

Analytical and machine learning methods for model discernment at CE ν NS experiments

Iain A. Bisset,^{1,*} Bhaskar Dutta^{1,†}, Doojin Kim^{2,‡}, Samiran Sinha^{3,§} and Joel W. Walker^{4,||}

¹*Mitchell Institute for Fundamental Physics and Astronomy, Department of Physics and Astronomy, Texas A&M University, College Station, Texas 77843, USA*

²*Department of Physics, University of South Dakota, Vermillion, South Dakota 57069, USA*

³*Department of Statistics, Texas A&M University, College Station, Texas 77843, USA*

⁴*Department of Physics and Astronomy, Sam Houston State University, Huntsville, Texas 77341, USA*



(Received 13 May 2026; accepted 9 September 2026; published 1 October 2026)

Neutrino experiments are currently limited by low statistics, sizable systematic uncertainties, and coarse observable binning, which can hinder discrimination among competing beyond-the-Standard-Model (BSM) explanations of anomalous signals. In particular, analyses based primarily on total event-rate differences are vulnerable to source-normalization uncertainties and to degeneracies among models that induce similar inclusive yields. Using stopped-pion coherent elastic neutrino-nucleus scattering (CE ν NS) as a benchmark environment, we study how much model-discrimination power can be obtained from correlations in baseline, recoil energy, and timing that are less sensitive to the total rate. As benchmark BSM scenarios, we consider a $3 + 1$ sterile-neutrino framework and neutral-current nonstandard neutrino interactions (NSI). We show with a likelihood-based analysis that these scenarios can be distinguished in nontrivial regions of parameter space once multidimensional shape information is retained. We further demonstrate with convolutional neural networks that substantial discrimination remains possible even after the total event rate is explicitly removed from the input, indicating that the relevant information is genuinely encoded in the shape of the CE ν NS distribution. Finally, through multiclass classification within the sterile parameter space, we show that in favorable regions the same observables can support approximate localization of the underlying sterile-neutrino benchmark point. Our results highlight the complementary roles of conventional and machine-learning-based inference in moving neutrino new-physics searches from anomaly detection to physics interpretation.

DOI: [10.1103/31x4-3z7h](https://doi.org/10.1103/31x4-3z7h)

I. INTRODUCTION

Predictions of the neutrino-sector of the Standard Model (SM) have been confronted with an increasingly broad program of precision measurements in oscillation, scattering, and source-based experiments [1,2]. As the number of precision tests in the neutrino sector continues to grow, it is becoming increasingly important to develop robust methods for model selection and parameter estimation tailored to the challenges inherent in these experiments. In

particular, the wide landscape of viable new-physics scenarios makes effective model discrimination an increasingly important task. New-physics signals associated with neutrinos across a wide range of experiments can arise from diverse mechanisms. These include, for example, oscillation-based scenarios involving sterile neutrinos [3,4] as well as nonoscillation explanations such as nonstandard neutrino interactions [5] or other modifications of neutrino production, propagation, and detection. In general, such scenarios can alter the observed data in several distinct ways: through a change in the overall event yield, through distortions in spectral, spatial, and/or temporal distributions, and through modified flavor structure. In practice, however, neutrino experiments often operate in regimes where limited statistics, imperfect flux knowledge, detector effects, and/or coarse binning reduce the extent to which these signatures can be cleanly disentangled. As a result, different new-physics models may generate qualitatively similar excesses or deficits at the level of inclusive rates, even when their underlying dynamical origins are very different.

*Contact author: ibisset@tamu.edu

†Contact author: dutta@physics.tamu.edu

‡Contact author: doojin.kim@usd.edu

§Contact author: sinha@stat.tamu.edu

||Contact author: jwalker@shsu.edu

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.

This general issue is especially relevant for stopped-pion neutrino experiments. In these experiments, a high intensity proton beam with energy ~ 1 GeV hits a fixed target and the detectors are kept at a distance ~ 20 meters from the target. In many analyses of such facilities, sensitivity has been driven by 2D likelihood analyses in energy and time [6–10]. Detailed statistical and phenomenological studies [4,11–14] as well as experimental analyses [10,15,16] have highlighted the usefulness of the intrinsic time structure of stopped-pion sources for new-physics searches: the prompt ν_μ component and the delayed $\nu_e/\bar{\nu}_\mu$ components populate different timing windows, so that timing information can be used to perform a partial flavor decomposition and thereby sharpen sensitivity to nonstandard scenarios. In this sense, the separation into prompt and delayed samples may already be viewed as a coarse form of shape-based analysis, built from correlations among the available observables. This observation motivates the present study, in which we ask how much further discriminatory power can be gained by treating baseline, recoil energy, and timing together within a more systematic multidimensional inference framework.

For this reason, the question of discovery is only part of a broader scientific problem. Once one goes beyond inclusive event counts and begins to exploit shape information encoded in the available observables, a more refined set of questions becomes accessible. If an anomalous excess or deficit is observed in a neutrino experiment, it is essential to determine not only whether the data prefer a nonstandard interpretation over the SM, but also which model class is favored and which region of parameter space is selected by the data. Such information is critical for connecting the observed anomaly to other measurements. A preferred parameter region inferred from one experiment can imply correlated signals, null tests, or exclusion targets in independent experiments, thus sharpening the program of confirmation and accelerating progress toward a consistent physical interpretation.

Apart from the motivation provided by possible nonstandard signals at future experiments, several past experiments have also exhibited anomalous results that can accommodate a range of different model-dependent interpretations, further necessitating the development of new model discernment methodologies. While many observations are well described within the standard three-neutrino framework, several longstanding anomalies and tensions continue to motivate the exploration of beyond-the-SM (BSM) physics. Representative examples arise from short-baseline, pion decay at rest, and reactor experiments. Notable anomalies include the Liquid Scintillator Neutrino Detector excess [17] and the MiniBooNE low-energy excess [18–20]. Results from Short Baseline experiments, such as BEST [21], when combined with SAGE [22] and GALLEX [23], provide evidence consistent with a $3 + 1$ sterile neutrino framework. However, the $3 + 1$ scenario is not preferred by the recent results from MicroBooNE [24].

Although the interpretation of these results remains unsettled, they provide concrete motivation for developing analysis strategies capable not only of identifying departures from standard expectations but also of determining which classes of new-physics scenarios may account for an observed signal.

Recent years have also seen the widespread application of machine-learning (ML) techniques to neutrino physics [25–35]. In many contexts, ML has demonstrated excellent performance in event classification, background rejection, and high-dimensional feature extraction (see e.g., [36–38]). This naturally raises an important question for neutrino-sector new-physics searches: beyond improving signal sensitivity or detecting anomalous behaviors, can ML methods also enhance our ability to identify the underlying model responsible for a deviation from the SM, and to infer the corresponding parameter point with meaningful precision? Addressing this question requires going beyond a binary signal-versus-background framework and instead treating the problem as one of structured inference in a multidimensional observable space. Related ML approaches have also recently been explored for model discrimination in collider settings, including at the LHC [39–41].

In this work, we investigate these issues in the context of sterile-neutrino-induced and/or nonstandard neutrino interaction-induced modifications to coherent elastic neutrino-nucleus scattering (CE ν NS) at stopped-pion neutrino sources. This environment is particularly well-suited to the present study. Because the neutrino production mechanism is time structured (charged-pion-induced prompt neutrino signals vs muon-induced delayed ones), and because detectors can in principle be deployed at multiple baselines, the signal is encoded not only in the total event rate but also in the correlated shapes of the recoil-energy, timing, and baseline distributions. These observables provide a richer arena for model discrimination than rate-only measurements and allow one to test how effectively different analysis strategies extract nonstandard information from realistic datasets.

Our study has two primary goals. First, we apply both a traditional likelihood-based analysis as well as a ML-based analysis to determine their ability to distinguish sterile-neutrino scenarios from alternative nonstandard interaction (NSI) scenarios. This allows us to assess model discrimination power using both conventional and high-dimensional methods within a common benchmark framework. Second, within the sterile-neutrino hypothesis itself for illustration, we study how well one can identify the underlying parameter point from the observed CE ν NS data, including a multiclass classification strategy for approximate localization within the sterile parameter space. In other words, we examine not only whether a nonstandard signal can be detected, but also how accurately the relevant model parameters can be reconstructed once such a signal is present.

The broader aim is to clarify the role that modern inference tools can play in moving neutrino new-physics searches from anomaly detection to physics interpretation and, in favorable cases, toward approximate parameter-space localization. If successful, such methods can provide a pathway from an observed deviation, to a favored model hypothesis, to concrete predictions for independent experimental confirmation. All of the techniques that we are developing can also be applied to distinguish light dark matter scenarios from those involving NSI and sterile neutrinos discussed so far.

The rest of this paper is structured as follows. In Sec. II, we introduce the benchmark physics setup used throughout this work, including CE ν NS at stopped-pion sources, the benchmark sterile-neutrino and NSI scenarios, and the statistical framework for constructing the binned event distributions. In Sec. III, we present a conventional likelihood-based analysis and quantify the resulting model-discrimination reach in the benchmark parameter space. In Sec. IV, we describe the common neural-network architecture and training procedure adopted for the ML studies. In Sec. V, we report the results of two complementary ML tasks: binary classification for model discrimination between the sterile-neutrino and best-fit NSI hypotheses, and multiclass classification for identifying the underlying region of sterile-neutrino parameter space. Finally, in Sec. VI, we summarize our main findings and discuss their implications for future neutrino and dark-sector searches.

II. BENCHMARK PHYSICS CASES

Before outlining the analysis framework and presenting our main results, we first specify the benchmark scenarios that will serve as the basis of our study. As briefly discussed in the Introduction, we focus on neutrino-induced CE ν NS signals at stopped-pion neutrino experiments and consider the possibility that observable deviations from the standard three-neutrino prediction originate from nonstandard neutrino physics. As representative examples, we study the $3 + 1$ sterile-neutrino oscillation scenario and scenarios with NSIs.

