

Efficient Monte Carlo event generation for neutrino-nucleus exclusive cross sections

Mathias El Baz^{1,*}, Federico Sánchez¹, Natalie Jachowicz², Kajetan Niewczas²,
Ashish Kumar Jha² and Alexis Nikolakopoulos^{3,4}

¹*Département de Physique Nucléaire et Corpusculaire, Université de Genève, Geneva, Switzerland*

²*Department of Physics and Astronomy, Ghent University, Proeftuinstraat 86, 9000 Gent, Belgium*

³*Fermi National Accelerator Laboratory, Batavia, Illinois, USA*

⁴*Department of Physics, University of Washington, Seattle, Washington 98195, USA*



(Received 28 February 2025; accepted 20 May 2025; published 6 June 2025)

Modern neutrino-nucleus cross section computations need to incorporate sophisticated nuclear models to achieve greater predictive precision. However, the computational complexity of these advanced models often limits their practicality for experimental analyses. To address this challenge, we introduce a new Monte Carlo method utilizing normalizing flows to generate surrogate cross sections that closely approximate those of the original model while significantly reducing computational overhead. As a case study, we built a Monte Carlo event generator for the neutrino-nucleus cross section model developed by the Ghent group. This model employs a Hartree-Fock procedure to establish a quantum mechanical framework in which both the bound and scattering nucleon states are solutions to the mean-field nuclear potential. The surrogate cross sections generated by our method demonstrate excellent accuracy with a relative effective sample size of more than 98.4%, providing a computationally efficient alternative to traditional Monte Carlo sampling methods for differential cross sections.

DOI: [10.1103/v7sz-vn3b](https://doi.org/10.1103/v7sz-vn3b)

I. INTRODUCTION

The study of neutrino interactions with nuclei is critical for reducing systematic uncertainties in modern neutrino oscillation experiments. Among the most important processes is charged-current quasielastic (CCQE) scattering, where a neutrino interacts with a nucleon target inside a nucleus, emitting a charged lepton while ejecting one nucleon. This interaction forms the backbone of event reconstruction in neutrino oscillation experiments like the Tokai to Kamioka experiment (T2K [1]) and the new Hyper-Kamiokande water Cherenkov detector [2]. However, the theoretical modeling of CCQE interactions is complex due to the need to account for the nuclear environment in which the nucleons are bound originally [3].

Moreover, these experiments often use neutrino beams that span a broad energy spectrum—from a few hundred MeV to several GeV—making Monte Carlo (MC) event generators indispensable. MC generators, such as NEUT [4], GENIE [5], NuWro [6] and ACHILLES [7], efficiently

sample different interaction processes over many possible final states, enabling exhaustive modeling across this wide energy range. Such modeling is essential for extracting oscillation parameters and minimizing systematic uncertainties in neutrino physics.

A variety of interaction models have been implemented in MC generators to describe neutrino-nucleus interactions. However, most of these models are limited to providing only inclusive cross sections and therefore cannot reliably predict final-state hadron kinematics. As a consequence, the latter are often generated in an *ad hoc* and unrealistic manner [8,9].

The only approaches currently used in neutrino generators that systematically provide final-state nucleon kinematics are based on the factorization obtained in the plane-wave impulse approximation (PWIA). In this picture, one assumes no final-state interactions, so the process can be factorized into (i) the primary neutrino-nucleon interaction and (ii) a nuclear model describing the bound nucleons. However, these approaches lack the effects of the nuclear medium on the outgoing nucleon.

In the simplest instance of PWIA, the nucleus is modeled as a relativistic Fermi gas [10], in which nucleons are treated as fermions uniformly occupying momentum states up to the Fermi momentum in a constant potential. This approach has been remarkably successful in describing general properties of intermediate-energy processes,

*Contact author: mathias.elbaz@unige.ch

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) 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.

especially in inclusive electron scattering. Yet, it neglects important nuclear features such as the shell structure and nucleon-nucleon correlations.

A more refined description of the nuclear medium still within the PWIA is provided by the spectral function (SF) model [11], which yields a probability distribution $S(p, E)$ for finding a nucleon with momentum p and removal energy E based on $(e, e'p)$ data and theoretical inputs. Although the SF model captures the spread in binding energies and some nucleon-nucleus correlations, its essential reliance on the PWIA means that correlations essential to accurately predict the final-state hadron kinematics cannot be fully taken into account.

While these approximations were adequate when the focus was largely on lepton kinematics, the increased precision of upcoming accelerator-based long-baseline neutrino oscillation experiments will enable much more detailed measurements of outgoing hadron kinematics. To achieve the few-percent level of uncertainty on neutrino interactions required by these next-generation neutrino oscillation studies, cross section models must go beyond PWIA to include final-state interactions.

To obtain exclusive differential cross sections, a theoretical framework that accurately describes the hadron system is required. One such approach is based on the distorted-wave impulse approximation (DWIA) in the mean-field approximation [12,13]. Mean-field-based models describe nucleons within an averaged nuclear potential generated by all other nucleons. This naturally leads to a shell structure, where each nucleon in the initial state is described by a bound wave function. Meanwhile, the distorted-wave treatment goes beyond the plane-wave description by also considering the final-state interactions of the outgoing nucleon: instead of emerging as a free particle, the outgoing nucleon propagates through the same mean-field potential and undergoes distortions due to its interactions with the residual nucleus. This approach has recently gained renewed attention in the context of event generator development, with the first implementation of a relativistic DWIA-based CCQE model in NEUT [14].

Despite their advantages, mean-field-based cross section calculations are computationally expensive compared to the factorized PWIA. The slow evaluation speed of these cross sections translates into slow sampling rates when traditional methods like accept-reject algorithms are used. To mitigate this, some current solutions [15–18] involve precomputing the hadron tensor components into heavy multidimensional grids, known as hadron tensor tables [8,19–21]. This method allows one to compute the cross section much faster and make an accept-reject sampling possible. However, this approach has significant drawbacks: it lacks flexibility, requiring recalculation for intricate change in theoretical parameters, and seem infeasible for multinucleon knockout exclusive cross sections due to the exponential scaling of table size with dimensionality.

Recent advancements in artificial intelligence, particularly in deep learning, offer promising solutions to these challenges [22–25]. In this work, we propose an alternative sampling method based on normalizing flows (NFs) [26]. Normalizing flows transform a simple base distribution, such as a Gaussian, into a complex target distribution through a series of learnable diffeomorphisms. This enables efficient sampling and evaluation of the target distribution—in this case, the neutrino-nucleus cross section. In this work, we demonstrate the capacity of NFs to build a Monte Carlo event generator for neutrino-induced cross sections. For our proof of principle we use the Hartree-Fock mean-field exclusive one-particle-one-hole ($1p1h$) cross section formalism developed by the Ghent group [13,27,28]. These cross sections exhibit all the features needed to demonstrate that our approach can pave the way for faster and more flexible event generation in neutrino physics.

The structure of this article is as follows: We first review the theoretical framework behind CCQE interactions within the mean-field approach adopted by the Ghent group. Following this, we introduce the normalizing flows technique and outline how it is integrated into the Monte Carlo event generation process. Finally, we present results demonstrating the efficiency and accuracy of the generator.

II. NEUTRINO-NUCLEUS EXCLUSIVE CCQE CROSS SECTION IN THE MEAN-FIELD FRAMEWORK

The aim of this study is to demonstrate the ability to generate events with precision and speed for single nucleon knockout distributed according to the Hartree-Fock calculations developed by the Ghent group. While this model can predict both single and multiple nucleon knockout, providing a Monte Carlo event generator for all processes in their full complexity lies beyond the scope of this work. Future steps for handling the other processes are discussed in Sec. V. Here, we focus on $1p1h$ interactions due to a single-nucleon operator. The goal of this section is to introduce the relevant kinematic structure of the cross section.

The process is the following:

$$\nu_\mu + {}^A_Z X \longrightarrow \mu^- + p + ({}^{A-1}_{Z-1} X)^*, \quad (1)$$

where

- (i) ν_μ is the incoming muon neutrino;
- (ii) ${}^A_Z X$ represents the target nucleus with atomic number Z and mass number A ;
- (iii) μ^- is the outgoing charged lepton;
- (iv) p is the ejected proton; and
- (v) $({}^{A-1}_{Z-1} X)^*$ is the residual nucleus in an excited state due to the interaction.

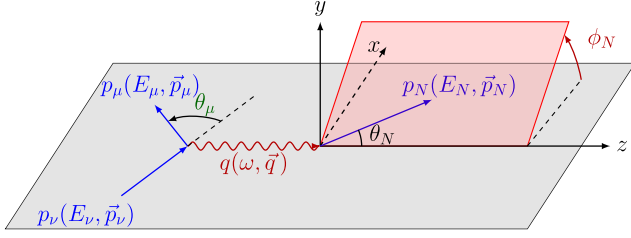


FIG. 1. $1p1h$ kinematics in the laboratory frame with the z axis along the momentum transfer.

The kinematics are displayed in Fig. 1, where the target nucleus is at rest, and the coordinate system is chosen so that the momentum transfer $\vec{q}$ defines the z axis.

The interaction between the neutrino and a nucleon within the nucleus by one-boson exchange is described using the contraction of the lepton and hadron tensors, denoted as $L_{\mu\nu}$ and $W^{\mu\nu}$, respectively,

$$\frac{d^6\sigma(E_\nu)}{d\omega d\Omega_\mu dE_N d\Omega_N} \propto L_{\mu\nu} W^{\mu\nu},$$

where $\omega = E_\nu - E_\mu$ is the energy transfer, E_N is the outgoing nucleon energy, and Ω are the solid angles for the muon and proton. The lepton tensor is independent of the nuclear model and is calculated in electroweak theory. As the cross section is not differential in neutrino energy, event generators typically treat E_ν as an external input, modeling the conditional cross section for a fixed neutrino energy.

Clearly, the cross section is invariant with respect to rotations around the z axis. This allows samples to be uniformly rotated by a ϕ angle in the range $[0, 2\pi]$ without altering the physical results. Consequently, the variables that need to be sampled at fixed E_ν are $(\omega, \theta_\mu, E_N, \Omega_N)$.

We make an additional simplification (without incurring any loss of information) to reduce the number of independent variables that is specific to the mean-field model. The hadron tensor $W^{\mu\nu}$ is a bilinear product of the hadron currents with an energy conserving δ function,

$$W^{\mu\nu} = \sum_{i,f} \langle i | \mathcal{J}^{\mu\dagger} | f \rangle \langle f | \mathcal{J}^\nu | i \rangle \delta(E_N + E_{A-1}^* - M_A - \omega), \quad (2)$$

where $|i\rangle$ and $|f\rangle$ are the initial and final hadron states. The summation is understood to run over all relevant final states and average over the initial states. $\mathcal{J}^\nu$ is the nuclear current operator.

