Efficient optimization of COHERENT detector design parameters with the Rare Event Surrogate Model (RESuM)

Brian Zhou Liu[1*] bzl002@ucsd.edu

Sonata Simonaitis-Boyd[1*] sonata@ucsd.edu

Ann-Kathrin Schuetz[2] aschuetz@lbl.gov

Aobo Li[1,3†] liaobo77@ucsd.edu

Zepeng Li[4†] zepengli@hawaii.edu

[1]Halıcıoğlu Data Science Institute, University of California, San Diego, La Jolla, CA, USA
[2]Nuclear Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[3]Department of Physics, UC San Diego, La Jolla, CA, USA
[4]Department of Physics & Astronomy, University of Hawai‘i at Mānoa, Honolulu, HI, USA
[*] Equal Contribution
[†] Corresponding Authors

Abstract

Coherent elastic neutrino-nucleus scattering (CEνNS) is a weak neutral-current process in which a neutrino scatters off of a nucleus as a whole. Following an initial observation by the COHERENT collaboration in 2017, the next-generation COH-Ar-750 detector is being developed to measure CEνNS with percent-level precision and probe for physics beyond the Standard Model. A primary design challenge is mitigating neutron backgrounds from the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory (ORNL), as neutron-induced nuclear recoils produce signals that are nearly indistinguishable from CEνNS. The detector's veto system combines passive shielding using lead and water blocks with active shielding using plastic scintillator panels. Because neutrons have a low probability of depositing energy in the active liquid argon volume, extensive Monte Carlo event simulations are required to optimize the arrangement of these shielding materials. In this work, we addressed this challenge with a Rare Event Surrogate Model (RESuM) to optimize the shielding design for COH-Ar-750. RESuM integrates a Conditional Neural Process (CNP) with a Multi-Fidelity Gaussian Process (MFGP) to reduce the need for expensive simulations. Our experimental results suggested thicker water shielding and thinner veto panels increase neutron rejection efficiency, with RESuM achieving a correlation coefficient of $r = 0.880$ and well-calibrated uncertainties. This work demonstrates that RESuM has the potential to accelerate the design optimization of a broad range of rare event search experiments.

1 Introduction

The Standard Model (SM) of particle physics represents the prevailing theory of how known elementary particles interact. In the SM, coherent elastic neutrino–nucleus scattering (CEνNS) is a process in which a neutrino transfers a small amount of momentum to an entire atomic nucleus via the neutral weak current, causing the nucleus to recoil without breaking apart or exciting internal states. CEνNS was first detected by the COHERENT experimental collaboration in 2017 [1]. Since then, COHERENT has been dedicated to installing more detectors to better understand CEνNS. The

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

newest planned detector is the COH-Ar-750 [2] detector, designed to deliver percent-level CEνNS measurements with liquid argon (LAr) detector technology.

A central challenge designing COH-Ar-750 is predicting and controlling neutron backgrounds from the Spallation Neutron Source (SNS) [3], an accelerator-based neutron production site at Oak Ridge National Laboratory (ORNL). Although the SNS produces a copious flux of neutrinos from pion decay, it also generates intense neutron fluxes. Since neutron-induced nuclear recoils in LAr are indistinguishable from CEνNS in a LAr detector, even small residual backgrounds can overshadow the signal of interest and reduce experimental sensitivity. To mitigate this, the experiment relies on both passive shielding (lead and water blocks) and active shielding (plastic scintillator veto panels). Optimizing the construction of these design parameters is critical for maximizing background rejection while keeping the experiment practical. Traditionally, optimization would involve Monte Carlo estimation; however, since neutrons have a low probability of depositing energy in the active LAr volume (i.e. as a rare event), extensive Monte Carlo event simulations would be needed to accurately find these optimal parameters.

The Rare Event Surrogate Model (RESuM) [4] was created by LEGEND [5] [6], a neutrinoless double-beta ($0\nu\beta\beta$) decay [7] experiment, to optimize their neutron shield design. RESuM combines a Conditional Neural Process (CNP) [8] with a Multi-Fidelity Gaussian Process (MFGP) [9] [10] to tackle rare event-limited design optimization. The CNP captures predictive structure from abundant low-fidelity simulations, while the MFGP fuses these predictions with sparse high-fidelity simulations to predict experimental performance across the design space. In this work[1], we show results of the application and validation of RESuM on COH-Ar-750 simulations, providing indications of optimized neutron rejection efficacy based on different water shielding and veto panel thickness configurations. The GitHub repository is available here.