A. CE ν NS at stopped-pion sources

To study model discrimination and parameter inference in a controlled setting, we consider benchmark CE ν NS signals induced by neutrinos from stopped-pion sources. Specifically, we envision a (hypothetical) experimental configuration consisting of multiple CsI detectors placed at different baselines from the neutrino source. This setup is motivated by existing and proposed stopped-pion neutrino programs, such as COHERENT at Oak Ridge National Laboratory [8], Coherent CAPTAIN-Mills (CCM) at Los Alamos National Laboratory [10], and the Search for Hidden Neutrinos at the European Spallation Source [42], as well as other forms of experiments designed to

detect CE ν NS [43,44]. In addition to its previous runs with CsI and LAr, the COHERENT collaboration is currently operating, or plans to operate NaI, Ge, Neon, CsI, and LAr detectors, all located at distances of approximately 19–28 meters [45]. The CCM experiment is likewise planning to deploy multiton segmented CsI detectors in addition to its ongoing LAr detector [46]. Our purpose here is not to model any one experiment in complete detail, but rather to define a realistic benchmark environment in which baseline, recoil-energy, and timing information can all be exploited simultaneously.

In stopped-pion neutrino experiments, intense GeV-scale proton beams strike a target and copiously produce charged pions. The π^- rapidly lose energy and are typically captured before decaying, while the π^+ come to rest and decay via $\pi^+ \rightarrow \mu^+ \nu_\mu$. The decay product μ^+ subsequently also stops and decays through $\mu^+ \rightarrow e^+ \nu_e \bar{\nu}_\mu$, thereby yielding the characteristic neutrino flux of stopped-pion sources. The two-body and three-body nature of π^+ and μ^+ decays, respectively, implies that the ν_μ spectrum is mono-energetic while the $\bar{\nu}_\mu$ and ν_e spectra span a continuous range of energies. Specifically, the differential fluxes for each source component are given as follows:

$$\frac{d\Phi_{\nu_\mu}}{dE_{\nu_\mu}} = \mathcal{N} \delta\left(E_{\nu_\mu} - \frac{m_\pi^2 - m_\mu^2}{2m_\pi}\right), \quad (1)$$

$$\frac{d\Phi_{\bar{\nu}_\mu}}{dE_{\bar{\nu}_\mu}} = \mathcal{N} \frac{16}{m_\mu^4} (3m_\mu E_{\bar{\nu}_\mu}^2 - 4E_{\bar{\nu}_\mu}^3), \quad (2)$$

$$\frac{d\Phi_{\nu_e}}{dE_{\nu_e}} = \mathcal{N} \frac{96}{m_\mu^4} (m_\mu E_{\nu_e}^2 - 2E_{\nu_e}^3), \quad (3)$$

where $\mathcal{N}$ is the normalization parameter given by $\mathcal{N} = r N_{\text{POT}} / (4\pi L^2)$ with r , N_{POT} , and L denoting the pion production fraction, the number of protons on target (POT), and the source-to-detector distance, respectively. For our benchmark scenario, we take $r = 0.05$ and $N_{\text{POT}} = 10^{22}$ in Sec. III and Sec. VA, while scaling up both N_{POT} and the detector masses by another factor of 10 each in Sec. VB relative to the values described here. For reference, the spallation neutron source (SNS) used by COHERENT runs at 60 Hz with 1.5×10^{14} POT per pulse [47], which yields 10^{22} POT in approximately 13 days of live-time. The baseline L varies for each detector in the experiment. We assume a setup consisting of several equal-mass 50 kg CsI detectors logarithmically distributed in length between 5m and 25m from the neutrino source.

Equally important for our purposes is the time structure of the neutrino source. Because the pion lifetime (~ 26 ns in the rest frame) is much shorter than the muon lifetime (~ 2.2 μ s in the rest frame), the prompt ν_μ component is temporally separated from the delayed ν_e and $\bar{\nu}_\mu$ components. We model the timing distributions as

$$\frac{d\Phi_{\nu_\mu}}{dt} = \frac{1}{b\tau_\pi\sqrt{2\pi}} \int_0^t \exp\left(-\frac{(t'-a)^2}{2b^2}\right) \exp\left(-\frac{(t-t')}{\tau_\pi}\right) dt', \quad (4)$$

$$\frac{d\Phi_{(\bar{\nu}_\mu, \nu_e)}}{dt} = \frac{1}{\tau_\mu} \int_0^t \frac{d\Phi_{\nu_\mu}}{dt'} \exp\left(-\frac{(t-t')}{\tau_\mu}\right) dt', \quad (5)$$

where τ_π and τ_μ denote the pion and muon lifetimes, while a and b parametrize the proton-pulse time profile. In our benchmark setup, these timing parameters are taken from the SNS timing structure, for consistency with the COHERENT analysis [8].

Since both π^+ and μ^+ decay at rest, the resulting neutrinos are emitted isotropically. They reach the CsI detector of interest, and scatter off the cesium and iodine nuclei primarily via the CE ν NS process, which involves the neutral current and is approximately flavor agnostic. [48,49]. The differential cross section with respect to the nuclear recoil energy E_r is given by [48–50]

$$\frac{d\sigma}{dE_r} = \frac{G_F^2 Q_V^2 M}{2\pi} F^2(q^2) \left[1 - \left(\frac{ME_r}{E_\nu^2} \right) + \left(1 - \frac{E_r}{E_\nu} \right)^2 \right], \quad (6)$$

where M is the total mass of the target nucleus, E_ν is the incoming neutrino energy, and $F(q^2)$ parametrizes the nuclear form factor that encodes the loss of coherence at finite momentum transfer q , for which the Helm form factor [51] is commonly adopted. The effective SM vector charge is $Q_V = (2g_u + g_d)Z + (g_u + 2g_d)N$, where g_u and g_d are the vector couplings to the u - and d -type quarks, and Z and N are respectively the proton and neutron numbers of the nucleus under consideration.

A key advantage of the CE ν NS process, and the detectors designed to measure it, is their ability to measure low-energy recoils while maintaining a significant detector mass. This is crucial for taking advantage of the cross section boost in the low-energy regime, which also scales quadratically (proportional to the neutron count squared) with the size of the target nucleus. However, one downside of such detectors is that they are unable to directly recover the incident neutrino energy, because the scattering angle is not measured. As such, if a neutrino source presents a continuous spectrum, there will be smearing of the oscillation response. However, this can be mitigated by binning in the recoil energy, and more energetic neutrinos are still likely to produce harder scatters, on average [52]. This also highlights the advantage of binning in time, since timing information can help disentangle the discrete (monochromatic) and continuous spectral components of the neutrino source. In particular, the oscillation response of the monochromatic component does not suffer from the smearing described above.

The key point is that this framework provides three complementary observables: recoil-energy spectra, timing

distributions, and variation with detectors placed across multiple length baselines. These observables respond differently to changes in neutrino flavor content, oscillation parameters, and interaction structure. Consequently, even when different new-physics scenarios generate similar modifications to the total event rate, they may still be distinguishable once the full multidimensional information is incorporated. This is precisely the feature that makes stopped-pion CE ν NS a useful benchmark for the likelihood-based and machine-learning-based analyses presented in the following sections.

B. Benchmark BSM scenarios

To study model discrimination and parameter reconstruction in a controlled setting, we consider benchmark datasets generated under the SM and two representative BSM scenarios.

1. Benchmark I: 3 + 1 sterile-neutrino oscillation scenario

As our first benchmark new-physics framework, we consider the 3 + 1 sterile-neutrino scenario. In this extension of the standard three-neutrino picture, the active neutrino sector is supplemented by one additional, predominantly sterile, mass eigenstate ν_4 . The corresponding leptonic mixing matrix is then enlarged from a 3×3 matrix to a 4×4 matrix. At short baselines, where the oscillation effects driven by Δm_{21}^2 and Δm_{31}^2 can be neglected, the phenomenology is controlled primarily by the larger mass-squared splitting

$$\Delta m_{41}^2 \equiv m_4^2 - m_1^2, \quad (7)$$

together with the mixing matrix elements $U_{\alpha 4}$ involving the fourth mass eigenstate, where $\alpha = e, \mu, \tau$.

In the short-baseline limit, the oscillation probabilities take a two-flavorlike form. For appearance channels ($\alpha \neq \beta$), one has

$$P_{\alpha\beta}(E_\nu, L) = \sin^2 2\theta_{\alpha\beta} \sin^2\left(\frac{\Delta m_{41}^2 L}{4E_\nu}\right), \quad (\alpha \neq \beta), \quad (8)$$

whereas for disappearance channels,

$$P_{\alpha\alpha}(E_\nu, L) = 1 - \sin^2 2\theta_{\alpha\alpha} \sin^2\left(\frac{\Delta m_{41}^2 L}{4E_\nu}\right). \quad (9)$$

The effective mixing amplitudes are given by

$$\sin^2 2\theta_{\alpha\beta} = 4|U_{\alpha 4}|^2 |U_{\beta 4}|^2, \quad (\alpha \neq \beta), \quad (10)$$

and

$$\sin^2 2\theta_{\alpha\alpha} = 4|U_{\alpha 4}|^2 (1 - |U_{\alpha 4}|^2). \quad (11)$$

In the stopped-pion CE ν NS setup considered here, the neutrino source consists of prompt ν_μ and delayed ν_e and $\bar{\nu}_\mu$ components. Since CE ν NS is a neutral-current process and is therefore not directly sensitive to the flavor identity among active neutrinos, the dominant observable effect of sterile mixing arises from the depletion of the active neutrino flux through oscillations into the sterile state. Consequently, the relevant parameters for our analysis are primarily Δm_{41}^2 , $|U_{e4}|^2$, and $|U_{\mu 4}|^2$. For simplicity, and because the source does not contain an initial ν_τ component, we set $|U_{\tau 4}|^2 = 0$ throughout this study.

