

Neutrino Physics and Machine Learning: Open Data Challenge Community Report

Manoj Adhikari¹, Saul Alonso-Monsalve², Omar Alterkait^{3,4}, Andrew Atta⁵, Muhammad Bilal Azam⁶, Marta Babicz⁷, Rishi Babu⁸, David Caratelli⁹, Bear Carlson¹⁰, Seokju Chung¹¹, Xuefeng Ding¹², Zelimir Djurcic¹³, Max Asa Albert Dornfest¹⁴, Francois Drielsma¹⁵, M. A. DuVernois¹⁶, Castaly Fan¹⁰, Terrance Figy¹⁷, Peter Gaemers¹⁵, Gabor Galgoczi¹⁸, K. M. Graczyk¹⁹, Pierre Granger²⁰, Gajendra Gurung^{21,20}, Nathan Higginbotham²², Aaron Higuera²³, Roger G. Huang²⁴, César Jesús-Valls²⁰, Daeun Jung²⁵, Christina Karagianni^{26,27}, Lukas Koch²⁸, Jiaoyang Li (J.-Y. Li)²⁹, Jiaxi Liu³⁰, Nived Puthumana Meleppattu³¹, Jun Nakane³², Subrahmanya Saicharan Pemmaraju³³, Nick W. Prouse³⁴, Geting Qin³⁵, Benjamin Quilain³⁶, Masooma Sarfraz³⁷, Hasung Song³⁸, Mohammed Sultan³⁹, Kazuhiro Terao¹⁵, Matt Toups²⁹, Ka Vang Tsang¹⁵, Hilary Utaegbulam⁴⁰, Ziao Wang⁴¹, Callum Wilkinson²⁴, Dowling Wong⁴², Junjie Xia¹⁵, Yikai Xu⁴³, Yongheng Xu⁴⁴, Samuel Young⁴⁵, and Larry Zhao³⁰

¹University of South Dakota

²ETH Zurich

³Tufts University

⁴The NSF AI Institute for Fundamental Interactions, MIT

⁵Monash University

⁶Illinois Institute of Technology

⁷University of Zurich

⁸Michigan State University

⁹University of California Santa Barbara

¹⁰University of Florida

¹¹Columbia University

¹²Institute of High Energy Physics, Chinese Academy of Sciences

¹³Argonne National Laboratory

¹⁴Department of Physics and Astronomy, University of Hawai'i at Mānoa, Honolulu, Hawai'i

¹⁵SLAC National Accelerator Laboratory

¹⁶University of Wisconsin-Madison

¹⁷Wichita State University

¹⁸Brookhaven National Laboratory

¹⁹Institute of Theoretical Physics, University of Wrocław

²⁰European Organization for Nuclear Research (CERN)

²¹University of Texas Arlington

²²University College London

²³Rice University

²⁴Lawrence Berkeley National Laboratory

²⁵Chonnam National University

²⁶University of Ferrara

²⁷INFN Ferrara
²⁸Johannes Gutenberg University Mainz
²⁹Fermi National Accelerator Laboratory
³⁰University of California Irvine
³¹IJCLab, Université Paris Saclay
³²Tohoku University
³³The University of Melbourne
³⁴Imperial College London
³⁵Duke University
³⁶ILANCE (CNRS-IN2P3/The University of Tokyo)
³⁷University of Delaware
³⁸University of California, Berkeley
³⁹University of Manchester
⁴⁰University of Rochester
⁴¹University of Science and Technology of China
⁴²Karlsruhe Institute of Technology
⁴³Kavli IPMU, University of Tokyo
⁴⁴University of Oslo
⁴⁵Stanford University

September 2026

Contents

1	Introduction	2
2	NPML Workshop Format	4
3	Expectations for Open Data Challenges	6
3.1	Definition of Open Data Challenges	6
3.2	Required Resources	7
4	Criteria for Judging Open Data Challenge Proposals	7
5	Near-Term Proposals	8
6	Mid-Term Proposals	10
7	Long-Term Proposals	13
8	Conclusions	15

1 Introduction

As neutrino experiments attempt to probe the finer aspects of neutrino behavior, they also face greater challenges in how they treat their data and extract results. These challenges cover aspects

of both simulation and data treatment, as well as all stages of the analysis chain, including signal processing, detector modeling, reconstruction, and high-level analysis. While no two experiments are exactly the same, many of these challenges are similar across experiments and would benefit from shared effort. In recent years, many of these problems in neutrino physics have been found to be well-suited for machine learning (ML) based solutions, which often have broad applications and require only minor modifications to tune to each specific use case.