The nuclear states $|i\rangle$ and $|f\rangle$ are obtained as Slater determinants, composed of nucleon solutions to the Schrödinger equation with the mean-field potential $V_{\text{MF}}^{(\alpha)}$,

$$[\hat{T}^{(\alpha)}(r) + V_{\text{MF}}^{(\alpha)}(r)]\phi_\alpha(r) = \epsilon_\alpha\phi_\alpha(r), \quad (3)$$

with the kinetic energy operator being

$$\hat{T}^{(\alpha)}(r) = -\frac{\hbar^2}{2} \left[\frac{1}{r^2} \frac{d}{dr} \left(\frac{r^2}{m^*(r)} \frac{d}{dr} \right) - \frac{\ell_\alpha(\ell_\alpha + 1)}{m^*(r)r^2} \right],$$

with α denoting the set of quantum numbers identifying the single-nucleon state and $m^*(r)$ the effective mass (since the Hartree-Fock Hamiltonian is nonlocal) [29]. The mean-field potential is derived via a Hartree-Fock procedure [30], starting from an effective Skyrme nucleon-nucleon potential [31] and extending it to a nucleon-nucleus potential.

The final state $|f\rangle$ is generally expressed as a sum of orthogonal continuum solutions to the same equation with the correct asymptotic behavior. The complexity of nuclear modeling is embedded in the summation over all possible final states, which constitutes the primary computational bottleneck due to the necessity of computing and summing over a large number of distorted-wave solutions of Eq. (3).

The ground state $|i\rangle$ is described as a Slater determinant of the A lowest-energy single-particle states of the mean-field potential. The states have quantum numbers $(\epsilon, n, l, j, m_j, i_z)$, i.e., respectively, energy, principal quantum number, orbital and total angular momentum, projection of total angular momentum, and isospin. Because of spherical symmetry, the states within a shell characterized by quantum numbers $\alpha = (\epsilon_\alpha, n_\alpha, l_\alpha, j_\alpha, i_z)$ are energy degenerate. Since we are considering nucleon knockout to the continuum, the residual system can then only be left in states with excitation energy given by the single-particle energies of the shells $\epsilon_\alpha = M_N + E_{A-1}^* - M_A$. Using this to integrate the δ function and summing over m_j , the cross section decomposes into a sum of partial cross sections,

$$\frac{d^4\sigma(E_\nu)}{d\omega d\theta_\mu d\Omega_N} = \sum_\alpha (2j_\alpha + 1) \frac{d^4\sigma_\alpha(E_\nu)}{d\omega d\theta_\mu d\Omega_N}, \quad (4)$$

where the sum runs over the shells with occupation $(2j_\alpha + 1)$. Each partial cross section produces a different spectrum of final-state nucleon energies determined by α .

We thus generate samples of the four independent continuous variables $(\theta_\mu, \omega, \theta_N, \phi_N)$ for each shell separately, with E_ν as an input. Among these variables, only the energy transfer ω has a more complex marginal distribution, particularly at low values where nuclear effects play a significant role. This is due to the Hartree-Fock method preserving the shell structure of the nucleus, which in return impacts the cross section. In contrast, the remaining variables are angular and exhibit relatively smooth distributions over well-defined, bounded ranges.

We finally note here that the dependence of the cross section on the azimuthal angle of the hadron plane, as shown in Fig. 1, can always be decomposed into five simple functions [32,33]. For nucleon knockout, the dependence on the nucleon's azimuthal angle can be written as

$$\frac{d^4\sigma_\alpha(E_\nu)}{d\omega d\theta_\mu d\Omega_N} = \frac{d^3\sigma_\alpha(E_\nu)}{d\omega d\theta_\mu d\cos\theta_N} + B\cos\phi_N + C\cos 2\phi_N + D\sin\phi_N + E\sin 2\phi_N, \quad (5)$$

in terms of the angle integrated cross section and the functions B, C, D, E that only depend on $E_\nu, \omega, \theta_\mu, \theta_N$, and the shell. This decomposition can be exploited to sample the ϕ_N dependence [34] and hence reduce the number of nontrivial independent variables. For this work, however, we sample the full kinematics directly from the partial cross sections of Eq. (4).

III. EVENT GENERATOR USING NORMALIZING FLOWS

A. Definition of normalizing flows

Normalizing flows are a class of machine learning models used to transform simple probability distributions into complex ones through a sequence of invertible and differentiable transformations of the probability space. They are particularly useful for efficiently sampling from high-dimensional and non-Gaussian distributions, making them highly suited for generating events from complex interaction models, such as neutrino-nucleus interactions. This contrasts with the accept-reject loop used in most neutrino event generators, which discards a large fraction of generated samples.

The core idea behind NFs is to start with a simple base distribution, typically a multidimensional Gaussian or uniform distribution, and apply a series of transformations that map this distribution into the more complex target distribution of interest. Mathematically, a typical NF transformation consists of a sequence of K diffeomorphisms of the probability space, each parameterized by a neural network, that progressively transform the base distribution $p_Z(z)$ into the target distribution $p_X(x)$. If z is a sample from the base distribution $p_Z(z)$, the final output $x = f_K \circ f_{K-1} \circ \dots \circ f_1(z)$ represents a sample from the target distribution $p_X(x)$.

To compute the likelihood $p_X(x)$, we use the change-of-variable formula,

$$p_X(x) = p_Z(z) \left| \det \frac{\partial f^{-1}}{\partial x} \right|,$$

where $z = f^{-1}(x)$ is the preimage of x under the inverse transformation, and $\det \frac{\partial f^{-1}}{\partial x}$ is the Jacobian determinant of the transformation. One can see the main advantage of NFs for event generation: they allow one to both sample and evaluate the probability of an event at the same time in an efficient and accurate way.

B. Normalizing flows for event generation

NFs have been successfully applied to event generation in high-energy physics, particularly at the Large Hadron Collider (LHC). They have been used as a more efficient alternative to traditional importance sampling algorithms like VEGAS [35] to reduce the variance of an integral estimate. Gao *et al.* [36,37] introduced an NF-based integrator called IFLOW to improve unweighting efficiency in Monte Carlo event generators, using Drell-Yan processes at the LHC as a case study. Bothmann *et al.* [38] proposed a similar NF architecture applied to top-quark pair production and gluon scattering into three- and four-gluon final states. Stienen and Verheyen [39] explored the use of autoregressive flows for efficient generation of particle collider events, performing experiments with leading-order top pair production events at an electron collider and next-to-leading-order top pair production events at the LHC. Further works, such as MadNIS developed by Heimel *et al.* [22,40], use NFs in a more advanced way [41,42]. Collectively, these works demonstrate that NFs significantly enhance unweighting efficiency in LHC event generation. Building upon these advancements, we aim to adapt and extend the use of NFs to neutrino-nucleus interactions, providing an efficient event generator for complex neutrino-nucleus cross sections.

Normalizing flows rely on invertible transformations that allow for efficient application of the change-of-variable formula. Two types of architectures show the best performances: autoregressive flows and coupling layers. Both approaches exploit triangular (or block triangular) Jacobians to ensure computational efficiency, with complexity scaling linearly with the dimensionality of the probability space. However, the choice between autoregressive flows and coupling layer-based flows ultimately depends on the application.

Autoregressive flows such as inverse autoregressive flow [43] or masked autoregressive flow (MAF) [44] are D times slower to invert than to evaluate, where D is the dimension of the probability space. On the other hand, flows based on coupling layers, such as nonlinear independent components estimation [45] or real-valued nonvolume preserving [46], have an analytic one-pass inverse. Sampling and evaluating the density at the same time requires one to evaluate both the forward and inverse transformations. Therefore, coupling layers allow one to both evaluate and sample a density fast. However, coupling layer-based flows are generally less expressive than autoregressive flows. Therefore, we chose to use autoregressive flows due to their higher expressiveness. A more recent comparative study between coupling and autoregressive flows by Cocco *et al.* [47] indicates that autoregressive flows stand out both in terms of accuracy and training speed. We will later also demonstrate that autoregressive flows offer sufficient evaluation and sampling speeds for the dimensionality of typical CCQE events.

In this study, we employ MAF, which use masked feedforward networks to parameterize the NF transformations while enforcing their autoregressive property. The masking mechanism involves applying a binary mask to the weight matrices of the network, effectively setting specific connections to zero. This ensures that the output for a given dimension depends only on its predecessors in the ordering, preserving the conditional dependency property required for autoregressive modeling. A permutation of the dimensions after each autoregressive transformation ensures that all dimensions are treated equally.

C. Circular rational quadratic neural spline flows

In addition to using autoregressive flows to implement our NF, we need to define the parametrization of the flow transformation. In this study, we use rational quadratic neural spline flows (RQ-NSFs) introduced by Durkan *et al.* [48], which effectively models complex, non-Gaussian target distributions by utilizing piecewise rational quadratic transformations, as parametrized by Delbourgo and Gregory [49]. RQ-NSFs are very expressive due to their infinite Taylor-series expansion while being defined by a small number of parameters. RQ-NSFs are widely used in recent applications of NFs for their state-of-the-art expressiveness. They have already been applied in the T2K Collaboration for tasks such as modeling simple cross sections [50] and posterior density estimation for near-detector fits [51].

However, standard RQ-NSFs, defined on Euclidean probability spaces, are suboptimal for modeling periodic variables, such as angular distributions. These distributions often exhibit sharp cut at boundaries when projected onto flat spaces, despite maintaining periodic boundary conditions. To address this, it is essential to model angles within their natural phase spaces—circles for angles and spheres for solid angles. In this work, we adopt the extension of RQ-NSFs to non-Euclidean probability spaces, such as tori and spheres, as developed by Rezende *et al.* [52]. By modifying rational quadratic spline flows to ensure periodicity in certain dimensions, we can appropriately model the periodic distributions through differentiable transformations of the probability space. In our specific $1p1h$ cross section model, the first two kinematics, θ_μ and ω , can be represented on a cylindrical manifold due to the periodicity of θ_μ . The remaining two variables, θ_N and ϕ_N , are naturally represented on a sphere. Therefore, the natural phase space $\mathcal{M}$ where the cross section is defined is a cylinder times a sphere.

