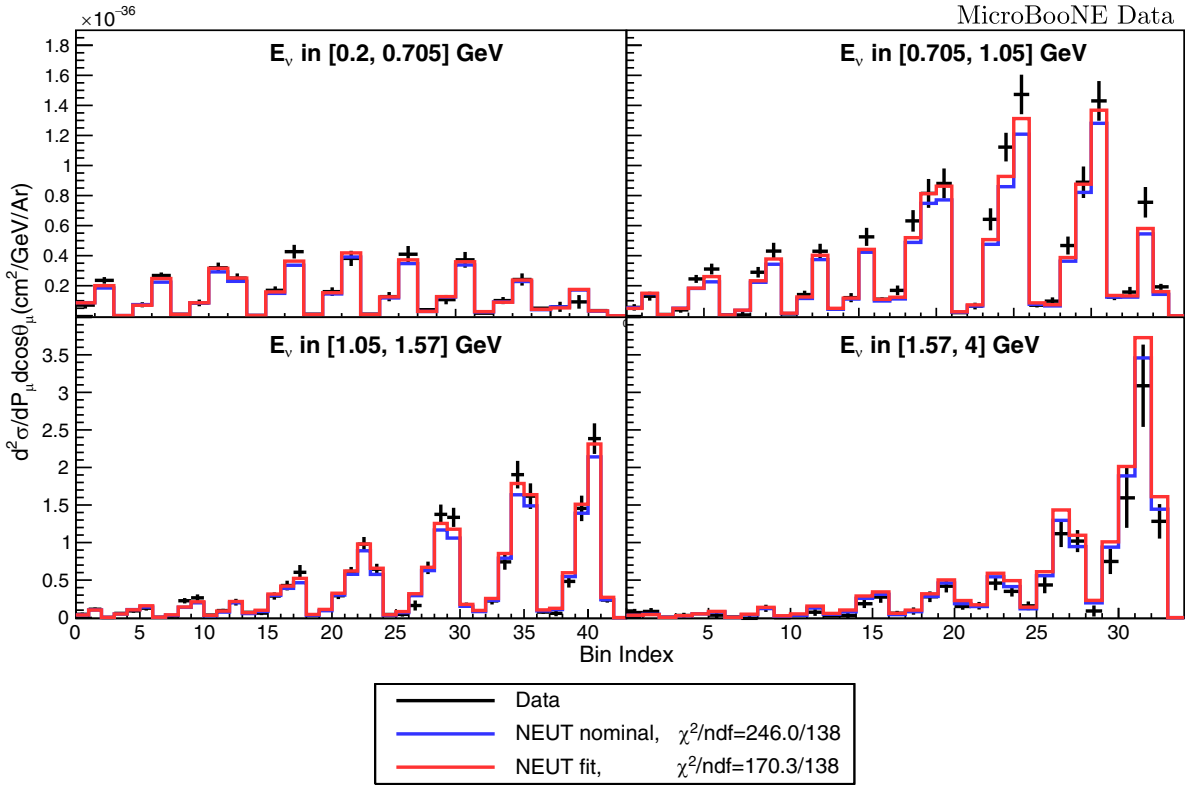


FIG. 13. Comparison of the measured MicroBooNE $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ cross section to the NEUT model predictions before (blue) and after (red) fitting with six free parameters.

distortions and serve as a basis for demonstrating how the QM approach can effectively address these issues.

As described in Sec. V, quantile mapping mitigates these PPP issues in the measurement by addressing the common symptom seen across many DoF. The mismatch between data and model prediction, magnified by a small covariance in some DoF, causes a naive fit attempt to focus on a few potentially insignificant diagonalized bins to the detriment of the model performance on other bins in the measurement space. The insufficiency of the nominal covariance matrix to describe the differences between data and nominal model prediction is shown in Fig. 14. The blue circles represent the nominal data-model σ tension in each bin in the diagonalized covariance matrix basis. The large number of points outside 2σ provides a clear visual indication of the insufficiency of the nominal covariance matrix. This issue is addressed in the QM approach by enlarging the uncertainties in the diagonalized covariance matrix basis, restoring a suitable overall χ^2 distribution as shown in Fig. 15. As a result, a new at tempt at fitting the model to the data after using QM is able to give more appropriate weight to the disagreement within each DoF.