2 Methods

2.1 COH-Ar-750 simulation

The training data for RESuM were generated through Monte Carlo simulations of the COH-Ar-750 detector geometry using particle simulation toolkit Geant4 [11] [12] [13], focusing on neutron backgrounds originating from the SNS. Each simulated event is defined by (i) a single design parameter θ, indicating the thickness of the veto panel[2], and (ii) event-level parameters including the incoming neutron's position, momentum, and energy, as well as energy depositions in the nuclear recoil volume and the active veto scintillator panels. The resulting event-level dataset therefore encodes both geometry-dependent shielding effects and stochastic neutron transport outcomes, providing the raw inputs for training the surrogate model.

When each neutron propagates within the Geant4 simulation framework, its final state can be parameterized by a binary target variable, `target_active`, which indicates rare but problematic outcomes where a neutron successfully enters the active LAr detector volume by penetrating the veto panels. Therefore, `target_active` is set to 1 when `veto_active` $= 0$ and `detector_active` $= 1$, corresponding to a veto miss coincident with a detector hit. Across 100 distinct shielding configurations θ_k—97 low-fidelity (LF) simulations each with $N = 42{,}000$ events and 3 high-fidelity (HF) simulations each with $N = 1{,}002{,}000$ events—the dataset comprises 7,080,000 simulated neutron events, of which 227,049 satisfy `target_active` $= 1$. Averaged over configurations, this corresponds to an overall rate of $\sim 3.21\%$, i.e. approximately 1.35×10^3 `target_active` events per LF simulation and 3.21×10^4 per HF simulation. Since `target_active` has a low occurrence rate as a rare event, training with RESuM allowed us to avoid the cost of extensive high-statistics simulations.

The design vector $\theta =$ (`veto_thickness_mm`) varies within $[10,\ 110]$ mm. For each θ, Geant4 produced neutron events with event-level parameters

$$\phi = (x, y, z,\ E,\ p_x, p_y, p_z,\ E_{\mathrm{NR}},\ E_{\mathrm{veto}}),$$

[1] This is not a COHERENT project, and is not endorsed by the COHERENT collaboration.

[2] Water shielding thickness (`water_shielding_mm`) and veto panel thickness (`veto_thickness_mm`) sum up to a fixed value (126.2 mm), therefore we only define one design parameter.

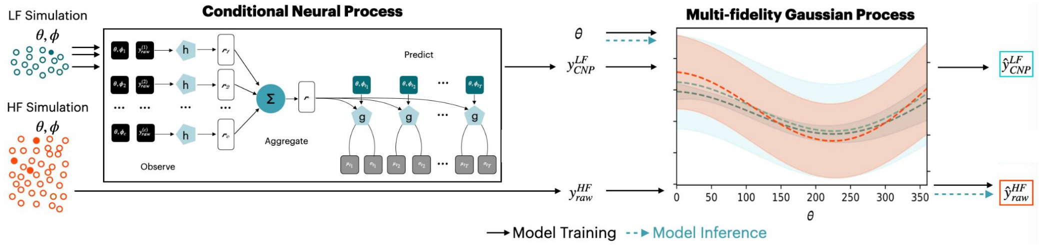


Figure 1: **Overview of RESUM framework**. The CNP trains on LF data to generate "probabilities" of signal, represented as $y_{\text{CNP}}^{\text{LF}}$ values. These predicted values, along with batched "probabilities" from HF simulation data $y_{\text{raw}}^{\text{HF}}$, are fed into the MFGP to estimate design metrics $\hat{y}_{\text{CNP}}^{\text{LF}}$, which are used for validation of $\hat{y}_{\text{raw}}$ for unseen LF configurations.

corresponding respectively to the neutron's generation position, kinetic energy, momentum, and energy depositions in the nuclear-recoil volume and veto panels. The binary target variable for an event i is $X_i \in \{0, 1\}$, where $X = 1$ indicates $\texttt{target\_active} = 1$.

2.2 RESUM overview

Similar to COH-Ar-750, RESUM was originally created to find the optimal design configuration of the LEGEND neutron shield. RESUM consists of three main algorithm parts: (i) an event-level CNP, (ii) a configuration-level MFGP, and (iii) an adaptive sampling technique for guiding additional HF simulation θs. The simulations produced for LEGEND were structured similarly to those described for COH-Ar-750 in the previous subsection, with design vectors θ, event-level parameter vectors ϕ, and binary target variables X, with data from both low- and high-fidelity simulations. The role of the CNP is to smooth out the highly-discrete, binary 0–1 target variable into a continuous probability of detection β for each event. These β values are then averaged over all events for a given θ to produce design metric y_{CNP}, which can be considered a smoothed approximation of y_{raw} at lower fidelity. Overall, three design parameter-specific metrics are calculated following the CNP step: y_{raw}, the fraction of events satisfying $\texttt{target\_active} = 1$ for a given θ_k in a HF simulation, calculated with the following definition:

$$y_{\text{raw}} = \frac{\sum X_i}{N},$$

and the ultimate design metric we want to minimize; $y_{\text{CNP}}^{\text{HF}}$, the CNP score calculated for HF events; and $y_{\text{CNP}}^{\text{LF}}$, the CNP score calculated for LF events. These three metrics are used to inform the MFGP.