D. Energy-dependent flows

In neutrino–nucleus interactions, the differential cross section depends on both the kinematic variables and the incoming neutrino energy E_ν . Neutrino event generators, such as NEUT [4], typically take E_ν as an input parameter

and generate samples weighted by the conditional cross section given E_ν , rather than modeling E_ν as an extra dimension. We condition the NFs directly on E_ν to mimic the functioning.

We model a family of probability distributions continuously parameterized by E_ν . This approach ensures that the NFs capture the underlying dependence of the cross section on energy without explicitly including E_ν as part of the flow’s dimensional space. Additionally, this eliminates the need to retrain the NFs for different values of E_ν .

For the specific case of $1p1h$ interactions, the goal is to model the conditional probability

$$p(\theta_\mu, \omega, \theta_N, \phi_N | E_\nu, \alpha),$$

over a wide range of E_ν . During the transformation, the flow modifies the kinematic variables $(\theta_\mu, \omega, \theta_N, \phi_N)$ in a way that captures the energy dependence implicitly. Here, we model each initial shell α separately.

The conditional probability is computed by normalizing the full 4D differential cross section for a given shell α by the total cross section,

$$p(\theta_\mu, \omega, \theta_N, \phi_N | E_\nu, \alpha) = \frac{d^4\sigma_\alpha(E_\nu)}{d\omega d\theta_\mu d\theta_N d\phi_N} \cdot \frac{1}{\sigma_\alpha(E_\nu)}.$$

To model $\sigma_\alpha(E_\nu)$, a standard approach is to integrate the cross section over the other variables and fit the resulting function using a polynomial approximation,

$$\sigma_\alpha(E_\nu) \approx P_\alpha(E_\nu).$$

E. Exhaustive training of the flow-based event generator

To train NF, we typically optimize the model by minimizing the Kullback-Leibler divergence (KL-D) between the true distribution p and the learned distribution q_θ (where θ refers to the neural net parametrization of the surrogate distribution). The KL-D measures the difference between two probability distributions, but it is inherently asymmetric, which means it leads to different behaviors depending on whether we minimize the forward or reverse KL-D. Minimizing the forward KL-D, expressed as

$$D_{KL}(p||q_\theta) = \mathbb{E}_{x \sim p} \left[\log \frac{p(x)}{q_\theta(x)} \right],$$

can lead to mean-seeking behavior, where the learned distribution q_θ covers the regions where the true distribution p is nonzero but may overestimate the distribution tails. Conversely, minimizing the reverse KL-D, expressed as

$$D_{KL}(q_\theta||p) = \mathbb{E}_{x \sim q_\theta} \left[\log \frac{q_\theta(x)}{p(x)} \right],$$

can result in mode-seeking behavior, where q_θ focuses on some peaks in the probability but may ignore other regions or even other peaks.

To balance these tendencies, we use the symmetric KL-D, which is the average of these two,

$$D_s(p, q_\theta) = \frac{1}{2} (D_{KL}(p \| q_\theta) + D_{KL}(q_\theta \| p)).$$

However, direct computation of the KL-D is impractical for our use case, as it would involve sampling from the true distribution p , which is computationally and time expensive. Furthermore, sampling from the learned distribution q_θ to compute the loss function introduces instability. This instability arises because, in the early training stages, the learned distribution may deviate significantly from the target, leading to unreliable loss estimates.

A significant improvement in our training process was achieved through the application of importance sampling to compute the loss function [53]. Importance sampling employs a proposal distribution that is simple to sample from yet effectively spans the target distribution p . In our case, we used a uniform distribution u defined over the manifold $\mathcal{M}$, as defined in Sec. III C, and sampled energy values uniformly within the chosen training range. This enabled us to compute a Monte Carlo estimate of the symmetric KL-D,

$$D_s(p, q_\theta) \approx \frac{1}{Z} \sum_{i,j} \left[(p(x_i | E_j) - q_\theta(x_i | E_j)) \log \frac{p(x_i | E_j)}{q_\theta(x_i | E_j)} \right],$$

where Z is a normalization constant, and x_i and E_j are sampled uniformly from $\mathcal{M}$ and the training energy range, respectively. This approach ensures robust and stable training by exhaustively covering the manifold, preventing the learned distribution q_θ from collapsing into narrow regions or missing areas.

To contextualize, a related approach was proposed by Pina-Otey *et al.* [50] in their exhaustive neural importance sampling framework. They employed a two-stage training process: an initial “warm-up” phase, during which the training points were uniformly sampled for 20% of the total training time, followed by direct sampling from the learned distribution to compute the loss function. The first stage allowed exhaustive exploration of the phase space, while the second stage accelerated convergence by focusing on regions where the model had already predicted probability density. Although the two-stage strategy is appealing, our results indicate that maintaining uniform sampling throughout the training process is not only simpler but also sufficiently fast for the dimensionalities involved in the exclusive $1p1h$ processes. Furthermore, in scenarios where cross section calculations are computationally expensive, this approach becomes impractical, as it would require evaluating the cross section repeatedly during training

for new samples of the predicted probability density. Exhaustivity remains a critical factor in cross section modeling, where the primary goal is to develop the most accurate and comprehensive surrogate for the cross section, rather than merely optimizing training time. However, for higher-dimensional processes like two-particle-two-hole ($2p2h$) interactions, uniformly sampling the kinematics may become infeasible. In such cases, an alternative approach could involve training the model using samples drawn from an already trained model for the semiexclusive cross section, where the kinematics of the subleading nucleon are marginalized. This semiexclusive distribution could then serve as a proposal distribution for training a normalizing flow to model the fully exclusive cross section.

IV. PERFORMANCES OF THE FLOW-BASED EVENT GENERATOR

A. Detail of implementation and training

We trained a flow-based Monte Carlo event generator for the CCQE $1p1h$ interaction between a neutrino and a carbon nucleus (^{12}C). The carbon nucleus consists of two shells, $1s_{1/2}$ and $1p_{3/2}$, with respective separation energies of 36.20 and 18.72 MeV. To enhance the flexibility of our event generation, we chose to model the interactions for each shell separately. This approach serves theoretical purposes, enabling the study of interactions with individual shells. A similar work can be applied to any nuclei, which would involve modeling more or fewer shells.

Additionally, we trained four separate models per shell, each corresponding to overlapping neutrino energy ranges in MeV: [180, 350], [300, 500], [450, 750], and [700, 1050]. Consequently, the event generator covers neutrino energies from 180 to 1050 MeV. The overlapping regions between these energy ranges ensure smooth transitions between models. In these overlap regions, we blend the two overlapping models by sampling from and evaluating the second model with a probability that increases linearly from 0 at the start of the overlap range to 1 at the end. ^{12}C having two shells, the total number of models is therefore eight.

Concerning the implementation of a single model, we used an implementation of autoregressive RQ-NSFs based on the NFLOWS PyTorch implementation of Durkan *et al.* [48]. Some modifications were made to accommodate the computation of the Symmetric Kullback-Leibler divergence using importance sampling.

We trained three types of models, each differing in size: a small model (S), a medium model (M), and a large model (L). All three models utilize nine bins in their spline transformations. Each RQ-NSF is parameterized by a masked autoregressive network composed of three hidden layers, with a residual connection from the first to the third layer. Model S includes 256 hidden nodes per hidden layer and 10 RQ-NSFs; model M uses 512 hidden nodes and

TABLE I. Comparison of model performance for different model sizes (S, M, L). The size is here given for the eight models (as discussed in Sec. IV A).

Model	S	M	L
Size (MB)	200	719	1797
Number of flows	10	10	25
Number of hidden nodes	256	512	512
Training time (h)	5	6	13
CPU speed (sec/ 1×10^6 samples)	141.8	235.6	669.3
GPU speed (sec/ 1×10^6 samples)	12.5	23.6	70.7
RESS, flat flux (%)	98.48	98.64	98.87
RESS, T2K flux (%)	98.48	98.64	98.85
rms _z ($1s_{1/2}$)	1.36	1.01	0.75
rms _z ($1p_{3/2}$)	1.86	1.52	0.88

10 RQ-NSFs; model L employs 512 hidden nodes and 25 RQ-NSFs. The performance of these three models is evaluated in Table I. The base distribution is a four-dimensional probability density with independent dimensions. The three angular dimensions are modeled as uniform distributions, while the energy transfer dimension follows a normal distribution.

We used a relatively low learning rate of 0.0001 with a cosine annealing scheduler with a minimum learning rate of 5×10^{-7} . We trained each model for 100,000 epochs. Each epoch computes the symmetric KL-D on a batch of 2^{14} samples with associated true cross section. A total of 10×10^6 samples were used to train each model, with each sample appearing multiple times throughout the training.

These samples and their cross sections were pregenerated prior to training—a potentially time-intensive step depending on the cross section computation speed, but required only once. The θ_μ and θ_N distributions are unfolded to $[-\pi, \pi]$ to ensure 2π periodicity. The training time for each model (on an NVIDIA RTX-4090 GPU) are given in Table I.

B. Comparison of the true cross section and its surrogate projecting in 1D and 2D spaces

In Monte Carlo event generators widely used in neutrino physics, such as NEUT, users can freely select any distribution of neutrino energy. As a result, our model must accurately represent all possible combinations of energy and, in our case, the interacting nucleon's shell. While it would be impractical to visually compare the true and surrogate cross sections for every such combination, Secs. IV C and IV D provide a more quantitative and exhaustive evaluation of the surrogate cross section's performance. This section offers a more visual and qualitative evaluation for a specific energy and shell combination.