This benchmark scenario provides a well-motivated and economical test case for our analysis framework. It generates characteristic distortions in the baseline and energy dependence of the CE ν NS event distribution, while the prompt-delayed timing structure of stopped-pion sources allows the effects associated with the ν_μ and $\nu_e/\bar{\nu}_\mu$ source components to be partially disentangled. These features make the $3 + 1$ framework a useful benchmark for assessing both model discrimination and parameter reconstruction.

2. Benchmark II: Neutral-current NSIs

As the second benchmark new-physics framework, we consider neutral-current NSIs of neutrinos with quarks. In a model-independent low-energy description, such effects can be parametrized by effective four-fermion operators that modify the neutrino-matter interaction strength relative to the SM. The corresponding effective Lagrangian is written as [5]

$$\mathcal{L}_{\text{NSI}} = -2\sqrt{2}G_F\epsilon_{\alpha\beta}^{fH}(\bar{\nu}_\alpha\gamma_\sigma P_L\nu_\beta)(\bar{f}\gamma^\sigma P_H f), \quad (12)$$

where $\alpha, \beta = e, \mu, \tau$ label neutrino flavors, f denotes the matter fermion, and $H = L, R$ specifies the chirality of the fermion current. For CE ν NS, the relevant target constituents are the light quarks u and d , and only the vector combinations $\epsilon_{\alpha\beta}^{fV} \equiv \epsilon_{\alpha\beta}^{fL} + \epsilon_{\alpha\beta}^{fR}$ contribute coherently at leading order. In this sense, CE ν NS provides a particularly clean probe of neutral-current NSI.

The presence of NSI modifies the effective weak charge of the nucleus and, consequently, the CE ν NS event rate. For an incident neutrino of flavor α , the squared effective coupling becomes

$$(Q_\alpha^{\text{NSI}})^2 = ([2(g_u + \epsilon_{\alpha\alpha}^{uV}) + (g_d + \epsilon_{\alpha\alpha}^{dV})]Z + [(g_u + \epsilon_{\alpha\alpha}^{uV}) + 2(g_d + \epsilon_{\alpha\alpha}^{dV})]N)^2 + \sum_{\beta \neq \alpha} [(2\epsilon_{\alpha\beta}^{uV} + \epsilon_{\alpha\beta}^{dV})Z + (\epsilon_{\alpha\beta}^{uV} + 2\epsilon_{\alpha\beta}^{dV})N]^2. \quad (13)$$

Here the diagonal couplings $\epsilon_{\alpha\alpha}^{fV}$ shift the flavor-conserving interaction strength, while the off-diagonal couplings $\epsilon_{\alpha\beta}^{fV}$ with $\alpha \neq \beta$ correspond to flavor-changing neutral-current NSI. In both cases, the observable consequence is a modification of the CE ν NS rate and its distribution across the multidimensional observable space used in our analysis.

In contrast to the sterile-neutrino scenario, where the leading CE ν NS signature is primarily a depletion of the active neutrino flux through oscillations into a sterile state, neutral-current NSI directly alter the interaction strength of active neutrinos with the detector material. As a result, NSI can generate distortions that differ in both normalization and flavor dependence. In the stopped-pion CE ν NS environment considered here, this is particularly relevant because the prompt ν_μ component and the delayed $\nu_e/\bar{\nu}_\mu$ components populate different regions of the timing distribution. Their contributions can therefore be weighted differently in the presence of flavor-dependent NSI couplings.

This makes neutral-current NSI a useful benchmark complementary to the $3 + 1$ sterile-neutrino scenario. While both frameworks can induce deviations from the SM expectation in CE ν NS observables, the working mechanisms are physically distinct: sterile mixing modifies the active neutrino flux reaching the detector, whereas NSI modify the scattering response of the detector to the incident active flux. The degree to which these mechanisms can be distinguished forms one of the central questions of the present study.

C. Statistical framework

We now specify the statistical framework used to generate and analyze the benchmark event distributions. For a given choice of model parameters—denoted by $\vec{s}$ for the $3 + 1$ sterile-neutrino scenario and by $\vec{e}$ for the NSI scenario—the predicted CE ν NS yield is computed in bins of detector baseline, reconstructed recoil energy, and event time. The corresponding expected number of events in each bin is written as

$$N_{ijk}^{\text{CE}\nu\text{NS}}(\vec{s}, \vec{e}) = \sum_\alpha \sum_\beta \sum_{\gamma=\nu_\mu, \bar{\nu}_\mu, \nu_e} \mathcal{N}_{i\beta} \int_k dt \int_j dE_{\text{rec}} \times \int dE_{\text{true}} \int dE_\gamma \epsilon_T(t) \epsilon_E(E_{\text{rec}}) P_E(E_{\text{rec}}, E_{\text{true}}) \times P_{\gamma\alpha}(\vec{s}, E_\gamma, L_i) \frac{d^2\Phi_\gamma}{dE_\gamma dt}(E_\gamma, t, L_i) \frac{d\sigma_{\alpha\beta}}{dE_{\text{true}}}(\vec{e}, E_{\text{true}}, E_\gamma). \quad (14)$$

The labels $\{i, j, k\}$ denote the binning in length, energy, and time, respectively. The numbers of bins in each dimension vary in this analysis, but binning in length is always distributed from 5m to 25m, while binning in recoil energy and timing are always distributed over the ranges defined in the COHERENT CsI experiment [8]. In this analysis, several scenarios are simulated for different binning choices, with length and energy each ranging from having 1 bin to being divided among 9 bins and time ranging from 1 to 3 bins. For the chosen range in length and the value of N_{POT} assumed in Sec. III and Sec. VA, the detectors closest to the source can have $\approx O(10) - O(100)$ counts in each bin and the detectors farthest from the source can have $\approx O(1) - O(10)$ counts in each bin depending on the granularity of the chosen binning. In Sec. VB, the detector mass and N_{POT} are each increased by a factor of 10, meaning these bin counts will be scaled up by a factor of 10^2 . The subscripts on the dt and dE_{rec} integrals indicate that the integrations are performed over the time and reconstructed-energy ranges corresponding to the k th timing bin and the j th E_{rec} bin, respectively. The integration over the true nuclear recoil energy, E_{true} , goes from 0 to the approximate maximum recoil energy of $2E_\gamma^2/M$ for nucleus mass M , with E_γ in the upper bound taken to be $m_\mu/2$. The integration over the neutrino energy E_γ goes from the approximate minimum neutrino energy required to induce a recoil $\sqrt{mE_{\text{true}}/2}$ to the maximum neutrino energy $\sim m_\mu/2$.¹

The quantity $\mathcal{N}_{i\beta}$ is the number of target nuclei of species β in detector i , and L_i is the corresponding source-to-detector distance. The factor $\frac{d^2\Phi_\gamma}{dE_\gamma dt}$ represents the differential neutrino flux from the source for initial flavor $\gamma \in \{\nu_\mu, \bar{\nu}_\mu, \nu_e\}$, including its characteristic energy and timing structure at stopped-pion facilities.

The oscillation factor $P_{\gamma\alpha}(\vec{s}, E_\gamma, L_i)$ gives the probability that a neutrino produced in flavor state γ is detected as flavor α after propagating a distance L_i . In the sterile-neutrino benchmark, this is the quantity that encodes the dependence on the parameters $\vec{s}$. By contrast, the differential cross section $\frac{d\sigma_{\alpha\beta}}{dE_{\text{true}}}$ contains the dependence on the NSI parameters $\vec{e}$ through the modified CE ν NS interaction strength discussed above. The functions $\epsilon_T(t)$ and $\epsilon_E(E_{\text{rec}})$ account for the detector efficiencies in timing and reconstructed energy. For simplicity, we simply take $\epsilon_T(t) = 1$

and take $\epsilon_E(E_{\text{rec}})$ to be a step function at 6.7 keV. Finally, $P_E(E_{\text{rec}}, E_{\text{true}})$ denotes the detector response function, i.e., the probability density for reconstructing an event with energy E_{rec} given a true recoil energy E_{true} . For this response function, we use the formulation given by COHERENT for their CsI CE ν NS detection [8].

Equation (14) therefore provides a unified description of the expected multidimensional event distribution in the presence of either sterile-neutrino oscillations or neutral-current NSI. In the SM limit, one simply recovers the corresponding expression by setting $P_{\gamma\alpha} = \delta_{\gamma\alpha}$, the Kronecker delta function, and replacing the cross section with its standard CE ν NS form. This binned event prediction serves as the basic input for both the conventional likelihood analysis and the ML-based inference discussed in the following sections.

III. LIKELIHOOD-BASED MODEL DISCRIMINATION

We first assess the extent to which the benchmark sterile-neutrino and NSI scenarios can be distinguished using a conventional likelihood-based analysis. To this end, we construct a binned likelihood for the CE ν NS event distribution and compare competing hypotheses using the Bayesian information criterion (BIC) [53]. This provides a baseline, non-ML approach against which the performance of more flexible inference strategies can later be compared.

A. Likelihood construction and benchmark setup

Our analysis is based on the binned CE ν NS event prediction introduced in Sec. II. For a given hypothesis, the expected number of events in bin (i, j, k) is denoted by

$$\lambda_{ijk} = (1 + \eta) N_{ijk}^{\text{CE}\nu\text{NS}}(\vec{s}, \vec{e}), \quad (15)$$

where η is a nuisance parameter that accounts primarily for the overall neutrino-flux normalization uncertainty. The corresponding likelihood for a dataset with observed counts N_{ijk}^{obs} is then written as

$$\mathcal{L}(\vec{s}, \vec{e}) = \int \prod_{(i,j,k)} \frac{e^{-\lambda_{ijk}} \lambda_{ijk}^{N_{ijk}^{\text{obs}}}}{N_{ijk}^{\text{obs}}!} \times \frac{e^{-\eta^2/(2\sigma_\eta^2)}}{\sqrt{2\pi\sigma_\eta^2}} d\eta, \quad (16)$$

where a mean-zero Gaussian prior with known standard deviation σ_η (derived from the experimental setup) is used on η and the product is over the total number of bins in length, energy, and time, which we denote as n_L , n_E , and n_T . In what follows, the observed counts are taken from simulated benchmark data generated under a chosen model hypothesis, represented either by $\vec{s}_{\text{choice}}$ for the sterile-neutrino scenario or by $\vec{e}_{\text{choice}}$ for the NSI scenario.