The role of the MFGP, built from the Emukit Python library [14], is to interpolate HF data with LF data, since the latter are comparatively computationally cheap to simulate and allow for coverage of a broader range of design parameters. The MFGP can not only emulate multiple fidelities but also output uncertainties for its emulation, and during inference, the model is expected to predict y_{raw} for out-of-sample datasets. The MFGP step is followed by the adaptive learning technique, an acquisition function allowing RESUM to converge on the most optimal design parameters with the goal of informing future HF simulation runs. See Appendix A.1 and Ref. [4] for details of the RESUM framework.

In this work, we inherited the same structure as Ref. [4], albeit without the application of active learning. This work also used Matérn 3/2 kernels instead of Radial Basis Function (RBF) kernels for the MFGP. In addition, our MFGP was trained only on $y_{\text{CNP}}^{\text{LF}}$ and $y_{\text{raw}}^{\text{HF}}$. The full end-to-end workflow is summarized in Figure 1. Training and validation details of RESUM in this work can be found in Appendix A.2.

3 Results & Discussion

Using 9 representative LF CNP configurations and 3 raw HF configurations, the MFGP learned to predict y_{raw} across 88 held-out LF θ configurations. Figure 2a compares the median LF rate $y_{\text{raw}}^{\text{LF}}$ with median CNP outputs $y_{\text{CNP}}^{\text{LF}}$ across all validation configurations. While $y_{\text{raw}}^{\text{LF}}$ fluctuates due to limited counts in the $0.02 - 0.06$ range, $y_{\text{CNP}}^{\text{LF}}$ displays significantly less fluctuation across a broader range of

Model	Dataset (#train, #val)	$1\hat{\sigma}$ [%]	$2\hat{\sigma}$ [%]	$3\hat{\sigma}$ [%]
COH-RESuM	(9 LF + 3 HF, 88 LF)	71.6	96.6	100
Proper statistical coverage		68.27	95.45	99.73

Table 1: Benchmarking result of RESuM applied to COH-Ar-750.

$0.14 - 0.28$. As a result, $y_{\mathrm{CNP}}^{\mathrm{LF}}$ reveals a clear overall trend where thicker water shielding and thinner veto panels reduce the neutron background probability. While this seems unexpected—the veto panels were designed to shield from neutrons, and in theory should do that better than water—Figure 2b clearly demonstrates that the RESuM model is correct and faithfully reproduces the overall trend of the Geant4 simulations. Although outside the scope of this project, it appears that the veto efficiency of the Geant4 simulations has a weak dependence on thickness, and the `detector_active` fraction has a strong dependence. Further investigation would be required to determine the reason behind these dependencies.

Figure 2b shows the predicted means with calibrated $\pm 1\hat{\sigma}$, $\pm 2\hat{\sigma}$, and $\pm 3\hat{\sigma}$ bands, the results of which are listed in Table 1. Statistical coverage was calculated by averaging y_{raw} for each of the 88 configurations and counting the means falling within the uncertainty bands, with $\hat{\sigma}$ predicted by the MFGP. The MFGP reproduced the aforementioned overall decreasing dependence of background probability on shielding geometry, with predictions correlating strongly with LF validation values ($r = 0.880$). Thus, the MFGP demonstrates reliable uncertainty quantification in addition to accurate means.

Furthermore, RESuM produced these results with significantly reduced data and computation costs. Without RESuM, traditional exploration of COH-Ar-750's shielding design space with the MFGP would require 12 HF simulations (12,024,000 events). With RESuM, 9 of those HF simulations were replaced by LF simulations; only 3 HF and 9 LF simulations were needed (3,384,000 events), saving 8,640,000 events ($\sim 72\%$ fewer).