Figure 2 illustrates the comparison between the true cross section and its surrogate modeled using normalizing flows for a fixed neutrino energy of $E_\nu = 200$ MeV, with the initial bound nucleon in the $1p_{3/2}$ shell. This particular configuration was chosen because nuclear effects have a significant impact on its cross section. At low-energy transfer, the cross section is especially sensitive to shell

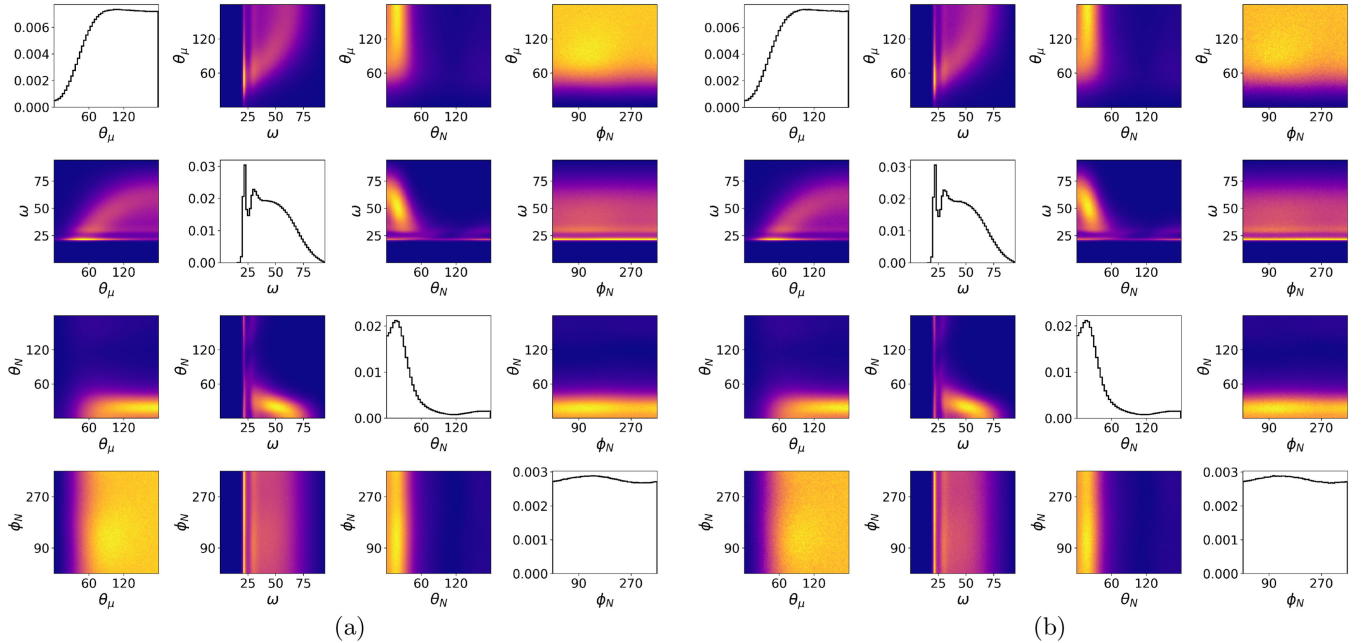


FIG. 2. 100×10^6 samples with weights from the true cross section (a) and 10×10^6 weighted samples from the predicted surrogate cross section (b) for a neutrino energy of $E_\nu = 200$ MeV and the $1p_{3/2}$ shell. The diagonal plots represent the marginal distribution of the four kinematic variables ($\theta_\mu, \omega, \theta_N, \phi_N$). The off-diagonal plots correspond to the 2D histograms of two different kinematic variables.

structure, a key feature captured by the Hartree-Fock mean-field approach. The angular distribution of the outgoing nucleon relative to the transferred momentum, θ_N , is strongly influenced by the properties of the initial bound state. Nevertheless, the NFs seem to model accurately this distribution from the shapes of the four marginal distributions to the correlations between two kinematic variables.

C. Comparison of the true cross section and its surrogate using density estimation

1. Point-by-point density comparison

To compare two multidimensional distributions, one often simplifies the problem by projecting the densities onto one or two dimensions at a time and/or by overlapping their binned histograms. These approaches, while visual, come with inherent limitations. By reducing the dimensionality, dependencies that might exist in higher-dimensional spaces are inevitably smoothed out. Similarly, binning locally averages out the densities and therefore can obscure fine-scale variations, introducing artificial agreement in regions where the true distributions may differ subtly.

Therefore, one needs a more holistic approach to comparing multidimensional distributions. In this work, we evaluate directly the weight, which is defined as the ratio between the true distribution p and its normalizing flows surrogate q_θ (where θ refers to the neural net parametrization of the surrogate distribution),

$$w(\theta_\mu, \omega, \theta_N, \phi_N, E_\nu, \alpha) = \frac{p(\theta_\mu, \omega, \theta_N, \phi_N | E_\nu, \alpha)}{q_\theta(\theta_\mu, \omega, \theta_N, \phi_N | E_\nu, \alpha)},$$

where α refers to the bound nucleon shell. This is the error one expects in a sample if the NF surrogate was used in its simplest form, and also the weight one would want to apply to an event generator sample to correct the imperfect surrogate to be closer to the true distribution.

The histogram of weights for the L model along with the histogram of weights for a uniform distribution defined on the manifold $\mathcal{M}$ are shown in Fig. 3. In the case of the surrogate cross section, the weights are tightly concentrated around 1, with a measured standard deviation of 10.7% and a 99.99th percentile value of 1.90. This represents a significant advantage when compared to a uniform distribution, where the weights span 13 orders of magnitude, leading to a large proportion of samples contributing minimally or not at all.

From these weights, we derive the relative effective sample size (RESS), which provides a quantitative measure of the quality of the surrogate distribution q_θ in approximating p . The RESS quantifies the proportion of effectively independent samples represented by the surrogate distribution when weighted by the importance weights. It is defined as

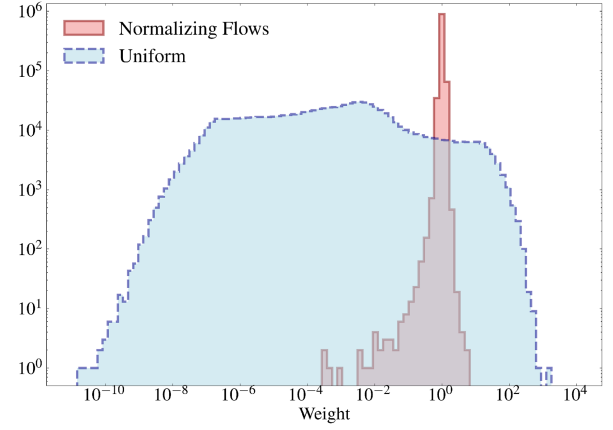


FIG. 3. Log-scale histogram of 1×10^6 weights for both shells of ^{12}C under a flat neutrino energy flux from 200 to 1000 MeV, generated using the surrogate cross section (red line) and a uniform distribution (blue dashed line) defined on the manifold $\mathcal{M}$.

$$\text{RESS} = \frac{1}{N} \frac{\left(\sum_{i=1}^N w_i \right)^2}{\sum_{i=1}^N w_i^2},$$

where w_i are the weights for the samples $i = 1, \dots, N$.

A RESS close to 1 indicates that the weights w_i are similar in magnitude, reflecting that the surrogate distribution q_θ closely approximates the true distribution p . In contrast, a RESS close to 0 suggests that the weights are spread out, implying a significant mismatch between q_θ and p . The performance of the models is evaluated using 1×10^6 samples drawn from q_θ across the full neutrino energy range [200, 1000] MeV and for both shells combined. The RESS values for the three model sizes are presented in Table I. These values are computed for two of neutrino fluxes: one with a flat distribution and another based on the truncated T2K near-detector flux [54].

2. Density comparison with finite resolution

Given that the RESS value obtained is slightly below 1, it is necessary to assess the potential impact of the observed weight variation. A central question is whether the largest deviations are concentrated in specific regions of the phase space, which could introduce systematic biases in the generated sample exceeding current experimental uncertainties. Alternatively, these variations may stem from localized fluctuations which, when integrated over sufficiently small kinematic bins, effectively average out and thus remain negligible for experimental observations.

To clarify this point, the true and surrogate cross sections are both evaluated on a grid for a monoenergetic neutrino beam at 650 MeV, summing both shell contributions. The grid spans different muon scattering angles with a step of 10° and outgoing proton angles, specifically, $\theta_N \in 1^\circ, 5^\circ, 10^\circ, 20^\circ, 30^\circ, 40^\circ, 50^\circ, 60^\circ$ with a fixed nucleon

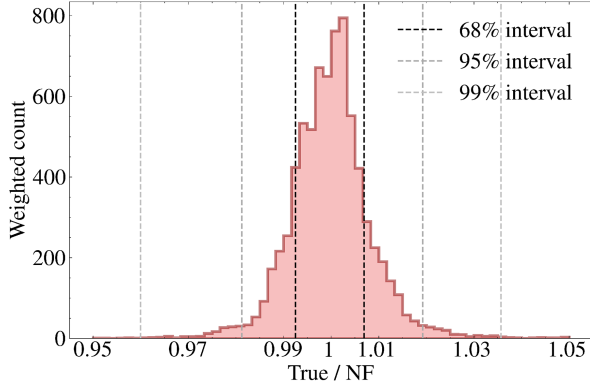


FIG. 4. Histogram of the ratio between true and surrogate cross sections weighted by the true cross section evaluated at a neutrino energy of 650 MeV across a fine kinematic grid and after local averaging. The dashed lines represent the confidence intervals: 68% CI = $(-0.75\%, +0.69\%)$, 95% CI = $(-1.88\%, +1.92\%)$, and 99% CI = $(-4.00\%, +3.57\%)$.

azimuthal angle of 180° . The energy transfer ω is sampled with a step of 2.5 MeV across its defined range. Importantly, the cross section values are averaged within localized four-dimensional kinematic bins centered at each point $(\omega, \theta_\mu, \theta_N, \phi_N)$. The resolution of each bin is 1 MeV in energy transfer and 1° in angular coordinates, which is significantly finer than what current experiments can resolve. Every bin includes 500 uniformly distributed samples, ensuring negligible statistical fluctuations in the averaged bin value.

Figure 4 presents the histogram of weights after averaging, of the given sampled points weighted by the true cross section values. The dashed lines represent the 68%, 95%, and 99% confidence intervals. The Appendix further illustrates the impact of these cross section deviations on the distribution of proton kinetic energy T_N at fixed kinematics, both with and without local averaging.

To quantify the residual discrepancies between the surrogate and true cross sections after local binning, we introduce the cross-section-weighted root mean squared error (rms_p), defined as

$$\text{rms}_p = \sqrt{\frac{\sum_{\mathbf{x}_i} p(\mathbf{x}_i) \Delta^2(\mathbf{x}_i)}{\sum_{\mathbf{x}_i} p(\mathbf{x}_i)}},$$

where

- (i) $\mathbf{x}_i$ denotes a point in the kinematic space;
- (ii) $\Delta(\mathbf{x}_i)$ is the relative difference between the surrogate and true cross section at point $\mathbf{x}_i$ after averaging; and
- (iii) $p(\mathbf{x}_i)$ is the true cross section at $\mathbf{x}_i$ after averaging.

We obtain an rms_p of 1.4%. This demonstrates that the local averaging procedure drastically reduces the modeling error compared to the initial spread of weights of 10.7%. The bias introduced by normalizing flows is therefore well below our experimental uncertainties. Furthermore, in

realistic experimental conditions, where measurements involve a convolution over a broad neutrino energy spectrum, such residual differences would be further smeared out, rendering them more negligible in practice.

