Bayesian analysis of a future β decay experiment's sensitivity to neutrino mass scale and ordering

A. Ashtari Esfahani,[1] M. Betancourt,[2] Z. Bogorad,[3] S. Böser,[4] N. Buzinsky,[3] R. Cervantes,[1] C. Claessens,[4] L. de Viveiros,[5] M. Fertl,[4] J. A. Formaggio,[3] L. Gladstone,[6] M. Grando,[7] M. Guigue,[8,*] J. Hartse,[1] K. M. Heeger,[9] X. Huyan,[7] J. Johnston,[3] A. M. Jones,[7] K. Kazkaz,[10] B. H. LaRoque,[7] A. Lindman,[4] R. Mohiuddin,[6] B. Monreal,[6] J. A. Nikkel,[9] E. Novitski,[1] N. S. Oblath,[7] M. Ottiger,[1] W. Pettus,[1,11] R. G. H. Robertson,[1] G. Rybka,[1] L. Saldaña,[9] M. Schram,[7] V. Sibille,[3] P. L. Slocum,[9] Y.-H. Sun,[6] P. T. Surukuchi,[9] J. R. Tedeschi,[7] A. B. Telles,[9] M. Thomas,[7] T. Thümmler,[12] L. Tvrznikova,[10,†] B. A. VanDevender,[1,7] T. E. Weiss,[3,9,‡] T. Wendler,[5,§] E. Zayas,[3] and A. Ziegler[5]

[1]*Center for Experimental Nuclear Physics and Astrophysics and Department of Physics, University of Washington, Seattle, Washington 98195, USA*

[2]*Symplectomorphic, LLC, New York, New York 10026, USA*

[3]*Laboratory for Nuclear Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

[4]*Institut für Physik, Johannes Gutenberg-Universität Mainz, 55128 Mainz, Germany*

[5]*Department of Physics, Pennsylvania State University, University Park, Pennsylvania 16802, USA*

[6]*Department of Physics, Case Western Reserve University, Cleveland, Ohio 44106, USA*

[7]*Pacific Northwest National Laboratory, Richland, Washington 99354, USA*

[8]*Laboratoire de Physique Nucléaire et de Hautes Énergies, Sorbonne Université, Université de Paris, CNRS/IN2P3, Paris, France*

[9]*Wright Laboratory, Department of Physics, Yale University, New Haven, Connecticut 06520, USA*

[10]*Lawrence Livermore National Laboratory, Livermore, California 94550, USA*

[11]*Department of Physics, Indiana University, Bloomington, Indiana 47405, USA*

[12]*Institute for Astroparticle Physics, Karlsruhe Institute of Technology, 76021 Karlsruhe, Germany*

(Received 24 December 2020; accepted 12 April 2021; published 1 June 2021)

Bayesian modeling techniques enable sensitivity analyses that incorporate detailed expectations regarding future experiments. A model-based approach also allows one to evaluate inferences and predicted outcomes, by calibrating (or measuring) the consequences incurred when certain results are reported. We present procedures for calibrating predictions of an experiment's sensitivity to both continuous and discrete parameters. Using these procedures and a new Bayesian model of the β-decay spectrum, we assess a high-precision β-decay experiment's sensitivity to the neutrino mass scale and ordering for one assumed design scenario. We find that such an experiment could measure the electron-weighted neutrino mass within $\sim$40 meV after 1 year (90% credibility). Neutrino masses $>$500 meV could be measured within $\approx$5 meV. Using only β decay and external reactor neutrino data, we find that next-generation β-decay experiments could potentially constrain the mass ordering using a two-neutrino spectral model analysis. By calibrating mass ordering results, we identify reporting criteria that can be tuned to suppress false ordering claims. In some cases, a two-neutrino analysis can reveal that the mass ordering is inverted, an unobtainable result for the traditional one-neutrino analysis approach.

DOI: 10.1103/PhysRevC.103.065501

I. INTRODUCTION

Model-based simulation is a standard tool for informing the design of physics experiments and predicting their outcomes [1]. Such model-based approaches allow one to incorporate detailed expectations regarding future data by performing pseudoexperiments that reflect the span of possible experimental and physical parameter values. In Bayesian sensitivity studies, specifically, those parameter values are weighted by prior probabilities. By contrast, computing and reporting predicted outcomes for "best guess" values ignores information by excluding regions of parameter space.

Moreover, inferential models lend themselves to procedures for investigating the consequences of assumptions made during analysis. Bayesian methods, in particular, illuminate the effects of assumptions underlying inference (i.e., extracting information from data) and decision making (i.e., claiming results based on inferences) by decoupling the two processes. Thus, when assessing an experiment's sensitivity, one can quantify, or *calibrate*, the expected success or accuracy of procedures that one plans to use to both analyze data and report results in a certain format. It is also possible to perform conditional Bayesian calibration by fixing one or more parameters before simulating data [2–5].

Here we employ Bayesian modeling to perform a sensitivity study for a physics experiment. Among physicists, *sensitivity* typically denotes the level of precision with which

*mguigue@lpnhe.in2p3.fr

†Present address: Waymo LLC, Mountain View, CA 94043, USA.

‡talia.weiss@yale.edu

§Present address: Pacific Northwest National Laboratory, Richland, WA 99354, USA.

 ©2021 American Physical Society

TABLE I. Definitions are consistent with Particle Data Group descriptions [1] with the exception of the two definitions of "sensitivity," which capture a common but less standard usage. The last row describes how "sensitivity" is used in this paper.

Term	Definition	Notes
Credibility	Fraction of Bayesian posterior probability mass that falls within a reported interval	Result of a *single* real or simulated experiment
Coverage	Fraction of likely experiments for which the reported interval contains the true parameter value, within model assumptions	Result of *multiple* simulated experiments
Confidence interval	Interval constructed to have a coverage that equals or exceeds a chosen probability (or "confidence level")	Frequentist term; not used in this analysis
Sensitivity analysis	Study of how result precision and accuracy change under reasonable variation of all parameters, within model assumptions	Requires simulated experiments (pseudodata)
Sensitivity *(to very small parameter)*	Upper limit on a parameter, to some confidence level	Usage by the KATRIN experiment [8]
Sensitivity *(to parameter of any magnitude)*	Width of a posterior interval with a chosen credibility	Definition in this paper

experimenters can expect to resolve a parameter of interest, assuming a reasonably accurate measurement. (We adopt that usage here, though among statisticians, sensitivity can refer to how a decision-making process's accuracy depends on model parameters [2,3].) For physics experiments, in particular, Bayesian sensitivity methods allow researchers to capitalize on their often extensive knowledge of experimental configurations, physical processes, and expected uncertainties to construct priors. More broadly, model-based analyses offer potential tools for physicists to collectively interpret results and judge whether discovery claims are warranted [2,6] (see Sec. II). These tools thus provide possible alternatives to a 5σ confidence requirement.

To assess sensitivity, we develop a model of an experiment's measurement process, then employ that model to repeatedly generate and analyze pseudodata—where "analyze" means "infer posterior distributions." Parameters assumed for data generation are sampled from priors. Next, expectations and intervals are computed from the posteriors, yielding sensitivity results. Finally, we calculate how often these results reflect "true" values underlying the generated data (a calibration). In doing so, we quantify the consequences of our modeling and reporting assumptions. For relevant statistical term definitions, see Table I.

The above procedure is applied here to assess sensitivity to the neutrino mass scale and ordering. Neutrinos are produced in one of three flavor states, each of which interacts with electrons, muons or tau leptons. The discovery of neutrino oscillations demonstrated that each flavor state can be represented as a superposition of mass states with eigenvalues m_1, m_2, and m_3, at least two of which are nonzero [7,9,10]. While nuclear and particle physics experiments as well as cosmological models have placed upper bounds on the masses and measured the squared mass differences [1], the absolute neutrino mass scale is unknown. In addition, two orderings of the mass spectrum are possible: If $m_1 < m_2 < m_3$, then the masses are said to obey a *normal ordering,* while if $m_3 < m_1 < m_2$, then they follow an *inverted ordering*. Although recent data are beginning to shed light on the ordering question, it remains unanswered to date. Sensitivity to the ordering in oscillation experiments is discussed in Qian *et al.* [11].

A promising approach to resolving the mass scale involves analyzing the shape of the electron spectrum produced when nuclei β decay. This "direct mass measurement" method is so named because it depends chiefly on decay kinematics imposed by energy conservation. Direct mass experiments probe the electron-weighted neutrino mass $m_\beta = \sqrt{\sum_{i=1}^{3} |U_{ei}|^2 m_i^2}$ (hereafter "neutrino mass"), where U_{ei} are Pontecorvo-Maki-Nakagawa-Sakata matrix elements.[1] The size of m_β corresponds to a shift in the electron's maximal kinetic energy and causes a distortion in the β spectrum shape. A precise m_β measurement would determine the mass scale, and as a by-product, it could constrain the ordering at masses $\lesssim$48 meV—the 95% lower limit on the inverted ordering mass [1]. Furthermore, the β-decay shape depends distinctly on each m_i [12]. Thus, we propose that, if a β-decay experiment is sensitive to the fractional contributions of individual neutrino masses to the full spectral shape, then such information might enable a clearer mass ordering determination. By modeling the shape of a β spectrum, one can thus assess a direct mass experiment's sensitivity to the ordering—accounting for both the magnitude of m_β and finer spectral features (see Fig. 1).

In this paper, we develop a β-decay spectral model suited to Bayesian inference. The model uses a two-neutrino approximation (motivated by the fact that $\Delta m_{21}^2 \ll |\Delta m_{13}^2|$) and formulates the mass ordering question in terms of a parameter η, the fractional contribution of the lighter mass to the spectrum. Constraints on η are most directly accessible via reactor neutrino experiments. Thus, for a β-decay experiment to potentially resolve the mass ordering, the only external data needed for the analysis are reactor data. Current as well as future direct mass experiments could employ this spectral model to examine their sensitivity to the neutrino mass scale and ordering. As a case study, we use the model to assess sensitivity to these neutrino mass parameters for one possible

[1]For either neutrino mass ordering, $m_\beta = m_1$ to 1% accuracy for $m_1 \gtrsim 0.05$ eV. Hence, with knowledge of the ordering and splittings, an m_β measurement determines all three masses.

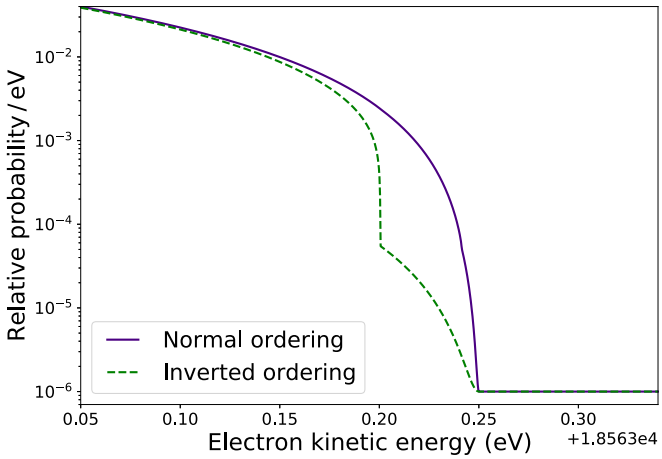

FIG. 1. A comparison of atomic tritium β-decay spectra for the two allowed neutrino mass orderings shows how a spectral shape analysis could be sensitive to the mass ordering. A background of 10^{-6}/eV is assumed.

design scenario of the Project 8 experiment, a high-precision β-decay experiment [13,14].