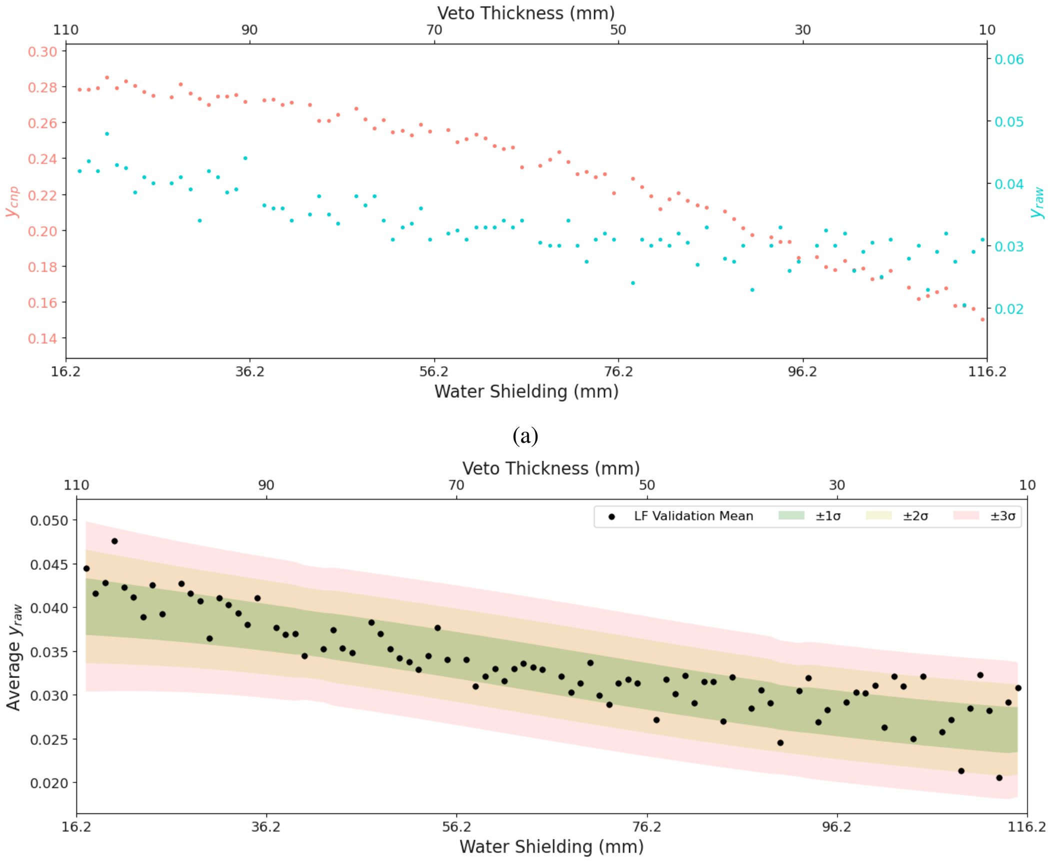
4 Conclusion

This study demonstrates that RESuM provides an efficient approach to optimizing the shielding configuration of the COH-Ar-750 detector. By training a CNP on abundant LF simulations, RESuM produced uncertainty-aware predictors that revealed an overall decreasing trend in neutron background as water shielding increases (and a corresponding increase with veto thickness under the fixed-thickness constraint). These features were then transferred to the MFGP, which fused them with sparse HF simulations to produce predictions for the configuration-level raw rate y_{raw}. The surrogate achieved a strong correlation with LF validation results ($r = 0.880$) and statistical coverage closely matching Gaussian expectations (71.6% within $1\hat{\sigma}$, 96.6% within $2\hat{\sigma}$, and 100% within $3\hat{\sigma}$), confirming predictive accuracy. RESuM also showed it could offer a more resource-efficient method of design optimization, requiring over 70% less data by utilizing both LF and HF simulations instead of only HF simulations.

Beyond this specific application to COH-Ar-750, RESuM demonstrates broader potential for detector design optimization across rare event search experiments. Future work will focus on (i) generating additional out-of-sample Geant4 simulations to further validate the trained model, (ii) conducting interpretability studies to understand the decreasing trends identified by RESuM, and (iii) extending the approach to other rare event detector design tasks such as LArTPC calibration source optimization. RESuM's ability to efficiently search large design spaces while providing uncertainty-aware metric predictions positions it as a powerful tool for maximizing the discovery potential of next-generation rare event search experiments.

References

[1] D. Akimov, J. B. Albert, P. An, C. Awe, P. S. Barbeau, B. Becker, V. Belov, A. Brown, A. Bolozdynya, B. Cabrera-Palmer, M. Cervantes, J. I. Collar, R. J. Cooper, R. L. Cooper, C. Cuesta, D. J. Dean, J. A. Detwiler, A. Eberhardt, Y. Efremenko, S. R. Elliott, E. M. Erkela, L. Fabris, M. Febbraro, N. E. Fields, W. Fox, Z. Fu, A. Galindo-Uribarri, M. P. Green, M. Hai, M. R. Heath, S. Hedges, D. Hornback, T. W. Hossbach, E. B. Iverson, L. J. Kaufman, S. Ki, S. R. Klein,



(a)

(b)

Figure 2: **RESuM results on COH-Ar-750 shielding.** (a) Conditional Neural Process (CNP) validation: median raw rates ($y_{\text{raw}}^{\text{LF}}$, cyan) fluctuate strongly due to rare event discreteness, while CNP outputs ($y_{\text{CNP}}^{\text{LF}}$, salmon) recover smooth and physically consistent trends across water and veto thickness. (b) Multi-Fidelity Gaussian Process (MFGP) predictions with uncertainty: black dots show LF validation means and shaded bands denote RESuM $\pm 1\hat{\sigma}$ (green), $\pm 2\hat{\sigma}$ (yellow), and $\pm 3\hat{\sigma}$ (red) intervals with coverage 71.6%, 96.6%, and 100%, respectively.