To compare competing hypotheses, we use the Bayesian information criterion,

¹For $\bar{\nu}_\mu$ and ν_e from the μ^+ decay at rest, the neutrino energy is bounded from above at $\frac{m_\mu^2 - m_e^2}{2m_\mu}$, which can be well approximated to $m_\mu/2$ in the $m_e \ll m_\mu$ limit. Using this approximation has no practical impact on our results. For ν_μ from the π^+ decay at rest, the neutrino energy is single-valued at $\frac{m_\pi^2 - m_\mu^2}{2m_\pi}$, which lies below this maximum energy. Since the ν_μ energy distribution is δ -functionlike as in Eq. (1), taking this integration range does not affect the results in practice.

$$\text{BIC} = k \ln N - 2 \ln \mathcal{L}_{\max}, \quad (17)$$

where k is the number of free model parameters, $N = n_L \times n_E \times n_T$ is the total number of bins in the observed dataset, and $\mathcal{L}_{\max}$ is the maximum of the likelihood given in (16), achieved under the hypothesis being tested. The preferred model is identified as the one with the smaller BIC value. For convenience, we also define the corresponding relative likelihood weight

$$\mathbb{L}(x) = \exp\left(\frac{\text{BIC}_{\min} - x}{2}\right), \quad (18)$$

where $\text{BIC}_{\min}$ denotes the smallest BIC value among the candidate hypotheses for the dataset under consideration, and x is the BIC value generated using the best-fit dataset of a given hypothesis. This quantity maps the BIC difference with respect to the best-fitting model onto the interval $(0, 1]$. In the present analysis, we consider two competing hypotheses, namely the sterile-neutrino and NSI scenarios, such that

$$\text{BIC}_{\min} = \min(\text{BIC}_{\text{sterile}}, \text{BIC}_{\text{NSI}}), \quad (19)$$

within which the SM is a special case with the BSM model parameters being zero. Therefore, if a pseudodataset is generated under the sterile-neutrino hypothesis, one typically has $\text{BIC}_{\min} = \text{BIC}_{\text{sterile}}$, which implies $\mathbb{L}_{\text{sterile}} \equiv \mathbb{L}(\text{BIC}_{\text{sterile}}) = 1$ and $\mathbb{L}_{\text{NSI}} \equiv \mathbb{L}(\text{BIC}_{\text{NSI}}) = \exp(-\Delta\text{BIC}/2)$ with $\Delta\text{BIC} \equiv |\text{BIC}_{\text{NSI}} - \text{BIC}_{\text{sterile}}|$. For reference, ΔBIC values of 2 and 5—corresponding to relative likelihoods below ~ 0.37 and ~ 0.08 —are indicative benchmarks for moderate and stronger levels of distinguishability [53].

Since one of the goals of this analysis is distinguish between models on the basis of shape information rather than total rate, one could consider defining a different likelihood than that of Eq. (16) to separate the likelihood from the total rate information. This could be done by instead defining a multinomial likelihood of the form

$$\mathcal{L}_{\text{mult}} \propto \prod_{(i,j,k)} \left(\frac{N_{ijk}^{\text{CElNS}}}{N_{\text{Total}}} \right)^{N_{ijk}^{\text{obs}}}, \quad (20)$$

where N_{Total} is the total event rate of the model. This removes the normalization of the model's total rate, which additionally removes the flux uncertainty represented by η in Eq. (16). Multinomial likelihoods are equivalent to Poisson likelihoods up to a model-dependent factor that is determined by the model's total rate. Calling this factor $\Lambda(N_{\text{Total}})$ and ignoring the flux uncertainty η , we have the relation $\mathcal{L} \approx \Lambda(N_{\text{Total}})\mathcal{L}_{\text{mult}}$. When calculating the BIC difference between models in the exponential of Eq. (18), the value of N_{Total} for each of the models considered in this analysis will always be very similar (as the pseudo-observed

datasets are generated from sterile neutrino hypotheses and NSI models are always capable of matching the resulting total rate deficit). The result is that the factors involving $\Lambda(N_{\text{Total}})$ cancel when calculating the relative likelihood, and the analysis performed here is equivalent regardless of whether one chooses Eq. (16) or Eq. (20) for the likelihood. The one difference that can have an effect is the inclusion of the flux uncertainty in Eq. (16), however, we have confirmed in our testing that the results are indistinguishable between either choice.

For the illustrative model-comparison study presented in this section, we restrict attention to two-dimensional parameter subsets of the benchmark new-physics scenarios introduced in Sec. II B. On the sterile-neutrino side, we consider

$$\vec{s}_{\text{choice}} = \{|U_{e4}|^2, \Delta m_{41}^2\}, \quad (21)$$

while on the NSI side we consider

$$\vec{e}_{\text{choice}} = \{\epsilon_{ee}^{uV}, \epsilon_{ee}^{dV}\}. \quad (22)$$

Both models therefore have a value of $k = 2$ in Eq. (17). This restricted setup is sufficient to illustrate how the likelihood-based framework responds to two physically distinct mechanisms for generating deviations from the SM expectation. For the purpose of the analyses in this study, we use a sterile-neutrino parameter space window of $10^{-2} - 1$ in $|U_{e4}|^2$ and $10^{-1} - 10^3$ in Δm_{41}^2 . For NSI, we use a window of -1 to 1 in both parameters.

For each point in the chosen sterile-neutrino parameter window, we generate the corresponding mean CE ν NS dataset, as given in Eq. (15) for $\eta = 0$, for a specified binning in baseline, recoil energy, and event time. We then fit the NSI hypothesis to this sterile-generated dataset using MultiNest [54–56], taking the flux-normalization uncertainty in Eq. (16) to be $\sigma_\eta = 0.10$ in accordance with the systematic flux uncertainty estimated by COHERENT [57]. From the selected sterile point and the best NSI hypothesis, we compute the corresponding BIC values and the relative likelihood defined in Eq. (18). By repeating this procedure across the scanned sterile-neutrino parameter space, we obtain a map of the regions in which the sterile benchmark can be distinguished from its best-fitting NSI counterpart for a given experimental configuration and binning choice.

B. Discrimination reach in benchmark parameter space

The resulting discrimination reach is shown in Fig. 1 for several representative binning choices. In each panel, the entries in braces indicate the numbers of bins used in baseline, recoil energy, and timing, respectively. In this case, the discrimination power is given by $\mathbb{L}_{\text{NSI}}$, which determines the color code in Fig. 1. The plotted relative likelihood weights therefore visualize how strongly a sterile-generated

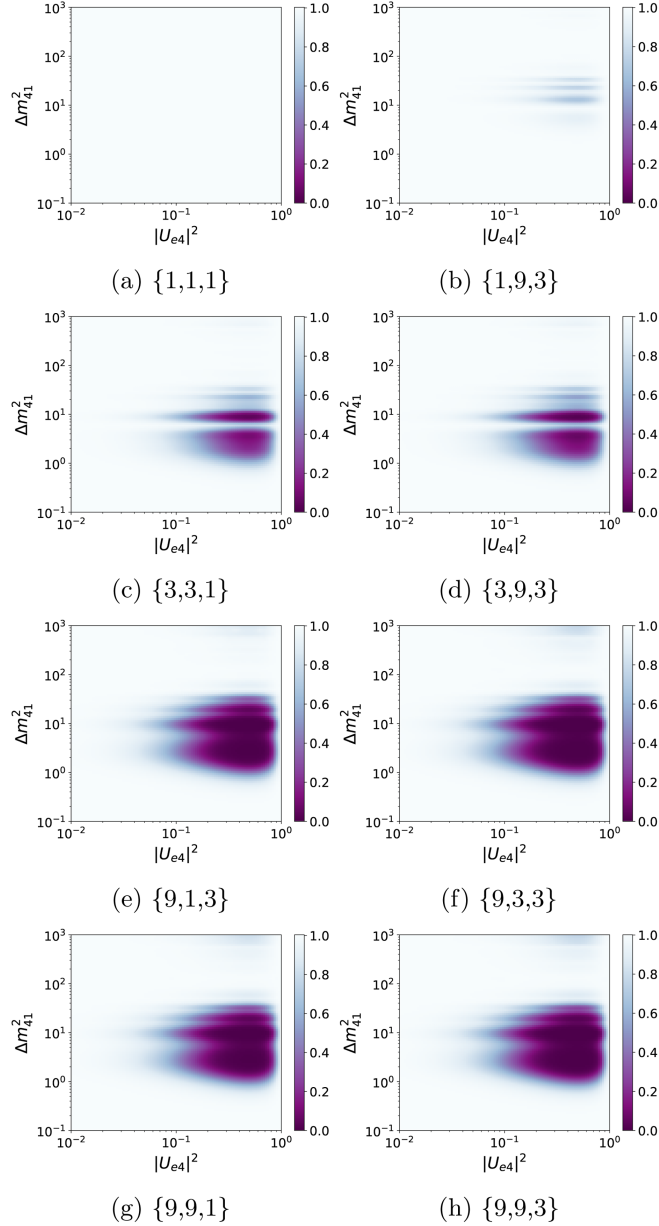


FIG. 1. Discrimination power between the NSI hypothesis and sterile-generated data in the ν_e -coupled sterile benchmark parameter space, shown for several binning configurations. The entries in braces indicate the numbers of bins in {Length, Energy, Time}.

dataset favors the sterile interpretation over the best-fitting NSI alternative across the benchmark parameter space. Equivalently, darker regions correspond to smaller values of $\mathbb{L}_{\text{NSI}}$ and hence to stronger discrimination against the NSI hypothesis.

