LEARNING PHYSICS, NOT ENERGY: A MULTIMODAL BENCHMARK FOR NEUTRINO DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

The experimental search for neutrinoless double-beta decay (NLDBD) would answer one of the most important questions in physics: Why is there more matter than antimatter in our universe? A key challenge in NLDBD searches is identifying NLDBD signal events from many unwanted background events. Deep learning classifiers have shown strong performance on this task, but they often exhibit energy shortcut learning: instead of learning event topology, which reflects the underlying physical interactions, they exploit event energy as an easier shortcut to make classification decision. In this work, we conducted a comprehensive multimodal benchmark for NLDBD event classification. This work makes three contributions: First, we coordinated the public release of four NLDBD datasets spanning three modalities, including time series, point clouds, and sparse 2D images. Second, we designed two complementary metrics to quantitatively characterize topology-based classification power and the severity of energy shortcut learning. Third, we designed 13 models to benchmark across four datasets, including both native-representation models and a unified Transformer backbone adapted through different tokenization and positional encoding. Our results identify the Region + MLP Transformer as the strongest overall model across all datasets, establishing a baseline architecture and evaluation framework toward multimodal foundation models for NLDBD experiments. The anonymized Github Repository is provided at [1]

1 INTRODUCTION

Why is there more matter than antimatter in our universe? This is one of the most fundamental questions in physics, yet it remains unsolved. One of the most promising milestone toward answering this question is the potential discovery of Neutrinoless Double-Beta Decay (NLDBD). If such a discovery happens, it would represent a major step toward explaining the matter-antimatter asymmetry and would undoubtedly be a Nobel Prize-worthy breakthrough in physics. Due to its utmost importance, the U.S. nuclear physics community has recommended the experimental search for NLDBD as its highest priority (NSAC, 2023).

NLDBD can be searched for using different detector technologies that produce substantially different data modalities, including waveforms, point clouds, particle tracks, and sparse images. Each data point corresponds to a physics **event**[2] recorded by the detector. A central challenge in NLDBD experiments is rejecting background events: physics events that are not NLDBD, such as radioactive decay or photon scattering. Background events leave detector signatures that differ from those of NLDBD events. We refer to these signature as **event topology**—the spatial and/or temporal pattern through which the underlying physical interaction is represented in the data modality. Deep learning classifiers have shown great success in learning such complex event topology directly from detector data. In practice, however, such classifier may instead exploit a much simpler scalar feature—event energy. This problem is defined as energy shortcut learning.

This work is the first multimodal benchmarking study across different NLDBD detector technologies. The goal is to identify models that learn the underlying physics—the event topology—rather

[1] https://anonymous.4open.science/r/0vBB-Benchmark-1F4B/

[2] We therefore use the terms "event" and "data point" interchangeably throughout this paper.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

than merely exploiting event energy. We made three key contributions: 1. We coordinated the release of 4 public, AI-ready datasets spanning different detector technologies and data modalities from the MAJORANA DEMONSTRATOR, NEXT, SuperNEMO, and EXO-200. 2. We introduce two complementary metrics: the energy-weighted AUC to evaluate classification power without energy shortcut, and the energy independence score I to characterize the severity of energy shortcut learning. 3. We benchmarked 13 models under two approaches—a native-representation approach and a unified Transformer approach—across all four datasets. The benchmarking result helped us identify a strong candidate that could serve as the baseline design for NLDBD multimodal foundation model in the future.

2 RELATED WORKS

Deep learning has been applied to NLDBD event classification on various detector technologies. In gaseous xenon TPCs, convolutional networks have been used to classify events from projected or three-dimensional track information(Renner et al., 2017; Qiao et al., 2018; Ai et al., 2018). In germanium detectors, pulse-shape discrimination is used to separate different classes of energy depositions, and machine-learning methods have been applied directly to detector waveforms (Alvis et al., 2019; Holl et al., 2019; Babicz et al., 2026). Neural-network classifiers have also been developed for EXO-200 and NEXT using detector waveforms or reconstructed event images (Anton et al., 2019b; Kekic et al., 2021). In liquid-scintillator detectors, KamNet uses the spatial and temporal patterns of PMT hits for background rejection in KamLAND-Zen (Li et al., 2023). Some of these studies already accounted for event energy when studying topology. For example, Ai et al. used samples with the same total energy, while EXO-200 used signal and background training samples with similar energy distributions. All of the studies above focus on a single experiment. To our knowledge, this work presents the first comprehensive multimodal benchmark spanning multiple NLDBD detector technologies and data modalities under a unified evaluation framework.

Deep learning classifier can classify by exploiting a predictive variable that is not the intended feature of interest. This behavior is often described as shortcut learning (Geirhos et al., 2020). In NLDBD event classification, event energy is usually the shortcut since it is much easier to extract compared to full event topology. Adjusting Receiver Operating Characteristic (ROC) analysis for a covariate is an established statistical problem. Placement-value methods and covariate-adjusted ROC curves account for changes in marker distributions with covariates (Pepe & Cai, 2004; Janes & Pepe, 2009; Janes et al., 2009). Liu and Zhou studied a standardized Area Under Curve (AUC) in which case and control marker distributions are standardized to a common reference covariate distribution before the AUC is evaluated (Liu & Zhou, 2013). Our energy-matched AUC applies the same principle to event energy by reweighting signal and background events to a common energy distribution before computing AUC. Because this metric measures classification power but does not directly quantify the degree of energy dependence in classification, we introduce a complementary energy independence score based on Jensen–Shannon divergence. This second metric measures how strongly the classifier score distribution varies with energy within each true class.

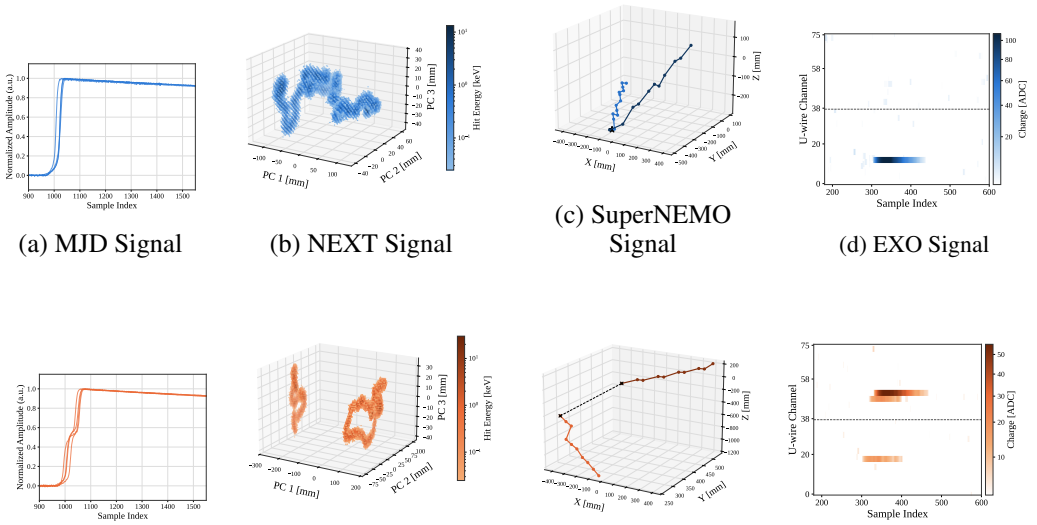
The datasets in our benchmark have different modalities/native representations, including point clouds, waveforms, and sparse detector images. There are existing work that handle each native representations. To name a few, PointNet++ and ParticleNet operate directly on unordered spatial points or particle constituents (Qi et al., 2017; Qu & Gouskos, 2020), while patch-based Transformers such as PatchTST divide continuous time series into local patches (Nie et al., 2023). Architectures such as Perceiver have also shown that inputs with different structures can be processed by a shared attention architecture (Jaegle et al., 2021). Our work adopts two modeling approaches—a native-representation approach and a unified Transformer approach—to design and benchmark 13 models across the four datasets.

3 DATASET DESCRIPTION

This work introduces the coordinated release of four public, AI-ready datasets including MAJORANA DEMONSTRATOR, NEXT, SuperNEMO, and EXO-200 experiments. The common benchmarking task across all four datasets is to distinguish signal-like from background-like events based on event topology. Signal-like NLDBD events originate from a single localized region in the de-

tector, producing a compact or localized topology, whereas background processes such as photon scattering generally deposit energy through multiple interactions at spatially separated locations, producing a more distributed or multi-part topology. These topological signatures — when represented in different detector — will manifest differently across data modalities as illustrated in Figure 1:

- **MJD Dataset** The MAJORANA DEMONSTRATOR (Arnquist et al., 2023a) is a high-purity germanium detector experiment. This dataset Arnquist et al. (2023b) consists of photon-calibration data with signal/background labels derived from physics analysis, where each event is represented as a 1D waveform with 3,800 time samples. Event topology is primarily encoded in the waveform's rising edge: signal-like events tend to exhibit a clean, single-step rise, whereas background-like events tend to exhibit multi-step structure.
- **NEXT Dataset** NEXT is a Gaseous Xenon Time Projection Chamber (TPC) experiment. This dataset (Mistry, 2026) contains simulated photon and NLDBD events, with each event represented as a sparse 3D point cloud in which each point corresponds to a localized energy deposition. Topology is encoded in the shape of point cloud: signal-like events tend to form a continuous track with more points clustered at both ends, while background-like events tend to exhibit only one cluster pre track or multiple tracks.
- **SuperNEMO Dataset** The SuperNEMO Petro et al. (2025) experiment is a tracker-calorimeter experiment. This dataset (Macko, 2026b) contains simulated photon and NLDBD events, represented as point-cloud tracks similar to NEXT but with points more strongly aligned along straight trajectories. Signal-like events tend to have two tracks starting from the same position, while background-like events tend to exhibit a gap between the two tracks.
- **EXO Dataset** EXO-200 (Anton et al., 2019a) is a liquid-xenon TPC experiment. This dataset EXO-200 Collaboration (2026) consists of photon-calibration data with signal/background labels derived from physics analysis. Each event in this dataset can be represented as a sparse 2D images, in which event topology appears as localized "bump" structure. Signal-like events tend to have only one such bump, while background-like events tend to have multiple bumps.



(a) MJD Signal (b) NEXT Signal (c) SuperNEMO Signal (d) EXO Signal

(e) MJD Bkg. (f) NEXT Bkg. (g) SuperNEMO Bkg. (h) EXO Bkg.

Figure 1: Signal (top) versus background (bottom) events from the four experiments, each in its native data modality. The top row contains the typical event topologies that represent signal-like events, while the bottom row contains the ones that represent background-like events. It is worth noting that the background events are deliberately chosen to showcase topological structure difference; the difference between signal and background could be much smaller and nuanced.

It is worth noting that the signal–background differences in Figure 1 may appear obvious, making the classification task seem trivial. However, these four events were deliberately selected to highlight clear topological differences. Within each dataset, signal-like and background-like topologies can be much more similar, and sometimes nearly indistinguishable. For example, in EXO-200, two energy bump from a background can occur close enough to partially or even fully merge into a single apparent bump. Distinguishing such ambiguous cases is the main classification challenge.

4 Benchmarking Metrics

4.1 Energy Shortcut Learning

When a deep learning classifier is trained on the datasets described in Section 3, it effectively has two paths to choose: it can choose the hard path to learn and encode the complex event topology, or it can take a shortcut to learn simpler and more obvious feature that already distinguishes signal from background. In NLDBD datasets, energy is such simple and obvious feature. As shown in Figure 2 (a), the two MJD waveforms correspond to two events with different energies: the taller waveform represents a higher-energy event, while the shorter waveform represents a lower-energy event. This energy difference is one of the most visually prominent features in the MJD dataset, and likewise among the most readily identifiable features across the other datasets.

The problem arises because signal-like and background-like events in the training set often follow different energy distributions. Figure 2 (b) shows the normalized energy distributions of the signal-like and background-like training samples. Without learning the event topology at all, a classifier can exploit this difference with a simple "energy cut": classifying events above a certain energy as signal-like and events below it as background-like. As shown in Figure 2 (c), this simple "energy cut" approach can already achieve an ROC AUC of 0.9706. In this case, strong classification performance does not necessarily indicate that the model has learned the underlying physics encoded in the event topology; instead, it may simply rely on energy—a simple feature that is often readily apparent by eye—and make its predictions primarily from this shortcut. We refer to this failure mode as **energy shortcut learning**.