A. Khromov, A. Konovalov, M. Kremer, A. Kumpan, C. Leadbetter, L. Li, W. Lu, K. Mann, D. M. Markoff, K. Miller, H. Moreno, P. E. Mueller, J. Newby, J. L. Orrell, C. T. Overman, D. S. Parno, S. Penttila, G. Perumpilly, H. Ray, J. Raybern, D. Reyna, G. C. Rich, D. Rimal, D. Rudik, K. Scholberg, B. J. Scholz, G. Sinev, W. M. Snow, V. Sosnovtsev, A. Shakirov, S. Suchyta, B. Suh, R. Tayloe, R. T. Thornton, I. Tolstukhin, J. Vanderwerp, R. L. Varner, C. J. Virtue, Z. Wan, J. Yoo, C.-H. Yu, A. Zawada, J. Zettlemoyer, A. M. Zderic, and COHERENT Collaboration. Observation of coherent elastic neutrino-nucleus scattering. *Science*, 357(6356):1123–1126, 2017.

[2] Jeremy Lu. Status of coh-ar from the coherent collaboration, 2025. Slideshow presented at Magnificent CEvNS 2025.

[3] T. E. Mason, T. A. Gabriel, R. K. Crawford, K. W. Herwig, F. Klose, and J. F. Ankner. The Spallation Neutron Source: A Powerful Tool for Materials Research, 2000.

[4] Ann-Kathrin Schuetz, Alan W. P. Poon, and Aobo Li. RESuM: Rare Event Surrogate Model for Physics Detector Design. Preprint for the International Conference on Learning Representations (ICLR), 2025. Code available at `https://github.com/annkasch/resum`.

[5] Large enriched germanium experiment for neutrinoless $\beta\beta$ decay. `https://legend-exp.org/`. Accessed: 2025-08-28.