The improvement in the fitted model under QM can also be seen by comparing the parameter values between different fit procedures. Taking the parameter values achieved through fitting the model prediction with the A_C matrix applied as the target outcome, we can compare the performances of the fits that omit the A_C matrix, both with and without using QM. Focusing on the GENIE parameters fitted in the MicroBooNE tune [21], the naive fit without use of the A_C matrix or QM yields fit parameters in tension with their counterparts from the proper fit using the A_C matrix at levels up to 6σ. In comparison, when QM is used tensions all drop to 2σ or lower, demonstrating how

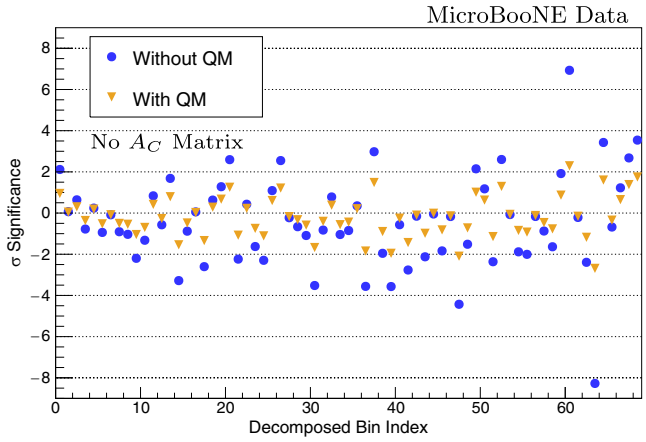


FIG. 14. Data-model σ tension in each bin in the diagonalized covariance matrix basis for the measured MicroBooNE double-differential cross section $d^2\sigma/dE_\mu d\cos\theta_\mu$ before (blue) and after (green) applying QM.

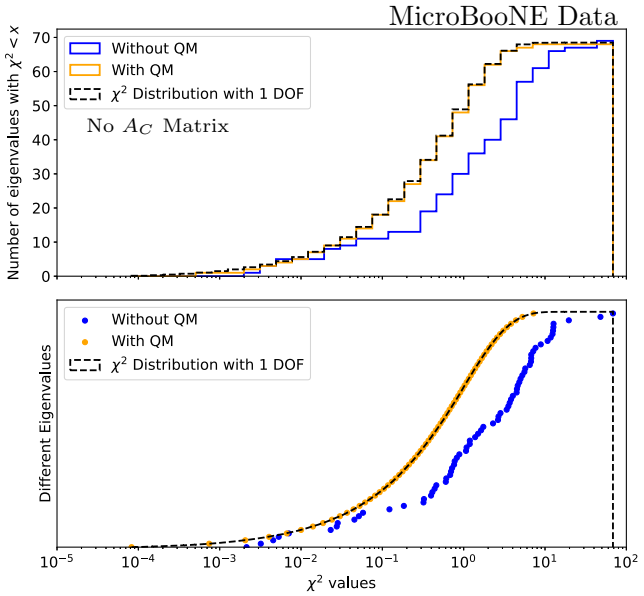


FIG. 15. Distribution of individual χ^2_i values for the measured MicroBooNE double-differential cross section $d^2\sigma/dE_\mu d\cos\theta_\mu$ before and after applying the QM procedure. The distribution after QM aligns more closely with the expected χ^2 distribution.

QM is able to enable a fit in the presence of a data-model mismatch with a physically consistent interpretation with respect to the correct fit.

Furthermore, QM increases the overall covariance so that the fit model prediction is able to describe the data within uncertainties, as shown in Fig. 16. This is especially useful when fitting to multiple data sets where one or more exhibit PPP symptoms, as this conservative approach allows for measurements that cannot be properly compared to a model prediction to be included in a joint fit without overwhelming the fit with PPP issues.