II. A MODEL-BASED APPROACH TO CALIBRATING SENSITIVITY RESULTS

Predictive analyses project whether, given some expected data, one will be able to report a particular result—for example, "m_β falls between 0.05 and 0.09 eV with 90% credibility" or "the mass ordering is normal." In Bayesian analysis, the *decision* of whether to claim a particular result occurs after the process of *inference*. Bayesian inference produces posterior distributions $\pi(\theta|y)$ for parameters θ given data y. Such inference exploits Bayes's rule $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$, where $\pi(y|\theta)$ is the likelihood of y given θ, and $\pi(\theta)$ are prior probability distributions on θ. Experimenters make claims about physics underlying their data by computing expectations (e.g., means and intervals) from posteriors.

In practice, there is no guarantee that the process by which one decides to claim a scientific result will perform well when faced with real data. To provide some assurance of the decision-making process' good performance, it is necessary to calibrate the process by evaluating it with respect to possible "model configurations," i.e., combinations of true parameter values.

Decision-making procedures are, in this context, cost-benefit analyses. A calibration requires an inferential loss (or utility) function that expresses the relative loss L incurred when an experimenter makes different reporting choices. In this work, L lies between 0 and 1. A common choice for L (used in Sec. II A) is a function that equals 0 if a credible or confidence interval obtained by analyzing a pseudodataset contains the true parameter value, or 1 if it does not. The expected loss is then estimated by finding the average loss $\bar{L}$ for a group of pseudodatasets. In the case just described, $\bar{L}$ would be the fraction of datasets for which the interval does *not* contain the true value. Given multiple reporting options, the experimenter should select the option with the smallest average loss over a group of pseudoexperiments [2,6]. (For example, this enables a decision of whether to report quantile or highest density intervals, as discussed further in Sec. II A.)

There is no one correct loss function for a given model, but the function should quantify the agreement or discrepancy between inputted and reported values. Given some loss function, model-based calibration then serves to compute how often one reports *accurate* results, across many pseudoexperiments with likely model configurations [2,6,15].

Frequentist calibration entails finding the *worst-case loss* over all model configurations. Such calibration requires tools like the Feldman-Cousins method, which addresses the fact that typical, Gaussian confidence intervals are inaccurate for bounded parameters, such as the positive neutrino mass [16,17]. This approach is too time-consuming to implement fully, as it requires that likelihood functions be computed and integrated for all reasonable parameter values (or a fine grid). While asymptotic approximations can make frequentist calibration computationally viable, they do not fully hold for the complex statistical models used in modern analyses [18,19].

Bayesian calibration, on the other hand, does not require that one determine the *worst-case loss*; instead, it entails finding the *expected loss with respect to the prior distribution.* This is a probabilistic calculation that can be readily implemented with sampling methods. In a Bayesian analysis, it is not necessary to consider all possible truths—only enough to accurately estimate expected losses [19]. Here we lay out Bayesian calibration procedures for sensitivity to the electron-weighted neutrino mass and mass ordering.

A. Calibrating neutrino mass sensitivity claims

The Bayesian result of a physics experiment will often be a posterior credible window—that is, the window within which some fraction of a parameter's posterior probability mass falls. This reporting scheme is sensible for continuous-domain parameters. If a posterior on m_β is inferred from a β spectrum, then experimenters can present their result as a credible window of neutrino masses (in eV). We call the width of this window "sensitivity to the neutrino mass." The reported mass window may consist of either an upper limit with a lower bound at zero, or a credible interval with upper and lower bounds. If posteriors are inferred for a large number of pseudodatasets, then one may predict an experiment's sensitivity by computing an expectation value (e.g., mean width or median width) from these credible windows. For a discussion of the frequentist and Bayesian perspectives underlying the use of confidence and credible intervals, respectively, see Ref. [2].

In the continuous-domain case, the loss function provides a method for computing the proportion of likely datasets for which a posterior interval contains the true parameter value. For an analysis of sensitivity to m_β, a calibration involves computing the fraction of pseudodatasets $C \equiv 1 - \bar{L}$ for which the credible window includes the true neutrino mass $\tilde{m}_\beta$, where $\bar{L}$ is the average loss for an ensemble of pseudoexperiments. ($\bar{L}$ serves to estimate the *expected loss* with respect to the prior distribution.) The fraction C is known as the Bayesian "model coverage," and it estimates the expected

accuracy of a sensitivity prediction.[2] We find C by repeatedly generating and analyzing data given an appropriate distribution of inputted $\tilde{m}_\beta$ values [2].

The calibration procedure is as follows:

(1) Develop a *generation* model of the data, and if necessary, a second *analysis* model. The latter may be approximate but is believed to adequately describe the data. Both models depend on a set of parameters θ (which includes m_β).
(2) Select "true" values $\tilde{\theta}$ by sampling from priors $\pi(\theta)$, which incorporate as much external knowledge as is reasonable.
(3) Generate spectral data $\tilde{y}$ using the generation model, with $\tilde{\theta}$ as inputs.
(4) Use the analysis model from #1 to infer a posterior $\pi(m_\beta|\tilde{y})$.
(5) Determine the posterior values ϑ that contain some fraction (credibility) α of the posterior probability mass on m_β. For a credible *interval*, calculate the loss function

$$L_{m_\beta} \equiv \begin{cases} 0, & \tilde{m}_\beta \in [\vartheta_{(1-\alpha)/2}, \vartheta_{(1+\alpha)/2}] \\ 1, & \text{otherwise} \end{cases}, \tag{1}$$

where upper and lower posterior bounds $\vartheta_{(1\pm\alpha)/2}$ are computed so that

$$\int_0^{\vartheta_{(1\pm\alpha)/2}} dm_\beta\, \pi(m_\beta|\tilde{y}) = \frac{1\pm\alpha}{2}. \tag{2}$$

That is, a fraction $(1\pm\alpha)/2$ of the posterior probability mass on m_β lies below the mass value $\vartheta_{(1\pm\alpha)/2}$. For a *limit*, the credible window is $[0, \vartheta_\alpha]$.
(6) Repeat steps 2–5 N_{trial} times. Each repetition constitutes a "pseudoexperiment."
(7) Compute C by subtracting the mean over resulting L_{m_β} values from 1. Potentially, adjust α to obtain a satisfying coverage—that is, to achieve an acceptable number of true- and false-positive results.

C may not equal α for all α; the relationship between these two values depends on the model and priors. The uncertainty on C is $\sqrt{C(1-C)/N_{\text{trial}}}$, assuming the number of true-positive results is binomially distributed.

A calibrated sensitivity result then consists of a projected (e.g., mean or median) credible window and its coverage. *It is necessary to sample all input values from priors before generation* (step 2). This creates an ensemble of many realistic datasets, where the probabilities of possible model configurations are weighted appropriately. If a model-based sensitivity analysis uses fixed generation inputs (or a grid of inputs, unweighted by prior probabilities), then it risks biasing results and under- or overestimating coverages. *It is also crucial to generate pseudodata that is as realistic as possible*, so that the coverage will reflect the potential consequences of all known assumptions made when devising the analysis model or choosing how to report results [2].

Note that expected fluctuations in the data itself (i.e., statistical uncertainties) are incorporated into priors used for both data generation and analysis—steps 3 and 4. By contrast, uncertainties representing a lack of clarity in one's knowledge of fixed parameters (i.e., systematic uncertainties) are incorporated into pregeneration sampling and analysis priors—steps 2 and 4.

In the above procedure, a choice of credibility α does not uniquely define a credible window, because one must also select the window's central value. A straightforward choice of window is the quantile interval, which contains an equal amount of probability mass above and below the posterior median [as in Eqs. (1) and (2)]. For asymmetric posteriors, however, highest density intervals (HDIs) may be preferable. An HDI is computed by finding all credible intervals for a given α and selecting the narrowest interval. For a continuous posterior, this is equivalent to lowering a horizontal line over the posterior until the outermost intersection points between the line and curve contain a fraction α of posterior probability mass [20]. For a particular ensemble of posteriors, assuming both of these interval types are qualitatively sensible, one can decide which to adopt by computing and comparing coverages for each.

When measuring a continuous parameter like m_β, physicists are often concerned not only with precision, but also with "discovery potential": the probability that the parameter is nonzero. While neutrinos have been found to be massive through oscillation experiments, a β-decay result distinguishing m_β from zero with high confidence or credibility would provide strong verification of physicists' interpretation of these oscillation data [8]. Here we claim a continuous parameter is nonzero if its highest density credible interval does not intersect with zero. (In practice, the m_β prior affects the outcome; see Sec. IV A 4.)

To verify that a scheme for assessing discovery potential is sound, a second calibration is required. This involves inputting a "true" mass value of zero for an ensemble of pseudoexperiments, then constructing HDIs with some credibility. Next, one confirms that the interval credibility approximately equals the fraction (coverage) of experiments for which the interval contains zero.

B. Calibrating mass ordering sensitivity claims

It is similarly possible to calibrate the process of claiming that the neutrino masses obey one ordering. This process is an example of result reporting for a discrete-domain parameter. In that case, we follow the above procedure through step 4, replacing m_β with a parameter that encodes mass ordering information. For our β spectral model, that parameter is η, the lighter mass' contribution to the spectrum. For normal and inverted orderings, respectively, η tends toward precisely known values η_N and η_I (see Sec. III). We claim a hypothetical ordering result when the posterior $\pi(\eta|\tilde{y})$ clusters near the predicted value for one ordering.

[2]Note that credible intervals do not guarantee any frequentist coverage. Constructing confidence intervals and computing frequentist coverages would require analyzing an ensemble of pseudoexperiments for a multidimensional grid of input parameter configurations. This becomes impractical in many dimensions, where the number of configurations on any reasonably sized grid grows exponentially fast [18].

Specifically, as a suggested decision-making scheme, we report a normal (inverted) ordering result when a posterior interval on η with credibility κ contains η_N (η_I) but not η_I (η_N). For a credible interval T on η, the associated loss functions for each ordering are

$$L_N \equiv \begin{cases} 0, & (\eta_N \in T) \text{ and } (\eta_I \notin T) \\ 1, & \text{otherwise} \end{cases},$$
$$L_I \equiv \begin{cases} 0, & (\eta_I \in T) \text{ and } (\eta_N \notin T) \\ 1, & \text{otherwise} \end{cases}, \quad (3)$$
$$T = [\phi_{(1-\kappa)/2}, \phi_{(1+\kappa)/2}],$$

where posterior bounds on η are computed so that

$$\int_0^{\phi_{(1\pm\kappa)/2}} d\eta\, \pi(\eta|\tilde{y}) = \frac{1 \pm \kappa}{2}.$$

These bounds may be selected using either a quantile or a highest density approach, depending on which yields higher coverage. If $L_N = L_I = 0$ or 1, then neither ordering is strongly favored and nothing can be claimed.