[6] H. Acharya, N. Ackermann, M. Agostini, A. Alexander, C. Andreoiu, G. R. Araujo, F. T. Avignone III, M. Babicz, W. Bae, A. Bakalyarov, M. Balata, A. S. Barabash, P. S. Barbeau, C. J. Barton, L. Baudis, C. Bauer, E. Bernieri, L. Bezrukov, K. H. Bhimani, V. Biancacci, E. Blalock, S. J. Borden, G. Borghi, F. Borra, B. Bos, A. Boston, V. Bothe, R. Bouabid, R. Brugnera, N. Burlac, M. Busch, S. Calgaro, L. Canonica, S. Capra, M. Carminati, R. M. D. Carney, C. Cattadori, R. Cesarano, Y. D. Chan, J. R. Chapman, A. Chernogorov, P. J. Chiu, C. D. Christofferson, M. L. Clark, A. I. Colon-Rivera, T. Comellato, V. D'Andrea, R. Deckert, J. A. Detwiler, A. Di Giacinto, N. Di Marco, T. Dixon, K. M. Dong, A. Drobizhev, G. Duran, Yu. Efremenko, S. R. Elliott, C. H. J. Emmanuel, E. Engelhardt, E. Esch, M. T. Febbraro, F. Ferella, D. E. Fields, C. Fiorini, M. Fomina, N. Fuad, R. Gala, A. Galindo-Uribarri, A. Gangapshev, A. Garfagnini, S. Gazzana, A. Geraci, L. Gessler, C. Ghiano, A. Gieb, S. Giri, M. Gold, C. Gooch, G. Grünauer, M. P. Green, J. Gruszko, I. Guinn, V. E. Guiseppe, V. Gurentsov, Y. Gurov, K. Gusev, B. Hackett, F. Hagemann, M. Haranczyk, F. Henkes, R. Henning, J. Herrera, D. Hervas Aguilar, J. Hinton, R. Hodák, H. F. R. Hoffmann, M. A. Howe, M. Huber, M. Hult, A. Ianni, K. Jędrzejczak, J. Jochum, R. W. L. Jones, D. S. Judson, M. Junker, J. Kaizer, V. Kazalov, M. F. Kidd, T. Kihm, K. Kilgus, A. Klimenko, K. T. Knöpfle, I. Kochanek, O. Kochetov, I. Kontul, L. L. Kormos, V. N. Kornoukhov, P. Krause, H. Krishnamoorthy, V. V. Kuzminov, K. Lang, M. Laubenstein, N. N. P. N. Lay, E. León, A. Leder, B. Lehnert, A. Leonhardt, N. Levashko, L. Y. Li, A. Li, Y. R. Lin, M. Lindner, I. Lippi, A. Love, A. Lubashevskiy, B. Lubsandorzhiev, N. Lusardi, C. Macolino, B. Majorovits, F. Mamedov, L. Manzanillas, G. G. Marshall, R. D. Martin, E. L. Martin, R. Massarczyk, A. Mazumdar, G. McDowell, D. M. Mei, S. P. Meireles, M. Menzel, S. Mertens, E. Miller, I. Mirza, M. Misiaszek, M. Morella, B. Morgan, T. Mroz, D. Muenstermann, C. J. Nave, I. Nemchenok, M. Neuberger, N. O'Briant, F. Paissan, L. Papp, L. S. Paudel, K. Pelczar, L. Pertoldi, W. Pettus, F. Piastra, M. Pichotta, P. Piseri, A. W. P. Poon, P. P. Povinec, M. Pruckner, A. Pullia, W. S. Quinn, D. C. Radford, Y. A. Ramachers, A. Razeto, M. Redchuk, A. L. Reine, S. Riboldi, K. Rielage, C. Romo-Luque, N. Rossi, S. Rozov, T. J. Ruland, N. Rumyantseva, J. Runge, R. Saakyan, S. Sailer, G. Salamanna, F. Salamida, G. Saleh, V. Sandukovsky, C. Savarese, S. Schönert, A. K. Schütz, D. C. Schaper, L. Schlüter, S. J. Schleich, O. Schulz, M. Schwarz, B. Schwingenheuer, C. Seibt, O. Selivanenko, G. Senatore, A. Serafini, K. Shakhov, E. Shevchik, M. Shirchenko, Y. Shitov, H. Simgen, F. Šimkovic, S. Simonaitis-Boyd, M. Skorokhvatov, M. Slavíčková, A. Smolnikov, J. A. Solomon, G. Song, A. C. Sousa, A. R. Sreekala, L. Steinhart, I. Štekl, T. Sterr, M. Stommel, S. A. Sullivan, R. R. Sumathi, K. Szczepaniec, L. Taffarello, D. Tagnani, D. J. Tedeschi, T. N. Thorpe, V. Tretyak, M. Turqueti, E. E. Van Nieuwenhuizen, L. J. Varriano, S. Vasilyev, A. Veresnikova, C. Vignoli, C. Vogl, K. von Sturm, A. Warren, D. Waters, S. L. Watkins, C. Wiesinger, J. F. Wilkerson, M. Willers, C. Wiseman, M. Wojcik, D. Xu, W. Xu, E. Yakushev, T. Ye, C. H. Yu, V. Yumatov, D. Zinatulina, K. Zuber, and G. Zuzel. First results on the search for lepton number violating neutrinoless double beta decay with the legend-200 experiment, 2025.

[7] Michelle J. Dolinski, Alan W.P. Poon, and Werner Rodejohann. Neutrinoless double-beta decay: Status and prospects. *Annual Review of Nuclear and Particle Science*, 69(Volume 69, 2019):219–251, 2019.

[8] Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional Neural Processes, Jul 2018.

[9] M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[10] Peter Z. Qian and C. F. Wu. Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50(2):192–204, May 2008.

[11] J. Allison, K. Amako, J. Apostolakis, P. Arce, M. Asai, T. Aso, E. Bagli, A. Bagulya, S. Banerjee, G. Barrand, B.R. Beck, A.G. Bogdanov, D. Brandt, J.M.C. Brown, H. Burkhardt, Ph. Canal, D. Cano-Ott, S. Chauvie, K. Cho, G.A.P. Cirrone, G. Cooperman, M.A. Cortés-Giraldo, G. Cosmo, G. Cuttone, G. Depaola, L. Desorgher, X. Dong, A. Dotti, V.D. Elvira, G. Folger, Z. Francis, A. Galoyan, L. Garnier, M. Gayer, K.L. Genser, V.M. Grichine, S. Guatelli, P. Guèye, P. Gumplinger, A.S. Howard, I. Hřivnáčová, S. Hwang, S. Incerti, A. Ivanchenko, V.N. Ivanchenko, F.W. Jones, S.Y. Jun, P. Kaitaniemi, N. Karakatsanis, M. Karamitros, M. Kelsey,