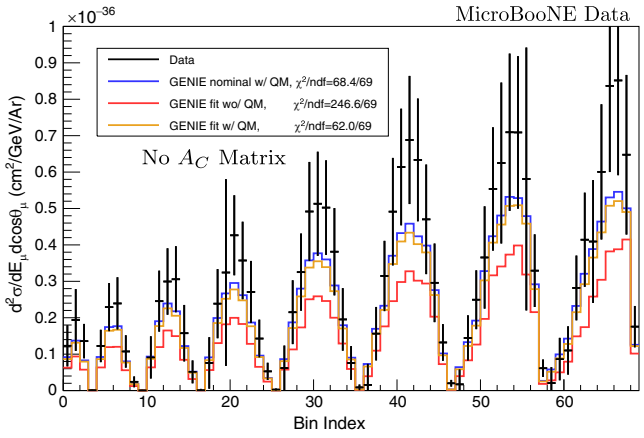


FIG. 16. Comparison of the measured MicroBooNE $d^2\sigma/dE_\mu d\cos\theta_\mu$ cross section data (black) with uncertainties enlarged through QM to the nominal GENIE model prediction without applying the regularization matrix (blue). Also plotted are fit results of the GENIE model without applying the regularization matrix tuned to data with (green) and without (red) using QM. In each case where QM is used in the fit it is also included in the reported χ^2 calculation, and when QM is omitted from the fit it is also omitted from the χ^2 calculation. The x-axis represents the bin index.

VII. DISCUSSIONS

In Sec. II, we explored several scenarios that can lead to data-model mismatches and provided recommendations for avoiding such discrepancies. In Sec. VI, we demonstrated these scenarios through a series of case studies. However, it is impractical to enumerate every possible way data-model mismatches might occur, particularly when miscommunication arises between model tuning teams and analyzers reporting cross sections. To address this challenge, it is highly beneficial for analyzers reporting cross sections to also perform model fitting exercises, ensuring greater consistency and alignment between the reported data and model predictions.

Currently, it is common in cross section papers to report the comparison between the extracted cross sections and central-value model predictions from theoretical models, such as event generators. Test statistics, such as χ^2, used to measure the discrepancies between data and central-value model predictions are often reported and used to draw qualitative conclusions regarding the model predictions. Building on this existing practice, we propose extending it to include uncertainties on the model predictions. This can be achieved using the NUISANCE package [46], which is commonly utilized for releasing extracted cross sections. Incorporating model prediction uncertainties not only enhances the ability to identify PPP-like issues but also facilitates a more thorough evaluation of the models' accuracy and validity. For this exercise, we recommend to adopt the method outlined in Sec. IV, which introduces global and differential goodness-of-fit metrics. These metrics provide a systematic framework for assessing data-model agreement, offering additional insights into potential mismatches and improving the overall robustness of the analysis. Additionally, when fitting models to data, we recommend using as an aid the covariance matrix fit approach with physical boundaries enforced, as described in Sec. III. This allows for computationally efficient model fitting that removes the need to limit the number of model parameters included in the fit.

Looking forward, avoiding mismatches between model predictions and reported cross sections is essential to ensure accurate analyses. However, existing results have already been affected by such mismatches, including missing A_C matrices and discrepancies between real and nominal neutrino flux. To address these challenges, we propose the QM technique, which introduces additional uncertainties to the data to mitigate inconsistencies between the data and model predictions. The QM approach conservatively reduces the weight of affected data and ensures that

potential mismatches do not unduly bias model fits and is particularly useful in combining multiple datasets. By allowing the inclusion of datasets with suspected mismatches while appropriately reducing their influence, QM facilitates robust joint analyses and meaningful comparisons even when inconsistencies are present.

VIII. SUMMARY

This paper addresses the critical challenge of improving neutrino-nuclei interaction modeling for high-precision neutrino oscillation experiments by investigating Peelle's Pertinent Puzzle, a persistent issue stemming from inconsistencies between experimental cross section data and theoretical models. Through case studies, the paper identifies key sources of these inconsistencies, including the omission of the regularization matrix A_C in data unfolding, discrepancies between real and nominal neutrino fluxes, and limitations in model coverage. These issues can lead to significant data-model mismatches, hindering the reliability of model predictions.

To mitigate these challenges, we recommend that analysts avoid the publication of measurements that necessitate known improper treatments such as the omission of the regularization matrix and the use of real-flux measurements without supporting flux details. We also recommend incorporating strategies such as model fitting exercises into cross section publications and leveraging the quantile mapping technique, which introduces additional uncertainties to account for inconsistencies while reducing their impact on model fits. The `NUISANCE` framework is highlighted as a valuable tool for performing systematic fits and identifying PPP-related challenges. Building on top of this, the paper introduces a new model fitting technique that minimizes a test statistic constructed from the data, model prediction, and their covariances. Describing model variations through a covariance matrix allows the simultaneous fitting of an arbitrarily large number of model parameters. By ensuring consistent treatment of data and model predictions, these methods facilitate robust joint analyses and meaningful comparisons across datasets. The findings emphasize the importance of addressing typical data-model inconsistencies and provide insights for advancing the accuracy and reliability of neutrino-nuclei interaction models, critical for the success of high-precision neutrino experiments.