D. Comparison between the true cross section and its surrogate through their respective datasets

In neutrino experiments such as T2K, we cannot achieve exact estimates of the kinematics we measure or infer. As a result, evaluating a model based on a direct point-by-point comparison of its distribution with the true one, as discussed in Sec. IV C 2, may be overly conservative. Instead, the goal in this work is to construct an event generator capable of populating the kinematic phase space in a manner consistent with a nonbiased accept-reject method. This must be achieved with fixed precision while accounting for the full dimensionality of the phase space.

Therefore, we compare four-dimensional binned histograms derived from

- (1) samples generated by the “surrogate cross section,” and
- (2) samples generated by an “accept-reject algorithm” with a uniform proposal distribution defined on the manifold $\mathcal{M}$.

We assess how closely these histograms align with a high-statistics, nonbiased “reference histogram” by computing the multinomial negative log-likelihood (MNLL),

$$\text{MNLL}(H^{\text{MC}}) = \frac{1}{N} \sum_{\text{bin}} \left[\log(H_{\text{bin}}^{\text{MC}}) - H_{\text{bin}}^{\text{MC}} \log\left(\frac{H_{\text{bin}}^{\text{ref}}}{N}\right) \right],$$

where the summation runs over all bins in the histogram and

- (i) H^{MC} is the histogram from the Monte Carlo dataset (either surrogate or accept-reject);
- (ii) H^{ref} is the high-statistics, nonbiased reference histogram rescaled to match the number of samples in H^{MC} ; and
- (iii) $N = \sum_{\text{bin}} H_{\text{bin}}^{\text{MC}}$ is the total number of observed samples.

To obtain statistical distributions of the MNLL for both the surrogate model and the accept-reject method, we repeat the procedure 100 times per method. Figure 5 illustrates how the MNLL evolves with neutrino energy for both approaches. In this context, the unbiased accept-reject MNLL serves as the optimal performance benchmark for the surrogate cross section.

This comparison is carried out at fixed neutrino energies of 200, 300, 400, 500, 600, 700, 800, 900, and 1000 MeV, considering the $1s_{1/2}$ and $1p_{3/2}$ shells separately. For each energy and shell, θ_μ and θ_N are binned with a width of 10° , while ϕ_N has a bin width of 30° . For the energy transfer, the bin width is 10 MeV from the separation energy up to 100 MeV, and then 50 MeV until reaching the maximum

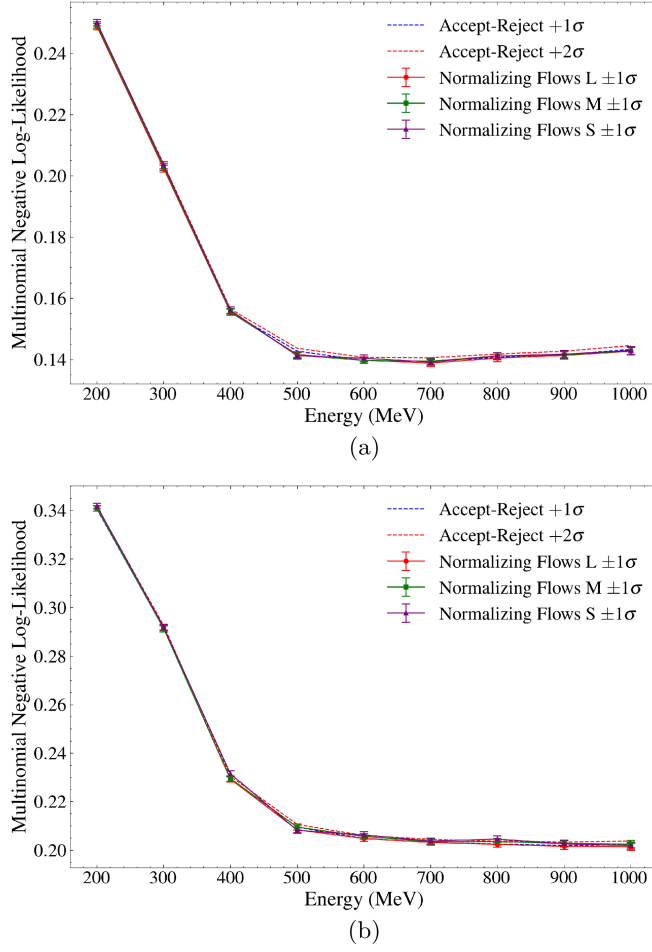


FIG. 5. Evolution of the multinomial negative log-likelihood with the neutrino energy for the $1s_{1/2}$ (a) and $1p_{3/2}$ (b) shell. Each plot shows both the MNLL for the three normalizing flow models' size. The $+1\sigma$ and $+2\sigma$ upper bounds of the MNLL using the accept-reject datasets are given for comparison.

energy transfer of $E_\nu - m_\mu$, where E_ν is the neutrino energy and $m_\mu = 105.66$ MeV is the muon rest mass. We choose a finer binning at lower-energy transfer to capture the complex structure inherited from the shell modeling. Each histogram H^{MC} uses 100,000 samples, while the reference histogram is built from 10×10^6 weighted samples.

To provide a measure of how the surrogate's MNLL values differ from those of the accept-reject method in average, we compute the rms of the Z score (rms_z) across all energies for a given shell,

$$\text{rms}_z = \sqrt{\frac{1}{N_e} \sum_{i=1}^{N_e} \left(\frac{\text{MNLL}(H_{E_\nu^{(i)}}^{\text{NF}}) - \text{MNLL}(H_{E_\nu^{(i)}}^{\text{AR}})}{\sigma_{E_\nu^{(i)}}^{\text{AR}}} \right)^2}.$$

where $\overline{\text{MNLL}(H_{E_\nu^{(i)}}^{\text{NF}})}$ is the mean MNLL for normalizing flow datasets at energy $E_\nu^{(i)}$, $\overline{\text{MNLL}(H_{E_\nu^{(i)}}^{\text{AR}})}$ is the mean

MNLL for accept-reject datasets at the same energy, and $\sigma_{E_\nu^{(i)}}^{\text{AR}}$ is the standard deviation of the accept-reject MNLL distribution at energy $E_\nu^{(i)}$. The rms_z for the three model sizes are given in Table I. The MNLL distributions for the surrogate and accept-reject histograms are statistically compatible within their $\text{rms}_z = 1$ range for the L model and within $\text{rms}_z = 2$ for the M and S models. In the case of the model size L, we find an average rms_z of 0.75σ for the $1s_{1/2}$ shell and 0.88σ for the $1p_{3/2}$ shell. In other words, two datasets of 100,000 samples, respectively, generated by accept-reject and by the normalizing flow surrogate are practically equivalent given the binning that we chose. Overall, these results demonstrate that the normalizing flow surrogate reproduces the cross section with high fidelity in *all* tested combinations of shell and energy and in the full kinematic space, while significantly reducing computational overhead compared to the accept-reject method.

E. Overall performance

The performances for the three model sizes are summarized in Table I. While accuracy improves slightly with increasing model size, this comes at the expense of reduced computational efficiency and larger model sizes. The sampling speeds are here given for a single neutrino energy and shell combination. As noted earlier, we prioritized modeling accuracy over sampling speed by not selecting the fastest RQ-NSF implementation. Nonetheless, even the slowest model, L, achieves a sampling rate of 10×10^6 samples in under 12 min on an NVIDIA RTX-4090 GPU and in 1 h 52 min on a single Intel Xeon CPU core, far surpassing the requirement of 10×10^6 samples per day. This encourages us to consider this implementation as a viable approach for higher-dimensional processes, such as $2p2h$ or charged-current single-pion production.

The results show that models with larger sizes provide higher accuracy, as indicated by a steady rise in RESS values from 98.48% (S) to 98.87% (L). However, this improvement is relatively small in terms of unweighting efficiency in the case of important sampling, suggesting that the smallest model might be sufficient when sampling speed or memory constraints outweigh minor accuracy gains in the proposal distribution. This trade-off between sampling speed and accuracy depends on how the surrogate cross section is utilized.

There are three ways to utilize the surrogate cross section:

- (1) Use the surrogate as a proposal distribution and apply an accept-reject algorithm to resample its outputs.
- (2) Sample from the surrogate and provide per-event weights by calling the true cross section once for each sampled event.
- (3) Sample directly from the surrogate without any further resampling. This option is by far the fastest

for cross sections that are computationally heavy. However, it is important to mention that no generative model can replace a true cross section without introducing bias. Therefore, for a strictly unbiased analysis, one should always rely on the importance weight. As mentioned in Sec. IV C, this is particularly crucial for studies focused on regions of very low cross section (e.g., low proton kinetic energy or backward-going protons with respect to the momentum transfer) or when very fine resolution is needed, such as angular resolutions better than 1° or energy resolutions finer than 1 MeV.

Nevertheless, using the surrogate cross section for accept-reject sampling can be beneficial in some situations. In cases where the true cross section can be evaluated quickly, it may be most efficient to adopt the smallest model (S) and, if necessary, reweight its samples with an accept-reject algorithm. However, if evaluating the true cross section is time consuming, the larger (L) model's higher accuracy can justify its additional computational and storage costs.

Moreover, our normalizing flow approach to build a proposal distribution already achieves greater accuracy than the previous normalizing flows applied to neutrino-nucleus cross section modeling: Pina-Otey *et al.* [50] reported a RESS of 91.40%, which is significantly lower than even our smallest model (S). The concentration of weights around 1 as shown in Fig. 3 highlights the efficiency of the surrogate cross section as a proposal distribution for an accept-reject algorithm. The low spread of weights ensures that most samples contribute effectively to the estimation of observables, minimizing the variance introduced by the importance weights. If the weight threshold is set at 1.90, corresponding to the 99.99th percentile of weights then only 1 in 10,000 samples surpasses this limit. The expected acceptance rate for the surrogate cross section with this weight cap, when used as a proposal distribution, is approximately $\frac{1}{1.90} \approx 53\%$.