In recognition of the growing interest in this field, the Neutrino Physics and Machine Learning (NPML) conference series was founded to provide a forum to share developments in ML techniques that are of interest to neutrino physicists. While this conference has grown alongside interest in the field, ML research across neutrino physics remains somewhat fragmented, with each experiment generally developing ad hoc solutions to their own problems and often repeating effort in foundational work.

This problem of coordinating research into new methods is not isolated to the field of neutrino physics. The benefits of open data, made freely and easily available to anyone who wishes to use them, are now widely acknowledged [1, 2, 3]. Some of these benefits comes from improving reproducibility of results, which is clearly desirable for a field of research. In addition, by generally improving community access, we can better establish community standards and best practices, as well as facilitate the creation of centralized knowledge bases [4, 5]. Newer students and researchers benefit from having accessible data with which to run methodological tutorials. Researchers developing new methods in the field benefit from having a common dataset to benchmark their techniques on for comparison against previous work, as well as standard references for what others have already attempted and the results they achieved. People outside the field benefit from having data representing a real problem of interest for them to explore and test their own work on. This simultaneous utility for training, research, and outreach is most pronounced when open datasets are associated with some *open data challenge*, where anyone is free to try to solve some specific problem associated with that open data.

There is already an established record of successful open data challenges in particle physics and astrophysics being run over the last decade. Many of these have centered on LHC-based experiments [6], which have particular computational challenges that are well-suited for this kind of approach, often featuring high-statistics, well-labeled, high-dimensional simulated data. These have included challenges for identification of Higgs events over some background [7], reconstruction of tracks from point-like hits [8, 9], anomaly detection [10], and fast calorimeter simulation [11]. The FAIR Universe initiative [12] has attempted to create a centralized platform for hosting scientific data and associated data challenges, including challenges on uncertainty-aware inference for Higgs physics [13, 14] and uncertainty-aware extraction of weak gravitational lensing maps from cosmological data [15].

Within neutrino physics, experiments like MicroBooNE, MINERvA, and T2K have released their data for general public use over the years, but without any particular challenge associated with them [16, 17]. Several neutrinoless double beta decay experiments, including CUORE, KamLAND-ZEN, MAJORANA, nEXO, NEXT, Project 8, and SuperNEMO, have recently released sets of open data for the purpose of facilitating tutorials, workshops, and other tests [18, 19, 20, 21, 22]. Recent years have also seen successful examples of explicit open data challenges in neutrino physics. The IceCube collaboration hosted a challenge focused on reconstruction of the direction of neutrino events, featuring cash prizes and attracting several hundred entries [23]. Tsinghua University has hosted a series of student-focused competitions around problems of interest to the JUNO experiment, which have run annually since 2019 [24].

The 4th NPML conference [25] hosted a satellite workshop to discuss the prospects for creating more open datasets and hosting a continuing series of open data challenges to build on this

momentum and address the gaps in neutrino ML research. Around 20 people participated in this workshop, including graduate students, postdoctoral researchers, faculty, and staff scientists. They represented a variety of backgrounds in neutrino physics, including (but not limited to) accelerator-based experiments, reactor-based experiments, astrophysical observatories, direct neutrino mass measurement experiments, and neutrinoless double beta decay searches. This coverage allowed for particular consideration of how to create open data challenges that solve problems that are not specific to just one experiment, as well as how to ensure they benefit people across a range of career stages. The workshop included discussion of how these challenges should be structured in order to provide maximal community benefit, what resources they would need, and how proposals for them should be evaluated to determine whether they are worth dedicating shared resources to. Lastly, all attendees were invited to draft concrete proposals for open data challenges across a range of research areas, which were then subjected to evaluation under these criteria during the workshop.