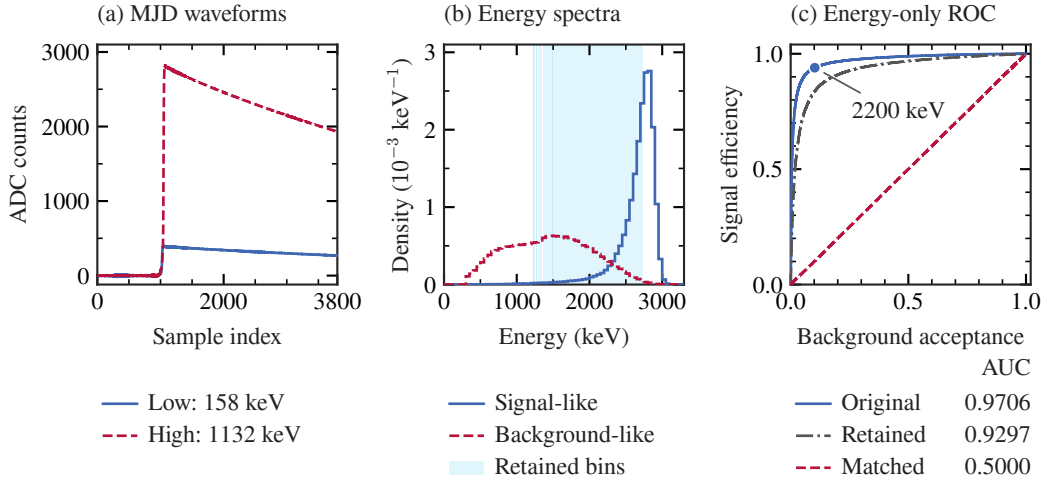


Figure 2: Figure (a) shows two representative MJD waveforms: the taller waveform corresponds to a higher-energy event, while the shorter waveform corresponds to a lower-energy event. Figure (b) shows the energy distributions of the signal-like and background-like SuperNEMO training samples; retained energy bins are highlighted by the blue shaded region. Figure (c) compares three ROC curves: the original ROC without energy-shortcut mitigation, the ROC evaluated only on events within the retained energy bins but without reweighting, and the energy-weighted ROC $\mathrm{AUC}_{\mathrm{w}}$ (Retained + Reweight).

4.2 ENERGY-AWARE METRIC DESIGN

The primary goal of this work is not to eliminate energy shortcut learning, but to provide a systematic benchmark to understand its impact across different NLDBD datasets and modalities. To characterize this effect, we introduce two complementary metrics: an *Energy-weighted AUC* that measures pooled classification power after balancing the signal and background energy spectra, and an *Energy Independence Score* that measures stability of the model score distribution with respect to energy within each true class.

We partition events into fixed 5 keV energy bins over 0–3000 keV, with dataset-specific treatment of energies outside this range detailed in Appendix A. An example histogram from the SuperNEMO dataset is shown in Figure 2(b). For energy-weighted AUC, we first restrict events to the *common energy support*: the overlap of the two classes' central 99% empirical energy ranges (0.5th–99.5th percentiles). This avoids comparing classes in energy regions without adequate representation from both. We then retain bins containing at least 20 events from each class within this overlap. For the independence score, the same count requirement applies separately within each class, without common-support restriction. The threshold of 20 is a pragmatic minimum-count rule intended to limit small-sample fluctuations in bin weights and score histograms, not a physical event-selection criterion or a guarantee of statistical precision. These restrictions limit the population described by each metric; excluded events remain in the inclusive AUC, and coverage diagnostics are given in Appendix A.

Energy-weighted AUC ($\mathrm{AUC_w}$) For event x_i, let $y_i \in \{0, 1\}$, E_i, and s_i denote its evaluation label, conditioning energy, and classifier score, respectively, with larger scores favoring $y_i = 1$. Within the common energy support of the two classes, we reweight events so that, in each retained energy bin, the signal and background classes have the same total weight. The target weight assigned to each bin is proportional to the smaller of the two class-normalized bin weights. Letting w_i denote the resulting event weights, normalized to sum to one within each class, the energy-weighted AUC is defined as

$$\mathrm{AUC_w} = \sum_{i:y_i=1} \sum_{j:y_j=0} w_i w_j \left[\mathbf{1}(s_i > s_j) + \frac{1}{2}\mathbf{1}(s_i = s_j) \right]. \tag{1}$$

This defines a single pooled weighted AUC, including cross-bin comparisons and half credit for ties (Hanley & McNeil, 1982); events are not paired at the same energy, nor are within-bin AUCs averaged. Reweighting changes only the evaluation weights, without retraining or random resampling. The metric measures signal–background discrimination across energies after balancing the two classes' binned energy distributions. Energy-dependent score shifts can still affect cross-bin rankings, so this is not a pure measure of topology discrimination at fixed energy. Residual within-bin energy differences and the retained-support restriction further limit this interpretation.

Figure 2(c) compares the ROC curves before and after energy reweighting for the SuperNEMO signal-like and background-like samples, using only event energy as the classification score (i.e., an "energy cut"). Using all events under their original energy distributions gives an inclusive AUC of 0.9706. Restricting the evaluation to the retained events without reweighting gives an AUC of 0.9297. Finally, reweighting these same events reduces the AUC to 0.5000 (rounded), showing that energy alone provides approximately chance-level discrimination in this illustration.

Energy Independence Score (I) The second metric asks whether the score distribution changes with energy within each true class. For class g, let P_{gk} denote its normalized score histogram in energy bin k, and let P_g denote the corresponding histogram pooled across all retained energy bins. Let π_{gk} denote the fraction of the retained event weight for class g that falls within energy bin k. Using Jensen-Shannon divergence in natural-log units (Lin, 1991),

$$I_g = 1 - \sqrt{\frac{\sum_k \pi_{gk}\, \mathrm{JS}(P_{gk}, P_g)}{\ln 2}}, \qquad I = \frac{\sum_g A_g I_g}{\sum_g A_g}, \tag{2}$$

where A_g is the total original weight, before sparse-bin exclusion, of events in class g with finite scores and energies within the specified energy range. Only classes for which I_g can be estimated are included in the average. This calculation uses the original event weights. Higher $I \in [0, 1]$ indicates greater stability of the classifier score distribution across energy, while lower I indicates stronger energy dependence. Importantly, I measures score stability rather than classification power:

a constant-score classifier attains $I = 1$ but an AUC of 0.5000. Energy-weighted AUC and I therefore provide complementary diagnostics, with the former measuring pooled discrimination after balancing the class energy distributions and the latter directly measuring how strongly the classifier score varies with energy within each class.

5 Model Design

We design two complementary modeling approaches for the NLDBD datasets: a *native-representation approach* and a *unified Transformer approach*. The first preserves each dataset as closely as possible to its native representation, applying only minimal preprocessing before benchmarking different model group. Each group provides different inductive biases—for example, RNNs emphasize sequential structure, while GNNs emphasize graph structure. Each dataset is then benchmarked across all architecture families to evaluate how different inductive biases affect different modalities. The second approach instead converts each event into a sequence of tokens through tokenization and positional encoding, after which all tokenized events are processed by a common Transformer backbone. This separates the representation of heterogeneous detector data from the architecture used to model it, allowing us to test whether a unified Transformer can operate effectively across substantially different detector modalities. We benchmark both approaches using $\mathrm{AUC_w}$ and I introduced in the previous section.

5.1 Native Representation Approach

We defined four model groups relevant to the NLDBD datasets at the beginning of model building stage: Projection and voxel model group operate on regular spatial grids using 2D or 3D convolutions, sparse convolutions, or projection-patch mixing; Point-cloud model group operate directly on unordered hits using shared point features and local geometric neighborhoods; Graph model group represent events through explicit neighborhood connectivity and message passing over fixed or learned graphs; sequence, state-space, topology, and hybrid model group capture ordered context, topological summaries, or combinations of multiple representations. We developed 23 classification models across these four groups and benchmarked them on the NEXT dataset, selecting the best-performing model from each group for the subsequent cross-dataset study. Details of this model-selection study are provided in Appendix E.1. The four selected models are:

- **Multi-view CNN.** A shared convolutional encoder processes three complementary views, whose embeddings are combined through late fusion. This architecture captures local patterns while preserving view-specific information, and the same backbone can operate on either spatial projections or waveform spectrograms.
- **Static GINE.** Static GINE performs edge-aware message passing over a fixed nearest-neighbor graph (Xu et al., 2019). For spatial data, the graph preserves local neighborhood relationships that may be obscured by projection; for time series data, it is constructed in pulse-shape coordinates.
- **BiGRU.** BiGRU processes two complementary Hilbert orderings using bidirectional recurrence (Cho et al., 2014), providing a sequential baseline for aggregating information across ordered event points.
- **PointMamba-lite.** PointMamba-lite applies input-dependent state-space updates along ordered event points, providing an alternative mechanism for long-range context aggregation (Gu & Dao, 2023; Liang et al., 2024). It is our reduced local implementation and is not intended as a reproduction of the official PointMamba architecture.

5.2 Unified Transformer Approach

The central idea of this approach is to use a common Transformer backbone (Vaswani et al., 2017) across all detector modalities. To enable this, each event must first be converted into a suitable sequence of tokens, together with positional or structural information added through positional encoding. The benchmarking goal is to understand how different combinations of tokenization strategy and positional encoding affect different detector modalities. For detector d and tokenization strategy

r, the tokenizer maps an event x to

$$\mathcal{T}_{d,r}(x) = (\boldsymbol{F}, \boldsymbol{P}, \boldsymbol{m}),$$

where $\boldsymbol{F}$ contains the token features, $\boldsymbol{P}$ contains the positional or structural descriptors supplied to the positional encoder, and $\boldsymbol{m}$ indicates valid tokens. A learned projection maps the token features into the Transformer embedding space. Separate $\boldsymbol{F}$ and $\boldsymbol{P}$ are trained for each dataset while the core Transformer architecture is held fixed across all experiments.

We compare three tokenization strategies that represent detector data at different levels of granularity. *Entity tokens* preserve individual detector measurements: NEXT and SuperNEMO use detector or tracker hits, MJD uses selected waveform samples, and EXO-200 uses sensor–time points balanced across its six sensor blocks. *Region tokens* aggregate nearby measurements while retaining local structure: NEXT and SuperNEMO use occupied three-dimensional regions, MJD uses non-overlapping waveform patches (Dosovitskiy et al., 2021), and EXO-200 uses regions defined jointly over sensor blocks, channel groups, and time intervals. *Summary tokens* apply stronger compression by replacing groups of measurements with descriptive statistics: NEXT and SuperNEMO use balanced Morton-ordered spatial groups, MJD summarizes temporal waveform segments, and EXO-200 summarizes sensor–time regions. Together, these tokenization tests the trade-off between preserving individual measurements, retaining local structure, and compressing detector information into higher-level summaries. Details are provided in Appendix G.

We additionally compare three positional encoding (PE) mechanisms: learned coordinate-MLP encoding, multiscale Fourier features (Tancik et al., 2020), and rotary positional encoding (RoPE) (Su et al., 2024). Learned coordinate-MLP encoding maps the positional or structural descriptors $\boldsymbol{P}$ through a small MLP into the embedding space, while multiscale Fourier features encode $\boldsymbol{P}$ using sinusoidal functions at multiple frequencies to capture structure across different spatial or temporal scales. In both cases, the resulting positional representation is added directly to the projected token features, $\mathrm{Proj}(\boldsymbol{F})$. RoPE instead uses $\boldsymbol{P}$ to apply descriptor-dependent rotations to the query and key vectors within self-attention.

The transformer backbone contains 64-dimensional embeddings, two Transformer encoder layers, four attention heads, a feed-forward dimension of 256, and masked mean pooling followed by a binary classification head. Although input projections and positional encodings can change the total parameter count, and representation-dependent sequence lengths affect computational cost, the core Transformer architecture remains fixed. This design isolates the effects of tokenization and positional encoding while maintaining a consistent backbone across detector modalities.

6 Experiment Results

Within each dataset, all benchmarked models use the same classification target, train/validation/test partition, and held-out test events. The NEXT and SuperNEMO datasets do not provide predefined train/validation/test splits, so we partition them approximately $80/10/10$. EXO-200 and MJD provide separate test sets; we therefore split their provided training data into training and validation subsets using a $90/10$ ratio.

The definitions of signal and background events are dataset-specific. For MJD, the provided binary labels directly identify signal-like and background-like waveforms. For EXO-200, the `n_charge_cluster` variable records the number of reconstructed charge clusters (i.e., "bumps"); we define events with `n_charge_cluster == 1` as signal-like and those with `n_charge_cluster > 1` as background-like. For NEXT, simulated NLDBD events are treated as signal-like and simulated photon events as background-like. For SuperNEMO, simulated two-neutrino double-beta decay (2NDBD) events are treated as signal-like and simulated photon events are treated as background.

All models are trained with the same random seed initialization. Native-representation models retain their architecture-specific training configurations, but are evaluated on the same test dataset and with the same metrics. Unified Transformer models use the same optimization procedure and backbone configuration across tokenization and positional-encoding variants. Models are trained with binary cross-entropy using AdamW, a cosine learning-rate schedule, batch size 64, an initial learning rate of 5×10^{-4}, weight decay 10^{-4}, gradient-norm clipping at 1, and early stopping with patience 5, for at most 50 epochs. Additional implementation details can be found in Appendix A.3.