A. Kimura, T. Koi, H. Kurashige, A. Lechner, S.B. Lee, F. Longo, M. Maire, D. Mancusi, A. Mantero, E. Mendoza, B. Morgan, K. Murakami, T. Nikitina, L. Pandola, P. Paprocki, J. Perl, I. Petrović, M.G. Pia, W. Pokorski, J.M. Quesada, M. Raine, M.A. Reis, A. Ribon, A. Ristić Fira, F. Romano, G. Russo, G. Santin, T. Sasaki, D. Sawkey, J.I. Shin, I.I. Strakovsky, A. Taborda, S. Tanaka, B. Tomé, T. Toshito, H.N. Tran, P.R. Truscott, L. Urban, V. Uzhinsky, J.M. Verbeke, M. Verderi, B.L. Wendt, H. Wenzel, D.H. Wright, D.M. Wright, T. Yamashita, J. Yarba, and H. Yoshida. Recent developments in geant4. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 835:186–225, 2016.

[12] J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce Dubois, M. Asai, G. Barrand, R. Capra, S. Chauvie, R. Chytracek, G.A.P. Cirrone, G. Cooperman, G. Cosmo, G. Cuttone, G.G. Daquino, M. Donszelmann, M. Dressel, G. Folger, F. Foppiano, J. Generowicz, V. Grichine, S. Guatelli, P. Gumplinger, A. Heikkinen, I. Hrivnacova, A. Howard, S. Incerti, V. Ivanchenko, T. Johnson, F. Jones, T. Koi, R. Kokoulin, M. Kossov, H. Kurashige, V. Lara, S. Larsson, F. Lei, O. Link, F. Longo, M. Maire, A. Mantero, B. Mascialino, I. McLaren, P. Mendez Lorenzo, K. Minamimoto, K. Murakami, P. Nieminen, L. Pandola, S. Parlati, L. Peralta, J. Perl, A. Pfeiffer, M.G. Pia, A. Ribon, P. Rodrigues, G. Russo, S. Sadilov, G. Santin, T. Sasaki, D. Smith, N. Starkov, S. Tanaka, E. Tcherniaev, B. Tome, A. Trindade, P. Truscott, L. Urban, M. Verderi, A. Walkden, J.P. Wellisch, D.C. Williams, D. Wright, and H. Yoshida. Geant4 developments and applications. *IEEE Transactions on Nuclear Science*, 53(1):270–278, 2006.

[13] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, F. Behner, L. Bellagamba, J. Boudreau, L. Broglia, A. Brunengo, H. Burkhardt, S. Chauvie, J. Chuma, R. Chytracek, G. Cooperman, G. Cosmo, P. Degtyarenko, A. Dell'Acqua, G. Depaola, D. Dietrich, R. Enami, A. Feliciello, C. Ferguson, H. Fesefeldt, G. Folger, F. Foppiano, A. Forti, S. Garelli, S. Giani, R. Giannitrapani, D. Gibin, J.J. Gómez Cadenas, I. González, G. Gracia Abril, G. Greeniaus, W. Greiner, V. Grichine, A. Grossheim, S. Guatelli, P. Gumplinger, R. Hamatsu, K. Hashimoto, H. Hasui, A. Heikkinen, A. Howard, V. Ivanchenko, A. Johnson, F.W. Jones, J. Kallenbach, N. Kanaya, M. Kawabata, Y. Kawabata, M. Kawaguti, S. Kelner, P. Kent, A. Kimura, T. Kodama, R. Kokoulin, M. Kossov, H. Kurashige, E. Lamanna, T. Lampén, V. Lara, V. Lefebure, F. Lei, M. Liendl, W. Lockman, F. Longo, S. Magni, M. Maire, E. Medernach, K. Minamimoto, P. Mora de Freitas, Y. Morita, K. Murakami, M. Nagamatu, R. Nartallo, P. Nieminen, T. Nishimura, K. Ohtsubo, M. Okamura, S. O'Neale, Y. Oohata, K. Paech, J. Perl, A. Pfeiffer, M.G. Pia, F. Ranjard, A. Rybin, S. Sadilov, E. Di Salvo, G. Santin, T. Sasaki, N. Savvas, Y. Sawada, S. Scherer, S. Sei, V. Sirotenko, D. Smith, N. Starkov, H. Stoecker, J. Sulkimo, M. Takahata, S. Tanaka, E. Tcherniaev, E. Safai Tehrani, M. Tropeano, P. Truscott, H. Uno, L. Urban, P. Urban, M. Verderi, A. Walkden, W. Wander, H. Weber, J.P. Wellisch, T. Wenaus, D.C. Williams, D. Wright, T. Yamada, H. Yoshida, and D. Zschiesche. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

[14] Andrei Paleyes, Maren Mahsereci, and Neil D. Lawrence. Emukit: A Python toolkit for decision making under uncertainty. *Proceedings of the Python in Science Conference*, 2023.