For each "true" mass ordering, given a series of pseudo-experiments, we then compute the rates at which we report correct and incorrect mass ordering results (see Sec. IV B). These true and false claim rates enable experimenters to select a credibility κ—i.e., to decide how stringent to make their reporting criterion. As in the continuous parameter case, this calibration of sensitivity to the mass ordering should be performed for a large number of model configurations sampled from priors. A similar calibration procedure would apply to accelerator, atmospheric and reactor experiments seeking to resolve the ordering [21], given an η-like parameter expressing mass ordering information.

We implement the above two procedures using the Stan software platform for Bayesian inference, which estimates posteriors by exploring a probability density parameter space using Markov-chain Monte Carlo methods (specifically, Hamiltonian Monte Carlo [22,23]). Stan is a valuable predictive analysis tool because it deals well with high dimensional problems and allows users to focus on modeling systems instead of developing computational architecture [24,25]. Along with Stan, we employ morpho, a python-based tool we developed to organize information input to and output from Stan. Morpho facilitates a Stan workflow involving convergence checks and analysis of posteriors, and it is designed to suit general Stan users [26].

III. MODEL FORMALISM FOR A β-DECAY EXPERIMENT

The differential spectrum predicted for β decay has a well understood analytic distribution, especially for superallowed transitions. The rate at which electrons are ejected as a function of their total energies is described by the equation

$$\frac{dN}{dE_e} = \left[\frac{G_F^2|V_{ud}|^2}{2\pi^3}|M_{\rm nuc}|^2 F(Z, p_e) p_e E_e\right] \times \left[\sum_{i=1}^{3} |U_{ei}|^2 \epsilon_\nu \sqrt{\epsilon_\nu^2 - m_i^2}\,\Theta(\epsilon_\nu - m_i)\right]. \quad (4)$$

In the electron phase space term (first bracketed term), G_F is the Fermi coupling constant, V_{ud} is the Cabbibo mixing angle, $M_{\rm nuc}$ is the nuclear matrix element, $E_e(p_e)$ is the outgoing electron energy (momentum), and $F(Z, p_e)$ is the Fermi function, for a daughter nucleus with charge Z. In the neutrino phase space term (second bracketed term), U_{ei} are the electron neutrino mixing matrix elements, ϵ_ν and $\sqrt{\epsilon_\nu^2 - m_i^2}$ represent the total energy and momenta of the released neutrino, and Θ is the Heaviside step function. We also define the kinetic energy of the electron, $K_e = E_e - m_e$.

In this section, we first justify our choice to hold the electron phase space term constant with respect to energy, allowing us to model spectral data by focusing on the second, neutrino-specific term. We then approximate and reparameterize the neutrino phase space, producing an analytic spectral form that both incorporates expected features of a real dataset and is suitable for Bayesian modeling.

A. Approximations to the β spectrum

For this model, we consider an eV-scale energy region near the high-energy end of a spectrum produced by β decay. For tritium decay, only superallowed transitions occur, so $M_{\rm nuc}$ is simply the sum of the vector (g_V) and axial vector (g_A) coupling constants:

$$|M_{\rm nuc}|^2 = g_V^2 + 3g_A^2.$$

$M_{\rm nuc}$ is therefore independent of electron energy.

The relativistic correction to the Fermi function is negligible at these energies, so the nonrelativistic form is used:

$$F(Z, p_e) = \frac{2\pi\alpha Z/\beta}{1 - e^{-2\pi\alpha Z/\beta}}, \quad (5)$$

where α is the fine structure constant and $\beta \equiv p_e/E_e$ is the electron's velocity. Since we confine our analysis to a region of width $\delta K_e \sim 10$ eV, and the variation in β is of order $\delta K_e/p_e \ll p_e^{\rm max}/E_e^{\rm max}$, β can be approximated as constant. Given Eq. (5), then, $F(Z, p_e) \simeq F(Z, p_e^{\rm max})$. Similarly, we treat $p_e E_e \simeq p_e^{\rm max} E_e^{\rm max}$ as constant, given that $\delta K_e \ll E_e^{\rm max}, m_e$. Thus, we can define a constant $A \equiv \frac{G_F^2|V_{ud}|^2}{2\pi^3}|M_{\rm nuc}|^2 F(Z, p_e^{\rm max}) p_e^{\rm max} E_e^{\rm max}$, representing the electron phase space.

In addition, the spectrum's neutrino-dependent term can be expressed in terms of the kinetic energy of the electron K_e. The neutrino phase space strongly depends on the final state of the daughter. When multiple final state configurations are possible—for example, in *molecular* tritium decay—all possible final state configurations need to be taken into account. In this case, however, we focus solely on *atomic* tritium (T) decay to singly-ionized $^3He^+$ (the process of interest for the Project 8 experiment [14]).

Assuming the decaying source is composed of nearly pure T, we need only consider a transition to one final state configuration of the helium-3 nucleus. Energy conservation then allows us to define ϵ_ν as

$$\epsilon_\nu \simeq \left(Q_T^0 + m_e - E_{\rm recoil} - E_e\right) \equiv (Q_T - K_e),$$
$$Q_T^0 \equiv M_i - M_f - m_e - \delta b,$$
$$E_{\rm recoil}^{\rm max} \simeq \frac{Q_T^0\left(Q_T^0 + 2m_e\right)}{2M_f Q_T^0},$$

where $M_{i(f)}$ is the parent (daughter) nucleus mass, δb is the difference in binding energy between the parent and daughter atoms, and E_{recoil} is the recoil energy of the decay nucleus (with maximum $E^{\max}_{\text{recoil}}$). The recoil energy varies by ∼0.5 eV over the spectrum's last 3.5 keV, so we approximate E_{recoil} as constant near the end of the spectrum [27]. This allows us to write the β spectrum in terms of an endpoint energy parameter that is assumed not to differ from decay-to-decay: $Q_T \equiv Q_T^0 - E^{\max}_{\text{recoil}}$. For atomic tritium, Q_T^0 has an experimentally determined mean value of 18566.66 eV, and $E^{\max}_{\text{recoil}}$ is 3.41 eV [27].

Putting this together with the constant electron phase space approximation, we formulate a spectral model $\mathcal{P}$:

$$\mathcal{P}(K_e) \equiv A \sum_i |U_{ei}|^2 (Q_T - K_e)\sqrt{(Q_T - K_e)^2 - m_i^2} \cdot \Theta(Q_T - K_e - m_i) \equiv \sum_i |U_{ei}|^2 \mathcal{P}_i(K_e). \quad (6)$$

Second-order effects are small compared with the overall spectral shape in and around our narrow analysis window. Hence, our analytic model ignores second-order corrections, including terms that account for finite nuclear radii and radiative corrections.

B. One- and two-neutrino spectral models with finite energy resolution

We must transform the function $\mathcal{P}(K_e)$ so that it includes features seen in experimental data, including an energy resolution, background events, and kinetic energy bounds. In performing these transformations, $\mathcal{P}(K_e)$ must meet two conditions to be suitable for Bayesian inference. First, we require that the function be normalizable, because Bayesian models are formulated as probability density functions (PDFs). Specifically, in Stan, one specifies features of a likelihood space by adding log PDFs to a total log probability. While strictly, the function's normalization need not be analytic because Stan provides for one-dimensional (1D) integration, inference with analytic PDFs is less computationally expensive. By incorporating smearing from an experimental energy resolution, we are able to formulate an analytically normalized version of $\mathcal{P}$. Second, to assess sensitivity to the mass ordering, our model must include a parameter η, as described in Sec. II B—or more generally, a variable that strongly depends on the ordering.

We consider two experimental factors: the uncertainty associated with reconstructing an energy spectrum and the presence of background events. As opposed to considering an integrating spectrometer (like the one used by KATRIN), we focus on differential spectrometers (used by Project 8, ECHO, and HOLMES) capable of measuring individual electron kinetic energies [28]. This allows us to assume that true kinetic energies are normally distributed around K. The mapping distribution is $\mathcal{N}(K_e|K,\sigma)$ for a standard deviation—that is, an energy resolution—σ.

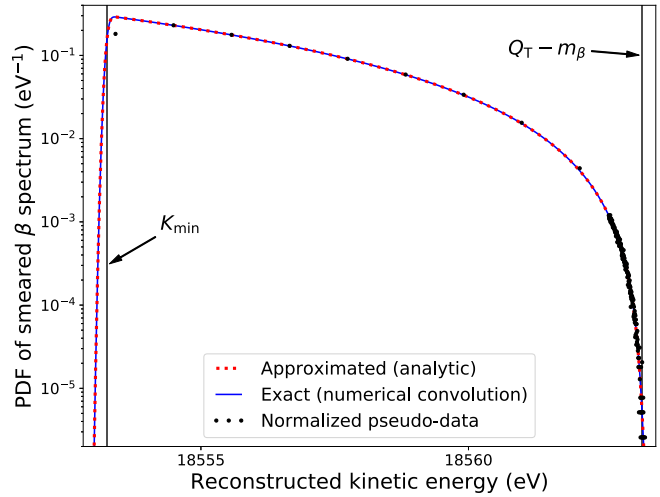

FIG. 2. Approximated spectral model [Eq. (7)] superimposed on a numerical convolution of a Gaussian with the exact T spectrum [Eq. (4)] and one year of data generated with the exact model. The signal activity is 1.7×10^8/yr in the analysis window, m_β = 8.5 meV, $K_{\min} = Q_T - m_\beta - 10$ eV, and σ = 54 meV (see Sec. IV A).

The convolution of the neutrino phase space term with $\mathcal{N}$ is not analytically integrable. We address this issue by expanding each neutrino mass term $\mathcal{P}_i$ within $\mathcal{P}$ [Eq. (6)] to first order in m_i^2:

$$\mathcal{P}_i(K_e) \simeq A\big[(Q_T - K_e)^2 - m_i^2/2\big]\Theta(Q_T - K_e - m_i).$$

This expansion is justified for $m_i^2 \ll (Q_T - K_e)^2$, which holds for all data points except those very close to the endpoint. Moreover, once the spectral shape is smeared by convolving it with $\mathcal{N}$, the exact and approximated curves appear very similar even near the endpoint, as seen in Fig. 2. When analyzing a full spectral shape, the expansion holds except for large quantities of data. (The count number at which the approximation breaks down depends on the analysis window and binning, among other factors.)

Given the expansion in m_i^2, we can define and integrate a reconstructed energy spectrum $\mathcal{F}_i$:

$$\mathcal{F}_i(K|Q_T, K_{\min}, m_i, \sigma) \equiv \mathcal{F}_i(K) \propto \int \mathcal{P}_i(K_e)\mathcal{N}(K_e|K,\sigma)\Theta(K_e - K_{\min})dK_e \to \frac{dN}{dK}$$
$$= \mathfrak{N}(m_i, Q_T - K_{\min})[\xi(K|Q_T, m_i, \sigma, m_i) - \xi(K|Q_T, m_i, \sigma, Q_T - K_{\min})]$$
$$\xi(K|Q_T, m_i, \sigma, t) = (Q_T - K + t)\sigma^2\mathcal{N}(Q_T - K|t,\sigma) + \frac{1}{2}\left[-\frac{m_i^2}{2} + (Q_T - K)^2 + \sigma^2\right]\text{Erfc}\left(\frac{t - Q_T + K}{\sqrt{2}\sigma}\right). \quad (7)$$

This model describes signal data in a kinetic energy window $[K_{\min}, Q_T]$. Its normalization term is defined based on the size δK_e of this window:

$$\mathfrak{N}(m_i, \delta K_e) = \frac{6}{2(\delta K_e)^3 - 3m_i^2 \delta K_e + m_i^3}.$$

The practical need to filter out events below some energy motivates our choice to include a minimum energy parameter. Because of the uncertainty σ associated with the reconstruction of $K_{\min}$, this lower bound is soft.