Although the surrogate can serve as an efficient proposal distribution, Sec. IV D shows that a dataset of events sampled directly from the normalizing flow surrogate is practically equivalent to an unbiased accept-reject dataset at our target precision. Section IV C and the Appendix further show that, once experimental conditions are taken into account, particularly a reasonable resolution in observables, the resulting uncertainties are already at the percent level for a monoenergetic neutrino beam. This residual error would be further reduced after convolution with a realistic neutrino flux, due to the additional smearing over neutrino energies. Importantly, the remaining modeling uncertainty is already well below the typical systematic uncertainties associated with experimental measurements, making the normalizing flows' surrogate sufficiently accurate for practical analyses. This is particularly advantageous for complex, theory-driven cross sections requiring lengthy

evaluations, since normalizing flows can reduce total sampling time from several days on a CPU (via accept-reject) to a few minutes on a GPU or CPU.

V. DISCUSSION

The aim of this study was to demonstrate the ability to efficiently and accurately generate $1p1h$ samples without relying on the full complexity of the Ghent model, which also accounts for two-body currents and the continuum random phase approximation (CRPA) [13]. These corrections are not expected to pose major challenges for the proposed method. Two-body currents, involving two-particle states, provide corrections to the impulse approximation, correcting the single nucleon knockout processes and giving rise to double nucleon knockout processes. The effect of two-body currents on the $1p1h$ cross section is relatively small and can be accounted for by fine-tuning a model already trained within the impulse approximation.

Similarly, although including more detailed physics features in the cross section, from a numerical point of view the inclusion of CRPA correlations is expected to simplify the energy transfer distribution by broadening the sharp features at low energy, smoothing out the δ -like structures associated with the missing energy from shell modeling, hence facilitating event generation using normalizing flows.

The greater challenge arises in describing $2p2h$ processes due to the increased number of dimensions in the phase space. These processes correspond to two-nucleon knockout, and their fully exclusive differential cross sections involve nine independent kinematic variables,

$$\frac{d^9\sigma(E_\nu)}{d\omega d\Omega_\mu d\Omega_{N_1} dE_{N_1} d\Omega_{N_2} dE_{N_2}} \propto L_{\mu\nu} W^{\mu\nu}.$$

As in the $1p1h$ case, we treat the neutrino energy as a conditional variable, the differential cross section does not depend on the muon's azimuthal angle ϕ , and we incoherently sum over the energy eigenstates of the target nucleons. Consequently, we must handle seven independent kinematic variables, three more than in the $1p1h$ scenario. The increased dimensionality should not affect the expressiveness of autoregressive normalizing flows as shown in [47]. However, a larger phase space requires more training samples, and this challenge is made even more difficult by the longer computation time needed for the $2p2h$ cross section.

In this context, our work on $1p1h$ modeling provides a solid foundation for extending the approach to semiexclusive cross sections, where the kinematics of the subleading nucleon are integrated out. Training a model on semiexclusive cross sections captures key features of the $2p2h$ process and allows for a more efficient training of fully exclusive $2p2h$ cross sections. Instead of using uniformly distributed samples, as done in this work, the training can

be performed using samples drawn from the surrogate semiexclusive cross section, reducing the number of required evaluations and improving training efficiency.

VI. CONCLUSION

In this work, we presented a normalizing-flow-based surrogate for generating neutrino-nucleus scattering events with high accuracy and efficiency. By comparing our surrogate’s output to that of an unbiased accept-reject method, we demonstrated that the two methods produce practically equivalent samples across a wide range of neutrino energies and nuclear shells. Even more importantly, we showed that the normalizing flow approach can generate millions of events in minutes on modern GPUs.

From a theoretical standpoint, such surrogates enable systematic studies of model variations and uncertainty quantification, giving insights into how different theoretical inputs manifest across multidimensional phase space.

Looking ahead, the flexibility of normalizing flows makes them well suited to other processes beyond the $1p1h$ process studied here. In particular, higher-dimensional processes such as $2p2h$ or single-pion production can benefit from a similar surrogate approach. Currently, neutrino experiments such as T2K rely on so-called “Frankenstein” models that are combinations of submodels drawn from different theoretical frameworks, each tuned to specific interaction channels and combined using *ad hoc* relative proportions. While practical, this approach makes it difficult to disentangle whether observed features in predictions arise from genuine physical effects, differences in underlying frameworks, or purely phenomenological adjustments. In contrast, a normalizing-flow-based surrogate can make a unified sampling scheme for *all* processes within a single theory-grounded cross section model possible. It could do so even in the case of a mean-field model, such as that developed by the Ghent group, which has been before too computationally intensive to integrate directly into event generators in its full complexity. Such a tool could be crucial for future neutrino experiments, delivering fast, accurate, and cohesive event generation that seamlessly bridges the gap between the experimental and theoretical frontiers.

ACKNOWLEDGMENTS

M. E. thanks Joshua Isaacson for the valuable discussion. M. E. acknowledges support from the Swiss National Foundation Grants No. 200021E_213196. A. N. is supported by the Neutrino Theory Network under Award No. DEAC02-07CH11359, the Department of Physics and Nuclear Theory Group at the University of Washington. K. N. acknowledges support from the Fund for Scientific Research Flanders (FWO) and Ghent University Special Research Fund.

DATA AVAILABILITY

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

APPENDIX: PROFILE OF THE PROTON KINETIC ENERGY FOR DIFFERENT ANGLES OF THE OUTGOING MUON AND PROTON

To complement the analysis presented in Sec. IV C 2, this appendix illustrates the impact of the biases between the true and surrogate cross sections on the distribution of proton kinetic energy T_N at fixed kinematics, both with and without local averaging. The study is performed using a monoenergetic neutrino beam with energy $E_\nu = 650$ MeV, and the total cross section is computed by summing the contributions from the two occupied shells, as described in Eq. (4). In the case of CCQE interaction in the impulse approximation, one can simply derive the kinetic energy by computing

$$T_N = \omega - E_b,$$

where E_b is the shell-dependent bounding energy.

Figures 6–8 show the comparisons between true and surrogate cross sections for $\theta_N = 5^\circ, 20^\circ$, and 50° , respectively. In each figure, the top panel displays the differential cross section as a function of T_N for various values of θ_μ , comparing the NF surrogate to the true differential cross section. The corresponding relative difference Δ is also shown. In the bottom panel, the same comparison is shown after averaging the cross section over a local 4D kinematic cube centered at the same point in $(T_N, \theta_\mu, \theta_N, \phi_N)$ space, using a bin size of 1 MeV in T_N and 1° in each angular variable.

The top panels reveal small but noticeable differences between the true cross section and the prediction from the normalizing flows, both in the tail and in the core of the distribution. However, most of these differences, particularly in the core, are significantly reduced when averaging over a small neighborhood in kinematic space as illustrated in the bottom plots. This supports the conclusion that the remaining differences primarily reflect localized fluctuations rather than global modeling errors. Smaller deviations persist after averaging particularly in regions of low cross section. For analyses that probe regions of very low cross section (e.g., backward-going protons with respect to the momentum transfer or low T_N) or require subdegree angular precision or sub-MeV energy resolution, the surrogate model may introduce non-negligible bias, making it necessary to reweight the events using the true cross section.

At the same time, the surrogate retains distinctive kinematic patterns predicted by the Ghent model, inherited from the shell modeling and the distorted-wave approach.