ACKNOWLEDGMENTS

We would like to acknowledge the support and valuable discussions provided during the [55] Generator Studies Workshop at the Pittsburgh Particle Physics, Astrophysics, and Cosmology Center (PittPACC). The workshop served as a highly collaborative platform that facilitated insightful exchanges and inspired advancements in this work. We deeply appreciate PittPACC's efforts in fostering a vibrant scientific community and enabling critical dialogues that contribute significantly to progress in the field. This work was supported by the U.S. Department of Energy, Office of Science, the University of Pittsburgh; the U.K. Science and Technology Facilities Council; the European Research Council; and the Japaneses Society for the Promotion of Science.

DATA AVAILABILITY

No data were created or analyzed in this study.

APPENDIX A: PARAMETER FIT VALUES

A selection of parameter fit values for each of the model fit attempts detailed above are presented in Tables II and III along with their uncorrelated uncertainties. Given the number of parameters fit in some instances, it is not practical to present all fit parameter, therefore only a subset

TABLE II. `GENIE` parameter physical values after fits to the CC inclusive `MicroBooNE` measurements [32,45] and the CC pionless `T2K` measurement [48].

Fit	M_A^{QE} (GeV)	RPA	CCMEC Norm	CCMEC Shape
Nominal	1.1 ± 0.1	0.85 ± 0.15	1.66 ± 0.5	1^{+0}_{-1}
`MicroBooNE` $d\sigma/dE_\mu$	1.18 ± 0.08	1.08 ± 0.36	1.37 ± 0.43	-0.13 ± 1.04
`MicroBooNE` $d\sigma/dE_\mu$ w/no A_C	1.15 ± 0.08	1.21 ± 0.34	1.29 ± 0.41	0.05 ± 1.00
`MicroBooNE` $d^2\sigma/dE_\mu d\cos\theta_\mu$	1.26 ± 0.07	0.72 ± 0.25	1.67 ± 0.38	0.09 ± 0.85
`MicroBooNE` $d^2\sigma/dE_\mu d\cos\theta_\mu$ w/no A_C	1.14 ± 0.04	0.12 ± 0.13	-0.85 ± 0.23	3.15 ± 0.56
`MicroBooNE` $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$	1.36 ± 0.05	0.14 ± 0.18	1.68 ± 0.34	0.05 ± 0.66
`MicroBooNE` $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ w/no A_C	0.66 ± 0.02	1.05 ± 0.07	0.25 ± 0.15	1.29 ± 0.52
`MicroBooNE` $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ w/alt flux	1.07 ± 0.05	0.69 ± 0.18	1.10 ± 0.28	1.33 ± 0.82
`MicroBooNE` $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ w/QM, no A_C	0.99 ± 0.09	0.76 ± 0.35	0.61 ± 0.44	1.51 ± 0.88
`MicroBooNE` $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$ w/QM, reordered, no A_C	1.02 ± 0.09	0.65 ± 0.34	0.98 ± 0.39	1.58 ± 0.88
`T2K` $d^2\sigma/dP_\mu d\cos\theta_\mu$	1.12 ± 0.10	0.75 ± 0.26	1.37 ± 0.43	0.48 ± 0.37
`T2K` $d^2\sigma/dP_\mu d\cos\theta_\mu$ w/argon model	0.53 ± 0.14	1.32 ± 0.23	1.37 ± 0.43	0.94 ± 0.71

TABLE III. NEUT parameter physical values after fits to the CC inclusive MicroBooNE measurement [45] and the CC pionless T2K measurement [48]. Due to the way the fits of the NEUT model were implemented, it is not feasible to accurately quote parameter uncertainties.