The background is assumed to be uniform in kinetic energy. If we include a smeared (i.e., convolved with $\mathcal{N}$) background $\mathcal{B}$, then the normalized spectral model for a single neutrino mass m_i is given by the master equation:

$$\begin{aligned}\mathcal{M}_i(K) &= f_s \mathcal{F}_i(K) + (1 - f_s)\mathcal{B}(K|K_{\min}, K_{\max}, \sigma) \\ \mathcal{B}(K|K_{\min}, K_{\max}, \sigma) &= \frac{\mathrm{Erf}\left(\frac{K_{\max}-K}{\sqrt{2}\sigma}\right) - \mathrm{Erf}\left(\frac{K_{\min}-K}{\sqrt{2}\sigma}\right)}{2(K_{\max} - K_{\min})}. \quad (8)\end{aligned}$$

Here f_s is the signal fraction of a dataset. Since $\mathcal{M}_i(K)$ is analytic and normalized, it can be formulated as a PDF and thus used for statistical inference in Stan. Moreover, Eq. (8) allows us to assess an experiment's sensitivity to the mass scale. Specifically, the calibration procedure in Sec. II A can be applied for posteriors on masses m_i inferred using this model.

Experimentally, K is constructed from some observed variable v_o—for example, in Project 8's case, an electron cyclotron frequency (see Sec. IV). The energy resolution derives in large part from statistical uncertainties on the quantities used to map $v_o \to K$. While these quantities and their errors are expected to be well known, a mapping bias that shifts the overall energy scale is possible. We model this bias by constructing a prior on $K_{\min}$ which allows the minimum energy to shift slightly relative to Q_T.

We bin data to reduce computation time, though unbinned analyses are possible in Stan. Details in the spectral shape on the order of a few meV only inform the neutrino mass measurement if they occur in the last $\approx$1 eV. Thus, data should be binned finely near the endpoint and coarsely (for computing efficiency) at lower energies. For narrow bins, the fraction of counts per bin can be fitted to the spectral rate at each bin center. However, modeling large bins [$\mathcal{O}$(1 eV) width] in this way biases m_β posteriors upward relative to inputs, due to the changing slope of the spectrum within each bin. To address this, we derive the cumulative distribution function $\mathcal{G}_i^{\mathrm{CDF}}(K)$ corresponding to the PDF model in Eq. (7), then set the number of events in a bin $[K_n, K_{n+1}]$ equal to $\mathcal{G}_i^{\mathrm{CDF}}(K_n) - \mathcal{G}_i^{\mathrm{CDF}}(K_{n+1})$. The CDF is provided in Appendix A.

To report mass ordering results based on inferred posteriors, we modify the spectral model in a second way. If one considers the smaller mass splitting ($\Delta m_{21}^2 \equiv m_2^2 - m_1^2$) to be negligible, then the signal [Eq. (7)] can be written in terms of only *two* neutrino masses, m_L and m_H. Here $m_H > m_L$, with a splitting $\Delta m_{ee}^2 \equiv m_H^2 - m_L^2 \simeq |\Delta m_{13}^2| \simeq |\Delta m_{23}^2|$. The signal is then simply a weighted sum of two spectra, corresponding to the two mass scales:

$$\begin{aligned}\mathcal{F}'(K) = \eta \mathcal{F}_L(K|Q_T, K_{\min}, m_L, \sigma) \\ +(1-\eta)\mathcal{F}_H(K|Q_T, K_{\min}, m_H, \sigma). \quad (9)\end{aligned}$$

As indicated previously, η is the fractional contribution of the lighter mass term to the spectral shape.

Since Δm_{ee}^2 is always positive, η is the only parameter in this model that depends on the ordering. Specifically, η should tend toward one value (η_N) if the ordering is normal and another (η_I) if it is inverted, where

$$\begin{aligned}\eta_N &\equiv |U_{e1}|^2 + |U_{e2}|^2 = \cos^2(\theta_{13}) \\ \eta_I &= 1 - \eta_N = |U_{e3}|^2 = \sin^2(\theta_{13}).\end{aligned}$$

The ordering question can thus be formulated solely in terms of the large mass splitting and θ_{13}, both of which are measured by reactor antineutrino disappearance experiments. Hence, the above model enables a mass ordering determination using only a β spectrum and reactor experiment results.

To perform a mass ordering sensitivity study, we substitute $\mathcal{F}_i(K) \to \mathcal{F}'(K)$ in Eq. (8). Then, by implementing the decision-making scheme in Sec. II B for posteriors on η, we can calibrate the analysis by estimating the expected accuracy of reporting different ordering results based on β spectra. Consequently, we have here developed a probability distribution that serves two key purposes: It acts as a likelihood function for Bayesian modeling, and it can be used to assess a direct mass experiment's sensitivity to the mass ordering.

IV. RESULTS

Our analysis seeks to determine how experimental parameters such as energy resolution and number of β-decay events affect sensitivity to m_β as well as the mass ordering. To construct concrete, realistic priors that reflect what parameter values an experiment might see, we incorporate information related to the Project 8 experiment. The Project 8 Collaboration developed the technique of Cyclotron Radiation Emission Spectroscopy (CRES) for obtaining a β spectrum at high precision, as originally proposed in Ref. [13]. CRES involves measuring the cyclotron frequencies of electrons in a magnetic field, then computing corresponding energies. In its final stage, Project 8 aims to measure the neutrino mass scale by analyzing a spectrum produced by atomic tritium β decay. The Collaboration is working to reach a neutrino mass sensitivity of about 40 meV [14].

A. Sensitivity to absolute neutrino mass scale

1. *Pseudodata generation and analysis*

This study follows the procedure for calibrating sensitivity claims described in Sec. II A. We perform 220 pseudoexperiments (that is, repetitions of steps 2–5 in the procedure), assuming a runtime $\Delta t = 1$ yr. For each experiment, data are generated with a β-spectrum model that is much more detailed than the inferential model, to reveal any biases arising from analysis assumptions. The generation model includes an energy-dependent relativistic Fermi function, as well as correction terms stemming from atomic physics

phenomena. These terms account for the emitted electron's recoiling charge distribution, radiative effects from real and virtual photons, three-body recoil effects from weak-magnetism and V-A interference, 1*s*-orbital electron interactions with the β and screening of the $^3He^+$ Coulomb field, and the $^3He^+$ nucleus' structure. The formulas for these corrections are taken from Ref. [29]. In this subsection, we generate data with a one-neutrino mass model and call that mass m_β.

To compose a full data generation model, the detailed β spectrum is broadened by numerically convolving it with a Gaussian of width σ. A nearly flat background [Eq. (8)] is then added to the spectrum. Before convolution, the data are confined within a ≈20-eV window centered on the mean energy at which the spectrum vanishes: $Q_T - m_\beta$, where Q_T is the mean T endpoint. The window's width varies modestly from spectrum-to-spectrum because its lower bound is sampled from a prior, as discussed below.

In Stan, we implement the one-neutrino spectral model $\mathcal{M}$ from Eq. (8), for $m_i \to m_\beta$. Each pseudospectrum is analyzed using this model. The data are histogrammed with 300 bins covering the 1 eV directly below the endpoint, nine ≈1-eV-wide bins at lower energies, and one bin containing any background events above the endpoint. For each of the 300 narrow bins bounded by $[K_n, K_{n+1}]$, we model the number of counts as a value sampled from a Poisson distribution with rate $\mathcal{M}(\frac{K_n+K_{n+1}}{2}) \times (K_{n+1} - K_n)$. For the nine wider signal bins, since the β spectrum decreases monotonically, the signal Poisson rate can be approximated as $\mathcal{G}^{\mathrm{CDF}}(K_n) - \mathcal{G}^{\mathrm{CDF}}(K_{n+1})$ (see Appendix A). To test the effect of bin size near the endpoint, a small analysis (40 pseudoexperiments) was performed with 500 bins in the eV below the endpoint. It yielded median m_β sensitivities and coverages consistent with those presented below, within statistical uncertainty.

2. *Selection of priors*

Each model parameter requires an associated prior, both for sampling "true" values (generator inputs) and for inferring posteriors from data. By sampling from these priors repeatedly, creating an ensemble of model configurations, we can approach an analysis that accounts for the full range of possible spectra—given anticipated statistical and systematic errors. To construct priors, we select functional forms with boundary conditions that accord with physical limits on parameters. For positive quantities, we therefore generally chose log-normal or gamma (γ) distributions—the former when likely values span multiple orders of magnitude, and the latter otherwise. See Table II for a summary of priors.

The one-neutrino model includes parameters m_β, Q_T, σ, K_{min}, and f_s. A γ prior on m_β was constructed so that 1% of its probability mass would fall below 0.008 eV, reflecting the lower bound from mass splitting measurements [1]. (This bound is not strict because of small uncertainties on those measurements.) Ten percent of the prior mass on m_β falls above 1.1 eV, the 90% confidence upper bound reported by KATRIN in 2019 [30].

We employ a normal prior on Q_T but define the parameter as positive in Stan, truncating a negligible negative portion of the normal distribution. The mean of the prior is the extrapolated tritium endpoint minus the electron mass, as calculated by Bodine *et al.* [27]. The largest contribution to the Q_T uncertainty is from the T-^{3}He mass difference, which has been measured in Penning traps [31]. That quantity serves as the Q_T prior standard deviation.

TABLE II. Priors for data generation and analysis using one- and two-neutrino models, denoted by "1" and "2," respectively. "Design" quantities reflect goals for Project 8, while "measured" ones derive from past experiments. Prior functions are defined in Appendix B.

	Prior	Model	Prior Source
Q_T	$\mathcal{N}$([18563.25, 0.07] eV)	1, 2	Measured
σ_{dopp}	γ(59.82, 2868 eV^{-1})	1, 2	Measured
σ_{inst}	$\mathcal{N}(\mu_{\mathrm{inst}}, \delta_{\mathrm{inst}})$	1, 2	Design
K_{min}	$\mathcal{N}$([$Q_T - m_{\beta,L}$–10, 0.01] eV)	1, 2	Design
A_b	lognorm(−27.31, 0.5678)	1, 2	Design
N_{atoms}	lognorm(44.07, 0.5677)	1, 2	Design
m_β	γ(1.135, 2.302 eV^{-1})	1	Measured
Δm^2_{ee}	γ(314.5, 122700 eV^{-2})	2	Measured
m_L	γ(2.186, 126.1 eV^{-1})	2	N/A

We consider two energy resolution effects, summed in quadrature to yield the total resolution σ: (1) Doppler broadening σ_{dopp} from translational motion of tritium atoms, and (2) Instrumental broadening σ_{inst} from the process of reconstructing kinetic energies. To select a γ prior on σ_{dopp}, we devised a Stan model that extracts posteriors for the mean expected energy spread due to thermal broadening (μ_{dopp}) and the uncertainty on that spread (δ_{dopp}), using the formulas in Ref. [27]. We set the mean ($\sqrt{\mathrm{variance}}$) of the σ_{dopp} prior equal to the mean of a Gaussian fit to the μ_{dopp} (δ_{dopp}) posterior, inferred for a 0.3000 ± 0.0015 Kelvin gas with negligible T_2 contamination.