Several qualitative trends are immediately apparent. As expected, the fully inclusive $\{1, 1, 1\}$ binning provides essentially no discriminatory power, since once all events are collapsed into a single counting bin, the sterile and NSI hypotheses can mimic one another modulo the level of the total event rate. In contrast, once multidimensional shape

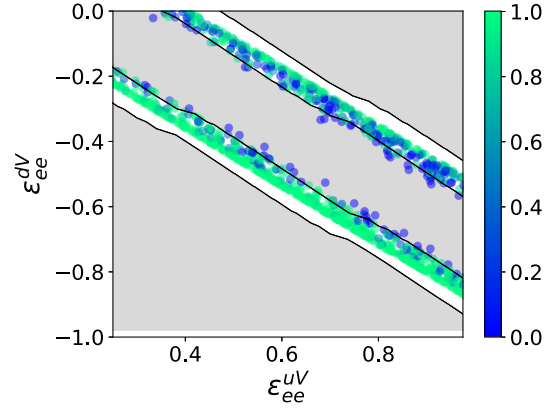


FIG. 2. Mappings of best-fit NSI points from the sterile-neutrino model points sampled in Fig. 1(h), colored according to their $\mathbb{L}_{\text{NSI}}$ values compared against their corresponding sterile-neutrino datasets. The 99% excluded regions from a parameter scan of a SM projected dataset are shown in gray for reference.

information is retained, the oscillatory structure characteristic of the sterile-neutrino scenario cannot, in general, be reproduced by the NSI benchmark, and substantial regions of parameter space become distinguishable.

The improvement is especially pronounced when the binning in baseline and recoil energy is refined. This behavior is physically well motivated: in the sterile-neutrino case, the signal is governed by an oscillation pattern that depends directly on L/E_ν , so resolving the baseline and energy dependence is essential for exposing the characteristic departures from the NSI hypothesis. By comparison, increasing the number of timing bins yields a more modest improvement. Timing information remains useful, particularly because it partially separates the prompt and delayed source components, but in the benchmark setup studied here it is not as powerful a discriminator as the baseline and energy dependence.

Figure 2 shows a complementary view of the likelihood-based analysis by showing the points of the NSI parameter space that are selected as best-fits to their corresponding sterile-neutrino model points displayed in Fig. 1(h). The points are colored according to their corresponding values of $\mathbb{L}_{\text{NSI}}$. The excluded regions for an NSI parameter scan under a SM projected dataset are shown in gray at 99% confidence, for reference. The excluded regions exhibit a characteristic double band structure, showing two allowed regions that lie around lines in the parameter space for which the resulting modifications in Eq. (13) give similar rates to the SM. For sterile-neutrino model points with insignificant oscillation effects and very little rate deficit relative to SM, the best-fit NSI solutions tend to lie close to the central trajectory of one of the two allowed NSI bands. For these points, distinguishability is naturally challenging as very little indication of nonstandard effects are present in the data. By contrast, for sterile-neutrino model points with statistically more pronounced oscillatory structures, the

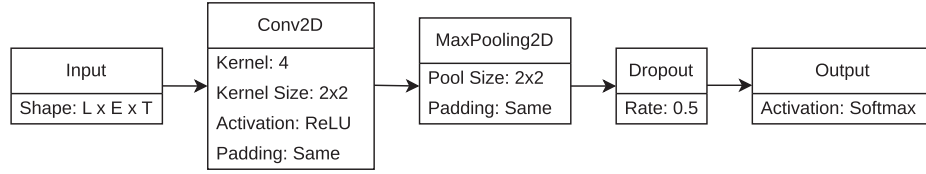


FIG. 3. The machine architecture used in this study. Batch normalization operations are also included between layers. The shape of the input layer is determined by the total number of bins, as indicated, while the shape of the output layer depends on the number of sorting classes.

best-fit NSI points are often displaced away from these central regions. This indicates that the parameter scans for those points are attempting to account for the total rate deficit imposed by the sterile-neutrino model via the NSI. The fact that many such points nevertheless yield very small values of $\mathbb{L}_{\text{NSI}}$ make clear that the distinguishability shown in Fig. 1 is not driven solely by inclusive rate effects, but also by shape information that the NSI cannot faithfully account for.

Taken together, these results show that even a conventional likelihood-based analysis can distinguish the two benchmark BSM scenarios in nontrivial regions of parameter space, provided that sufficient multidimensional information is retained. This establishes a useful baseline for the ML-based analyses presented in the following sections, where we examine whether more flexible inference methods can further enhance both model discrimination and parameter reconstruction.

IV. NEURAL NETWORK ARCHITECTURE AND TRAINING

The ML analyses that will be presented in Sec. VA and Sec. VB are based on a common neural-network architecture and training procedure, which we summarize here. In our study, the binned $\text{CE}\nu\text{NS}$ event distributions may be viewed as imagelike objects: the events are organized in bins of baseline, reconstructed recoil energy, and timing, which play roles analogous to the two pixel dimensions and three color channels of an image. This representation makes convolutional neural networks (CNNs) a natural choice for the classification tasks considered here, in close analogy with their widespread use in neutrino-event classification and reconstruction [26–33].²

For a given binning choice $\{n_L, n_E, n_T\}$ in baseline, energy, and time, each pseudodataset is represented as an array of shape $n_L \times n_E \times n_T$. The baseline-energy plane is treated as a two-dimensional image, while the timing bins are interpreted as color channels. To process these inputs, we employ a shallow CNN implemented in the *Keras*

deep-learning Application Programming Interface [58]. The network begins with an input layer matching the dimensions of the binned $\text{CE}\nu\text{NS}$ tensor, followed by a two-dimensional convolutional layer with 4 filters and a kernel size of 2×2 , using the ReLU activation function. Batch-normalization layers are inserted between successive stages to stabilize training. The convolutional stage is followed by a two-dimensional max-pooling layer, a dropout layer with rate 0.5 to curtail overtraining, and a final classification layer. For the binary-classification task in Sec. VA, the output layer contains two output nodes with softmax activation, corresponding to the sterile and NSI classes. For the multiclass task in Sec. VB, the number of output nodes is chosen to match the number of classes and a softmax activation is used. To ensure compatibility across different binning choices, padding (empty cells) is applied to the input when necessary so that the same network architecture can be used throughout the analysis. A schematic of the chosen architecture is shown in Fig. 3.

The training data are constructed from pseudodatasets generated around the mean event predictions of the benchmark model points. Starting from the mean $\text{CE}\nu\text{NS}$ distribution for a given point in parameter space, we generate large samples of fluctuated binned datasets, which are then labeled according to the underlying model hypothesis. In the current implementation, the fluctuations are applied independently to each bin using Poisson distributions. To focus the network on differences in multidimensional shape rather than on inclusive normalization, all pseudodatasets are renormalized to a common total event rate before being passed to the classifier. The ML analysis is therefore intentionally insensitive to total-rate information and instead probes the discriminating power contained in the baseline-energy-timing structure of the $\text{CE}\nu\text{NS}$ signal. In particular, we emphasize that this is not intended as performance comparison relative to the likelihood-based analysis, but rather as a complementary method that is demonstrated in this different context. The purpose is to emphasize what discrimination power is available to the neural network above and beyond standard approaches based on total rate.

Because the dominant flux-normalization uncertainty is largely removed by this normalization, the pseudodatasets are fluctuated according to Poisson distributions only, with the mean of the distribution used for each bin determined by the number of counts in each bin before renormalization

²Other network architectures (e.g., graph neural networks) could also be employed with appropriately processed data. However, comparing the performance of different network architectures is beyond the scope of this study.

to a common event rate. In this sense, the total event rate still has an influence on the neural network results via the significance of these bin-by-bin fluctuations, although model distinguishability is still being driven by shape effects and correlations in baseline, energy, and time.

The resulting ensemble of labeled pseudodatasets is then used for supervised training. The network is trained for 5 epochs in Sec. VA and 10 epochs in Sec. VB, with a validation split of 15%. We use Keras’s categorical cross-entropy loss function and the Adam optimizer during training. After training, the network performance is evaluated on an independently generated test sample of fluctuated pseudodatasets that is not used during training or validation. This is done to limit possible overfitting or bias. As mentioned earlier, the same architecture and training strategy are applied to the binary and multiclass classification problems described in Sec. VA and Sec. VB, respectively.

V. MACHINE-LEARNING-BASED MODEL DISCRIMINATION

We now present the results of the ML-based analyses introduced in the previous section. As discussed above, the binned CE ν NS event distributions are treated as imagelike inputs to a CNN, allowing the classifier to learn discriminating features directly from the multidimensional baseline-energy-timing structure of the signal. We consider two complementary tasks: binary classification between the sterile-neutrino and best-fit NSI hypotheses, and multiclass classification aimed at resolving different regions of the underlying parameter space.

A. Machine-learning-based binary classification

We begin with the binary-classification task, in which the network is trained to distinguish sterile-generated CE ν NS datasets from those generated under the corresponding best-fit NSI hypothesis. For each point in the scanned sterile-neutrino parameter space, an independent classifier is trained and tested, allowing the classification performance to be mapped directly onto the benchmark parameter space. In this way, the binary-classification analysis provides a measure of how readily the sterile and NSI scenarios can be separated using the shape information contained in the binned event distributions alone.