	M_A^{QE} (GeV)	Optical Potential	High-Q^2 Norm	2p2h Norm	M_A^{RES} (GeV)	C_5^A
Nominal	1.21	0.00	1.00	1.00	0.95	1.01
T2K $d^2\sigma/dP_\mu d\cos\theta_\mu$	0.56	1.00	1.96	0.13	1.08	1.10
T2K $d^2\sigma/dP_\mu d\cos\theta_\mu$ expanded	0.70	0.90	1.51	0.54	0.44	2.00
T2K $d^2\sigma/dP_\mu d\cos\theta_\mu$ w/Nieves	0.79	0.00	1.19	0.57	0.86	1.30
MicroBooNE $d^2\sigma(E_\nu)/dP_\mu d\cos\theta_\mu$	1.45	0.00	1.00	1.00	0.95	1.00

are shown. Specifically, in fits of GENIE predictions, only the four GENIE parameters used in the tuning of the MicroBooNE model [21] are shown. Note that fits performed by minimizing a χ^2 test statistic computed with the covariance matrix do not inherently constrain parameter values within their physical bounds. As a result, some postfit parameter values are nonphysical. This issue is only significantly observed in cases where improper treatments are artificially added, and could be avoided through a more sophisticated test statistic minimization that incorporated parameter boundaries, as discussed in Sec. III.

Fit parameters are shown only to give further insight into the degree of PPP symptoms observed, not to promote a particular model tune for any physics goal. For example, in some cases where inconsistencies are introduced nonphysical parameter fit values are observed. These fit parameters are intended as illustrative rather than prescriptive for multiple reasons, including their incomplete listing, nonphysical fitted parameter values, the limitation of only presenting uncorrelated uncertainties rather than the full covariance between parameters, and the fact that correlations between cross section models used to unfold some measurements and fitted models were not considered in fitting.

APPENDIX B: QUANTILE MAPPING DETAILS

QM considers a given random variable X with probability distribution function (PDF) f_X and CDF g_X. The goal of QM is to employ a mapping that transforms the PDF f_X into a chosen target distribution f_Y with CDF g_Y. This is achieved by ensuring that g_X maps to g_Y, which guarantees agreement between the distributions f_X and f_Y as well. Consider the mapping $X \to \tilde{X}$:

$$\tilde{X} = g_Y^{-1}(g_X(X)). \tag{B1}$$

Since $g_X(X)$ and therefore $g_Y(\tilde{X})$ are uniformly distributed on [0, 1], $\tilde{X}$ has the same CDF g_Y as Y. Therefore, for any well-behaved f_X, a mapping can be constructed to yield the desired PDF f_Y so long as g_Y^{-1} and g_X are known.

In this application, the given random variable of interest is the uncertainty-weighted difference between measurement and prediction in the diagonalized basis of uncorrelated bins. This allows each χ_i^2 term to be taken as a throw from the PDF of X:

$$f_X(x) = k(x)/n, \tag{B2}$$

$$g_X(x) = K(x)/n, \tag{B3}$$

where (x) counts the number of bins with a χ_i^2 value of x, $K(x)$ counts the number of bins with a χ_i^2 value less than or equal to x, and the domain $\mathbb{R} \geq 0$ includes all possible χ_i^2 values. Given the finite number of bins, $k(x)$ will necessarily be 0 for most values of x. We are interested in constructing a covariance $\tilde{\Sigma}_D$ that is able to describe the observed discrepancy D between measurement and prediction, therefore individual χ_i^2 terms under $\tilde{\Sigma}_D$ should follow a χ^2 distribution with one DoF, and we can set

$$f_Y(x) = \chi^2(x, 1). \tag{B4}$$

Rather than computing g_Y^{-1} from f_Y, it is easier to determine its form by inspecting the mapping performed by g_Y. Specifically, note that g_Y maps the $\chi^2(x, 1)$ distribution to p, where p represents the p-value of sampling a χ^2 value of x or smaller. Conversely, g_Y^{-1} maps a p-value to its corresponding χ^2 value. For one DoF, the χ^2 distribution is generated from the square of a single normally distributed variable, and thus the relation between p-value and χ^2 value can be computed from the error function Erf:

$$g_Y^{-1}(p) = 2(Erf^{-1}(p))^2. \tag{B5}$$

It is now possible to determine the required mapping $X \to \tilde{X}$ so that $\tilde{X}$ will be χ^2 distributed with one DoF. To achieve this transformation the covariance $\Sigma_{D,ii}$ of each bin is scaled by $\chi_i^2/g_Y^{-1}(g(\chi_i^2))$ while the measured and predicted cross sections, and therefore D_i, are left unchanged. Since the initial issue was an insufficient covariance to describe the difference between measurement and prediction, this will naturally lead to an overall enlargement of the covariance describing the comparison, which can be thought of as an enlargement of the uncertainty in the data measurement. Furthermore, one can deviate from a strict implementation of QM by declining to reduce the uncertainty of any bins whenever QM calls for such a reduction.