Table 1 summarizes weighted AUC, energy independence I, and trainable parameter counts for both modeling approaches. Overall, the unified Transformer is competitive with or stronger than the native-representation models across detector modalities. The unified Transformer also attains the highest I on all four datasets. Notably, these Transformer variants use approximately 106–125K trainable parameters, compared with 317–808K for the native-representation models.

Table 1: Cross-detector classification results. Higher values are better for energy-weighted AUC ($\mathrm{AUC_w}$) and energy independence (I). Blue marks the highest AUC and I across all models and **bold** marks represent the highest within its approach. ‡MJD GINE uses additional PSD-label supervision and is excluded from best-result highlighting. Evaluation details are provided in Appendix A.

		NEXT		MJD		EXO-200		SuperNEMO	
Model	#Params (K)	$\mathrm{AUC_w}$	I	$\mathrm{AUC_w}$	I	$\mathrm{AUC_w}$	I	$\mathrm{AUC_w}$	I
Native Representation Approach									
Multi-view CNN	808	0.977	0.971	0.970	0.760	0.863	0.752	0.586	0.910
Static GINE	480	**0.995**	0.970	0.969‡	0.797‡	**0.949**	**0.801**	**0.590**	0.913
BiGRU	734	0.979	**0.971**	**0.975**	**0.775**	0.929	0.789	0.579	0.921
PointMamba-lite	317	0.925	0.971	0.960	0.773	0.934	0.786	0.579	**0.922**
Unified Transformer Approach									
Entity + MLP	111	0.975	**0.975**	0.929	**0.812**	0.941	0.737	0.591	0.975
Entity + Fourier	113–116	0.963	0.974	0.941	0.802	0.947	0.729	0.589	0.974
Entity + RoPE	106–113	0.497	0.512	0.793	0.749	0.946	0.715	0.551	0.865
Region + MLP	111–121	**0.988**	0.973	0.973	0.798	0.970	0.801	0.591	0.976
Region + Fourier	113–125	0.979	0.973	**0.981**	0.788	**0.972**	0.790	0.590	0.977
Region + RoPE	107–116	0.512	0.483	0.916	0.768	0.971	0.789	**0.602**	0.953
Summary + MLP	111	0.987	0.971	0.965	0.799	0.947	0.776	0.590	0.975
Summary + Fourier	113–116	0.983	0.973	0.969	0.782	0.960	**0.802**	0.589	**0.977**
Summary + RoPE	106–114	0.602	0.532	0.878	0.777	0.955	0.773	0.582	0.956

Our goal is not only to identify a best-performing model on each dataset, but to identify a model that consistently performs well across different datasets and modalities. We thereby define *Top-3 count*: for each dataset, a model receives one count if its $\mathrm{AUC_w}$ ranks among the top three of all 13 benchmarked models. Since there are four datasets, the maximum possible Top-3 count is four.

As shown in Figure 3 (left), among all native-representation models, only Static GINE and BiGRU has one Top-3 count when architectural inductive bias matches data modality: Static GINE for the NEXT point-cloud data and BiGRU for MJD time series data. In contrast, the unified Transformer models exhibit a much stronger cross-dataset pattern. Figure 3 (middle) shows that region tokenization produces more Top-3 counts than entity or summary tokenization, while MLP positional encoding outperforms other two PE strategy. Their combination, Region+MLP, reached Top-3 performance on every benchmarked dataset. This study suggests that, while native-representation models can achieve top-tier performance when their inductive biases align well with the data modality, the unified Transformer with Region+MLP performs consistently well across all datasets.

Figure 3 (right) shows the distribution of the energy independence score I across the four datasets. The distributions suggest that variation in I across datasets is substantially larger than the variation across models within a given dataset, indicating that dataset variation play a more important role than model variation. In particular, EXO-200 and MJD exhibit similar I distributions centered around approximately 0.8, whereas NEXT and SuperNEMO occupy a higher-I regime around 0.9–1.0. The two low-I outliers in NEXT correspond to failed RoPE configurations and are not representative of the remaining NEXT models.

One possible explanation is the difference in energy distributions across the datasets. In EXO-200 and MJD, photon calibration events labeled as signal and background have similar energy distributions. In NEXT and SuperNEMO, the simulated NLDBD events labeled as signal are more concentrated at high energies than the photon background events. Label construction may also play

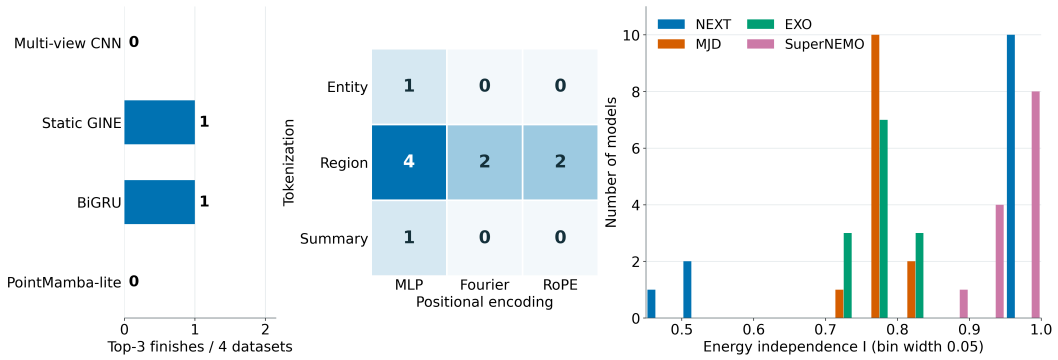


Figure 3: Benchmarking result analysis according to Table 1. The left and middle figure shows the top-3 count of native representation approach and unified transformer approach respectively; the right figure shows the distribution of energy independence score over the 4 datasets respectively.

an important role. EXO-200 and MJD use real detector data with labels assigned through physics-analysis procedures, while NEXT and SuperNEMO use simulated events labeled from simulation ground truth. The real-data labels may contain additional noise or imperfect correspondence to the underlying event topology, which could contribute to stronger apparent energy dependence. Further study is needed to understand the underlying cause for the difference in independence score I across the four datasets.

7 Limitations

This benchmark is based on datasets curated specifically for AI/ML research, with two of the four datasets generated using simulation tools. The reported results therefore should not be interpreted as representative of full real-detector performance, nor used to assess the validity or scientific potential of the underlying experiments. In addition, because the classification targets, available detector information, and evaluated representations differ across datasets, cross-detector comparisons should not be used to rank detector technologies or task difficulty, or interpreted as the best achievable performance for any experiment.

Our energy-aware metrics focus specifically on energy shortcut learning. Energy matching balances the signal and background energy distributions across the retained bins, but finite bin widths can leave residual energy variation within each bin. Moreover, other detector- or event-level variables may remain correlated with the classification target. Energy-matched AUC and the Energy Independence Score should therefore be interpreted as complementary diagnostics of energy dependence, rather than as proof that a model relies exclusively on event topology. Finally, the reported model results are based on a single training seed, and the native-representation models retain architecture-specific training configurations. We therefore emphasize broad cross-dataset trends and consistent performance patterns rather than small numerical differences between individual models.

8 Conclusion

This work introduced the first multimodal benchmark for NLDBD event classification based on four AI-Ready datasets from the Majorana Demonstrator, NEXT, SuperNEMO, and EXO-200 experiments. Beyond conventional classification metrics, we used two complementary metrics—Energy-matched AUC and Energy Independence Score—to characterize energy shortcut learning during classifier training. We benchmarked 13 models spanning native-representation architectures and tokenized Transformer-based models, and identified the Region + MLP Transformer as the strongest overall model across all datasets. Future work will build toward a multimodal NLDBD foundation model based on this architecture, while expanding the benchmarking pipeline to additional AI-ready datasets and tasks to better assess the foundation model's generalization ability.

AI USE STATEMENT

Generative AI use differed between the two modeling approaches, so we describe each separately.

Unified Transformer approach (Section 5.2, Appendix G). We used generative AI tools for two tasks that require disclosure: implementing methods and providing feedback on research methodology and experiments. AI tools assisted with the design, implementation, debugging, and testing of the entity, region, and summary tokenizers, the MLP, Fourier, and RoPE positional encodings, the dataset-specific input adapters, and the training workflow. They were also used to give feedback on the experimental methodology and to investigate inconsistencies between the evaluation protocols applied to different datasets (Appendix A).

Native-representation approach (Section 5.1, Appendices E–F). The model architectures and training code were written by the authors without generative AI. AI tools were used only to assist with debugging and to launch and manage runs of the existing training programs across the four datasets. They did not modify model code, input adapters, or training configurations.

Shared benchmark and analysis. The shared benchmark evaluation code that computes the energy-weighted AUC and the Energy Independence Score for all models (Section 4, Appendix A) was developed with AI assistance. AI tools were also used to write analysis scripts that aggregate results, such as the tables and Top-3 counts in Table 1 and Figure 3. The interpretation of results and the conclusions drawn from them were made by the authors.

Tasks not using AI. We did not use generative AI tools to generate synthetic data, formulate hypotheses, develop the theoretical or conceptual framework, formulate mathematical claims, or interpret results. Translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: creating and editing software code, formatting tables, improving the visual presentation of some figures without modifying the underlying data, suggesting the structure of the manuscript, and drafting and editing portions of the text for clarity.

We have reviewed all AI-assisted work. All AI-assisted code, including the shared evaluators, was reviewed and tested by the authors. The authors inspected the resulting prediction files and evaluation artifacts, checked the evaluation procedures against the recorded datasets, test partitions, and benchmark protocols, and verified that AI-assisted figure edits did not alter the plotted data. All methodological decisions, analyses, and conclusions were made by the authors. Generative AI was not used to produce any reported numerical result, and it was not used to generate, select, or verify citations. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses four datasets released by the MAJORANA DEMONSTRATOR, NEXT, SuperNEMO, and EXO-200 collaborations through public archival repositories. The MJD and EXO-200 datasets contain detector data (for EXO-200, a subset of calibration data), and the NEXT and SuperNEMO datasets contain simulated events. None of the data involve human subjects, personal information, or other sensitive attributes. All datasets were released with the approval of their respective collaborations and are used in accordance with their stated terms. The work concerns fundamental physics, and we do not foresee direct harmful applications or dual-use concerns. All experiments were run on single GPUs, so the computational and environmental footprint is modest. To support research integrity, Appendix A documents which reported results were re-evaluated under the current evaluation protocol and which retain earlier protocol versions, together with all changes from historical implementations. As discussed in Section 7, the cross-detector comparisons should not be read as a ranking of detector technologies or experiments.

REPRODUCIBILITY STATEMENT

All four datasets are publicly available, and their sources and structure are described in Section 3. The classification targets and label definitions are given in Section 6 and Appendix A.3, and the train/validation/test partitions for each dataset are described in Section 6 and Appendix A.3. The two benchmark metrics are defined in Section 4. Full definitions, including binning, common-

support estimation, reporting gates, degenerate cases, and exact weighted quantiles, are given in Appendix A. Diagnostics that allow the reported values to be checked are in Tables 3–7. The unified Transformer backbone, optimizer, and training hyperparameters are listed in Sections 5.2 and 6, and the dataset-specific tokenizers are specified in Appendix G (Table 11). The native-representation architectures, input adapters, recorded training settings, and software and hardware environment are documented in Appendix E.1 (Tables 8-9), and the complete NEXT model-selection inventory is in Appendix F (Table 10). The supplementary material contains [*an anonymized code repository with*] the frozen evaluators for each protocol version. All results use a single training seed (42), and several configurations use mixed precision without enforced deterministic algorithms. Retraining is therefore not expected to be bitwise identical, but re-running the evaluators on the released predictions reproduces the reported metrics exactly.

REFERENCES

P. Ai, D. Wang, G. Huang, and X. Sun. Three-dimensional convolutional neural networks for neutrinoless double-beta decay signal/background discrimination in high-pressure gaseous time projection chamber. *Journal of Instrumentation*, 13(08):P08015, 2018. doi: 10.1088/1748-0221/13/08/P08015. URL https://doi.org/10.1088/1748-0221/13/08/P08015.

S. I. Alvis et al. Multisite event discrimination for the Majorana Demonstrator. *Physical Review C*, 99:065501, 2019. doi: 10.1103/PhysRevC.99.065501. URL https://doi.org/10.1103/PhysRevC.99.065501.

G. Anton et al. Search for neutrinoless double-β decay with the complete EXO-200 dataset. *Physical Review Letters*, 123:161802, 2019a.

G. Anton et al. Search for neutrinoless double-beta decay with the complete EXO-200 dataset. *Physical Review Letters*, 123:161802, 2019b. doi: 10.1103/PhysRevLett.123.161802. URL https://doi.org/10.1103/PhysRevLett.123.161802.