The primary two expected contributions to the instrumental resolution are (a) a cyclotron frequency measurement error and (b) an uncertainty on the field value in the frequency to energy conversion. We construct a σ_{inst} prior assuming that the field error $\Delta B/B \sim 10^{-7}$ is the larger contribution [14]. In this case, $\sigma_{\mathrm{inst}} \sim 0.05$ eV. As Project 8 is considering multiple energy calibration schemes, the uncertainty on σ_{inst} could reasonably fall anywhere in the large range of ≈0.5–10%. Accordingly, the σ_{inst} prior's "true" mean and standard deviation (μ_{inst}, δ_{inst}) are sampled from distributions before data generation, then fixed to their sampled values during inference. The σ_{inst} prior is then $\mathcal{N}(\mu_{\mathrm{inst}}, \delta_{\mathrm{inst}})$. The μ_{inst} distribution for pregeneration sampling is $\gamma(25.0, 2 \times 10^{-3}\ \mathrm{eV}^{-1})$, with mean 0.05 eV and $\sqrt{\mathrm{variance}} = 0.01$ eV. The δ_{inst} distribution is $\gamma(1.583, 809.7\ \mathrm{eV}^{-1})$, selected so that 5% of its probability mass would fall below (above) 2.5×10^{-4} eV (5×10^{-3} eV). Combining the two sources of broadening, the mean σ is 0.054 eV.

Experimenters can select K_{min} before analysis by filtering out events above some cyclotron frequency. If the conversion (σ) to K were known exactly, then K_{min} could be fixed during inference at a value computed from that frequency. Instead, to allow for a systematic shift in K on the order of 0.01 eV, we employ a normal prior on K_{min} with that standard deviation.

We also incorporated priors associated with the spectral signal fraction. While external information does not directly inform a prior on f_s, it pertains more directly to signal and background activities A_s and A_b. Here A_s (A_b) is the number of events per second generated by $\mathcal{F}(K)$ ($\mathcal{B}(K)$) in the window $[K_{\min}, Q_T]$ ($[K_{\min}, K_{\max}]$). We thus model the signal fraction as $f_s = S/(S+B)$, where $S = \Delta t A_s$ and $B = \Delta t A_b$ are signal and background Poisson event rates.

To inform the prior on A_s, a possible expected signal activity in the unconvolved spectrum's last electronvolt can be expressed in terms of both experiment-specific quantities (atomic source density n; effective source volume V_{eff}) and physical parameters (T half-life $\tau_{1/2}$; fraction f_{eV} of counts between $Q_T - m_\beta - 1\,\text{eV}$ and $Q_T - m_\beta$ for $\sigma \to 0$). Following the approach in Ref. [32],

$$A_s \text{ in the last eV} = nV_{\text{eff}} \frac{\ln(2)}{\tau_{1/2}} f_{\text{eV}}. \tag{10}$$

Here the fraction of counts f_{eV} in the last eV takes into account that all events observed in the last electronvolt are produced by decays to the $^3\text{He}^+$ electronic ground state [27], which comprise 70.06% of the total tritium decay width [33]. The detailed spectral model we developed for data generation enabled a new, precise calculation of f_{eV}, a quantity that has historically been central to projecting the activities of tritium-based neutrino mass experiments [8,32]. Assuming $m_\beta = 0$, we find $f_{\text{eV}} = 1.69 \times 10^{-13}$ for T_2 and 2.06×10^{-13} for T.

For a number density $n = 10^{18}$ atoms/m^3 and $V_{\text{eff}} = 10$ m^3, target values for an experimental design scenario considered by the Project 8 Collaboration [14], the experiment would detect $\approx 1.2 \times 10^5$ events per year above $Q_T - m_\beta - 1$ eV, and a factor of 1000 more above $Q_T - m_\beta - 10$ eV. We employed a log-normal prior on $N_{\text{atoms}} \equiv nV_{\text{eff}}$ for this scenario, setting its mode and standard deviation equal to 10^{19} atoms. For a given apparatus, this allows for some variation in source density and detection efficiency. A_s was then computed from N_{atoms}.

The A_b prior is informed by the Project 8 Collaboration's goal for its dominant source of background to be cosmic rays passing through the tritium gas. Since the expected cosmic ray activity is approximately 10^{-12}/eV/s for the n and V_{eff} values assumed above, and the activity varies with those parameters [14], the A_b prior distribution is chosen to have mode and standard deviation equal to 10^{-12}/s for each 1-eV-wide bin of data.

3. *Neutrino mass scale sensitivity results*

A close correspondence between "true" neutrino masses and m_β posteriors indicates that each β spectrum strongly informs a neutrino mass determination (see Fig. 3). Each posterior standard deviation on m_β is at least 22 times smaller than the corresponding prior spread. See Refs. [2,34] for more information on posterior shrinkage and evaluating model performance.

Table III summarizes credible interval width results for m_β. Highest density credible intervals (C.I.s) were computed for $\alpha = 0.6826$, 0.9, and 0.95 [see Eq. (1)], and standard deviations were computed by halving the first of these. The

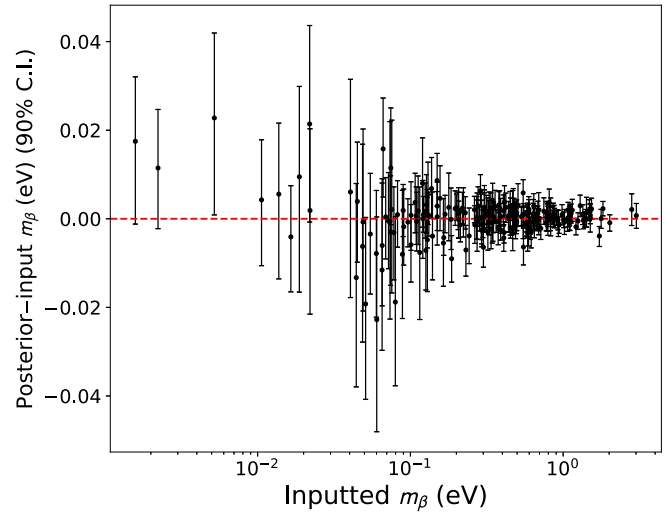

FIG. 3. Neutrino mass posterior means and 90% credible intervals as a function of inputted m_β, for a one-neutrino model and the assumed experimental design. Interval widths ("sensitivities") decrease with m_β, asymptoting at ∼5 meV.

HDI approach produces higher coverages than do quantile intervals. To enable reliable C.I. estimation, we required the effective size of each posterior array (as computed by PyStan [25]) to exceed 6000, so that at least 150 effective samples fall outside each bound.

We can verify that the process of inference itself was successful: As expected, posterior means for Q_T, σ_{inst}, σ_{dopp}, $K_{\min}$, A_s, and A_b track with input values. During all 220 analyses, the five Stan convergence diagnostics—$\hat{R}$, effective sample size ratio, E-BFMI, tree depth, and divergences [23,35,36]—showed no signs of pathological behavior. Moreover, the coverage of 90% credible intervals is between 85% and 99% for all parameters.

For true $m_\beta > 0.5$ eV, the mean 90% C.I. width is 0.005 eV. The reported coverage uncertainties are $\sqrt{C(1-C)/N_{\text{trial}}}$.

The left plot of Fig. 4 shows that mass sensitivity depends weakly on σ_{inst}, because the scenario considered here is relatively statistics limited and the range in σ_{inst} is small. However, for this scenario, smaller uncertainties on σ_{inst} noticeably improve sensitivity (see Sec. IV A 4 for an instance of this). We would also expect increasing the effective volume to improve neutrino mass sensitivity. Indeed, for an ensemble with fixed energy resolution and a wide range in V_{eff} values, the widths of m_β credible intervals depend strongly on V_{eff}, as seen in

TABLE III. Sensitivity to m_β after 1 yr with coverages of credible intervals.

Interval	Sensitivity (eV)			Coverage
	Median	Mean	Maximum	
90% C.I.	0.0071	0.0112	0.0493	$(90.0 \pm 2.0)\%$
95% C.I.	0.0084	0.0133	0.0598	$(93.2 \pm 1.7)\%$
Stdev.	0.0022	0.0034	0.0158	$(70.1 \pm 3.1)\%$

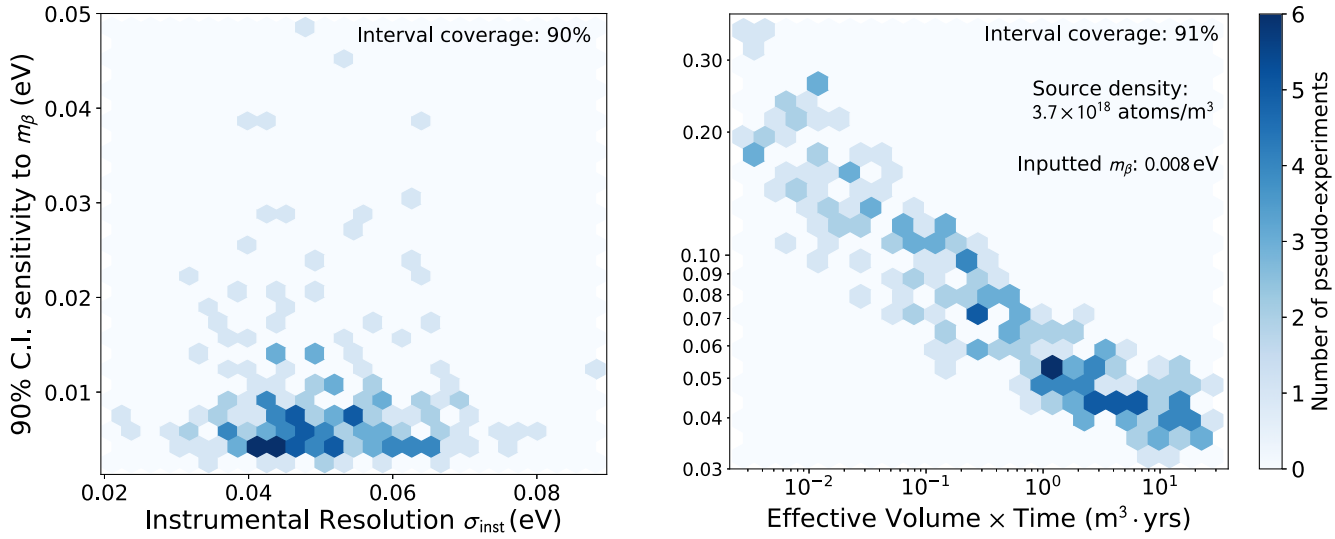

FIG. 4. Dependence of mass sensitivity (width of 90% credible intervals) on σ_{inst} and volume×efficiency×time. The left plot assumes the design scenario described in this section. The right plot shows a larger range in signal exposure, for an alternate scenario where n (3.7×10^{18} m^{-3}) and σ (115 ± 2 meV) are chosen to minimize m_β uncertainty, given a trade-off between frequency reconstruction error and exposure. The right plot "pessimistically" assumes $m_\beta = 0.008$ meV.

the right plot in Fig. 4. These results could inform how future direct mass experiments prioritize their efforts to improve the expected energy resolution, resolution uncertainty, and statistical yield of an apparatus design.