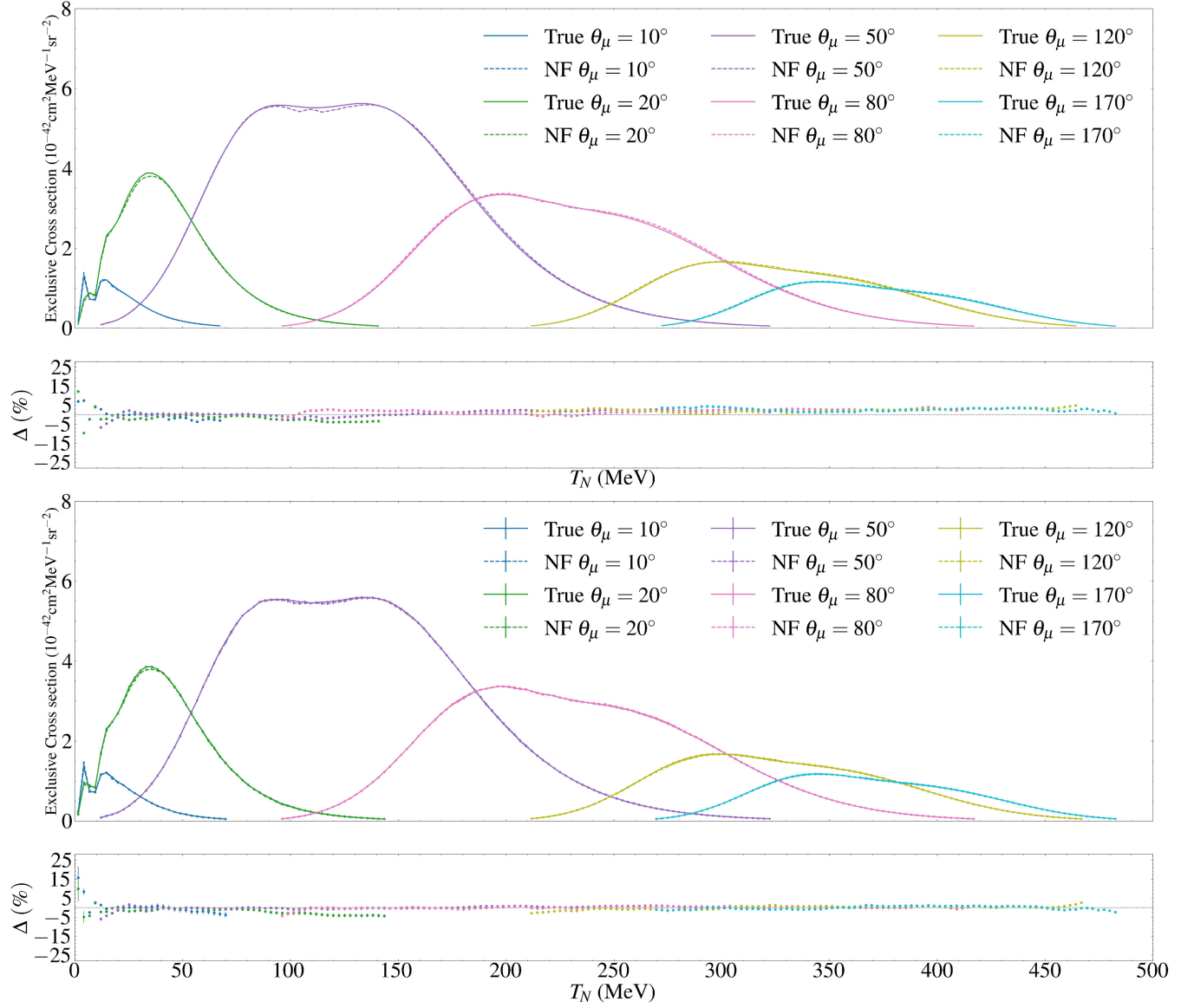
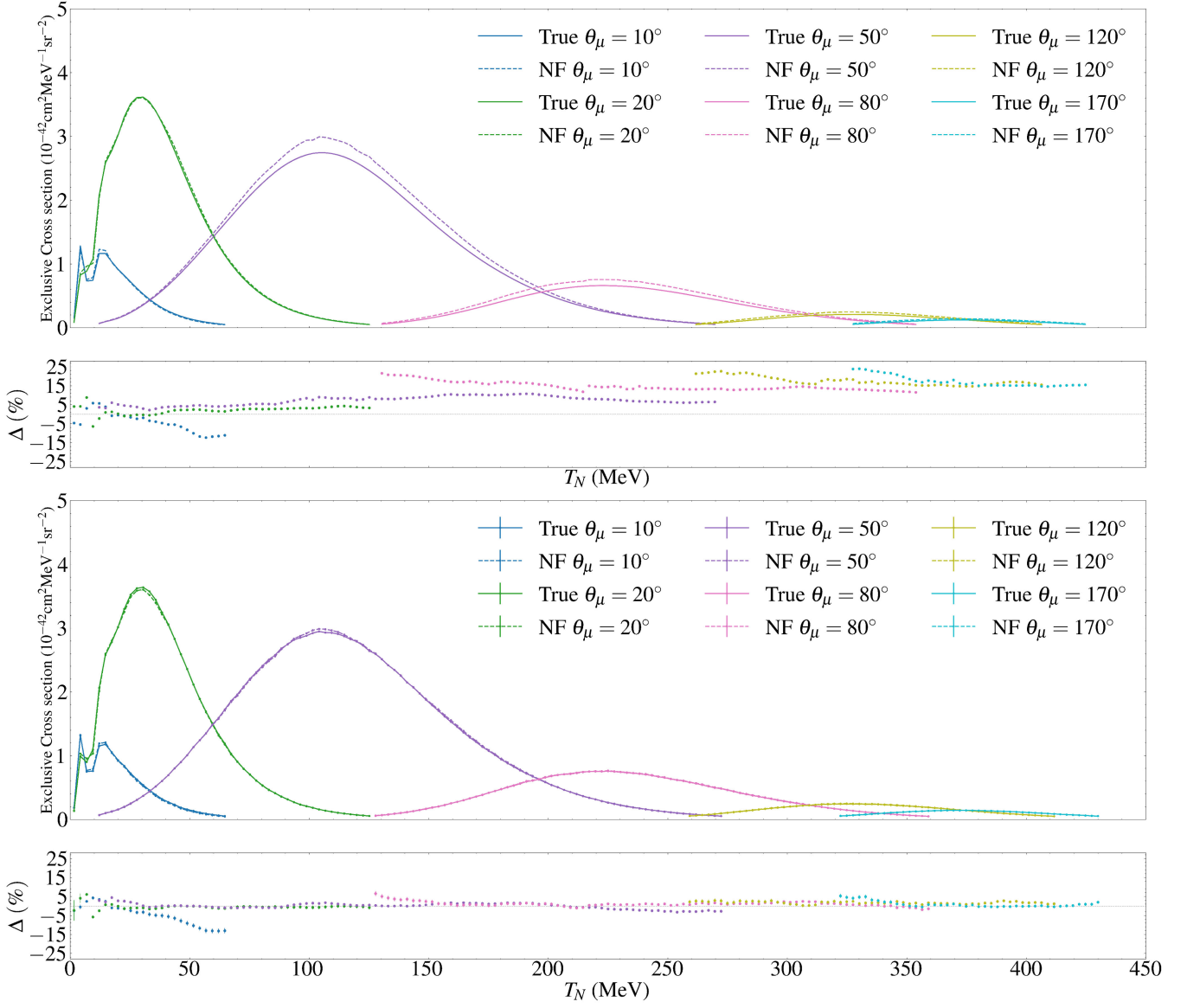
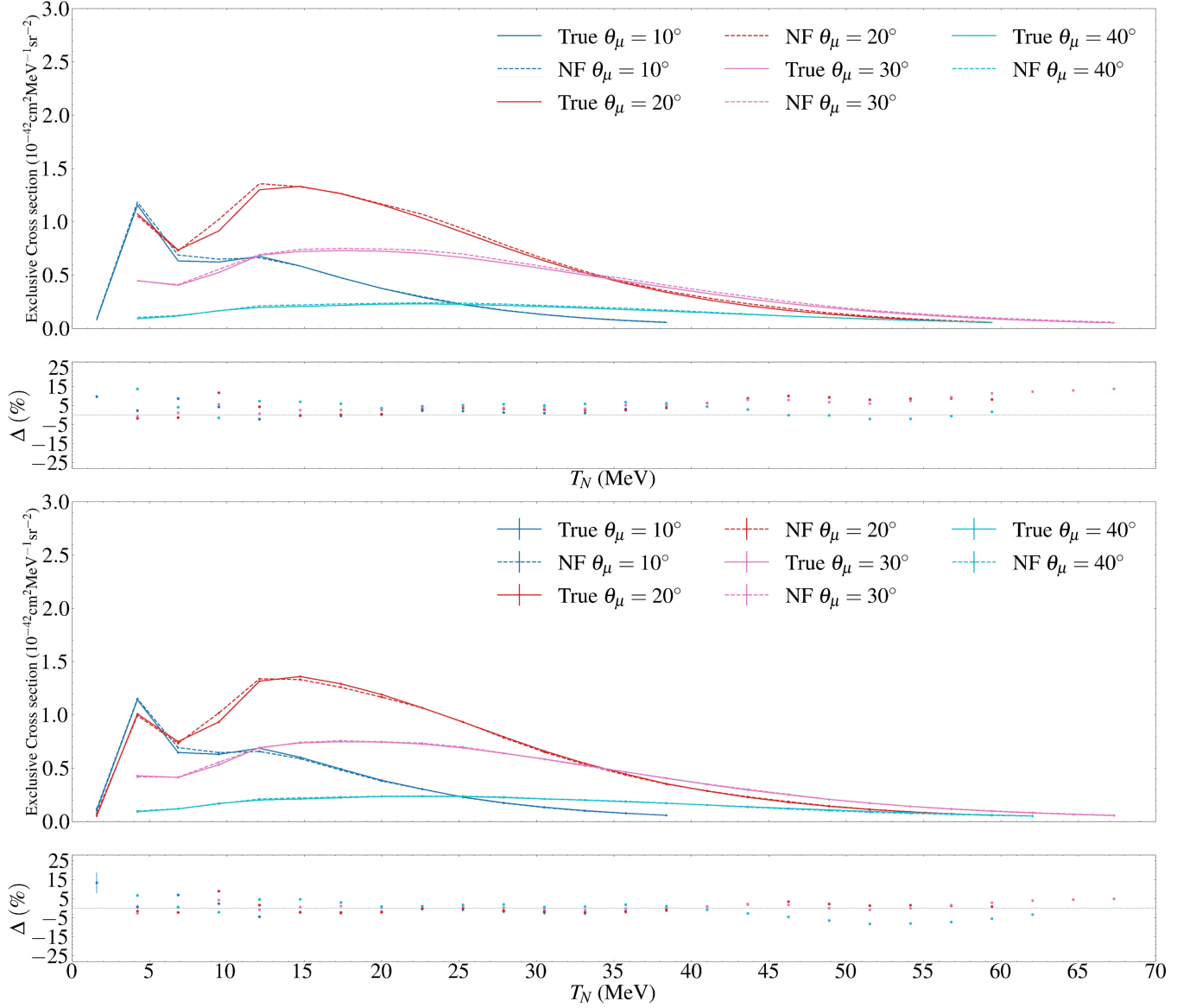


FIG. 6. Profile of the proton kinetic energy distribution T_N for fixed values of the muon scattering angle $\theta_\mu \in \{10^\circ, 20^\circ, 50^\circ, 80^\circ, 120^\circ, 170^\circ\}$ and for fixed outgoing nucleon angles $(\theta_N, \phi_N) = (5^\circ, 180^\circ)$. The cross section is computed for a monoenergetic neutrino beam of 650 MeV and for both shells combined. Top: the differential cross section as a function of T_N for each θ_μ value, comparing the output of the NF surrogate to the true model. The NF prediction is shown with dashed lines, and the true model with solid lines. The relative difference Δ between the two curves is also shown. Bottom: the same comparison is shown after averaging the cross section over a local 4D kinematic cube centered at the same $(T_N, \theta_\mu, \theta_N, \phi_N)$ point. The bin resolution of this cube is 1 MeV in T_N and 1° in each angular dimension. Each cube contains 500 uniformly distributed events sampled from the NF or from the true cross section.

Using a generator whose truth distribution is matched to that surrogate model will therefore show these signatures in simulated samples, increasing the chances that experiments with near- 4π angular acceptance and high-granularity

detectors, with good sensitivity to low-momentum protons, such as the SuperFGD at T2K [55], will be able to observe some of them, if they are not already visible in existing datasets.

FIG. 7. $\theta_N = 20^\circ$.

FIG. 8. $\theta_N = 50^\circ$.