I. J. Arnquist et al. Final result of the MAJORANA DEMONSTRATOR's search for neutrinoless double-β decay in ^{76}ge. *Physical Review Letters*, 130:062501, 2023a.

I. J. Arnquist et al. Majorana Demonstrator data release for AI/ML applications, 2023b. URL https://arxiv.org/abs/2308.10856.

Marta Babicz, Saúl Alonso-Monsalve, Alain Fauquex, and Laura Baudis. Transformer-based pulse shape discrimination in HPGe detectors with masked autoencoder pre-training. *arXiv preprint arXiv:2603.06192*, 2026. doi: 10.48550/arXiv.2603.06192. URL https://arxiv.org/abs/2603.06192.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

EXO-200 Collaboration. EXO-200 Public Data Release for AI/ML Applications, 2026. [Dataset]. https://doi.org/10.5281/zenodo.22155860.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL https://arxiv.org/abs/2312.00752.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747. URL https://www.med.mcgill.ca/epidemiology/hanley/software/hanley_mcneil_radiology_82.pdf.

P. Holl, L. Hauertmann, B. Majorovits, O. Schulz, M. Schuster, and A. J. Zsigmond. Deep learning based pulse shape discrimination for germanium detectors. *The European Physical Journal C*, 79:450, 2019. doi: 10.1140/epjc/s10052-019-6869-2. URL https://doi.org/10.1140/epjc/s10052-019-6869-2.

Andrew Jaegle et al. Perceiver: General perception with iterative attention. *Proceedings of Machine Learning Research*, 139:4651–4664, 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Holly Janes and Margaret S. Pepe. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*, 96(2):371–382, 2009. doi: 10.1093/biomet/asp002. URL https://doi.org/10.1093/biomet/asp002.

Holly Janes, Gary Longton, and Margaret S. Pepe. Accommodating covariates in ROC analysis. *The Stata Journal*, 9(1):17–39, 2009. doi: 10.1177/1536867X0900900102. URL https://doi.org/10.1177/1536867X0900900102.

M. Kekic et al. Demonstration of background rejection using deep convolutional neural networks in the NEXT experiment. *Journal of High Energy Physics*, 2021(01):189, 2021. doi: 10.1007/JHEP01(2021)189. URL https://doi.org/10.1007/JHEP01(2021)189.

Aobo Li, Zepeng Fu, Christopher Grant, Hiroshi Ozaki, Itaru Shimizu, Hasung Song, Atsuto Takeuchi, and Lindley A. Winslow. Kamnet: An integrated spatiotemporal deep neural network for rare event searches in kamland-zen. *Physical Review C*, 107(1):014323, 2023. doi: 10.1103/PhysRevC.107.014323. URL https://doi.org/10.1103/PhysRevC.107.014323.

Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/395371f778ebd4854b88521100af30ad-Abstract-Conference.html.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115. URL https://ieeexplore.ieee.org/document/61115/.

Danping Liu and Xiao-Hua Zhou. ROC analysis in biomarker combination with covariate adjustment. *Academic Radiology*, 20(7):874–882, 2013. doi: 10.1016/j.acra.2013.03.009. URL https://doi.org/10.1016/j.acra.2013.03.009.

Miroslav Macko. SuperNEMO simulation dataset for $0\nu\beta\beta$ AI summer school at University of California, Irvine 2026, June 2026a. Dataset release documentation, 18 June 2026, on behalf of the SuperNEMO Collaboration.

Miroslav Macko. SuperNEMO Dataset for 0vbb AI Summer School - UC Irvine, USA (20-21 June 2026), 2026b. [Dataset]. https://doi.org/10.5281/zenodo.20698789.

Krishan Mistry. NEXT Simulation Dataset for AI Summer School UC Irvine 2026. *arXiv preprint arXiv:2605.07566*, 2026.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

NSAC. A new era of discovery: The 2023 long range plan for nuclear science. October 2023. https://nuclearsciencefuture.org/wp-content/uploads/2024/02/23-G06476-2024-LRP-17x11-pcg-1.24.24.pdf.

Margaret S. Pepe and Tianxi Cai. The analysis of placement values for evaluating discriminatory measures. *Biometrics*, 60(2):528–535, 2004. doi: 10.1111/j.0006-341X.2004.00200.x. URL https://doi.org/10.1111/j.0006-341X.2004.00200.x.

M. Petro et al. Status of the SuperNEMO experiment and sensitivity estimates to $0\nu\beta\beta$ and beyond. *PoS*, MEDEX2025:034, 2025.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html.

H. Qiao, C. Lu, X. Chen, K. Han, X. Ji, and S. Wang. Signal-background discrimination with convolutional neural networks in the PandaX-III experiment using MC simulation. *Science China Physics, Mechanics & Astronomy*, 61:101007, 2018. doi: 10.1007/s11433-018-9233-5. URL https://doi.org/10.1007/s11433-018-9233-5.

Huilin Qu and Loukas Gouskos. Jet tagging via particle clouds. *Physical Review D*, 101: 056019, 2020. doi: 10.1103/PhysRevD.101.056019. URL https://doi.org/10.1103/PhysRevD.101.056019.

J. Renner et al. Background rejection in NEXT using deep neural networks. *Journal of Instrumentation*, 12(01):T01004, 2017. doi: 10.1088/1748-0221/12/01/T01004. URL https://doi.org/10.1088/1748-0221/12/01/T01004.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2019. URL https://arxiv.org/abs/1810.00826.

A Full Metric Definitions and Evaluation Protocols

The 0–3000 keV protocol below applies to the reevaluated MJD and EXO-200 classic models, nine EXO-200 Transformer variants, and six MJD Transformer variants with MLP or Fourier position encoding. The NEXT and SuperNEMO classic results and the separate SuperNEMO energy-only illustration retain their preceding evaluation profile, specified below. The three MJD RoPE variants and the NEXT and SuperNEMO Transformer entries retain their historical reported values outside this reevaluation.

A.1 ENERGY-WEIGHTED AUC

For event x_i, let $y_i \in \{0, 1\}$ denote its evaluation label, E_i its conditioning energy, s_i its scalar classifier score, and $a_i \geq 0$ its base event weight. Scores are oriented so that larger values favor $y_i = 1$; this convention need not coincide with the stored training label (Table 2). The selected evaluations use unit base weights. We first describe the weighted form implemented by the evaluators. Only finite scores with positive base weight contribute to inclusive AUC; energy-dependent metrics additionally require finite physical energy satisfying $0 \leq E_i \leq 3000$ keV. This range restriction precedes support estimation.

The common support $C = [L, U]$ is the intersection of the two classes' central energy ranges:

$$L = \max_{c\in\{0,1\}} Q_c(\alpha), \qquad U = \min_{c\in\{0,1\}} Q_c(1-\alpha), \qquad \alpha = 0.005, \tag{3}$$

where Q_c is the base-weighted empirical energy quantile of finite-score, finite-energy, positive-weight events in class c within 0–3000 keV. No additional matching region of interest is specified in these runs. For energy bins B_k, define the count n_{ck} and base mass m_{ck} of class-c events in $C \cap B_k$, and let $p_{ck} = m_{ck}/\sum_\ell m_{c\ell}$. A bin is retained only if it contains at least 20 positive-weight events from each class. Writing $\mathcal{V}$ for these bins, the implemented overlap target and event weights are

$$q_k = \frac{1\{k \in \mathcal{V}\}\min(p_{0k}, p_{1k})}{\sum_{\ell\in\mathcal{V}} \min(p_{0\ell}, p_{1\ell})}, \tag{4}$$

$$w_i = a_i \frac{q_k}{m_{y_i k}} \quad \text{if } E_i \in C \cap B_k,\ k \in \mathcal{V}; \qquad w_i = 0 \text{ otherwise.} \tag{5}$$

Thus each class has unit total matched weight and assigns mass q_k to bin k. The target favors energy regions represented in both classes and is defined by the minimum of their class-normalized bin masses.

All retained events are pooled into one weighted ROC curve. Equivalently, with $H(u) = 1\{u > 0\} + \frac{1}{2}1\{u = 0\}$,

$$\mathrm{AUC_w} = \frac{\displaystyle\sum_{i:y_i=1}\sum_{j:y_j=0} w_i w_j H(s_i - s_j)}{\left(\displaystyle\sum_{i:y_i=1} w_i\right)\left(\displaystyle\sum_{j:y_j=0} w_j\right)}. \tag{6}$$

This is the weighted pairwise ranking interpretation of AUC (Hanley & McNeil, 1982); equal scores receive half credit. The code implements the equivalent weighted pairwise calculation by sorting scores and processing ties together. It does not average within-bin AUCs: pairs from different energy bins also contribute to Equation 6. Inclusive AUC uses the same ranking calculation with base weights and without energy restriction.

Matching is a deterministic evaluation operation on the supplied split; it does not resample events or alter the fitted classifier. For the reported test results, support and weights are estimated on the test events themselves. The formal weighted AUC is withheld if either class retains less than 50% of its original finite-score/finite-energy class base mass, or fewer than two bins survive. Exactly constant common energy exempts only the two-bin requirement; the per-class event-count and coverage gates still apply. Empty, single-class, and insufficient-count bins receive zero matched weight.

Both energy metrics in the MJD and EXO-200 reevaluation use the same global grid after conversion to keV:

$$B_k = [5k, 5(k+1))\ \mathrm{keV}, \quad k = 0, \ldots, 598, \tag{7}$$

$$B_{599} = [2995, 3000]\ \mathrm{keV}. \tag{8}$$

There are exactly 600 energy bins. Exactly 3000 keV belongs to the last bin; energies below zero or above 3000 keV are excluded and reported separately. Energies are never clipped. Matching additionally requires $L \leq E_i \leq U$, so support boundaries may cut through an energy bin.

Matching equalizes bin masses, not continuous energy distributions within bins. Finite width, finite statistics, trimming, and sparse-bin exclusion limit the interpretation to the retained support and resolution. A score that is solely a function of the matched bin index has AUC $1/2$, because its two weighted class distributions are identical; a score depending on within-bin energy need not. This is a property of the weighting construction, not an empirical claim about the trained models.

A.2 ENERGY INDEPENDENCE SCORE

Energy matching measures discrimination after balancing class spectra. The complementary independence score asks whether the *distribution of classifier scores* changes with energy within a true class. Conditioning on class avoids interpreting class separation itself as unwanted dependence. The evaluator groups by supplied physical category, falling back to the binary label when category metadata are absent. Here each reported task has two such groups. This calculation uses the original base weights, without matching or common-support trimming.

For group g, let P_{gk} be the normalized score histogram in an eligible energy bin, P_g the group's pooled score histogram, and π_{gk} the fraction of retained base weight in that energy bin. Score bins are shared within a group and obtained from up to 20 weighted score-quantile bins. Using natural logarithms, define

$$\mathrm{JS}(P,Q) = \tfrac{1}{2}\sum_r P_r \log\frac{P_r}{M_r} + \tfrac{1}{2}\sum_r Q_r \log\frac{Q_r}{M_r}, \qquad M_r = \tfrac{1}{2}(P_r + Q_r), \tag{9}$$

$$D_g = \sum_k \pi_{gk}\,\mathrm{JS}(P_{gk}, P_g), \qquad I_g = \mathrm{clip}_{[0,1]}\left(1 - \sqrt{D_g/\log 2}\right), \tag{10}$$

$$I = \frac{\sum_{g\in\mathcal{G}_*} A_g I_g}{\sum_{g\in\mathcal{G}_*} A_g}. \tag{11}$$

Here r indexes score-histogram bins, zero-probability terms contribute zero, $\mathcal{G}_*$ contains estimable groups, and A_g is their base weight after finite-pair and 0–3000 keV range filtering, but before sparse-bin exclusion. These MJD and EXO-200 evaluations use the retained-event pooling and eligibility rules specified below. The Jensen–Shannon divergence is bounded by $\log 2$ (Lin, 1991), giving $I_g, I \in [0,1]$ whenever estimable. Larger values indicate less detected dependence at the chosen binning resolution. The report also retains $\min_g I_g$ to expose a poorly behaving group that the weighted mean could obscure.

This is a histogram-based dependence proxy, not Pearson correlation, distance correlation, mutual information, or a test certifying statistical independence. Its ideal value is one when all compared histograms coincide. Constant predictions also attain one while providing AUC $1/2$; a single populated energy bin makes the score uninformative about energy dependence. Finite samples can produce nonzero histogram divergences even under population independence, and coarse bins can hide dependence. The score depends on energy coverage, score discretization, and group composition, so it should be compared under a common protocol. No threshold or operating-point diagnostic enters Equations 9–11.

Joint interpretation. High weighted AUC indicates ranking ability after the specified spectral balancing; high I indicates stable within-class score distributions across the inspected energy bins. Neither implies the other. Inclusive AUC supplies the complementary performance under the original sampled spectra. Its change after matching is descriptive, rather than a causal decomposition of energy use: both the support and the event weights can change. We report the three quantities together.