4. Claiming m_β is inconsistent with zero

We also evaluate the ability of an experiment with the design described here to distinguish the electron-weighted neutrino mass from zero. As introduced in Sec. II A, for a given β spectrum, it is possible to claim that the neutrino mass is nonzero with credibility α if the lower bound of a posterior highest density α-credible interval exceeds zero. The m_β prior in Table II is in conflict with this test, as that prior assumes that it is highly improbable for the mass to be zero, considering the lower bound from oscillations measurements. When Project 8 analyzes real data, its main mass scale analysis can include an m_β prior with an oscillations-based lower bound. However, to assess consistency with zero, the data will need to be reanalyzed with an oscillations-bound-free prior.

As an example sensitivity study, we perform 75 pseudo-experiments with 10% of the neutrino mass prior probability falling below 0.005 eV and 10% above 0.1 eV. Resulting posterior credible intervals on m_β are shown in Fig. 5. The neutrino mass can be distinguished from zero with 90% credibility in 65 of these analyses. It is possible to claim the mass is inconsistent from zero for true $m_\beta \gtrsim 0.04$ eV, with two outliers caused by an underestimation of the true mass, combined with poor m_β precision due to large inputted uncertainties (i.e., prior widths) on σ_{inst}.

How can one be confident that this method will not produce frequent false claims? We may perform another calibration: For β spectra produced given a true neutrino mass of zero, we should rarely claim that m_β is distinguishable from zero. Indeed, when we analyze 150 such spectra, the mass is judged to be consistent with zero 93% of the time ($\alpha = 0.9$).

B. Sensitivity to neutrino mass ordering

The analysis in this section follows the procedure described in Sec. II B for calibrating sensitivity claims to discrete parameters. Pseudodata are generated with the same detailed spectral model as in Sec. IV A, but with two neutrino masses instead of one. Similarly, for inference in Stan, we now employ a two-neutrino model—Eq. (8), with a spectral signal $\mathcal{F}'(K)$ [Eq. (9)]—to analyze data in the approximate window

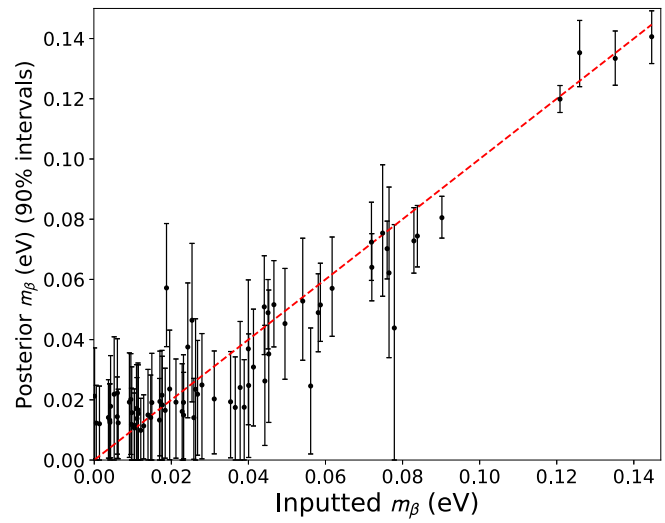

FIG. 5. Mass posterior means and 90% credible intervals for inputted m_β near zero. It is possible to distinguish the mass from zero for true $m_\beta \gtrsim 0.04$ eV, with outliers characterized by large uncertainties δ_{inst} (energy broadening standard deviation).

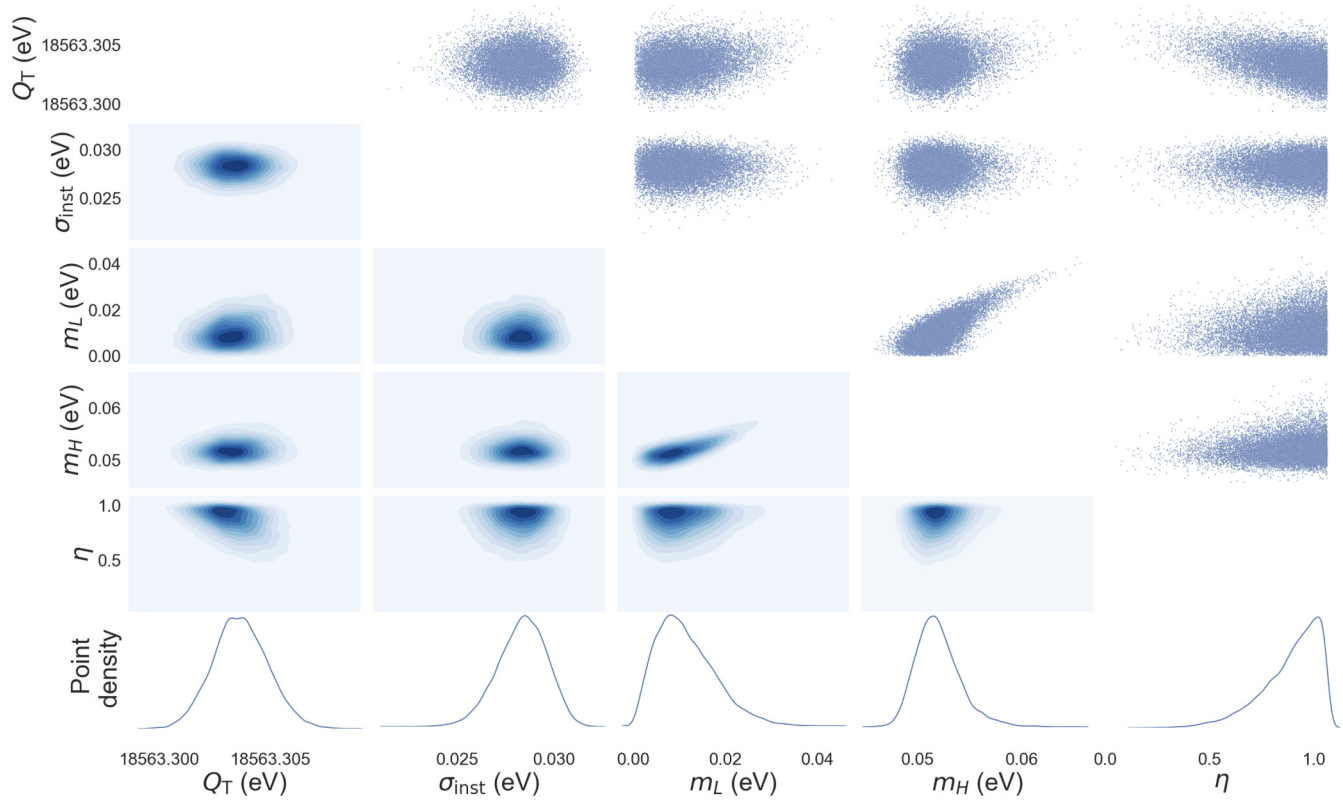

FIG. 6. For one pseudoexperiment, example posterior probability density plots and 2D histograms (in both contour and scatter plot form) for parameters in the two-neutrino spectral model. Posteriors were obtained by analyzing data ($\Delta t = 1$ yr) that was generated assuming a normal mass ordering.

$[Q_T - m_L - 1\text{ eV}, Q_T - m_L + 10\text{ eV}]$. This region extends only 1 eV below the endpoint so that the likelihood will be strongly informed by fine-grained mass ordering-dependent features near Q_T. To help constrain the overall mass scale, data in the next eV below $Q_T - m_L - 1$ eV are fitted to a one-neutrino m_β model, with the requirement $m_\beta^2 = \eta m_L^2 + (1-\eta)m_H^2$.

We repeat this two-neutrino analysis for $\Delta t = 1$ yr and 2 yrs with at least 170 pseudoexperiments per runtime, producing coverage uncertainties of 1–5%. Again, data are binned after generation and then analyzed assuming Poisson-distributed events. The Stan model includes the same priors on parameters Q_T, σ_{dopp}, σ_{inst}, N_{atoms}, and A_b as in the one-neutrino case. The prior on K_{min} is similar, with its mean dependent on m_L instead of m_β. We also constructed priors on Δm_{ee}^2 and m_L (see Table II), while m_H required no prior, as it was modeled by transforming those parameters.[3]

A γ prior on Δm_{ee}^2 was formulated by extracting a 90% confidence interval from a global fit of three reactor neutrino experiments: $[2.38, 2.75] \times 10^{-3}\,\text{eV}^2$ [37]. At the time when we began the analysis, this was the most up-to-date global fit of reactor data. As these bounds differ slightly according to mass ordering, to be conservative, we selected each bound (either the normal or inverted ordering limit) so as to obtain a wider prior. Ten percent of the prior mass on Δm_{ee}^2 falls outside each bound. In addition, before generation, either η_N or η_I was sampled from a Gaussian prior, depending on the "true" ordering. Prior parameters were determined based on the mean of $\cos^2\theta_{13}$ (0.979) and error on that mixing parameter (0.001), as measured by reactor experiments [37]. Posteriors extracted from one of the two-neutrino model fits are shown in Fig. 6, and Fig. 7 compares pseudodatasets for the normal and inverted orderings.

For the prior on m_L, we avoided computing soft bounds using current limits on the mass scale from particle physics experiments, as those constraints do not translate easily to bounds on individual masses [1,38]. Instead, we envision a scenario in which m_L is restricted below $\approx$0.05 eV, potentially based on future cosmological constraints on the sum of the three neutrino masses. Specifically, the prior for pregeneration sampling and inference is γ-shaped with 10% of its mass below 5 meV and 5% above 40 meV, resulting in $m_L < 0.08$ eV for all pseudoexperiments. (The distribution peaks near zero, since there is no oscillations-based lower bound on m_L for the inverted ordering and a very small lower bound for the normal case.) For true masses above 0.08 eV, one rarely claims to have resolved the mass ordering using our reporting scheme. Hence, by choosing a prior localized in a low-mass region, we proportionally inflate true and false ordering claim rates. This

[3]To avoid noninvertible transforms and the need for Jacobian adjustments, in Stan, we define a "`positive_ordered`" `transformed parameter m`, with `m[1]` $= m_L$ and `m[2]` $= \sqrt{m_L^2 + \Delta m_{ee}^2}$ (see Sec. 22 of Ref. [25]). The entries of `m` then serve as inputs to the spectral log probability density function.