This white paper provides a summary of the outcomes of this NPML satellite workshop, which we hope will provide a foundation for open data challenges in neutrino physics going forward. Section 2 provides a summary of the format of the NPML workshop, and the rest of this white paper summarizes the points discussed and agreed upon during this workshop. Section 3 describes the resulting vision for future open data challenges and resources that are expected to be needed for them. Section 4 lays out the criteria developed from the workshop discussion that were used to evaluate data challenge proposals from participants, following this vision. Applying these criteria, section 5 summarizes the proposals we judged to be viable in the near-term with minimal additional preparation, section 6 summarizes the ones that seemed viable in the mid-term with a moderate amount of further preparation, and section 7 summarizes the ones that require significant additional preparation.

2 NPML Workshop Format

The NPML open data challenge workshop was organized as an “unconference”, where the agenda is driven by alternating small-group and large-group discussions instead of pre-organized talks and presentations. In this format, all participants are asked to contribute to discussions, with no passive listeners. We started each session by introducing the topic we wanted to address and briefly discussing among the entire group what questions we want to cover within that session. Everyone was then randomly assigned to a group (different for each session, so that people do not end up forming teams), with each group consisting of 3–5 people. Each group discussed the topic among themselves for 45 minutes, guided by “quad charts” that outlined the questions we wanted answered. These quad charts served as single-slide summaries of the outcome of each group’s discussion, which were then presented to the entire workshop for discussion among the larger group. An example of one of these quad charts is shown in figure 1. Groups were asked to keep the presentation of their summaries to within a couple minutes, but we did not impose time limits on large-group discussion, so that each topic could be fully explored.

Discussion sessions were broadly divided into five topics. The first was on big-picture goals and design best-practices for open data challenges, with the questions presented to participants being:

1. What are the interesting research challenges for AI/ML methods development?
2. What kinds of datasets do we need? How can we obtain and maintain them?
3. How shall we work across themes (e.g. science, detector types, AI methods)?
4. How can we lower barriers to entry and incentivize participation?

Scientific Challenges	What Datasets? How to obtain/maintain?
- Providing sensible metrics/labels so the models are trying to learn interesting/useful/relevant information - Made more difficult by application of same models to multiple detectors? - How to break out of the neutrino bubble? - Getting people to compare against the same data (avoid space race mentality), even within the same experiment - Interpretability and robustness to baked in physics assumptions – uncertainty quantification	- Generic open-source datasets – how to ensure these are sophisticated enough? - How to ensure similar across detector types? - Experiments may also want to have “internal” datasets for internal benchmarking - Finding shared resources everyone can access is challenging within a collaboration, also tough outside
Scopes (science/detectors/AI methods)	Increase Participation
- Finding complementarity between research subfields - But maybe also good to not force it, and recognize that sometimes factorization is sensible - Is it possible to find a level of abstraction for comparing multiple detector types? Or is that just FMs?	- Providing data in a friendly format: flat ROOT trees might be okay, but custom ROOT classes a no-no - Documentation of standards – consistent choices between open datasets - If a challenge is so large that you need to download and process millions of large events, it might be difficult both from a local storage and processing point of view

Figure 1: Example of a quad chart that one group filled out in response to questions on the first discussion topic, focused on big-picture goals of open data challenges and how to design them. Each quadrant answers one of the four focus questions presented to participants.

The second, third, and fourth sessions were on challenges in neutrino physics associated with reconstruction, physics modeling, and high-level analysis respectively. For each one, groups would brainstorm specific prospective open data challenges within the theme of the session, answering questions on:

1. What are the science goals and current challenges associated with this problem?
2. What datasets would be needed to study it?
3. What are tasks and metrics that would be presented to people interested in tackling this problem?
4. How would we run this challenge (e.g. preparation, release, evaluation, duration, etc.)?

Lastly, the fifth topic was on the structure and technical details of running open data challenges. Groups addressed the questions:

1. How can we successfully engage different communities, such as experimentalists and AI/ML experts, with our challenges?
2. Where can effort be shared in these challenges, such as with dataset preparation, needed infrastructure, and the actual solutions?
3. How shall we prioritize the datasets and challenges?
4. What do we need for NPML to successfully organize this?

The results of the discussion of the first and final topics are summarized in sections 3 and 4. These cover the envisioned nature and purpose of open data challenges, which are tied to the

resources needed for running them effectively and to the criteria by which they can be prioritized. The results of the other discussion sessions are summarized in sections 5, 6, and 7, covering the specific open data challenge proposals that were proposed.