This makes the application of QM a conservative strategy for mitigating tensions between data and model.

APPENDIX C: UNCERTAINTY REORDERING

In practice, QM proves to be effective at reducing the impact of an improper treatment and can prevent a PPP normalization issue, but does not always lead to a fit result that significantly improves on the prefit model prediction. One way to understand this outcome is to realize that QM merely rescales each χ_i^2 value without changing their ordering from smallest to largest. As a result, bins that previously exhibited extreme tension from an improper treatment will still exhibit large tension after QM but to a reduced degree. This tension may still be significant enough to dis-incentivize a fit from addressing disagreements in other DoF. The result is that a fit attempt, even after performing QM, may be biased by nonphysical discrepancies resulting from an improper treatment rather than focusing on real physics information present in a measurement.

To resolve this issue, an analyst may apply their own intuition to give preference to the fitting of certain DoFs over others. In general this is a difficult task as it requires deciding which discrepancies in the diagonalized space are most physically relevant. However, it may be reasonable to increase the significance of a large eigenvalue DoF such as the normalization. Mechanically, this can be achieved by reducing the uncertainty for one or more DoF before applying QM. Then QM will ensure that the overall covariance is able to describe the data-model discrepancy, but will place any preferenced DoF in higher tension than otherwise, at the expense of other DoF.

We demonstrate an implementation of uncertainty reordering for the example of the previously mentioned inclusive ν_μCC double differential cross section $d^2\sigma/dE_\mu d\cos\theta_\mu$ in the case where the regularization matrix is omitted from the model prediction. Previously, this omission was demonstrated to lead to a severe PPP issue in the overall normalization of the postfit model. Furthermore, the use of QM was found to reduce this PPP issue, but visually the postfit model did not significantly improve the comparison to the data compared to the nominal model prediction. In this example the largest eigenvalue DoF (referred to here as normalization) is prioritized by halving its uncertainty, in effect increasing the corresponding χ_i^2 by a factor of four. Quantile mapping is then applied, which increases the covariance in many DoF including the normalization, but keeps the newly established relative order among DoF. As a result, the fit gives higher significance to preserving a good normalization agreement, shown in Fig. 17. A better visual agreement between postfit model and data is observed in this instance, approaching the visual fit quality observed when the regularization matrix is included. In this example the use of uncertainty reordering was straightforward as only the normalization was addressed, however in principle this method may be more difficult to successfully apply based on the complexity of the measurement and the nature of the underlying improper treatment between the data and model.

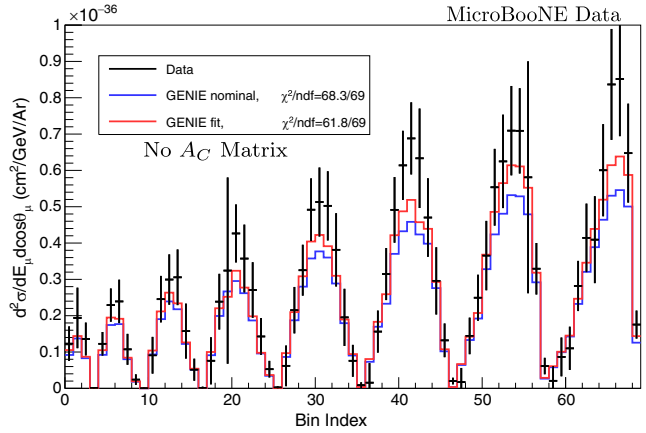


FIG. 17. Comparison of the measured MicroBooNE $d^2\sigma/dE_\mu d\cos\theta_\mu$ cross section data (black) to the nominal model prediction without applying the regularization matrix (blue), and the fit result without applying the regularization matrix (red). The x-axis represents the bin index. The PPP-mitigation strategy combining uncertainty reordering and quantile mapping allows for a somewhat successful fit despite the mismatch between data and model.

[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), Expanding neutrino oscillation parameter measurements in NOvA using a Bayesian approach, Phys. Rev. D **110**, 012005 (2024).

[3] B. Abi *et al.* (DUNE Collaboration), Long-baseline neutrino oscillation physics potential of the DUNE experiment, Eur. Phys. J. C **80**, 978 (2020).