A.3 IMPLEMENTATION DETAILS

label selection The stored NEXT labels assign 1 to $0\nu\beta\beta$ and 0 to ^{214}Bi, while SuperNEMO assigns 1 to $2\nu\beta\beta$ and 0 to ^{214}Bi. EXO-200 stores label 1 for events with more than one charge cluster and label 0 for single-cluster events; during evaluation, we reverse the label and score orientation so that single-cluster events form the positive class, following Table 2. MJD assigns label 1 only when all four reference pulse-shape-discrimination criteria pass. The evaluation-positive class therefore has a dataset-specific physical interpretation. NEXT uses an approximately $80/10/10$ training/validation/test partition based on class-wise event counts, and different portions of a source file may enter different splits. SuperNEMO uses the same approximate proportions while assigning blocks of complete events within each source category. EXO-200 keeps acquisition runs disjoint across splits. MJD preserves its official test files and divides the provided training data $90/10$ into training and validation subsets. Each model for a given dataset reuses the same recorded partition.

Table 2: Conditioning variables for classic models and the reevaluated EXO-200 and MJD Transformers. The MJD and EXO-200 reevaluation uses 600 fixed 5 keV bins on 0–3000 keV and excludes energies outside this range (v3). NEXT and SuperNEMO classic results retain 600 regular bins plus one overflow bin (v2). The two metrics retain distinct statistical definitions.

Dataset	Evaluation-positive class	Physical conditioning energy
NEXT	$0\nu\beta\beta$	Summed deposited-hit energy; MeV converted to keV
MJD	Clean pulse	Original calibrated float64 energy (keV), recovered from test-event metadata
EXO-200	Stored label 0 ($n_{\mathrm{CCL}} = 1$)	Reconstructed rotated energy (keV); negate background-oriented logit
SuperNEMO	$2\nu\beta\beta$	Reconstructed calorimeter $E_1 + E_2$ (keV)

Evaluation inputs and orientation. The MJD and EXO-200 reevaluation reuses saved test predictions without training or checkpoint substitution. The unchanged NEXT classic results retain stored scalar classifier scores and convert summed deposited-hit energy from MeV to keV. MJD uses calibrated physical energy recovered in official test-shard/row order. Event identities and the saved labels and energies are verified against the original metadata. The classic adapter's clipping is removed; Transformer energy caches are restored from float32 to the original float64 values. The three classic binary classifiers and all six reevaluated Transformers retain their single clean-oriented scores, while classic GINE retains the recorded aggregation of four PSD outputs (Appendix E). EXO-200's stored label 0 is positive ($n_{\mathrm{CCL}} = 1$): we set $y = 1 - y_{\mathrm{stored}}$ and negate the stored background-oriented logit for both metrics, including all nine reevaluated Transformers. The unchanged SuperNEMO classic logits favor $2\nu\beta\beta$ and use reconstructed calorimeter $E_1 + E_2$ in keV. The separate $0\nu\beta\beta$ illustration uses $s = E_1 + E_2$ directly. All selected files have unit base weights. No sigmoid is added to scalar logits: reproducing I requires retaining the score representation as well as the ranking direction.

Exact weighted quantiles. The evaluator sorts finite observations $z_{(j)}$ with positive weights $b_{(j)}$ and assigns cumulative center positions

$$t_j = \frac{\sum_{h \le j} b_{(h)} - b_{(j)}/2}{\sum_h b_{(h)}}. \tag{12}$$

The empirical quantile $Q(u)$ is linearly interpolated between $(t_j, z_{(j)})$, with endpoint values used outside $[t_1, t_n]$. Stable sorting preserves input order for tied observations. These are the quantiles in Equation 3 and in the histogram construction below. Setting the trim fraction to zero would use each class's minimum and maximum; the reported configurations instead use 0.005.

Units, boundaries, and population. Physical energy is converted to float64 keV before bin assignment. To avoid a unit-dependent one-step floating-point shift at a theoretical edge, an input MeV value exactly equal to a canonical keV edge divided by 1000 maps to that exact edge. Other nearby observations are not snapped by a tolerance. Internal boundaries enter the bin starting at that boundary, with the explicit 3000 keV exception in Equation 8. Original energies are retained in the audit, including MJD's 67 events above 3000 keV (66 nonclean and one clean), excluded from the two energy metrics rather than clipped into the 2995–3000 keV bin. Inclusive AUC always uses the original finite-score, positive-weight population, regardless of energy.

Matching calculation and diagnostics. Reproduction first retains finite score/energy pairs with $0 \le E_i \le 3000$ keV and positive base weight, then estimates C and assigns the shared 600 bins. Event counts and base masses are computed within $C \cap B_k$. The remaining steps reject bins with fewer than 20 events of either class, compute Equations 4–5, and evaluate one pooled weighted ROC. The class fractions p_{ck} are normalized over all common-support bins *before* invalid bins are removed. Sparse bins are not merged. The formal result is unavailable if support is disjoint, statistics

are inadequate, or a reporting gate fails. Coverage for class c is

$$\gamma_c = \frac{\sum_{i:y_i=c,\ w_i>0} a_i}{\sum_{i:y_i=c,\ s_i,E_i \text{ finite},\ a_i>0} a_i}. \tag{13}$$

This is retained *base* mass, not the matched mass, which is normalized to one in each class. Its denominator precedes both lower and upper energy range exclusions; range filtering cannot inflate the reported coverage. Range loss, subsequent support loss and sparse-bin loss are recorded separately. Both $\gamma_c \geq 0.5$ and at least two valid bins are required. Exactly constant common energy exempts only the two-bin requirement; the coverage and per-class event-count gates still apply. A diagnostic weighted AUC may be saved even when the formal reporting gate fails; it is not substituted for a missing formal result in our table. The effective sample size for any weights v_i is $(\sum_i v_i)^2 / \sum_i v_i^2$. Table 3 reports support and coverage for the reference CNN prediction sets.

Table 3: Matching diagnostics for the multi-view CNN prediction sets under their dataset-specific profiles (Table 2). All reevaluated classic models within each displayed test population use identical test event IDs, labels, physical energies, and base weights. Coverage γ_c uses the original finite score/energy class base mass before energy-range exclusion; ESS uses matched weights.

Dataset	Common support (keV)	Bins	γ_1	γ_0	ESS_1	ESS_0
NEXT	[2434.84, 2480.35]	11	0.9882	0.9700	55195	42260
MJD	[18.15, 2613.79]	493	0.9874	0.9631	141205	213947
EXO-200	[538.18, 2662.57]	426	0.9457	0.9414	51924	61627
SuperNEMO	[350.21, 2218.25]	374	0.9900	0.8988	312506	207113

Table 4: Disjoint matching losses and the separate independence sparse-bin loss, reported as positive/negative event counts. N_1/N_0 is the original test population. Range loss precedes support estimation, then matching sparse-bin exclusion. For MJD and EXO-200, range loss includes energies below 0 or above 3000 keV; the $E > 3000$ column is a subset of that loss. For retained NEXT and SuperNEMO results, only negative energies are excluded and $E > 3000$ counts are retained overflow diagnostics. The separate I sparse loss is measured after the applicable range filtering. Counts are shared by reevaluated classic models within each displayed test population.

Dataset	N_1/N_0	Range loss	$E > 3000$	I sparse loss	Support loss	Matching sparse loss
NEXT	64879/51670	0/0	0/0	0/0	767/1550	0/0
MJD	148251/241749	1/66	1/66	137/214	1746/7378	121/1481
EXO-200	66062/74321	3/1423	3/1423	152/257	3587/2932	0/0
SuperNEMO	327680/262144	0/0	0/40	333/271	3275/26530	0/0

No resampling, new model fitting, bootstrap interval, or multiple-seed average is introduced. Support restriction and matching weights can both alter the AUC relative to the inclusive population; their difference is descriptive, not a causal attribution to energy.

Fixed-grid independence score. The reevaluated MJD and EXO-200 records group events by physical category when supplied, otherwise by binary label. Within each group the evaluator removes nonfinite score/energy pairs, energies outside 0–3000 keV, and zero-weight events, retains only energy bins with at least 20 events, and computes weighted score quantiles using the retained events. Duplicate score edges are removed. For nonconstant scores the extreme edges are moved outward by one floating-point step; constant scores receive one interval centered at that value with half-width $\max(|s|10^{-9}, 10^{-12})$. Score histograms use left-closed/right-open bins with the last right edge included. They are normalized probability masses, not densities. No pseudocount is added.

Let R_g be the retained events, R_{gk} those in bin k, and S_{gr} score-histogram interval r. Then

$$P_{gk,r} = \frac{\sum_{i \in R_{gk}} a_i 1\{s_i \in S_{gr}\}}{\sum_{i \in R_{gk}} a_i}, \qquad P_{g,r} = \frac{\sum_{i \in R_g} a_i 1\{s_i \in S_{gr}\}}{\sum_{i \in R_g} a_i}, \tag{14}$$

$$\pi_{gk} = \frac{\sum_{i \in R_{gk}} a_i}{\sum_{i \in R_g} a_i}. \tag{15}$$

Equations 9–10 operate on these histograms. The final group weight A_g includes all finite-pair, positive-weight events in that group within 0–3000 keV, including events from rejected sparse bins. Thus the overall score is not an unweighted class mean and not a mean weighted only by retained events. A group with fewer than 20 finite events or no eligible energy bin is excluded; if no group is estimable, the score is missing. One eligible bin is sufficient for this implementation, unlike the usual two-bin reporting gate for weighted AUC.

Table 5: Class-level independence diagnostics under the profiles in Table 2. Retained fractions divide eligible-bin base mass by original finite score/energy class base mass, before energy-range exclusion. Score histograms retain up to 20 weighted score-quantile bins. The audit files include all extended NEXT configurations and full precision.

Dataset / model	I_1	I_0	$\min_g I_g$	$\text{Bins}_{1/0}$	$\text{Retained}_{1/0}$
NEXT / Multi-view CNN	0.9763	0.9643	0.9643	11/11	1.0000/1.0000
NEXT / Static GINE	0.9778	0.9609	0.9609	11/11	1.0000/1.0000
NEXT / BiGRU	0.9753	0.9657	0.9657	11/11	1.0000/1.0000
NEXT / PointMamba-lite	0.9776	0.9632	0.9632	11/11	1.0000/1.0000
MJD / Multi-view CNN	0.7325	0.7763	0.7325	496/524	0.9991/0.9988
MJD / Static GINE	0.8057	0.7911	0.7911	496/524	0.9991/0.9988
MJD / BiGRU	0.7791	0.7730	0.7730	496/524	0.9991/0.9988
MJD / PointMamba-lite	0.7519	0.7865	0.7519	496/524	0.9991/0.9988
EXO-200 / Multi-view CNN	0.7315	0.7709	0.7315	438/484	0.9977/0.9774
EXO-200 / Static GINE	0.7925	0.8093	0.7925	438/484	0.9977/0.9774
EXO-200 / BiGRU	0.7739	0.8033	0.7739	438/484	0.9977/0.9774
EXO-200 / PointMamba-lite	0.7854	0.7863	0.7854	438/484	0.9977/0.9774
SuperNEMO / Multi-view CNN	0.9213	0.8962	0.8962	417/514	0.9990/0.9990
SuperNEMO / Static GINE	0.9241	0.8981	0.8981	417/514	0.9990/0.9990
SuperNEMO / BiGRU	0.9292	0.9105	0.9105	417/514	0.9990/0.9990
SuperNEMO / PointMamba-lite	0.9292	0.9119	0.9119	417/514	0.9990/0.9990

Changes from historical implementations. The historical EXO-200 evaluations, including the nine Transformer variants, used six class-balanced pooled energy-quantile bins for matching, and a separate eight per-category energy-quantile construction for I. Both energy constructions are replaced by Equation 8; the up-to-20 score-quantile bins remain. Their old I implementation could compare a local energy bin containing only two events and pooled all finite events. The common reference instead requires 20 events per group/bin and computes score edges and the pooled histogram only on retained events. Bin weights use retained group mass; final group weights use eligible population mass before sparse exclusion. These choices follow the former NEXT/MJD fixed-grid estimator and preserve the JS-then-square-root and final group-aggregation definitions. The six MJD Transformer caches used 836 fixed 5 keV bins extending to 4180 keV, without clipping. They now use Equation 8 and the same range exclusions as the other reevaluated models. Their original float64 physical energies are restored before binning. These changes can alter both score-histogram populations and support quantiles, even though the historical MJD Transformer bin width was already 5 keV.

Retained NEXT and SuperNEMO profile. The previously reported NEXT and SuperNEMO classic results, including NEXT's separate exploratory batches, and the SuperNEMO energy-only illustration retain `EnergyBench-unified-5keV-overflow-v2.0.0`. This profile uses the 600 regular bins in Equation 8 plus $B_{600} = (3000, \infty)$ keV. Negative energies are excluded, while

Table 6: Diagnostics for the 15 reevaluated MJD and EXO-200 Transformers using 600 fixed 5 keV bins on 0–3000 keV. All match the CNN test identities, class partitions, physical energies and unit weights within their dataset, so common support and matching losses equal Tables 3–4. Retained fractions and coverage use original finite-pair class mass before range exclusion. The minimum group score is the smaller of the two displayed I_g values. All formal weighted AUCs pass the reporting gates.