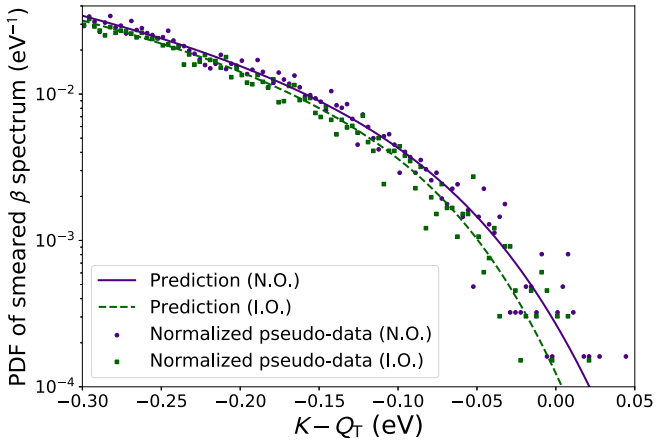

FIG. 7. Example pseudospectra overlaid on predicted curves [Eq. (4) numerically convolved with a Gaussian] for normal and inverted orderings, with $m_L = 0$ eV and a 2-yr runtime. Spectra are plotted as a function of the difference between reconstructed energy and the T endpoint.

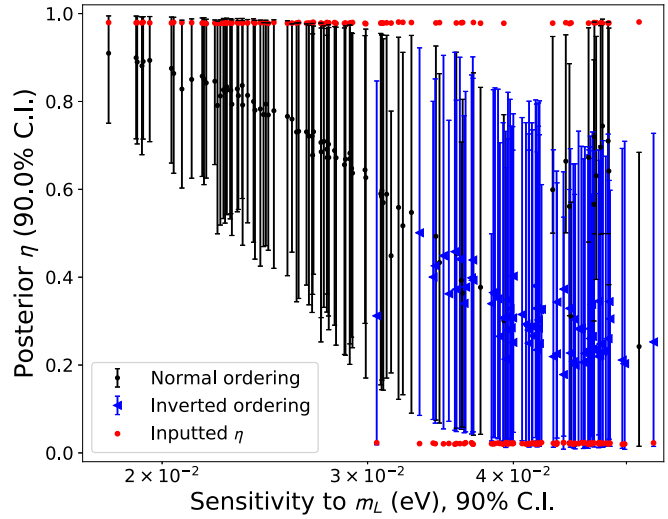

FIG. 8. For 2 yrs of data assuming normal (dots) and inverted (triangles) orderings, posterior means and intervals on η as a function of potential sensitivity to m_L, defined as C.I. width.

makes the process of selecting ideal reporting criteria κ based on claim rates more statistically reliable than it would be for a wider m_L prior.

As in the one-neutrino case, posterior means track with input values for all parameters. During analysis of most spectra, no Stan MCMC diagnostics indicated a failure to converge. However, a quarter of runs exhibited signs of incomplete convergence [35]: Fifteen percent showed a small number of diverging iterations (1–10 of 15,000), and 10% failed at least one other check. Mass ordering sensitivity results are robust despite this, since observing and minimizing false-positive rates ultimately validates the analysis. Still, a more consistently converging model might improve sensitivity.

Table IV summarizes results for calibration of sensitivity to the mass ordering. Uncertainties on 0% claim rates represent 68.3% confidence limits derived from a binomial probability law. (Given the ensemble's finite size, the actual probability of a false claim is not exactly zero.) The loss functions L_N and L_I in Eq. (3) dictated whether an ordering result should be reported for each pseudoexperiment. That is, a normal (inverted) ordering claim was made if a posterior interval on η of credibility κ contained $\eta_N = \cos^2\theta_{13}$ ($\eta_I = 1 - \eta_N$) but not η_I (η_N) (see Fig. 8). Given the small experimental error on $\cos^2\theta_{13}$, we assumed a known value $\eta_N = 0.978$. The credibility κ acts as a reporting criterion, and modifying κ affects the rates at which we *correctly* and *incorrectly* claim to have resolved the neutrino mass ordering (see Fig. 9).

TABLE IV. Assuming either a normal or inverted true ordering, percentages of pseudoexperiments for which the three possible reporting outcomes ("normal," "inverted," or "no claim") occur. To minimize false claims, different reporting criteria κ are used for each ensemble and observed ordering.

$\Delta t = 2$ yrs, $m_L \lesssim 0.05$ eV		
	Claim N	Claim I
Optimal κ	0.985	0.855
Truth: N	86.8% ± 3.5%	0.0% (+1.3%)
Truth: I	0.0% (+1.5%)	21.8% ± 4.7%
$\Delta t = 1$ yr, $m_L \lesssim 0.05$ eV		
	Claim N	Claim I
Optimal κ	0.925	0.875
Truth: N	45.6% ± 5.2%	0.0% (+1.3%)
Truth: I	0.0% (+1.2%)	23.5% ± 4.3%

We recommend an "optimal κ" by selecting the value for which the relevant correct claim rate is maximized, given a minimal incorrect rate—which can be zero, in this study. Values of κ are considered in 0.5% increments. We observe that, for both 1 yr and 2 yrs of data, false inverted claims begin to occur for κ values above a *lower* number than do false normal claims. In fact, Fig. 9 shows that false normal claims are never made for $\Delta t = 2$ yrs, for these pseudodatasets. Using that knowledge, for real data, it is possible to boost the probability of a correct ordering claim without increasing the risk of a false claim by applying the following procedure:

(a) Check what result would be reported using the optimal κ for normal ordering true/false claims (as predicted with pseudoexperiments)—here 0.925 (0.985) for 1 (2) yr(s).
(b) If the result is "normal" or "no claim," then report it.
(c) If the result is "inverted," then it could be a false positive. Reduce κ to the inverted optimal value—0.875 (0.855) for 1 (2) yr(s)—to determine if to report "inverted" or nothing.

This procedure accounts for the fact that it is easier to claim a normal than an inverted ordering result for our model. The procedure enables false claim rates of 0% for the pseudoexperiments performed here, with true rates reaching 87% (22%) for the normal (inverted) ordering after 2 yrs. We see here

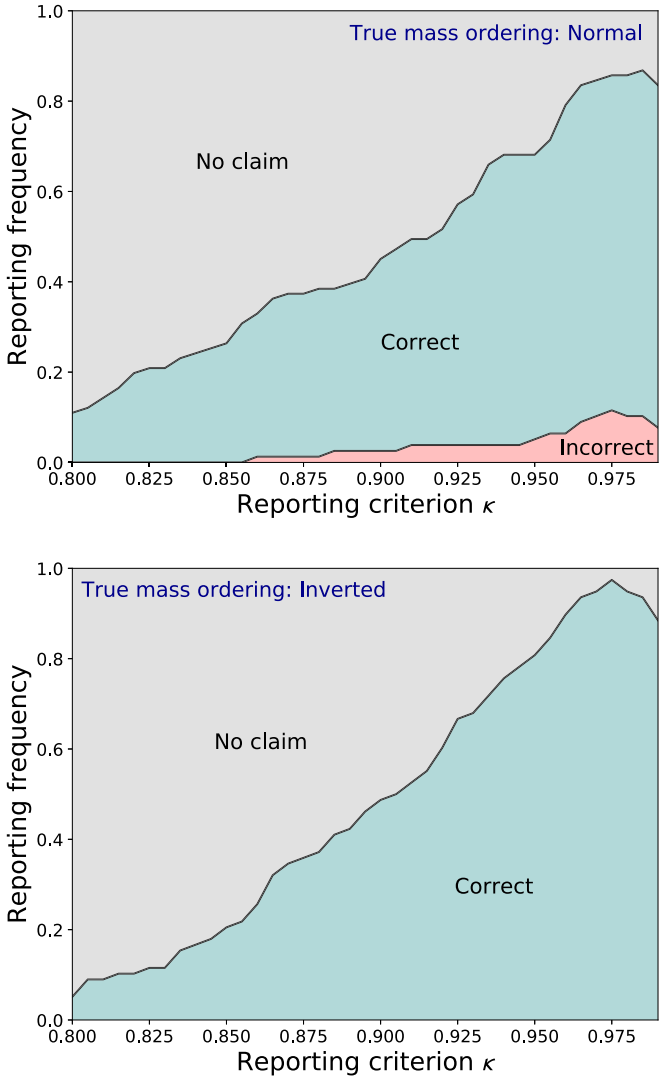

FIG. 9. Mass ordering reporting frequencies for $\Delta t = 2$ yrs as a function of κ, the credibility of the η interval [see Eq. (3)]. To obtain the rates in Table IV, different κ values are chosen depending on whether the initially favored result is normal or inverted. For the upper plot, this adjustment enables one to reduce the incorrect claim rate.

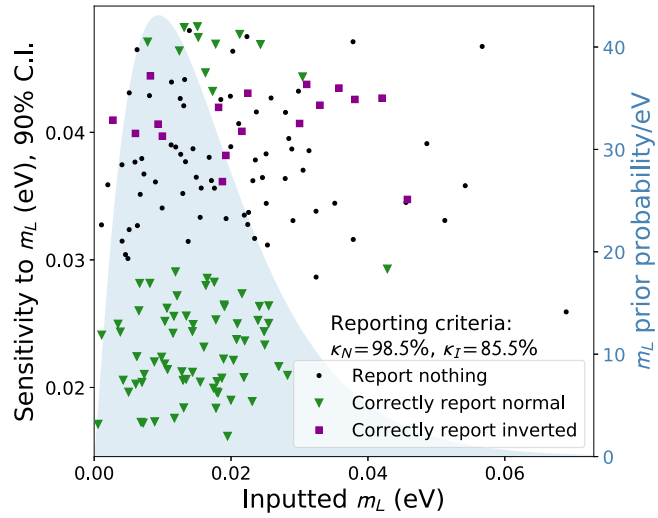

FIG. 10. Distribution of mass ordering results with respect to m_L and mass sensitivity ($\Delta t = 2$ yrs). The normal (inverted) ordering was reported when a 98.5% (85.5%) η credible interval excluded one ordering and was consistent with the other.

that as statistical power improves over time, sensitivity to the normal ordering improves.

These results indicate that a Project 8-like neutrino mass experiment could resolve the mass ordering for various likely combinations of physical and experimental parameter values. If the neutrinos obey a normal ordering and the lightest mass is constrained below ≈0.05 eV, then this analysis predicts there is a high chance of resolving the ordering after 2 yrs of data taking. We observe that a direct mass experiment would resolve the normal ordering especially often in the low-m_L, small mass sensitivity region (see Fig. 10).

For this study, we chose to employ a *two-neutrino* spectral model, as opposed to constraining the ordering based on a *single mass* measurement—for which $m_\beta \lesssim 48$ meV rules out an inverted ordering. While it would be impossible to resolve the inverted ordering using a one-neutrino model, a two-neutrino analysis can enable an inverted ordering determination. In other words, the process of inference is sensitive to fine structure near the endpoint of the spectrum produced by individual neutrino mass eigenstates.

V. CONCLUSIONS

In this paper, we presented a Bayesian approach to analyzing sensitivity to the neutrino mass scale and ordering. That approach included a calibration, which quantified the performance of two processes: inferring information and reporting results. Our sensitivity and calibration procedures are applicable to any experiment that produces information regarding the mass scale and ordering. These procedures also serve as templates for sensitivity studies by other physics experiments—whether they measure continuous or discrete parameters. As design planning for Project 8's final phase advances, future work will include a detailed analysis of systematic features to inform more precise priors in a Project 8-specific study.