[4] K. Abe *et al.* (Hyper-Kamiokande Collaboration), Hyper-Kamiokande Design Report, arXiv:1805.04163.

[5] K. Abe *et al.* (T2K Collaboration), Constraint on the matter-antimatter symmetry-violating phase in neutrino oscillations, Nature (London) **580**, 339 (2020).

[6] X. Qian and P. Vogel, Neutrino mass hierarchy, Prog. Part. Nucl. Phys. **83**, 1 (2015).
[7] K. N. Abazajian *et al.*, Light sterile neutrinos: A white paper, arXiv:1204.5379.
[8] F. Di Lodovico, R. B. Patterson, M. Shiozawa, and E. Worcester, Experimental considerations in long-baseline neutrino oscillation measurements, Annu. Rev. Nucl. Part. Sci. **73**, 69 (2023).
[9] R. L. Workman *et al.* (Particle Data Group), Review of particle physics, Prog. Theor. Exp. Phys. **2022**, 083C01 (2022).
[10] J. A. Formaggio and G. P. Zeller, From eV to EeV: Neutrino cross sections across energy scales, Rev. Mod. Phys. **84**, 1307 (2012).
[11] U. Mosel, Neutrino interactions with nucleons and nuclei: Importance for long-baseline experiments, Annu. Rev. Nucl. Part. Sci. **66**, 171 (2016).
[12] F. Gross *et al.*, 50 years of quantum chromodynamics, Eur. Phys. J. C **83** (2023).
[13] L. Alvarez-Ruso *et al.* (GENIE Collaboration), Recent highlights from GENIE v3, Eur. Phys. J. Special Topics **230**, 4449 (2021).
[14] Y. Hayato and L. Pickering, The NEUT neutrino interaction simulation program library, Eur. Phys. J. Special Topics **230**, 4469 (2021).
[15] T. Golan, J. T. Sobczyk, and J. Zmuda, NuWro: The Wroclaw Monte Carlo generator of neutrino interactions, Nucl. Phys. B, Proc. Suppl. **229–232**, 499 (2012).
[16] 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).
[17] J. Isaacson, W. I. Jay, A. Lovato, P. A. N. Machado, and N. Rocco, Introducing a novel event generator for electron-nucleus and neutrino-nucleus scattering, Phys. Rev. D **107**, 033007 (2023).
[18] H. Gallagher, Event generator tuning in the resonance region for the minos experiment, Nucl. Phys. B, Proc. Suppl. **159**, 229 (2006).
[19] M. A. Acero *et al.* (NOvA, R. Group Collaborations), Adjusting neutrino interaction models and evaluating uncertainties using NOvA near detector data, Eur. Phys. J. C **80**, 1119 (2020).
[20] P. Stowell *et al.* (MINERvA Collaboration), Tuning the GENIE pion production model with MINERνA data, Phys. Rev. D **100**, 072005 (2019).
[21] P. Abratenko *et al.* (MicroBooNE Collaboration), New $CC0\pi$ GENIE model tune for MicroBooNE, Phys. Rev. D **105**, 072001 (2022).
[22] R. Frühwirth, D. Neudecker, and H. Leeb, Peelle's pertinent puzzle and its solution, EPJ Web Conf. **27**, 00008 (2012).
[23] G. Mention, M. Fechner, T. Lasserre, T. A. Mueller, D. Lhuillier, M. Cribier, and A. Letourneau, The reactor antineutrino anomaly, Phys. Rev. D **83**, 073006 (2011).
[24] C. Zhang, X. Qian, and M. Fallot, Reactor antineutrino flux and anomaly, Prog. Part. Nucl. Phys. **136**, 104106 (2024).
[25] C. Zhang, X. Qian, and P. Vogel, Reactor antineutrino anomaly with known θ_{13}, Phys. Rev. D **87**, 073018 (2013).
[26] C. Giunti, Y. Li, C. Ternes, and Z. Xin, Reactor antineutrino anomaly in light of recent flux model refinements, Phys. Lett. B **829**, 137054 (2022).
[27] J. Chakrani *et al.*, Parametrized uncertainties in the spectral function model of neutrino charged-current quasielastic interactions for oscillation analyses, Phys. Rev. D **109**, 072006 (2024).
[28] G. D'Agostini, On the use of the covariance matrix to fit correlated data, Nucl. Instrum. Methods Phys. Res., Sect. A **346**, 306 (1994).
[29] R. Radev and S. Dolan, Flow matching mitigates gaussian error approximations in neutrino cross-section measurements, *Presented at International Workshop on Neutrino Cross Sections for Astrophysics and Oscillations* (NuSTEC, Milan, 2024).
[30] P. Abratenko *et al.* (MicroBooNE Collaboration), Data-driven model validation for neutrino-nucleus cross section measurements, Phys. Rev. D **111**, 092010 (2025).
[31] 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).
[32] P. Abratenko *et al.* (MicroBooNE Collaboration), Measurement of three-dimensional inclusive muon-neutrino charged-current cross sections on argon with the MicroBooNE detector, Phys. Lett. B **870**, 139939 (2025).
[33] 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).
[34] W. Tang, X. Li, X. Qian, H. Wei, and C. Zhang, Data unfolding with Wiener-SVD method, J. Instrum. **12**, P10002 (2017).
[35] S. Gardiner, Mathematical methods for neutrino cross-section extraction, arXiv:2401.04065.
[36] L. Koch and S. Dolan, Treatment of flux shape uncertainties in unfolded, flux-averaged neutrino cross-section measurements, Phys. Rev. D **102**, 113012 (2020).
[37] K. Abe *et al.* (The T2K Collaboration), Characterization of nuclear effects in muon-neutrino scattering on hydrocarbon with a measurement of final-state kinematics and correlations in charged-current pionless interactions at T2K, Phys. Rev. D **98**, 032003 (2018).
[38] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA, 2006).
[39] F. James and M. Winkler, MINUIT User's Guide (2004), https://inspirehep.net/literature/1258345.
[40] L. Cooper-Troendle, Conditional constraint fitting (2025), https://github.com/London2/Conditional_Constraint_Fitting.
[41] E. Gross and O. Vitells, Trial factors for the look elsewhere effect in high energy physics, Eur. Phys. J. C **70**, 525 (2010).
[42] Z. Sidák, Rectangular confidence regions for the means of multivariate normal distributions, J. Am. Stat. Assoc. **62**, 626 (1967).
[43] F. N. David and N. L. Johnson, The probability integral transformation when parameters are estimated from the sample, Biometrika **35**, 182 (1948).