A Appendix

A.1 RESUM framework

Batch level (CNP). In the simulations, each shielding configuration is indexed by k, and within that configuration each neutron event is indexed by i. The binary outcome $X_{ki} \in \{0, 1\}$ indicates whether the i-th event under shielding design θ_k both enters the detector and escapes veto detection. Directly modeling these rare Bernoulli outcomes produces high variance at small N. To stabilize statistics, we partition each LF simulation into fixed-size batches indexed by B, and compute batch-averaged event-level features together with the corresponding average of the raw binary outcomes. These batch-level summaries are then used as inputs to the CNP. The CNP encoder aggregates across context batches, and the decoder maps a batch summary to a Gaussian-distributed output, from which we obtain one probability per batch:

$$y_{\mathrm{CNP}}^{\mathrm{LF}} \;\approx\; \Pr(X = 1 \mid \theta, B).$$

This representation smooths the discrete LF outcomes while preserving sensitivity to shielding design parameters.

Configuration level (MFGP). At the next stage, the surrogate learns how LF CNP batch probabilities transfer to HF outcomes. Let $f_L(\theta)$ denote the latent LF function corresponding to $y_{\mathrm{CNP}}^{\mathrm{LF}}$ and $f_H(\theta)$ the HF function corresponding to the raw HF tally $y_{\mathrm{raw}}^{\mathrm{HF}} = m/N$, for N total background events and m number of background events satisfying $\texttt{target\_active} = 1$. We fit a two-level co-kriging model with the standard autoregressive form

$$f_H(\theta) \;=\; \rho\, f_L(\theta) \;+\; \delta(\theta),$$

following [4] with $f_L \sim \mathcal{GP}(0, k_L)$, $\delta \sim \mathcal{GP}(0, k_\delta)$. It should be noted that while the Radial Basis Function (RBF) kernel was used for k_L and k_δ in the original study, Matérn 3/2 kernels were chosen for this experiment, as during initial model tuning the Matérn kernels produced better $\hat{y}_{\texttt{raw}}$ and $\hat{\sigma}$ predictions due to greater exploration of the parameter space. Later tuning also revealed that the coverage produced by the Matérn 3/2 kernel was closer to proper statistical coverage compared to that by the RBF kernel. The training datasets are abundant LF pairs $\{(\theta_k, y_{\mathrm{CNP}}^{\mathrm{LF}}\}$ and sparse HF pairs $\{(\theta_k, y_{\mathrm{raw}}^{\mathrm{HF}}\}$, with each pair representing a batch. Conditioning the joint GP then yields posterior means and variances for unseen LF configurations, providing calibrated HF-level predictions with quantified uncertainty.

A.2 RESUM training and validation

CNP training and feature extraction. Since y_{raw} has high statistical variance by being aggregated from discrete rare-event outcomes, the CNP "converts" these values to continuous y_{CNP} scores, substantially reducing variance while preserving the target variable's dependence on θ. The CNP model was trained on the event-level data from only the 9 designated LF simulations (chosen evenly across the design space) and then used as a feature extractor. For the simulations designated for MFGP training (9 LF and 3 HF), these batch-level scores were used directly. For the 88 held-out simulations, $y_{\mathrm{CNP}}^{\mathrm{LF}}$ were generated to create the graph in Figure 2a.

MFGP training and validation. The MFGP was trained to map the smoothed LF features to the true HF outcomes using fine-grained data from the training simulations. It was trained on two datasets at the batch level:

- **Low-fidelity data:** The set of all batch-level pairs $\{(\theta_k, y_{\mathrm{CNP}}^{\mathrm{LF}})\}$ from the 9 LF training simulations.
- **High-fidelity data:** The set of all batch-level pairs $\{(\theta_k, y_{\mathrm{raw}}^{\mathrm{HF}})\}$, where $y_{\mathrm{raw}}^{\mathrm{HF}}$ is the ground-truth background rate for each batch from the 3 expensive HF simulations.

For validation, the model's predictive performance was assessed at the configuration level. The MFGP predicted the final background rates for the 88 held-out LF configurations, using their respective θ_k as input. These predictions, along with their corresponding $\pm 1\hat{\sigma}$, $\pm 2\hat{\sigma}$, and $\pm 3\hat{\sigma}$ uncertainty bands, were then compared to the mean of all batch-level $y_{\mathrm{raw}}^{\mathrm{LF}}$ in each LF configuration to generate the final validation plots and assess the model's statistical coverage.