Dataset / Transformer	I_1/I_0	$\text{Bins}_{I,1/0}$	$\text{Retained}_{I,1/0}$
EXO-200 / entity–mlp	0.6868/0.7828	438/484	0.9977/0.9774
EXO-200 / entity–fourier	0.6545/0.7969	438/484	0.9977/0.9774
EXO-200 / entity–rope	0.6583/0.7671	438/484	0.9977/0.9774
EXO-200 / region–mlp	0.7893/0.8115	438/484	0.9977/0.9774
EXO-200 / region–fourier	0.7646/0.8131	438/484	0.9977/0.9774
EXO-200 / region–rope	0.7502/0.8246	438/484	0.9977/0.9774
EXO-200 / summary–mlp	0.7296/0.8183	438/484	0.9977/0.9774
EXO-200 / summary–fourier	0.7894/0.8132	438/484	0.9977/0.9774
EXO-200 / summary–rope	0.7713/0.7740	438/484	0.9977/0.9774
MJD / entity–mlp	0.8388/0.7950	496/524	0.9991/0.9988
MJD / entity–fourier	0.8196/0.7913	496/524	0.9991/0.9988
MJD / region–mlp	0.7935/0.8013	496/524	0.9991/0.9988
MJD / region–fourier	0.7869/0.7881	496/524	0.9991/0.9988
MJD / summary–mlp	0.8019/0.7973	496/524	0.9991/0.9988
MJD / summary–fourier	0.7914/0.7759	496/524	0.9991/0.9988

Dataset / Transformer	K_{match}	γ_1/γ_0	$\text{ESS}_{1/0}$
EXO-200 / entity–mlp	426	0.9457/0.9414	51924/61627
EXO-200 / entity–fourier	426	0.9457/0.9414	51924/61627
EXO-200 / entity–rope	426	0.9457/0.9414	51924/61627
EXO-200 / region–mlp	426	0.9457/0.9414	51924/61627
EXO-200 / region–fourier	426	0.9457/0.9414	51924/61627
EXO-200 / region–rope	426	0.9457/0.9414	51924/61627
EXO-200 / summary–mlp	426	0.9457/0.9414	51924/61627
EXO-200 / summary–fourier	426	0.9457/0.9414	51924/61627
EXO-200 / summary–rope	426	0.9457/0.9414	51924/61627
MJD / entity–mlp	493	0.9874/0.9631	141205/213947
MJD / entity–fourier	493	0.9874/0.9631	141205/213947
MJD / region–mlp	493	0.9874/0.9631	141205/213947
MJD / region–fourier	493	0.9874/0.9631	141205/213947
MJD / summary–mlp	493	0.9874/0.9631	141205/213947
MJD / summary–fourier	493	0.9874/0.9631	141205/213947

all finite nonnegative energies enter the support-quantile calculation and the pre-sparse group masses A_g. The overflow bin is subject to the same event-count rules as a regular bin. The score-quantile histograms, overlap target, weighted AUC and original finite-pair coverage denominators use the same definitions as above. These retained results and their figures are not recomputed under the MJD/EXO-200 range restriction.

Reproduction and validation. The MJD/EXO-200 configuration is `EnergyBench-unified-5keV-range-v3.0.0`. The preceding v2 configuration remains available for the retained NEXT and SuperNEMO results; each result records its applicable profile and SHA256 fingerprint. The reproducibility bundle contains the corresponding frozen evaluators, event-input manifests and hashes, full-precision metrics, class-level bin counts and losses, and a table generator that reads these results. Numerical checks cover unit equivalence at all 601 finite edges, adjacent floating-point values, exclusion outside 0–3000 keV, constants, ties, missing and sparse inputs, weighted pair enumeration, and replay of the historical fixed-grid

estimator on identical supported inputs. Independent recalculation checks the metrics, coverage gates, and original event sources.

Degeneracies and scope. The unified classification entry point rejects labels outside $\{0, 1\}$ on the entire input array. It also rejects negative or nonfinite weights and misaligned arrays; entirely zero weight yields unestimable metrics. Nonfinite scores are excluded from ROC, and missing energy is additionally excluded from energy-dependent statistics. An absent ROC class makes AUC unavailable; I can still be computed for estimable groups. If no category array is supplied, binary labels define the groups; missing entries in a supplied category array are rejected. The dependence routine filters nonfinite pairs; no correlation coefficient enters I. Constant scores produce identical histograms and $I = 1$. Constant conditioning energy likewise supplies no meaningful test of energy dependence.

The score in Equation 10 averages pairwise Jensen–Shannon divergences *before* taking a square root; it is not an average of Jensen–Shannon distances and not the multi-distribution Jensen–Shannon divergence. We use a descriptive benchmark name for this implemented transformation without claiming that it is a standard test or an original statistical construction.

B MJD Waveform Evidence for Energy-Aware Evaluation

Task, population, and selection. The local MJD release contains six test shards with 390,000 events, including 148,251 clean events. Clean means that all four reference PSD labels—low AvsE, high AvsE, DCR, and LQ—equal one; the formal task is clean versus non-clean, not low versus high energy. Calibrated energy is provided by `energy_label` in keV, and raw waveforms contain 3800 ADC samples (Arnquist et al., 2023b). We choose the detector/run pair with the most clean test events, breaking count ties by detector and run identifier. This gives detector 133, run 50928, with 410 events. Selection uses metadata and labels, before waveform inspection.

The lower and upper energy quartile boundaries are 226.871 and 471.248 keV. The corresponding cohorts contain 103 events each, covering 5.7–226.9 and 471.2–2616.4 keV; the middle group contains 204. For the event displays, we choose the event nearest each cohort's median energy, with event identifier and row index breaking ties. The selected events have energies 157.975 and 1131.515 keV and identifiers 2407207 and 2585508. This rule does not inspect pulse shape or noise.

Actual waveform preprocessing. For raw waveform samples v_{it}, with $t = 0, \ldots, 3799$, define

$$b_i = \frac{1}{200} \sum_{t=0}^{199} v_{it}, \qquad \alpha_i = \max_t |v_{it} - b_i|, \qquad u_{it} = \frac{v_{it} - b_i}{\alpha_i}. \tag{16}$$

The implementation reads float32 samples, computes the baseline mean in float64 and casts it to float32, then subtracts and normalizes in float32. Division is skipped when $\alpha_i = 0$; no selected event has zero amplitude. This is the common waveform preparation before the STFT or point representation, not an assertion that every classifier directly receives the plotted one-dimensional array. The time axis is sample index; no sampling interval is assumed.

We summarize residual baseline fluctuations by

$$r_i = \left(\frac{1}{200} \sum_{t=0}^{199} u_{it}^2 \right)^{1/2}. \tag{17}$$

Up to float32 rounding, r_i is baseline RMS in ADC counts divided by α_i. It is a descriptive input statistic, distinct from the benchmark's energy independence score. The lowest/highest energy cohorts have median r_i values 0.006720 and 0.000889, a ratio of 7.56. Their absolute baseline RMS medians are much closer, 2.233 and 2.166 ADC counts. Across the 410 events, energy has Spearman correlation 0.993 with α_i, -0.012 with absolute baseline RMS, and -0.826 with r_i. These are single-cohort descriptive statistics, without uncertainty intervals or a claim that they hold for every detector/run. Neither correlation nor these event displays determines the strategy of a trained classifier.

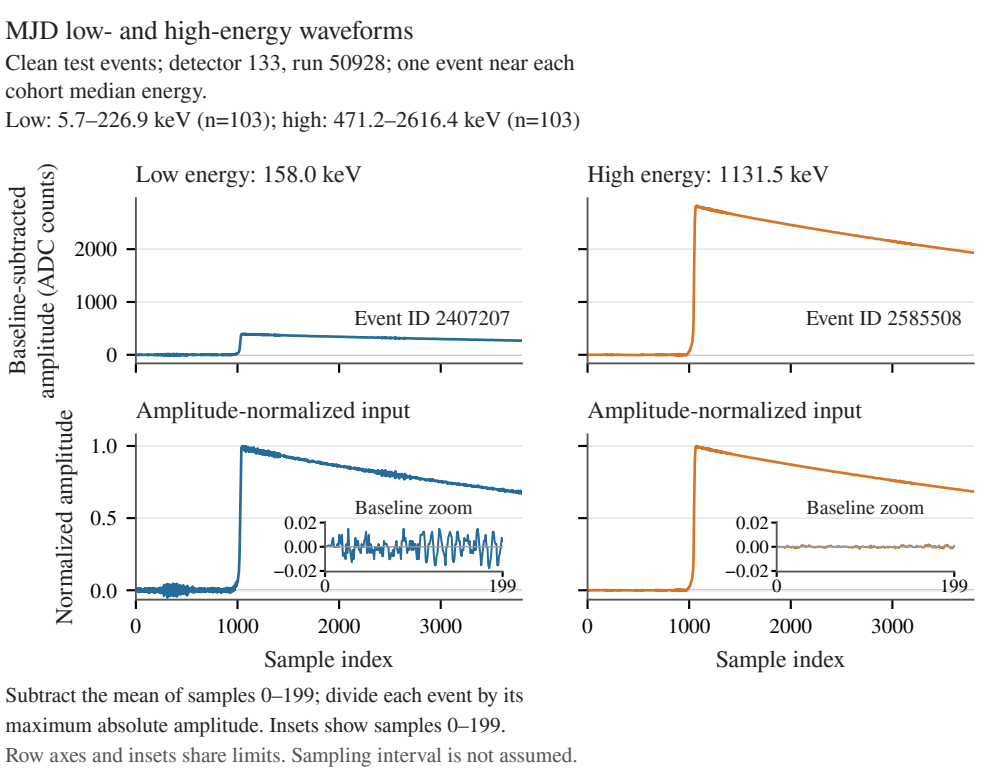


Figure 4: Detailed MJD event displays from the fixed detector/run cohort. Top: baseline-subtracted ADC amplitudes. Bottom: normalized waveforms before model-specific representation, with enlarged first-200-sample baselines. The two columns share vertical limits within each row, and both baseline insets use the same scale. Events were selected by energy alone using the stated median rule.

Connection to the saved benchmark. Raw test labels and event IDs match the saved CNN evaluation population. Calibrated energies are recovered in float64 from the original test shards; the old adapter's float32 conversion and clipping are not used. The 67 events above 3000 keV (66 non-clean and one clean) are excluded from energy evaluation before support estimation; they remain in the original coverage denominator. The common support is $[18.151, 2613.790]$ keV with 493 valid 5 keV bins, retaining 146,383 clean and 232,824 non-clean events. The CNN's inclusive and weighted AUCs are 0.971283 and 0.969851, respectively, while $I = 0.759637$. The small AUC difference does not establish energy-dominated classification. Matching controls the class energy mixture; the independence score asks a separate question about the predictions within each class.

C Energy-Only SuperNEMO Diagnostic

Population and operating point. This diagnostic uses all 196,608 simulated $0\nu\beta\beta$ and 262,144 ^{214}Bi events in the existing shared test partitions. It is separate from the trained $2\nu\beta\beta$ classification task. Energy is the reconstructed calorimeter sum $E_1 + E_2$ in keV, extracted once per event; the identical event-level fields repeated on tracker rows are not counted as separate observations. The released two-electron topology selection applies, with no additional energy cut or random sampling for the spectra. Class normalization removes differences in simulation sample counts; it does not estimate exposure-normalized physical rates.