[44] G. D'Agostini, A Multidimensional unfolding method based on Bayes' theorem, Nucl. Instrum. Methods Phys. Res., Sect. A **362**, 487 (1995).
[45] 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).
[46] P. Stowell *et al.*, NUISANCE: A neutrino cross-section generator tuning and comparison framework, J. Instrum. **12**, P01016 (2016).
[47] T2K Collaboration (T2K Collaboration), Measurement of double-differential muon neutrino charged-current interactions on C_8H_8 without pions in the final state using the T2K off-axis beam, Phys. Rev. D **93**, 112012 (2016).
[48] K. Abe *et al.* (T2K Collaboration), Simultaneous measurement of the muon neutrino charged-current cross section on oxygen and carbon without pions in the final state at T2K, Phys. Rev. D **101**, 112004 (2020).
[49] O. Benhar, A. Fabrocini, S. Fantoni, and I. Sick, Spectral function of finite nuclei and scattering of GeV electrons, Nucl. Phys. **A579**, 493 (1994).
[50] A. M. Ankowski, O. Benhar, and M. Sakuda, Improving the accuracy of neutrino energy reconstruction in charged-current quasielastic scattering off nuclear targets, Phys. Rev. D **91**, 033005 (2015).
[51] A. M. Ankowski and J. T. Sobczyk, Construction of spectral functions for medium-mass nuclei, Phys. Rev. C **77**, 044311 (2008).
[52] K. M. Graczyk and J. T. Sobczyk, Form factors in the quark resonance model, Phys. Rev. D **77**, 053001 (2008); Phys. Rev. D **79**, 079903(E) (2009).
[53] J. Nieves, I. Ruiz Simo, and M. J. Vicente Vacas, Inclusive charged–current neutrino–nucleus reactions, Phys. Rev. C **83**, 045501 (2011).
[54] K. Abe *et al.* (T2K Collaboration), T2K neutrino flux prediction, Phys. Rev. D **87**, 012001 (2013).
[55] https://indico.fnal.gov/event/64511/.