Using the β spectrum model developed here, and given the experimental expectations in Sec. IV A, we find that a high-precision direct mass experiment could resolve the electron-weighted neutrino mass $m_\beta \approx m_1$ in a 90% credible interval, with a "true claim rate" or coverage of $(90.0 \pm 2.0)\%$. For very small m_β, the width of this interval approaches 40 meV, and for $m_\beta > 0.5$ eV, the average width is only 5 meV. A similar analysis may be employed to search for and measure the mass(es) of sterile neutrino states, each of which would produce one kink in the β spectrum.

This study also investigates the tritium β-decay technique's sensitivity to the neutrino mass ordering. We emphasize that, by using a utility function to judge whether to report an ordering result, it is possible not only to predict the probability of a false ordering claim, but also to determine a reporting

tolerance (here, the η interval credibility) that minimizes the risk of false claims. For the experimental parameters assumed here and a two-year runtime, we would recommend reporting a normal ordering result when a 98.5% posterior credible interval on the light-mass fraction η contains $|U_{e1}|^2 + |U_{e2}|^2$ but not $|U_{e3}|^2$. To report an inverted ordering determination, the opposite should hold for an 85.5% interval around η. Those reporting criteria enable the normal (inverted) ordering to be resolved ≈87% (22%) of the time, with a ≈0% false claim rate. It is also possible to infer posteriors on individual neutrino masses. When sensitivity to the lightest mass is better than 0.03 eV, it is nearly always possible to resolve the mass ordering.

These results demonstrate that we can access more information by modeling the full spectral shape than would be possible using a one-neutrino model in terms of m_β. As more events are detected, the spectral shape method becomes increasingly sensitive to count rate kinks that inform inferences about individual neutrino masses and their ordering. Direct mass experiments thus offer a unique potential probe of individual $|U_{ei}|$ matrix elements, complementary to oscillations-based probes of their products.

ACKNOWLEDGMENTS

The authors thank André de Gouvêa for insight and discussions. This material is based on work supported by the following sources: the US Department of Energy Office of Science, Office of Nuclear Physics, under Award No. DE-SC0020433 to Case Western Reserve University (CWRU), under Award No. DE-SC0011091 to the Massachusetts Institute of Technology (MIT), under the Early Career Research Program to Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by Battelle for the US Department of Energy under Contract No. DE-AC05-76RL01830, under Early Career Award No. DE-SC0019088 to Pennsylvania State University, under Award No. DE-FG02-97ER41020 to the University of Washington, and under Award No. DE-SC0012654 to Yale University; the National Science Foundation under Award No. PHY-1205100 to MIT; the Cluster of Excellence "Precision Physics, Fundamental Interactions, and Structure of Matter" (PRISMA+ EXC 2118/1) funded by the German Research Foundation (DFG) within the German Excellence Strategy (Project ID 39083149); the Laboratory Directed Research and Development (LDRD) 18-ERD-028 at Lawrence Livermore National Laboratory (LLNL), prepared by LLNL under Contracts No. DE-AC52-07NA27344 and No. LLNL-JRNL-817667; the LDRD program at PNNL; the University of Washington Royalty Research Foundation; Yale University; and the Karlsruhe Institute of Technology (KIT) Center Elementary Particle and Astroparticle Physics (KCETA). A portion of the research was performed using the Engaging cluster at the MGHPCC facility.

APPENDIX A: APPROXIMATE SPECTRAL MODEL

The approximate β spectral model for Bayesian inference in Eq. (7) has a corresponding cumulative distribution function. It is given by

$$\mathcal{G}_i^{\mathrm{CDF}}(K) = \int_K^\infty \mathcal{F}_i(K')dK' = [G_A(K|m_i, Q_T, \sigma) - G_B(K|m_i, Q_T, \sigma, K_{\min})]/C,$$

where

$$\begin{aligned}
G_A &= \mathcal{N}(Q_T - K|m_i, \sigma)2\sigma^2\left[4\sigma^2 - m_i^2 + 2m_i(Q_T - K) + 2(Q_T - K)^2\right] \\
&\quad + \mathrm{Erfc}\left(\frac{m_i - Q_T + K}{\sqrt{2}\sigma}\right)\left\{m_i^3 + (Q_T - K)\left[6\sigma^2 - 3m_i^2 + 2(Q_T - K)^2\right]\right\} \\
G_B &= \mathcal{N}(Q_T - K|Q_T - K_{\min}, \sigma)2\sigma^2\left\{4\sigma^2 - 3m_i^2 + 2[(Q_T - K_{\min})^2 - (Q_T - K_{\min})(Q_T - K) + (Q_T - K)^2]\right\} \\
&\quad + \mathrm{Erfc}\left(\frac{K - K_{\min}}{\sqrt{2}\sigma}\right)\left\{(Q_T - K_{\min})\left[3m_i^2 - 2(Q_T - K_{\min})^2\right] + (Q_T - K)\left[6\sigma^2 - 3m_i^2 + 2(Q_T - K)^2\right]\right\} \\
C &= [G_{\mathrm{high}}(K) - G_{\mathrm{low}}(K)]\big|_0^\infty.
\end{aligned}$$

We implemented this function in Stan and employed it to analyze fake spectra.

APPENDIX B: PRIOR DISTRIBUTION DEFINITIONS

The prior distributions used in this paper are defined as follows. Each distribution is implemented via a Stan function that outputs the log of the probability density of a parameter y [25].

(1) Normal distribution:

$$\mathcal{N}(\mu, \sigma) \equiv \mathcal{N}(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y - \mu}{\sigma}\right)^2\right].$$

(2) Gamma distribution:

$$\gamma(\alpha,\beta) \equiv \gamma(y|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1}\exp(-\beta y), \quad \text{where} \quad \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx.$$

(3) Log-normal distribution:

$$\text{lognorm}(\mu,\sigma) \equiv \text{lognorm}(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma y}\exp\left[-\frac{1}{2}\left(\frac{\log y-\mu}{\sigma}\right)^2\right].$$

[1] P. A. Zyla *et al.* (Particle Data Group), Prog. Theor. Exp. Phys. **2020**, 083C01 (2020).
[2] M. Betancourt, arXiv:1803.08393.
[3] R. J. Little, Am. Stat. **60**, 213 (2006).
[4] T. Sellke, M. J. Bayarri, and J. O. Berger, Am. Stat. **55**, 62 (2001).
[5] S. Lacoste-Julien, F. Huszár, and Z. Ghahramani, Proc. Mach. Learn. Res. **15**, 416 (2011).
[6] P. Gustafson and S. Greenland, Stat. Sci. **24**, 328 (2009).
[7] K. Abe *et al.* (Super-Kamiokande Collaboration), Phys. Rev. D **83**, 052010 (2011).
[8] J. Angrik *et al.* (KATRIN Collaboration), KATRIN Design Report (2004), 10.5445/IR/270060419.
[9] B. Aharmim *et al.* (SNO Collaboration), Phys. Rev. C **81**, 055504 (2010).
[10] K. Eguchi *et al.* (KamLAND Collaboration), Phys. Rev. Lett. **90**, 021802 (2003).
[11] X. Qian, A. Tan, W. Wang, J. J. Ling, R. D. McKeown, and C. Zhang, Phys. Rev. D **86**, 113011 (2012).
[12] J. A. Formaggio, Phys. Dark Univ. **4**, 75 (2014).
[13] B. Monreal and J. A. Formaggio, Phys. Rev. D **80**, 051301(R) (2009).
[14] A. Ashtari Esfahani *et al.* (Project 8 Collaboration), J. Phys. G **44**, 054004 (2017).
[15] S. Talts *et al.*, arXiv:1804.06788 [stat.ME].
[16] G. J. Feldman and R. D. Cousins, Phys. Rev. D **57**, 3873 (1998).
[17] J. Neyman and H. Jeffreys, Philos. Trans. R. Soc. Lond. Ser, A: Math. Phys. Sci. **236**, 333 (1937).
[18] M. J. Bayarri and J. O. Berger, Stat. Sci. **19**, 58 (2004).
[19] M. Betancourt, Probabilistic modeling and statistical inference (2019), https://github.com/betanalpha/knitr_case_studies/tree/master/modeling_and_inference, commit b474ec1a5a79347f7c9634376c866fe3294d657a.
[20] R. Hyndman *et al.*, Am. Stat. **50**, 120 (1996).
[21] X. Qian and P. Vogel, Prog. Part. Nucl. Phys. **83**, 1 (2015).
[22] R. Neal, Mcmc using hamiltonian dynamics, in *Handbook of Markov Chain Monte Carlo*, edited by Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng (Chapman & Hall/CRC, Boca Raton, FL, 2011), pp. 116–162.
[23] M. Betancourt, arXiv:1701.02434.
[24] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, J. Stat. Softw. **76**, 1 (2017).
[25] Stan modeling language users guide and reference manual, version 2.25 (Stan Development Team, 2020).
[26] M. Guigue, J. Formaggio, T. Weiss, J. Johnston, N. Oblath, and B. LaRoque, morpho (2020), github package, 10.5281/zenodo.3371729.
[27] L. I. Bodine, D. S. Parno, and R. G. H. Robertson, Phys. Rev. C **91**, 035505 (2015).
[28] M. Fertl, Hyperfine Interact. **239**, 52 (2018).
[29] M. Kleesiek, J. Behrens, G. Drexlin, K. Eitel, M. Erhard, J. A. Formaggio, F. Glück, S. Groh, M. Hötzel, S. Mertens, A. W. P. Poon, C. Weinheimer, and K. Valerius, Eur. Phys. J. C **79**, 204 (2019).
[30] M. Aker *et al.* (KATRIN Collaboration), Phys. Rev. Lett. **123**, 221802 (2019).
[31] E. G. Myers, A. Wagner, H. Kracke, and B. A. Wesson, Phys. Rev. Lett. **114**, 013003 (2015).
[32] P. J. Doe *et al.* (Project-8 Collaboration), arXiv:1309.7093 [nucl-ex].
[33] R. D. Williams and S. E. Koonin, Phys. Rev. C **27**, 1815 (1983).
[34] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. (Chapman & Hall/CRC, Boca Raton, FL, 2013).
[35] M. Betancourt, Toward a principled bayesian workflow (pystan) (2018), https://github.com/betanalpha/jupyter_case_studies/tree/master/principled_bayesian_workflow, commit 2580fdeac38f77859b2d1c60f9c4a37237864e63.
[36] M. Betancourt, Markov chain monte carlo (2020), https://github.com/betanalpha/knitr_case_studies/tree/master/markov_chain_monte_carlo, commit b474ec1a5a79347f7c9634376c866fe3294d657a.
[37] P. F. de Salas, D. V. Forero, S. Gariazzo, P. Martínez-Miravé, O. Mena, C. A. Ternes, M. Tórtola, and J. W. F. Valle, J. High Energy Phys. 02 (2021) 071.
[38] K. N. Abazajian *et al.*, Astropart. Phys. **35**, 177 (2011).