At the illustrative threshold $t = 2200\,\text{keV}$, the rule $\widehat{y}_i = 1\{E_i \geq t\}$ accepts 184,727 of 196,608 signal events and 26,793 of 262,144 background events. Hence

$$\text{TPR} = 0.939570, \qquad \text{FPR} = 0.102207, \qquad \text{BA} = \tfrac{1}{2}[\text{TPR} + (1 - \text{FPR})] = 0.918681, \quad (18)$$

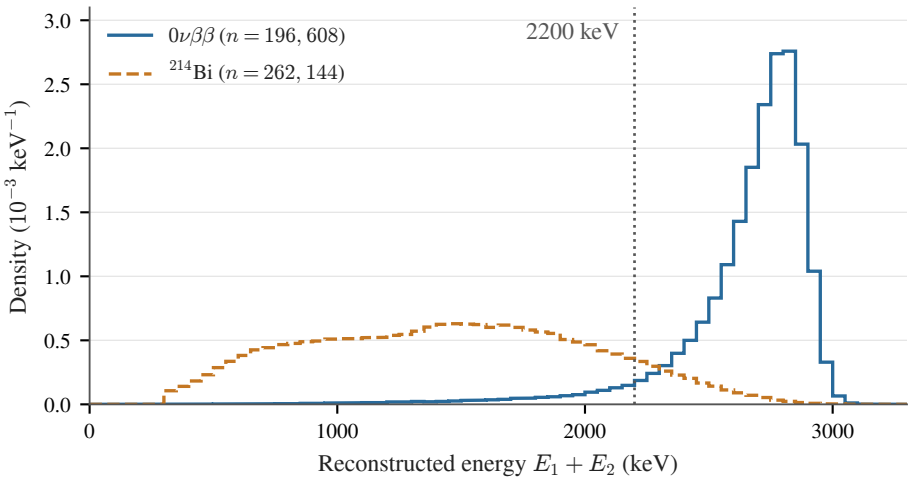


Figure 5: The motivating energy-spectrum difference: 196,608 simulated $0\nu\beta\beta$ and 262,144 ^{214}Bi events from shared SuperNEMO test partitions. The 50 keV display bins have unit area per class, so sample counts do not determine their relative heights. The released two-electron selection applies (Macko, 2026a); no additional energy cut is applied to the plotted spectra. The vertical line marks the illustrative 2200 keV threshold, not a test-optimized operating point.

where BA is balanced accuracy, assigning equal importance to both classes. Figure 6 shows the operating-point trade-off. The threshold was specified for illustration rather than selected by maximizing test performance. Neither this cut nor the score $s_i = E_i$ requires training or event topology.

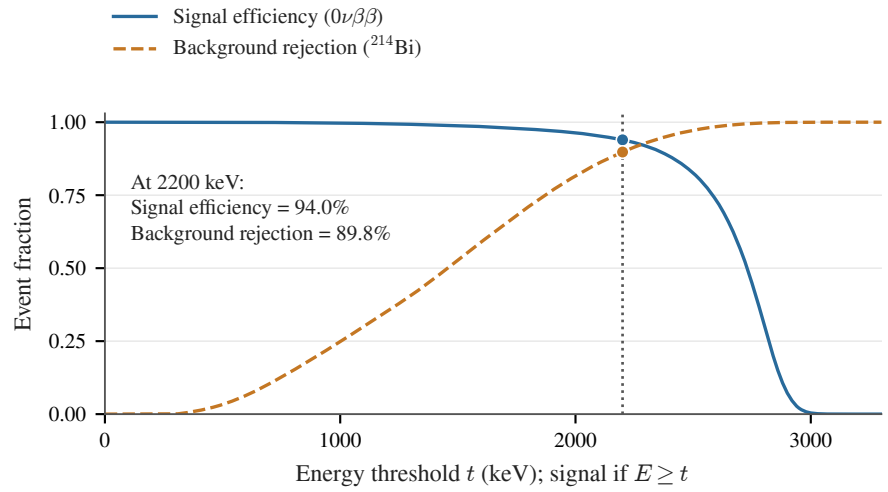


Figure 6: Signal efficiency and background rejection as an energy-only threshold varies in the same SuperNEMO $0\nu\beta\beta$–^{214}Bi test partitions. Curves use all events; the display samples thresholds every 10 keV. Markers identify the illustrative 2200 keV cut, with rates calculated directly from event energies.

Matching and coverage. We apply the shared 600 regular 5 keV bins plus one overflow bin to this diagnostic pair, with class-wise 0.5% tail trimming, overlap matching, and at least 20 events of each class per bin. Common support remains $C = [1169.390, 2734.195]$ keV. The slice of the zero-anchored grid intersecting C has edges

$$(1165,\ 1170,\ 1175,\ \ldots,\ 2730,\ 2735)\ \text{keV}. \tag{19}$$

Events must also satisfy $E \in C$, so the boundary bins are partially occupied. Of 314 candidate bins, 283 pass the count requirement; sparse bins are excluded without merging. This retains 99,139 signal and 150,591 background events, or fractions 0.504247 and 0.574459, respectively. Both satisfy the minimum coverage of 0.5, with signal coverage close to the boundary. The matched result therefore describes the retained energy population, not the full signal peak. Reweighting gives effective sample sizes of approximately 34,127 and 64,317. The 50 keV display bins in Figure 5 do not set the matching resolution.

The inclusive ROC uses all original energies. The shared energy grid is assigned before support trimming, including 758 signal and 40 background events with $E > 3000$ keV in one overflow bin. All energies are finite and nonnegative. Here C lies below 3000 keV, so overflow events receive zero matched weight; they remain in the coverage denominators and in the separate independence calculation whenever the group/bin count rule passes. No energies are clipped. The trained $2\nu\beta\beta$ task uses the same evaluation rules on its own distinct test population.

Table 7: Energy-only AUC under successive evaluation conditions. Retained event counts are unweighted; only the last row uses matching weights. The populations and weights differ, so these rows are not competing models.

Evaluation condition	Signal events	Background events	AUC
Original test populations	196,608	262,144	0.970613
Common support, unit weights	99,595	172,757	0.934642
Retained 5 keV bins, unit weights	99,139	150,591	0.929715
Retained 5 keV bins, matched weights	99,139	150,591	0.500024

Residual within-bin ordering. For the increasing score $s = E$, let q_k denote the shared matched mass of ordered energy bin k, and let A_k be the energy-only AUC within that bin, with half credit for ties. Cross-bin comparisons give

$$\mathrm{AUC_w}(E) = \sum_{k>\ell} q_k q_\ell + \sum_k q_k^2 A_k = \frac{1}{2} + \sum_k q_k^2 \left(A_k - \frac{1}{2} \right). \tag{20}$$

Equal bin masses balance cross-bin pairs, while within-bin energy differences remain. This identity gives AUC 0.500024 for the fixed 5 keV matching bins; it does not redefine the metric as an average of per-bin AUCs. A score depending only on bin identity would instead have weighted AUC exactly $1/2$. The independence calculation uses the same energy grid with original base weights and up to 20 score-quantile bins, yielding $I = 0.079210$ for $s = E$. Its class values are $I_{0\nu} = 0.082192$ and $I_{\mathrm{Bi}} = 0.076972$. The 337 and 514 eligible group bins retain 195,169 signal and 261,873 background events, respectively, including their eligible overflow bins. The energy-only score still varies within each 5 keV bin; balanced bin masses therefore do not imply exact continuous-score AUC $1/2$. These are deterministic diagnostic values, without bootstrap intervals or multiple-seed averages. They demonstrate the availability of an energy-only decision rule, not the internal strategy of any trained model or a measured failure under distribution shift.

D SuperNEMO Tracker Extent and Energy

For an event with h_i raw tracker-hit positions $\vec{r}_{ij}$ in millimeters, we measure spatial extent by

$$R_{g,i} = \left[\frac{1}{h_i} \sum_j \|\vec{r}_{ij} - \bar{\vec{r}}_i\|^2 \right]^{1/2}, \qquad \bar{\vec{r}}_i = \frac{1}{h_i} \sum_j \vec{r}_{ij}. \tag{21}$$

This radius of gyration uses all stored tracker hit centers with equal weights; it is not a fitted track length. Reconstructed energy is the calorimeter sum $E_1 + E_2$, counted once per event. We retain all 262,144 ^{214}Bi and 196,608 $0\nu\beta\beta$ test events, with no missing energies or coordinate summaries.

Figure 7 compares complete populations. Each 100 keV energy column is normalized over all 20 mm extent bins. Columns with fewer than 20 events are masked, affecting one ^{214}Bi and 35

$0\nu\beta\beta$ events. Extents above 800 mm remain in the normalization but are outside the heatmap view; the cumulative plots include their full tails. These display bins are separate from EnergyBench matching bins.

To compare balanced cohorts, events are ranked by energy and then event ordinal within each physical class. We take the first and last quarter: 65,536 events per cohort for ^{214}Bi and 49,152 for $0\nu\beta\beta$. Exact energy ties are split by ordinal, not geometry. The ^{214}Bi quartile boundaries are 1003.0 and 1870.25 keV; the $0\nu\beta\beta$ boundaries are 2583.125 and 2821.5 keV. These relative energy cohorts are not common absolute windows across classes.

Median R_g decreases from 335.97 to 291.14 mm for the illustrative $0\nu\beta\beta$ sample and changes from 281.97 to 294.04 mm for ^{214}Bi. The distributions overlap substantially. Geometry can therefore contain energy-associated information, but these observations do not establish its use by a trained classifier or a universal monotonic size–energy relation. No specific physical or selection mechanism is inferred. The formal SuperNEMO benchmark remains $2\nu\beta\beta$ versus ^{214}Bi.

The current point adapters center events, quantize at 15 mm, and divide coordinates by a fixed 1000 mm scale; CNNs use a fixed 44 mm occupancy grid. They exclude calorimeter energies and do not normalize each event to unit spatial extent. Quantization and aggregation nevertheless change the input, so the raw extent statistic is not a claim about the learned representation.

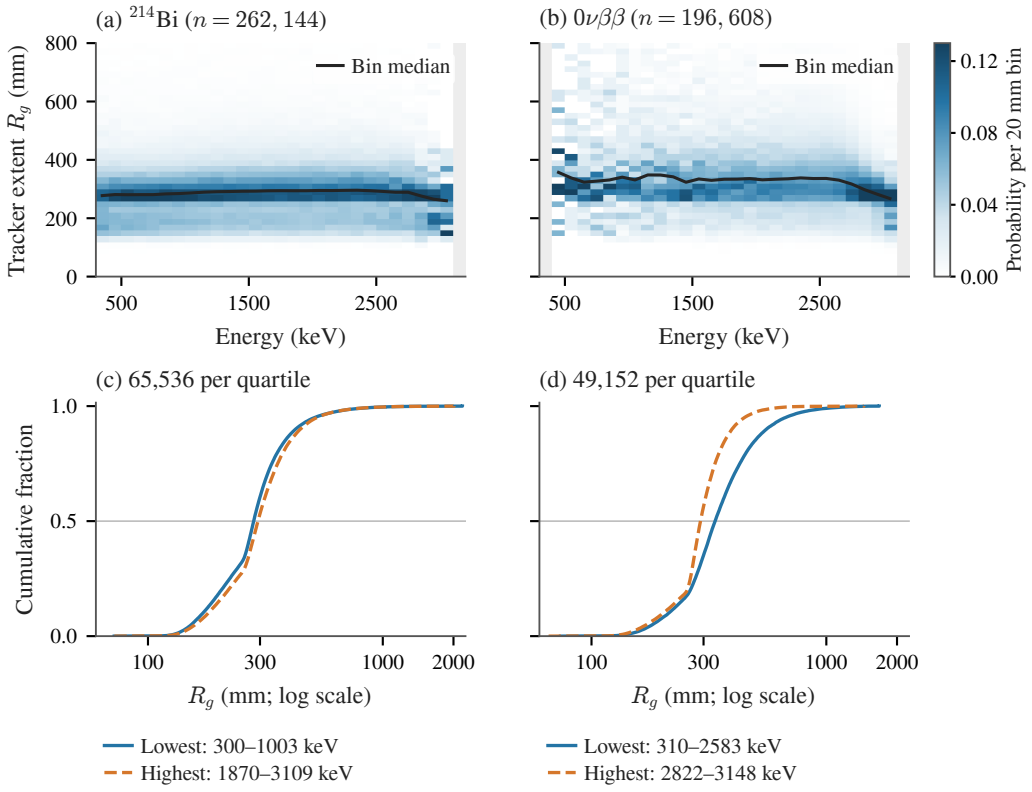


Figure 7: SuperNEMO tracker extent versus reconstructed energy. Top: conditional distributions using 100 keV by 20 mm bins; each energy column is normalized over all extents. Black curves show medians; gray columns contain fewer than 20 events. Extents above 800 mm (0.99% of ^{214}Bi and 1.10% of $0\nu\beta\beta$) are excluded from view only. Bottom: complete cumulative distributions for the lowest and highest energy-ranked quartiles within each class, on a logarithmic extent axis. Cohort sizes and energy ranges are given in the panels.

Table 8: Input adaptations for the four non-Transformer architectures. GINE, BiGRU, and PointMamba-lite share each dataset's point representation. Model dimensions and representation details are given in Appendix E.

Dataset	MV-CNN input	Graph and sequence input
NEXT	Three normalized spatial energy projections	Centered 15 mm voxels; energy fraction and log hit count
SuperNEMO	Three binary tracker projections	Centered 15 mm tracker cells; occupancy fraction and log hit count
MJD	Three-resolution waveform STFT	512 pooled waveform segments with mean and RMS features
EXO-200	STFT of serialized 226×300 waveform	Same waveform construction after channel-major serialization

E Non-Transformer Architectures and Run Protocols

E.1 Evidence Supporting the Four-Model Comparison

Unified reevaluation of earlier NEXT checkpoints provides retrospective context for the representative architectures. In the alternative-model batch, the multi-view CNN has weighted AUC 0.955804, compared with 0.947782 for the multiscale CNN and 0.928273 for the dense 3D CNN; Static GINE leads the graph candidates in that batch at 0.981008. In the subsequent candidate batch, BiGRU exceeds the tested dilated TCN (0.933864 versus 0.926085). PointMamba-lite, the only recorded state-space candidate, has weighted AUC 0.925386 and $I = 0.971387$; its inclusion broadens the architectural comparison rather than following an overall accuracy ranking.