- [1] K. Abe *et al.* (T2K Collaboration), The T2K experiment, *Nucl. Instrum. Methods Phys. Res., Sect. A* **659**, 106 (2011).
- [2] H.-K. Proto-Collaboration, Hyper-kamiokande design report, [arXiv:1805.04163](https://arxiv.org/abs/1805.04163).
- [3] L. Alvarez-Ruso *et al.* (NuSTEC Collaboration), NuSTEC white paper: Status and challenges of neutrino-nucleus scattering, *Prog. Part. Nucl. Phys.* **100**, 1 (2018).
- [4] Y. Hayato, A neutrino interaction simulation program library NEUT, *Acta Phys. Pol. B* **40**, 2477 (2009), <https://inspirehep.net/literature/844435>.
- [5] C. Andreopoulos *et al.*, The GENIE neutrino Monte Carlo generator, *Nucl. Instrum. Methods Phys. Res., Sect. A* **614**, 87 (2010).
- [6] T. Golan, C. Juszczak, and J. T. Sobczyk, Final state interactions effects in neutrino-nucleus interactions, *Phys. Rev. C* **86**, 015505 (2012).
- [7] J. Isaacson, W. I. Jay, A. Lovato, P. A. Machado, and N. Rocco, Introducing a novel event generator for electron-nucleus and neutrino-nucleus scattering, *Phys. Rev. D* **107**, 033007 (2023).
- [8] S. Dolan, G. Megias, and S. Bolognesi, Implementation of the susav2-meson exchange current 1p1h and 2p2h models in genie and analysis of nuclear effects in T2K measurements, *Phys. Rev. D* **101**, 033003 (2020).
- [9] A. Nikolakopoulos, S. Gardiner, A. Papadopolou, S. Dolan, and R. González-Jiménez, From inclusive to

- semi-inclusive one-nucleon knockout in neutrino event generators, *Phys. Sci. Forum.* **8**, 10 (2023).
- [10] A. V. Butkevich and S. P. Mikheyev, Test of Fermi gas model and plane-wave impulse approximation against electron-nucleus scattering data, *Phys. Rev. C* **72**, 025501 (2005).
- [11] O. Benhar, A. Fabrocini, S. Fantoni, and I. Sick, Spectral function of finite nuclei and scattering of GeV electrons, *Nucl. Phys.* **A579**, 493 (1994).
- [12] P. Ring, Relativistic mean field theory in finite nuclei, *Prog. Part. Nucl. Phys.* **37**, 193 (1996).
- [13] N. Jachowicz, K. Heyde, J. Ryckebusch, and S. Rombouts, Continuum random phase approximation approach to charged-current neutrino-nucleus scattering, *Phys. Rev. C* **65**, 025501 (2002).
- [14] J. McKean, R. González-Jiménez, M. Kabirnezhad, J. M. Udías, and Y. Uchida, Implementation of a relativistic distorted wave impulse approximation model into the neut event generator, [arXiv:2502.10629](#).
- [15] A. Nikolakopoulos, R. González-Jiménez, N. Jachowicz, K. Niewczas, F. Sánchez, and J. M. Udías, Benchmarking intranuclear cascade models for neutrino scattering with relativistic optical potentials, *Phys. Rev. C* **105**, 054603 (2022).
- [16] A. Nikolakopoulos, A. Ershova, R. González-Jiménez, J. Isaacson, A. M. Kelly, K. Niewczas, N. Rocco, and F. Sánchez, Final-state interactions in neutrino-induced proton knockout from argon in microboone, *Phys. Rev. C* **110**, 054611 (2024).
- [17] D. Douqa, S. Bordoni, L. Giannessi, F. Sánchez, C. Schloesser, and A. Nikolakopoulos, Superscaling variable and neutrino energy reconstruction from theoretical predictions to experimental limitations, *Phys. Rev. D* **109**, 073001 (2024).
- [18] R. González-Jiménez, M. B. Barbaro, J. A. Caballero, T. W. Donnelly, N. Jachowicz, G. D. Megias, K. Niewczas, A. Nikolakopoulos, J. W. Van Orden, and J. M. Udías, Neutrino energy reconstruction from semi-inclusive samples, *Phys. Rev. C* **105**, 025502 (2022).
- [19] S. Dolan, A. Nikolakopoulos, O. Page, S. Gardiner, N. Jachowicz, and V. Pandey, Implementation of the continuum random phase approximation model in the GENIE generator and an analysis of nuclear effects in low-energy transfer neutrino interactions, *Phys. Rev. D* **106**, 073001 (2022).
- [20] J. Schwehr, D. Cherdack, and R. Gran, GENIE implementation of IFIC valencia model for QE-like 2p2h neutrino-nucleus cross section, [arXiv:1601.02038](#).
- [21] H. Prasad, J. T. Sobczyk, A. M. Ankowski, J. L. Bonilla, R. D. Banerjee, K. M. Graczyk, and B. E. Kowal, New multinucleon knockout model in the NuWro Monte Carlo generator, *Phys. Rev. D* **111**, 036032 (2025).
- [22] T. Heimel, N. Huetsch, F. Maltoni, O. Mattelaer, T. Plehn, and R. Winterhalder, The madnis reloaded, *SciPost Phys.* **17**, 023 (2024).
- [23] K. M. Graczyk, B. E. Kowal, A. M. Ankowski, R. D. Banerjee, J. L. Bonilla, H. Prasad, and J. T. Sobczyk, Electron-nucleus cross sections from transfer learning, [arXiv:2408.09936](#).
- [24] B. *et al.*, Machine learning and LHC event generation, *SciPost Phys.* **14**, 079 (2023).
- [25] J. L. Bonilla, K. M. Graczyk, A. M. Ankowski, R. D. Banerjee, B. E. Kowal, H. Prasad, and J. T. Sobczyk, Generative adversarial neural networks for simulating neutrino interactions, [arXiv:2502.20244](#).
- [26] D. J. Rezende and S. Mohamed, Variational inference with normalizing flows, [arXiv:1505.05770](#).
- [27] J. Ryckebusch, M. Vanderhaeghen, L. Machenil, and M. Waroquier, Effects of the final-state interaction in (γ, pn) and (γ, pp) processes, *Nucl. Phys.* **A568**, 828 (1994).
- [28] V. Van der Sluys, J. Ryckebusch, and M. Waroquier, Two-body currents in inclusive electron scattering, *Phys. Rev. C* **51**, 2664 (1995).
- [29] D. Vautherin and D. M. Brink, Hartree-Fock calculations with Skyrme's interaction. I. Spherical nuclei, *Phys. Rev. C* **5**, 626 (1972).
- [30] K. Heyde, *The Nuclear Shell Model*, Springer Series in Nuclear and Particle Physics (Springer, Berlin, Heidelberg, 2012).
- [31] M. Waroquier, G. Wenes, and K. Heyde, An effective skyrme-type interaction for nuclear structure calculations: (II). Excited-state properties, *Nucl. Phys.* **A404**, 298 (1983).
- [32] O. Moreno, T. W. Donnelly, J. W. Van Orden, and W. P. Ford, Semi-inclusive charged-current neutrino-nucleus reactions, *Phys. Rev. D* **90**, 013014 (2014).
- [33] J. E. Sobczyk, E. Hernández, S. X. Nakamura, J. Nieves, and T. Sato, Angular distributions in electroweak pion production off nucleons: Odd parity hadron terms, strong relative phases, and model dependence, *Phys. Rev. D* **98**, 073001 (2018).
- [34] K. Niewczas, A. Nikolakopoulos, J. T. Sobczyk, N. Jachowicz, and R. González-Jiménez, Angular distributions in Monte Carlo event generation of weak single-pion production, *Phys. Rev. D* **103**, 053003 (2021).
- [35] G. Peter Lepage, A new algorithm for adaptive multidimensional integration, *J. Comput. Phys.* **27**, 192 (1978).
- [36] C. Gao, S. Höche, J. Isaacson, C. Krause, and H. Schulz, Event generation with normalizing flows, *Phys. Rev. D* **101**, 076002 (2020).
- [37] C. Gao, J. Isaacson, and C. Krause, i-flow: High-dimensional integration and sampling with normalizing flows, *Mach. Learn.* **1**, 045023 (2020).
- [38] E. Bothmann, T. Janßen, M. Knobbe, T. Schmale, and S. Schumann, Exploring phase space with neural importance sampling, *SciPost Phys.* **8**, 069 (2020).
- [39] B. Stienen and R. Verheyen, Phase space sampling and inference from weighted events with autoregressive flows, *SciPost Phys.* **10**, 038 (2021).
- [40] T. Heimel, R. Winterhalder, A. Butter, J. Isaacson, C. Krause, F. Maltoni, O. Mattelaer, and T. Plehn, Madnis—neural multi-channel importance sampling, *SciPost Phys.* **15**, 141 (2023).
- [41] N. Deutschmann and N. Götz, Accelerating HEP simulations with neural importance sampling, *J. High Energy Phys.* **03** (2024) 083.
- [42] A. Kofler, V. Stimper, M. Mikhasenko, M. Kagan, and L. Heinrich, Flow annealed importance sampling bootstrap meets differentiable particle physics, in *Proceedings of the*

- 38th Conference on Neural Information Processing Systems (2024), [arXiv:2411.16234](https://arxiv.org/abs/2411.16234).
- [43] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, Improving variational inference with inverse autoregressive flow, [arXiv:1606.04934](https://arxiv.org/abs/1606.04934).
 - [44] G. Papamakarios, T. Pavlakou, and I. Murray, Masked autoregressive flow for density estimation, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), Vol. 30, pp. 2338–2347, <https://proceedings.neurips.cc/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf>.
 - [45] L. Dinh, D. Krueger, and Y. Bengio, Nice: Non-linear independent components estimation, [arXiv:1410.8516](https://arxiv.org/abs/1410.8516).
 - [46] L. Dinh, J. Sohl-Dickstein, and S. Bengio, Density estimation using real NVP, [arXiv:1605.08803](https://arxiv.org/abs/1605.08803).
 - [47] A. Cocco, M. Letizia, H. Reyes-González, and R. Torre, Comparison of affine and rational quadratic spline coupling and autoregressive flows through robust statistical tests, *Symmetry* **16**, 942 (2024).
 - [48] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Neural spline flows, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019), Vol. 32, pp. 7511–7522, <https://proceedings.neurips.cc/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf>.
 - [49] R. Delbourgo and J. A. Gregory, Piecewise rational quadratic interpolation to monotonic data, *IMA J. Numer. Anal.* **3**, 141 (1983).
 - [50] S. Pina-Otey, F. Sánchez, T. Lux, and V. Gaitan, Exhaustive neural importance sampling applied to Monte Carlo event generation, *Phys. Rev. D* **102**, 013003 (2020).
 - [51] M. El Baz and F. Sánchez, Fast posterior probability sampling with normalizing flows and its applicability in bayesian analysis in particle physics, *Phys. Rev. D* **109**, 032008 (2024).
 - [52] D. J. Rezende, G. Papamakarios, S. Racanière, M. S. Albergo, G. Kanwar, P. E. Shanahan, and K. Cranmer, Normalizing flows on tori and spheres, [arXiv:2002.02428](https://arxiv.org/abs/2002.02428).
 - [53] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák, Neural importance sampling, [arXiv:1808.03856](https://arxiv.org/abs/1808.03856).
 - [54] K. Abe *et al.* (T2K Collaboration), Measurements of neutrino oscillation in appearance and disappearance channels by the T2K experiment with 6.6×10^{20} protons on target, *Phys. Rev. D* **91**, 072010 (2015).
 - [55] A. Blondel *et al.*, The superfgd prototype charged particle beam tests, *J. Instrum.* **15**, P12003 (2020).