For each point in the scanned sterile-neutrino parameter space, we first identify the corresponding best-fit NSI point again using MultiNest [54–56]. Following the pseudodata generation procedure described in Sec. IV, we then construct training samples for the binary task from 10^4 fluctuated datasets generated under the sterile benchmark point and an equal number generated under its best-fit NSI counterpart, yielding a balanced set of training samples. A separate neural network is trained for each sterile benchmark point. The trained network is evaluated on an

independently generated held-out test sample consisting of equal numbers of sterile-generated and NSI-generated pseudodatasets. The classification accuracy is then defined in the standard way as the fraction of correctly classified examples in this balanced binary test set,

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{test}}} = \frac{N_{\text{sterile} \rightarrow \text{sterile}} + N_{\text{NSI} \rightarrow \text{NSI}}}{N_{\text{sterile}} + N_{\text{NSI}}}, \quad (23)$$

where $N_{\text{sterile} \rightarrow \text{sterile}}$ and $N_{\text{NSI} \rightarrow \text{NSI}}$ denote the numbers of correctly classified sterile and NSI test samples, respectively. In the present setup, the held-out test sample contains 100 pseudodatasets from each class, so that $N_{\text{test}} = 200$. An accuracy of 0.5 therefore corresponds to no effective discrimination on the balanced test set (i.e., random guessing), while an accuracy of 1.0 corresponds to perfect separation between the sterile and NSI hypotheses. Higher accuracy therefore indicates that the trained network more reliably distinguishes sterile-generated datasets from those generated under the best-fit NSI alternative, (or vice versa). The resulting classification accuracies across the sterile parameter space are shown in Fig. 4 for several representative binning choices, labeled by the numbers of bins in {baseline, energy, time}. The classification accuracy shown by the gray color scale ranges from 0.5 to 1.0.

Overall, the binary-classification maps exhibit qualitative trends similar to those found in the likelihood-based analysis, providing a useful cross-check that the network is responding to the same underlying physical structure of the signal. More importantly, however, the present results demonstrate that substantial model-discrimination power remains even after the total event rate has been explicitly removed from the input. The CNN is therefore not relying on inclusive normalization differences, but is instead learning features associated with the multidimensional shape of the CE ν NS distribution.

This point is particularly significant in the present context. In the sterile-neutrino benchmark, the dominant departures from the NSI hypothesis arise through correlated distortions in the baseline and recoil-energy dependence of the signal, with timing information playing a secondary but still useful role. The fact that the classifier achieves high accuracy over substantial regions of parameter space indicates that these oscillation-induced shape features are sufficiently distinctive to support robust model separation on their own. From this perspective, the ML-based binary analysis should be viewed not merely as a repetition of the likelihood-based study, but as evidence that the discriminatory information is genuinely encoded in the shape of the data. This is encouraging for applications in which normalization uncertainties are large or challenging to control (for example, at reactor neutrino sources), since it suggests that rate-insensitive inference strategies may still provide meaningful separation between competing new-physics scenarios.

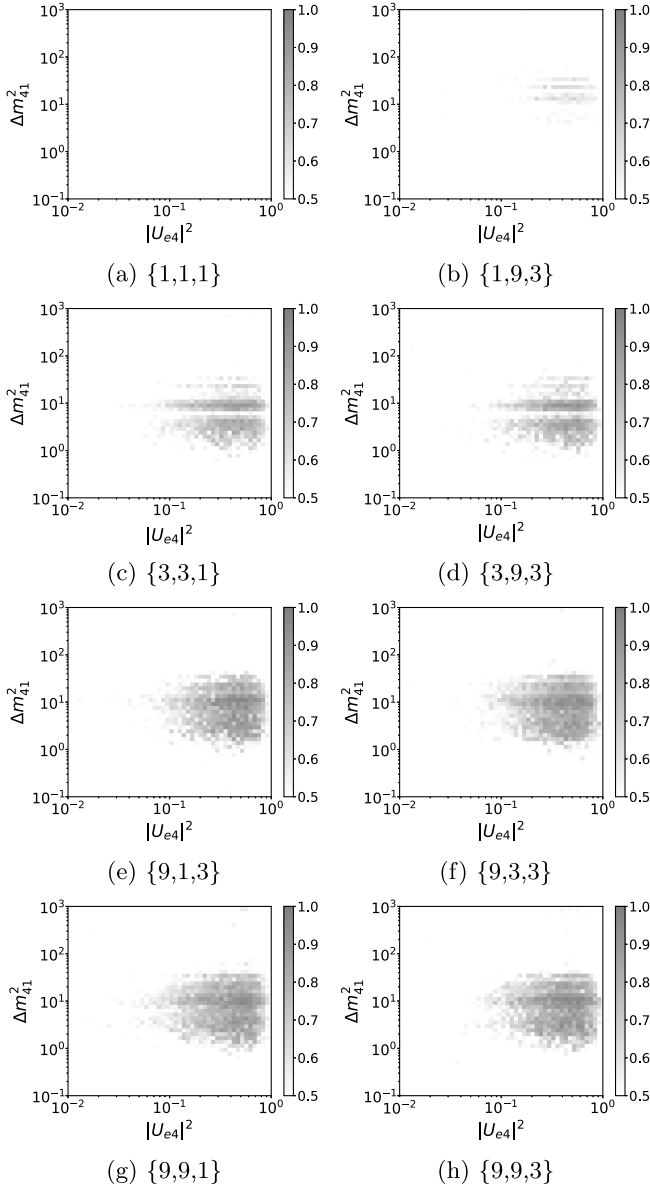


FIG. 4. Classification accuracy for the binary classification between NSI data and data from points of the ν_e -coupled sterile model space. Several binning configurations are shown, indicating the number of bins in each dimension as {Length, Energy, Time}. The accuracy is shown ranging from 0.5 (random guessing) to 1.0 (perfect accuracy).

One may also consider the reverse construction, namely generating pseudodatasets from selected points in the NSI parameter space, determining the corresponding best-fit sterile-neutrino points, and then training the network to discriminate between fluctuations of each NSI benchmark and its sterile counterpart. We do not pursue that analysis in this study. In the restricted benchmark setup adopted here, the reverse exercise is expected to be less informative than the sterile-generated study shown in Fig. 4. This is because for many of the NSI points considered here, the dominant observable effect is a change in the effective ν_e scattering

rate, which can often be approximately mimicked by a sterile-neutrino hypothesis if the oscillation pattern is sufficiently washed out over the experimentally accessible range. In such cases, the reverse scan would be dominated by regions with limited shape-based discrimination, while requiring a substantially larger computational cost. For this reason, we focus on the sterile-generated construction, where the physically distinctive oscillation structure provides the more informative benchmark for assessing model discernment. This should be understood as a practical limitation of the present benchmark study, rather than a general statement about arbitrary sterile-neutrino and NSI scenarios.

B. Machine multiclassification

We next consider a more demanding neural-network task, namely multiclass classification within the sterile-neutrino parameter space itself. As in the binary-classification study, pseudodatasets are generated from benchmark points in the chosen sterile parameter window following the procedure described in Sec. IV. The essential difference is that, rather than training the network to distinguish the sterile hypothesis from the best-fit NSI alternative, we now train a single network to distinguish among multiple cells within the sterile parameter space. In this way, the analysis moves beyond model discrimination and addresses a more refined question: whether the multidimensional CE ν NS event distribution contains enough information for the network to identify the approximate location of the underlying sterile-neutrino parameter point.

To carry out the multiclassification study, we divide the chosen sterile-neutrino parameter window into a 20×20 grid and take the center of each cell as a representative benchmark point. For every cell, we generate 10^4 fluctuated pseudodatasets for training, and after training, we evaluate the network on an independently generated test sample containing 10^3 pseudodatasets for that same cell. In testing, we have found that parameter reconstruction requires greater statistics than the model selection of Sec. VA. To allow for effective selection between similar oscillation signals, a lower fractional fluctuation magnitude of the pseudodataset bins is required than in the case of selection between oscillated and unoscillated signals as in Sec. VA. Since the bins are fluctuated according to Poisson distributions, this is achieved by increasing the exposure in this scenario. Compared to Sec. III and Sec. VA, in this scenario we increase N_{POT} and the detector masses by a factor of 10 each.

In contrast to the binary-classification task discussed in Sec. VA, the purpose here is not simply to decide between two competing hypotheses, but rather to determine how well the network can identify the approximate location of the underlying sterile-neutrino parameter point. Accordingly, the performance measure is based on the average distance between the true cell and the cell selected

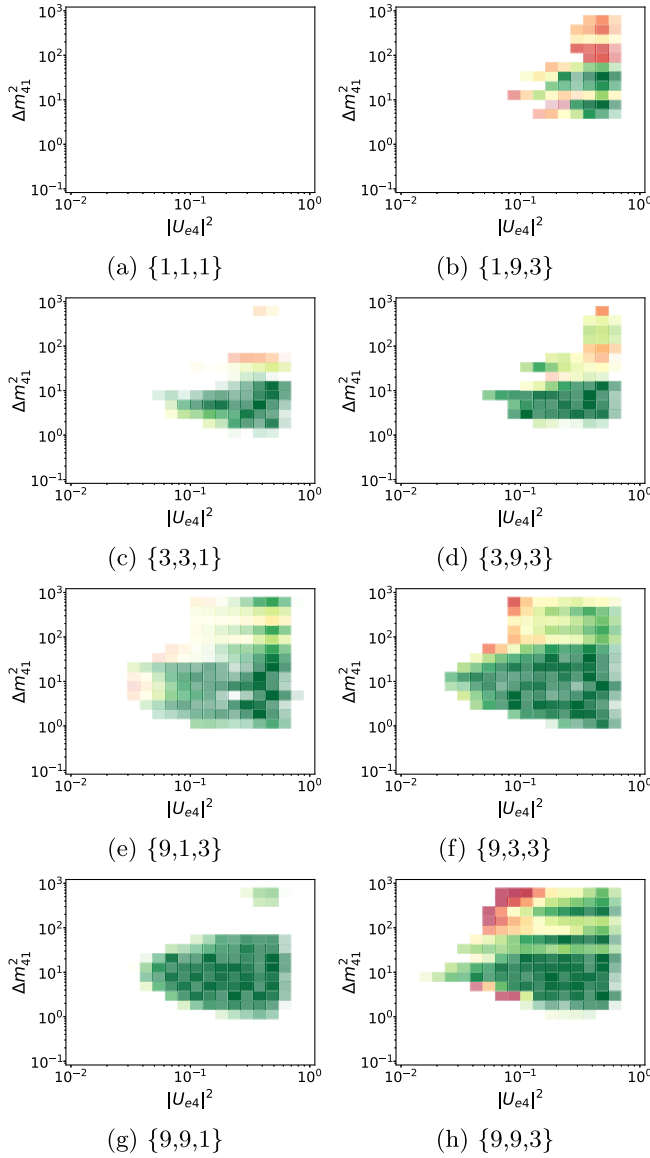


FIG. 5. Multiclass localization accuracy in the ν_e -coupled sterile-neutrino parameter space, with opacity indicating horizontal error in $|U_{e4}|^2$ and color indicating vertical error in Δm_{41}^2 (shown for cells with average activation > 0.05). Several binning configurations are shown, indicating the number of bins in each dimension as $\{\text{Length, Energy, Time}\}$.

by the network, computed separately along the horizontal axis ($|U_{e4}|^2$) and the vertical axis (Δm_{41}^2).