These are held-out test comparisons on 115,499 events from earlier exploratory batches. They are distinct from the 116,549-event reevaluation used for the available NEXT classic results in Table 1. PointMamba-lite is included in Table 10. The original exploratory batches used different evaluator-code fingerprints; the numbers above now use the same regular-plus-overflow protocol as the current classic results while retaining the earlier checkpoints and test set. Their distinct test population is not pooled with the retained campaign. The four search groups are recorded in the model-family presentation and are reproduced in Table 10. They organize the candidate inventory; the surviving records do not document a prospective validation-based rule selecting one winner from each group. The retained architectures span three of the four groups, including two distinct mechanisms from the sequence/state-space group. The structural rationale and early test comparisons therefore support the comparison set without establishing a recovered formal selection rule.

Parameter-count provenance. Parameter counts in Table 10 are the saved counts of trainable neural parameters, checked against the corresponding run summaries and reevaluated metrics. Neural parameter count is inapplicable to XGBoost. The earlier NEXT PointMamba-lite run summary records 316,993 trainable parameters, consistent with `Benchmarking table.xlsx`; this is displayed as 0.3170 M in Table 10 and 317 K in Table 1. Its retained-campaign predictions remain unavailable.

Spatial representations. NEXT MV-CNN uses three 128×128 projections with 30 mm bins. For point-based models, hits are centered by the complete-event energy centroid and merged into 15 mm cells. Voxel centers are recentered and divided by 1,000 mm; the two features are the voxel's fraction of complete-event energy and $\log(1 + n)$, where n is its hit count. Events exceeding 512 occupied cells retain the highest-energy cells with deterministic tie breaking, without renormalizing the retained energy fractions. SuperNEMO instead uses tracker coordinates: its CNN projects binary occupancy on 128×128 grids with 44 mm bins. Its point adapter centers the tracker hits geometrically and merges 15 mm cells, replacing the unavailable hit-energy fractions with occupancy fractions. It raises an error above 512 cells instead of truncating. The non-Transformer SuperNEMO inputs exclude E_1, E_2, and tracker drift radii. These input choices restrict direct energy access but do not establish independence from energy.

Table 9: Specified common protocol versus recorded training settings for the audited non-Transformer models. Every listed configuration has an upper limit of 50 epochs, weight decay 10^{-4}, gradient-norm limit 1, and seed 42. NEXT settings refer to the three available retained-campaign runs. "Other datasets" means EXO-200, MJD, and SuperNEMO.

Scope	Architecture	Batch	Initial learning rate	Patience
Specified protocol	Neural models	64	5×10^{-4}	5
NEXT recorded	MV-CNN, GINE, BiGRU	64	5×10^{-4}	5
Other datasets	MV-CNN	16	10^{-3}	12
Other datasets	GINE, BiGRU	16	5×10^{-4}	12
Other datasets	PointMamba-lite	4	3×10^{-4}	12

Waveform representations. MJD subtracts the mean of the first 200 waveform samples and divides by the maximum absolute event amplitude when nonzero. EXO-200 subtracts a baseline separately for each channel using its first 200 time samples, divides by the event-wide maximum absolute amplitude, and serializes the 226×300 array in channel-major order. For these datasets, MV-CNN uses Hann-window STFTs with FFT sizes $(64, 128, 256)$ and hops $(16, 32, 64)$, followed by a $\log(1 + |\mathrm{STFT}|)$ transform and resizing to 128×128. The point adapter uses adaptive pooling to form 512 segment means and root-mean-square amplitudes. Its three coordinates are normalized time, amplitude-normalized segment mean, and the first difference of that normalized mean; its two features are the segment mean and RMS. These are pulse-shape coordinates, not calibrated detector positions.

Backbone details. MV-CNN has four residual stages with two blocks each, base width 16, and a 256-dimensional fusion layer. GINE uses hidden width 128, five message-passing layers, $k = 12$ neighbors, and a 160-dimensional classifier. BiGRU uses a 96-dimensional input embedding, two bidirectional layers with hidden size 128, and a 256-dimensional classifier. The two Hilbert orders quantize each event's point bounding box to 10 bits per axis; the second swaps the first two axes before applying the same ordering. PointMamba-lite uses three blocks with model width 128, inner width 192, state dimension 16, and 32-position scan chunks. Its guarded PyTorch scan is not the official fused Mamba implementation. The backbones use dropout 0.1. MJD GINE has four output logits z_{ij}, one per reference label, and uses

$$q_i = \prod_{j=1}^{4} \sigma(z_{ij}), \qquad s_i = \log q_i - \log(1 - q_i), \tag{22}$$

where σ is the logistic sigmoid and s_i is its evaluated clean-event score. This product is the implemented score construction; it is not a demonstrated factorization of the labels' joint conditional distribution. The computation uses summed log-sigmoids and clamps the log probability below zero for numerical stability. EXO-200 uses the same GINE backbone with one output.

Run completion and randomness. The audited non-Transformer classification loaders retain the sampled class proportions; their binary losses do not use class weights, and these paths contain no synthetic data augmentation. NEXT shuffles training slices and buffered events, waveform loaders shuffle training events, and SuperNEMO records proportional sampling without replacement. Input normalization and ordering are dataset-specific preprocessing rather than energy matching of training events.

In the NEXT campaign reported here, MV-CNN, GINE, and BiGRU finish after 19, 36, and 26 epochs, selecting epochs 14, 31, and 21, respectively, by validation inclusive AUC. The loader batch size is 64; the wrapper evaluates these backbones in chunks of 16, concatenates their logits, and applies one batch loss and optimizer step. The corresponding run summaries record Python 3.11.15, PyTorch 2.11.0+cu128, an NVIDIA GeForce RTX 5090, and bfloat16 mixed precision. This environment evidence applies to those runs rather than automatically to every dataset. The audited non-Transformer training paths add neither class-weighted losses nor data augmentation; the current SuperNEMO stream retains the source class proportions without replacement.

For SuperNEMO, the workbook records 21, 50, 22, and 9 training epochs for MV-CNN, GINE, BiGRU, and PointMamba-lite, with selected epochs 16, 43, 17, and 4, respectively. It records 47 epochs and selected epoch 43 for MJD PointMamba-lite. Earlier logs were incomplete; the current results and epoch counts follow the workbook, whose evaluation status does not establish training completion. The audited non-Transformer configurations set seed 42 and disable enforced deterministic algorithms. Mixed precision is enabled in these configurations except for MJD BiGRU. Seeds therefore do not imply bitwise reproducibility.

Screening scope. The current NEXT registry contains three legacy CNNs and twenty alternative classifiers, including point, graph, sequence, state-space, sparse-convolution, hybrid, mixer, persistence, and boosted-tree approaches. A classification entry need not correspond to a completed retained-campaign run: PointMamba-lite has an earlier exploratory result but remains unavailable in that campaign. The search taxonomy groups the three legacy projection CNNs together with projection, patch-mixer, dense-voxel, and sparse-voxel alternatives. Point-set and graph methods form separate groups; ordered sequence methods, the state-space model, topological classifiers, and the CNN–GNN hybrid form the fourth. This representation-based organization differs from the finer model-family labels in the run registry. Checkpoint selection is documented within completed runs, while a prospective rule for choosing the cross-dataset architectures is not. The boosted-tree classifier in the extended NEXT inventory uses up to 500 estimators, learning rate 0.04, maximum depth 4, subsampling 0.85, column subsampling 0.9, early-stopping rounds 12, and random seed 42. Its estimator settings, rather than the neural epoch, optimizer, and clipping fields in the generic run configuration, determine its training procedure.

F EXTENDED NEXT CLASSIFICATION RESULTS

Table 10: NEXT classification candidates reevaluated with 600 regular 5 keV bins on 0–3000 keV plus one overflow bin above 3000 keV. Params (M) counts trainable neural parameters in millions. A dash denotes an inapplicable parameter count. †Architecture instantiated across datasets; no new model is fitted.

Model	Params (M)	Inclusive AUC ↑	Weighted AUC ↑	I ↑
Projection and voxel models				
Two-conv CNN	0.0014	0.953280	0.952992	0.966472
Energy-skip CNN	0.0014	0.923394	0.922735	0.967952
Residual spatial CNN	1.2288	0.979030	0.978966	0.969531
Multi-view CNN†	0.8077	0.976636	0.976587	0.970993
Multi-scale CNN	0.8659	0.979270	0.979046	0.973145
Dense 3D ResNet	0.6884	0.984115	0.983956	0.971101
Projection MLP mixer	0.5391	0.950352	0.950086	0.972549
Submanifold ResNet	0.2982	0.991574	0.991414	0.967824
Point-set models				
DeepSets	0.0756	0.969660	0.969459	0.972602
PointNet++	0.1645	0.985459	0.985354	0.972546
PointMLP	0.4435	0.994033	0.993814	0.967686
Rigid KPConv	0.3910	0.986654	0.986341	0.969283
Graph models				
Static GINE†	0.4799	0.994801	0.994625	0.970265
ParticleNet-lite	0.1505	0.992161	0.991971	0.968044
EGNN	0.6405	0.890742	0.889878	0.968064
GravNet	0.2936	0.991316	0.991070	0.969637
DimeNet-lite	0.1696	0.992939	0.992716	0.968001
Sequence, state-space, topology and hybrid models				
BiGRU†	0.7340	0.979563	0.979377	0.971051
Dilated TCN	0.6945	0.973450	0.973183	0.974533
PointMamba-lite†	0.3170	0.925742	0.925386	0.971387
Persistence/PersLay	0.0633	0.918258	0.917657	0.968522
Topology XGBoost	—	0.937279	0.936938	0.970396
CNN–GNN	0.3420	0.991417	0.991260	0.970276

G TRANSFORMER REPRESENTATION DETAILS

Table 11 records the concrete dataset-specific implementations of the three tokenization families used in Section 5.2. Here, d_f is the token content width and d_p is the coordinate width. A token count preceded by $\leq$ denotes a maximum sequence length; shorter events are masked and padded only for batching.

NEXT and SuperNEMO use three-dimensional detector position coordinates. MJD uses normalized time together with normalized waveform level and local temporal change. EXO-200 uses normalized time, within-block channel position, detector side, and indicators for the U-wire, V-wire, and APD sensor families.

Table 11: Dataset-specific Transformer tokenization configurations. Internal feature padding, such as the channel padding used by EXO-200 raw patches, does not introduce additional Transformer tokens.

Dataset	Family	Tokens	d_f/d_p	Construction
NEXT	Entity	≤ 512	$2/3$	Individual detector hits selected under a fixed token cap. Token content describes the local hit measurement, while centered three-dimensional coordinates provide position.
NEXT	Region	≤ 512	$2/3$	Hits are aggregated within occupied three-dimensional voxels of width 15 detector-coordinate units. If necessary, the most occupied voxels are retained under the token cap.
NEXT	Summary	≤ 512	$4/3$	Hits are ordered using a three-dimensional Morton ordering and partitioned into balanced spatial groups. Each token summarizes the measurements within one group.
MJD	Entity	500	$2/3$	The length-3800 waveform is represented by a deterministic mixture of uniformly spaced and high-importance samples. Local context spans nine waveform samples.
MJD	Region	500	$20/3$	The waveform is divided into 512 non-overlapping patches containing 20 raw amplitudes each.
MJD	Summary	500	$2/3$	The same 512 temporal regions are represented by their segment mean and root-mean-square amplitude.
EXO-200	Entity	504	$2/6$	Entities are selected independently within each of six sensor blocks using equal proportions of uniform and waveform-salient samples. Each token stores amplitude and nine-sample local RMS.
EXO-200	Region	504	$150/6$	Six sensor blocks are divided into seven channel regions and twelve time regions. Each token retains up to 6×25 raw waveform values; smaller channel regions are padded only within the token feature.
EXO-200	Summary	504	$4/6$	The same sensor–temporal regions are described by mean, RMS, maximum absolute amplitude, and the difference between the second- and first-half means.
SuperNEMO	Entity	≤ 512	$4/3$	Individual tracker hits form tokens under a deterministic event-level cap, with centered three-dimensional detector coordinates.
SuperNEMO	Region	≤ 512	$8/3$	Hits are aggregated into occupied cubic spatial regions with width $88\,\mathrm{mm}$. Occupancy-based selection is used if an event exceeds the token cap.
SuperNEMO	Summary	≤ 16	$8/3$	Hits are Morton ordered and divided into at most sixteen balanced spatial groups, each represented by aggregated event statistics.