To display the multiclassification results in a compact and interpretable way, we retain only those cells for which the average node activation of the network's preferred class exceeds a minimum threshold. In practice, this threshold is introduced as a visualization criterion to suppress low-confidence assignments that would otherwise populate the map with noisy or weakly informative cells. We choose a value of 0.05, which was found empirically to provide a reasonable balance between retaining coverage of the

parameter space and highlighting regions where the classifier exhibits meaningful localization performance.³ Each surviving cell is then assigned an opacity and a color that summarize the localization performance. The opacity encodes the average horizontal displacement: full opacity corresponds to perfect reconstruction in $|U_{e4}|^2$, while zero opacity indicates an average error of three cells or more. The color similarly encodes the average vertical displacement in Δm_{41}^2 , with green indicating perfect reconstruction and red indicating an average error of three cells or more.

Results of this analysis for a suite of different binning choices are shown in Fig. 5. Several qualitative features are immediately apparent. First, once baseline and recoil-energy binning are introduced, nontrivial regions of the sterile parameter space begin to show useful localization performance. In particular, the opacity patterns indicate that the network can recover the horizontal direction, corresponding to $|U_{e4}|^2$, with reasonable fidelity over substantial regions of parameter space, while the cell colors show that the vertical direction, corresponding to Δm_{41}^2 , is in general more difficult to reconstruct. This behavior is physically sensible. The mixing parameter primarily controls the overall strength of the oscillation-induced distortion, whereas the mass-squared splitting governs the oscillation scale itself; the latter is more easily degraded by recoil-energy smearing, finite binning, and partial averaging of the oscillation pattern.

Second, the most successful reconstruction is obtained in intermediate and relatively large-mixing regions, where the sterile-induced distortions are strong enough to imprint a recognizable multidimensional pattern on the CE ν NS distribution. By contrast, in regions with smaller mixing or in parts of the large- Δm_{41}^2 regime, the network often fails to achieve confident localization, as reflected either by reduced opacity or by the absence of plotted cells after the activation threshold is applied. In this sense, the blank or weakly shaded regions should be interpreted not merely as poorer classification performance, but as regions where the available observables do not support robust parameter-space identification with the present setup.

Finally, the overall trend with binning again highlights the importance of multidimensional shape information. Increasing the granularity in baseline and recoil energy leads to the most significant gains, while timing information provides an additional but subleading improvement. The multiclassification results therefore reinforce the broader conclusion of this work: the stopped-pion CE ν NS signal contains sufficient structured information not only to discriminate between benchmark BSM scenarios, but also, in favorable regions of parameter space,

³The qualitative features of the multiclassification results remain stable under modest variations of this visualization threshold around the chosen value.

to localize the underlying sterile-neutrino parameters themselves.

VI. DISCUSSION AND CONCLUSION

In this work, we have investigated how stopped-pion CE ν NS experiments can be used not only to detect deviations from the standard three-neutrino expectation, but also to discriminate among candidate new-physics explanations and to infer the underlying parameter region responsible for an observed signal. Using the $3 + 1$ sterile-neutrino scenario and neutral-current NSI as benchmark frameworks, we have shown that the correlated baseline, recoil-energy, and timing structure of the CE ν NS signal provides nontrivial discriminatory power beyond what can be obtained from inclusive event rates alone.

At the level of model discrimination, our likelihood-based analysis established a useful baseline by quantifying the regions of sterile parameter space for which sterile-generated datasets can be distinguished from their best-fit NSI counterparts. As expected, discrimination is weak when the data are reduced to a single counting bin, but becomes substantially stronger once multidimensional shape information is retained. In particular, the baseline and recoil-energy structure of the signal play the dominant role, while timing provides an additional but comparatively more modest handle.

The ML analyses reinforce and extend this picture. In the binary classification study, we found that a CNN trained on rate-renormalized pseudodatasets is still able to separate the sterile and NSI benchmarks over substantial regions of parameter space. This is an important observation, because it shows that the distinguishing information is not merely a consequence of normalization differences, but is genuinely encoded in the multidimensional shape distortions of the CE ν NS event distribution. In this sense, the ML approach should be viewed not simply as a cross-check of the likelihood analysis, but as evidence that rate-insensitive inference strategies can retain significant discriminatory power even in the presence of large normalization uncertainties.

The multiclassification analysis further demonstrates that, in favorable regions of sterile parameter space, the CE ν NS signal contains enough structured information not only for model discrimination but also for approximate parameter-space localization. Although this reconstruction is necessarily coarser and more challenging than binary separation, the results show that the same baseline-energy-timing correlations that support model discernment can also be used to identify the approximate location of the underlying sterile benchmark point. The reconstruction is generally more robust in the $|U_{e4}|^2$ direction than in the Δm_{41}^2 direction, which is physically consistent with the differing

ways in which these parameters control the strength and scale of the oscillation pattern.

Taken together, these results provide a proof of concept that ongoing CE ν NS experiments with multiple detectors at stopped-pion sources offer a promising arena not only for testing the presence of nonstandard neutrino-sector physics, but also for interpreting its origin. More broadly, the methods developed here illustrate how conventional likelihood-based analyses and ML-based inference can play complementary roles in moving from anomaly detection to physics interpretation. Such strategies may become increasingly valuable in future neutrino and dark-sector searches, where low statistics, sizable systematic uncertainties, and partial degeneracies among candidate models can otherwise obscure the underlying physics.

There are several natural directions for future work. On the phenomenological side, the benchmark study considered here can be extended to broader parameter spaces and to additional classes of new-physics scenarios, including light dark matter signals that may mimic or compete with sterile-neutrino and NSI interpretations. Within the restricted benchmark setup adopted here, we have focused on sterile-generated pseudodatasets, since the reverse construction based on NSI-generated samples is expected to be less informative while incurring substantially greater computational cost; a more symmetric treatment of both directions would nevertheless be worthwhile in broader future studies. On the methodological side, it will be important to incorporate more realistic detector effects, systematic-uncertainty treatments, and alternative network architectures, as well as to explore direct parameter-regression strategies beyond the classification-based framework adopted here. Finally, because the present analysis makes clear which observables most strongly enhance model discernment, these methods may also help inform the design of the ongoing and proposed CE ν NS detector configurations by clarifying the detector capabilities and source characteristics needed to achieve robust discrimination between competing new-physics hypotheses.

ACKNOWLEDGMENTS

The work of I. A. B. and B. D. is supported by the U.S. Department of Energy Grant No. DE-SC0010813. The work of J. W. W. is supported by the National Science Foundation Grant No. PHY-2412995.

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

- [1] M. S. Athar, S. W. Barwick, T. Brunner, J. Cao, M. Danilov, K. Inoue *et al.*, Status and perspectives of neutrino physics, *Prog. Part. Nucl. Phys.* **124**, 103947 (2022).
- [2] X. Qian and J.-C. Peng, Physics with reactor neutrinos, *Rep. Prog. Phys.* **82**, 036201 (2019).
- [3] M. A. Acero *et al.*, White paper on light sterile neutrino searches and related phenomenology, *J. Phys. G* **51**, 120501 (2024).
- [4] I. A. Bisset, B. Dutta, W.-C. Huang, and L. E. Strigari, Short baseline neutrino anomalies at stopped pion experiments, *J. High Energy Phys.* **10** (2024) 003.
- [5] P. S. Dev *et al.*, Neutrino non-standard interactions: A status report, *SciPost Phys. Proc.* **2**, 001 (2019).
- [6] D. Akimov *et al.* (COHERENT Collaboration), Observation of coherent elastic Neutrino-Nucleus scattering, *Science* **357**, 1123 (2017).
- [7] D. Akimov *et al.* (COHERENT Collaboration), First measurement of coherent elastic neutrino-nucleus scattering on argon, *Phys. Rev. Lett.* **126**, 012002 (2021).
- [8] D. Akimov *et al.* (COHERENT Collaboration), Measurement of the coherent elastic neutrino-nucleus scattering cross section on CsI by COHERENT, *Phys. Rev. Lett.* **129**, 081801 (2022).
- [9] S. Adamski *et al.* (COHERENT Collaboration), Evidence of coherent elastic neutrino-nucleus scattering with coherent's germanium array, *Phys. Rev. Lett.* **134**, 231801 (2025).
- [10] A. A. Aguilar-Arevalo *et al.* (CCM Collaboration), First dark matter search results from coherent CAPTAIN-mills, *Phys. Rev. D* **106**, 012001 (2022).
- [11] B. Dutta, D. Kim, S. Liao, J.-C. Park, S. Shin, and L. E. Strigari, Dark matter signals from timing spectra at neutrino experiments, *Phys. Rev. Lett.* **124**, 121802 (2020).
- [12] B. Dutta, D. Kim, S. Liao, J.-C. Park, S. Shin, L. E. Strigari, and A. Thompson, Searching for dark matter signals in timing spectra at neutrino experiments, *J. High Energy Phys.* **01** (2022) 144.
- [13] P. Coloma, P. B. Denton, M. C. Gonzalez-Garcia, M. Maltoni, and T. Schwetz, Curtailing the dark side in non-standard neutrino interactions, *J. High Energy Phys.* **04** (2017) 116.
- [14] P. B. Denton and J. Gehrlein, A statistical analysis of the coherent data and applications to new physics, *J. High Energy Phys.* **04** (2021) 266.
- [15] D. Akimov *et al.* (COHERENT Collaboration), First probe of Sub-GeV dark matter beyond the cosmological expectation with the COHERENT CsI detector at the SNS, *Phys. Rev. Lett.* **130**, 051803 (2023).
- [16] A. A. Aguilar-Arevalo *et al.* (CCM Collaboration), First leptophobic dark matter search from the coherent-CAPTAIN-Mills liquid argon detector, *Phys. Rev. Lett.* **129**, 021801 (2022).
- [17] A. Aguilar *et al.* (LSND Collaboration), Evidence for neutrino oscillations from the observation of $\bar{\nu}_e$ appearance in a $\bar{\nu}_\mu$ beam, *Phys. Rev. D* **64**, 112007 (2001).
- [18] A. A. Aguilar-Arevalo *et al.* (MiniBooNE Collaboration), Unexplained excess of electronlike events from a 1-GeV neutrino beam, *Phys. Rev. Lett.* **102**, 101802 (2009).
- [19] A. A. Aguilar-Arevalo *et al.* (MiniBooNE Collaboration), Significant excess of ElectronLike events in the MiniBooNE Short-Baseline neutrino experiment, *Phys. Rev. Lett.* **121**, 221801 (2018).
- [20] A. A. Aguilar-Arevalo *et al.* (MiniBooNE Collaboration), Updated MiniBooNE neutrino oscillation results with increased data and new background studies, *Phys. Rev. D* **103**, 052002 (2021).
- [21] V. V. Barinov *et al.*, Search for electron-neutrino transitions to sterile states in the BEST experiment, *Phys. Rev. C* **105**, 065502 (2022).
- [22] J. N. Abdurashitov *et al.* (SAGE Collaboration), Measurement of the response of a gallium metal solar neutrino experiment to neutrinos from a ^{51}Cr source, *Phys. Rev. C* **59**, 2246 (1999).
- [23] W. Hampel *et al.* (GALLEX Collaboration), Final results of the ^{51}Cr neutrino source experiments in GALLEX, *Phys. Lett. B* **420**, 114 (1998).
- [24] P. Abratenko *et al.* (MicroBooNE Collaboration), Search for light sterile neutrinos with two neutrino beams at MicroBooNE, *Nature (London)* **648**, 64 (2025).
- [25] C. Backhouse and R. B. Patterson, Library event matching event classification algorithm for electron neutrino interactions in the NO ν A detectors, *Nucl. Instrum. Methods Phys. Res., Sect. A* **778**, 31 (2015).
- [26] A. Aurisano, A. Radovic, D. Rocco, A. Himmel, M. D. Messier, E. Niner, G. Pawloski, F. Psihas, A. Sousa, and P. Vahle, A convolutional neural network neutrino event classifier, *J. Instrum.* **11**, P09001 (2016).
- [27] R. Acciarri *et al.* (MicroBooNE Collaboration), Convolutional neural networks applied to neutrino events in a liquid argon time projection chamber, *J. Instrum.* **12**, P03011 (2017).
- [28] C. Adams *et al.* (MicroBooNE Collaboration), Deep neural network for pixel-level electromagnetic particle identification in the MicroBooNE liquid argon time projection chamber, *Phys. Rev. D* **99**, 092001 (2019).
- [29] P. Baldi, J. Bian, L. Hertel, and L. Li, Improved energy reconstruction in NO ν A with regression convolutional neural networks, *Phys. Rev. D* **99**, 012011 (2019).
- [30] S. Aiello *et al.* (KM3NeT Collaboration), Event reconstruction for KM3NeT/ORCA using convolutional neural networks, *J. Instrum.* **15**, P10005 (2020).
- [31] B. Abi *et al.* (DUNE Collaboration), Neutrino interaction classification with a convolutional neural network in the DUNE far detector, *Phys. Rev. D* **102**, 092003 (2020).
- [32] P. Abratenko *et al.* (MicroBooNE Collaboration), A convolutional neural network for multiple particle identification in the MicroBooNE liquid argon time projection chamber, *Phys. Rev. D* **103**, 092003 (2021).
- [33] P. Abratenko *et al.* (MicroBooNE Collaboration), Semantic segmentation with sparse convolutional neural network for event reconstruction in MicroBooNE, *Phys. Rev. D* **103**, 052012 (2021).
- [34] IceCube Collaboration, Search for an eV-scale sterile neutrino using improved high-energy ν_μ event reconstruction in IceCube, *Phys. Rev. Lett.* **133** (2024).
- [35] J. Villarreal, J. Woodward, J. Hardin, and J. Conrad, Machine learning-informed 3 + 1 sterile neutrino global fits using posterior density estimation of electron disappearance data, *Eur. Phys. J. C* **86**, 326 (2026).

- [36] F. Psihas, M. Groh, C. Tunnell, and K. Warburton, A review on machine learning for neutrino experiments, *Int. J. Mod. Phys. A* **35**, 2043005 (2020).
- [37] D. Kim, K. Kong, K. T. Matchev, M. Park, and P. Shyamsundar, Deep-learned event variables for collider phenomenology, *Phys. Rev. D* **107**, L031904 (2023).
- [38] R. Franceschini, D. Kim, K. Kong, K. T. Matchev, M. Park, and P. Shyamsundar, Kinematic variables and feature engineering for particle phenomenology, *Rev. Mod. Phys.* **95**, 045004 (2023).
- [39] P. Agrawal, N. Craig, A. Madden, and I. V. Lombera, The FERMIACC: Agents for particle theory, [arXiv:2603.22538](https://arxiv.org/abs/2603.22538).
- [40] E. A. Moreno, S. Bright-Thonney, A. Novak, D. Garcia, and P. Harris, AI agents can already autonomously perform experimental high energy physics, [arXiv:2603.20179](https://arxiv.org/abs/2603.20179).
- [41] S. Qiu, Z. Cai, J. Wei, Z. Li, Y. Yin, Q.-H. Cao, C. Liu, M.-x. Luo, X.-B. Yuan, and H. X. Zhu, An End-to-end architecture for collider physics and beyond, [arXiv:2603.14553](https://arxiv.org/abs/2603.14553).
- [42] S. R. Soleti, P. Coloma, J. J. Gómez Cadenas and A. Cabrera (SHiNESS Collaboration), Search for hidden neutrinos at the European Spallation Source: The SHiNESS experiment, *J. High Energy Phys.* **03** (2024) 148.
- [43] M. Mirzakhani *et al.*, MINER reactor based search for axionlike particles using sapphire (Al₂O₃) detectors, *Phys. Rev. D* **112**, 032013 (2025).
- [44] V. Iyer *et al.*, Large mass single electron resolution detector for dark matter and neutrino elastic interaction searches, *Nucl. Instrum. Methods Phys. Res., Sect. A* **1010**, 165489 (2021).
- [45] M. Adhikari *et al.* (COHERENT Collaboration), The COHERENT experiment: 2026 update, [arXiv:2602.15652](https://arxiv.org/abs/2602.15652).
- [46] R. Van de Water, Probing the dark sector with accelerators: New opportunities!, *Talk at Interplay of Nuclear, Neutrino and BSM Physics at Low-Energies, INT, Seattle* (2023), <https://www.int.washington.edu/program/schedule/1205>.
- [47] Y. Y. W. Blokland and M. Radaideh, Machine learning to improve sns operations, https://frib.msu.edu/sites/default/files/_files/pdfs/science_ase_aset_seminars_Blokland.pdf.
- [48] D. Z. Freedman, Coherent effects of a weak neutral current, *Phys. Rev. D* **9**, 1389 (1974).
- [49] K. Scholberg, Prospects for measuring coherent neutrino-nucleus elastic scattering at a stopped-pion neutrino source, *Phys. Rev. D* **73**, 033005 (2006).
- [50] J. A. Formaggio and G. P. Zeller, From eV to EeV: Neutrino cross sections across energy scales, *Rev. Mod. Phys.* **84**, 1307 (2012).
- [51] R. H. Helm, Inelastic and elastic scattering of 187-Mev electrons from selected even-even nuclei, *Phys. Rev.* **104**, 1466 (1956).
- [52] B. Dutta, Y. Gao, R. Mahapatra, N. Mirabolfathi, L. E. Strigari, and J. W. Walker, Sensitivity to oscillation with a sterile fourth generation neutrino from ultra-low threshold neutrino-nucleus coherent scattering, *Phys. Rev. D* **94**, 093002 (2016).
- [53] A. E. Raftery, Bayesian model selection in social research, *Sociol. Methodol.* **25**, 111 (1995).
- [54] F. Feroz and M. P. Hobson, Multimodal nested sampling: An efficient and robust alternative to MCMC methods for astronomical data analysis, *Mon. Not. R. Astron. Soc.* **384**, 449 (2008).
- [55] F. Feroz, M. P. Hobson, and M. Bridges, MultiNest: An efficient and robust Bayesian inference tool for cosmology and particle physics, *Mon. Not. R. Astron. Soc.* **398**, 1601 (2009).
- [56] F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt, Importance nested sampling and the MultiNest algorithm, *Open J. Astrophys.* **2**, 10 (2019).
- [57] D. Akimov *et al.* (COHERENT Collaboration Collaboration), Simulating the neutrino flux from the spallation neutron source for the coherent experiment, *Phys. Rev. D* **106**, 032003 (2022).
- [58] F. Chollet *et al.*, Keras, <https://keras.io> (2015).