3 Expectations for Open Data Challenges

3.1 Definition of Open Data Challenges

We define an open data challenge as any setup that contains both:

1. A set of open data, freely available to anyone who wishes to use it.
2. Standardized performance metrics on that data, which define a problem of interest and provide a means to compare different solutions. Such metrics should generally be accompanied by benchmark values to indicate what performance levels are currently achievable or are otherwise considered “good”.

While datasets only satisfying condition (1) are certainly useful [1, 2], it is the addition of (2) that brings much of the direct benefit to the research community. Without a shared set of performance metrics, people who are not already familiar with the data in question are unlikely to engage with it. Having these metrics allows more people to understand the problem associated with the dataset and the quality of existing solutions, even without a background in the field. This brings in a greater variety of expertise and increases the probability of obtaining innovative solutions.

In the context of neutrino physics, the open dataset can be either simulated data or actual observed data from an experiment, which both bring their own benefits. Real data clearly provides the most realistic test environment, but it can also introduce additional complications that distract from the primary question of interest. For instance, it can be more difficult to create clear performance metrics, which often require proper labels, compared to simulated data. In addition, many open problems in the field are associated with upcoming experiments, where real data is not yet available and we have no choice but to work with simulated data.

Entrants to an open data challenge should be asked to publicly provide the code they used to create their solution, including data handling and processing, model architecture, model training, and the final weights for their trained model. Winning entries would also be invited to contribute papers with more detailed descriptions of their approaches. All of this material from the entries would be compiled in a publicly and easily accessible manner over the course of the open data challenge. This way, in addition to establishing the current state-of-the-art solutions for future researchers to build off of, this material can be used for teaching and training students or newer researchers. The creation of this corpus of knowledge and tools should be considered one of our primary goals in creating open data challenges related to neutrino physics. While we would not be able to enforce requiring public releases of solutions as part of any entry, we can enforce them as part of the requirement to claim a winning prize, and we believe offering at least a nominal prize is worthwhile to further incentivize serious participation in any case. In such cases where prizes are being offered, we expect many entrants will want to temporarily maintain some degree of competitive secrecy, but this is easily addressed by postponing public release of the details of their solutions to after the winners of the challenge have been determined.

We note a few salient features of past open data challenges that should be considered when deciding on the details of future ones in neutrino physics. Many past challenges were hosted on well-known platforms like Kaggle and CodaLab, both to make use of existing infrastructure and

to attract broader participation. Some of them featured monetary prizes, with the ones involving monetary prizes generally attracting much greater interest. Some of them were not designed as formal competitions, so that while there were still metrics for evaluating performance, there was not necessarily a clearly defined winner. The vast majority of them made use of only simulated data, which allowed for easier definitions of target metrics for evaluating solutions. In all cases, details of the leading solutions were published, allowing researchers to understand and continue building off of them in the future.

3.2 Required Resources

Besides the resources needed for actually generating the data that would be part of an open data release, open data challenges will need:

1. Servers to store the open data and allow anyone to download it for themselves.
2. Infrastructure for receiving submissions. Platforms like Kaggle and Codalab already exist for this, so in many cases this would mostly entail arranging the challenge into a format compatible with them.
3. Documentation on the data and the nature of the challenge. Wherever this is hosted, it can also be used to host papers and links on the submissions and winning solutions after the challenge is completed.
4. Computational infrastructure for evaluating submissions. For some challenges, the needed computing power for this may be negligible. However, we expect it may be fairly significant in other cases, depending on the nature of the dataset and metrics in use.

In an ideal world, a challenge would also make some amount of computational resources available to participants for training their models, which would promote access by those without easy access to powerful computing clusters. However, we recognize there are many practical difficulties with this, and so leave this as a bonus to be considered instead of a requirement.

We decided to take the demands on these resources into consideration when evaluating the viability of different open data challenge proposals, as these practical matters affect the ease with which such a challenge can be prepared. For the first challenges, we expect to handle submissions with a platform similar to that used by the FAIR Universe challenges [12], which use CodaBench as an interface to computing resources provided by the NERSC cluster at Lawrence Berkeley National Laboratory [26]. This cluster can also serve as the data storage and distribution platform. These first challenges would be selected to not have severe demands on the dataset size and computational resources needed for evaluation; later challenges can aim to scale up on these demands.

4 Criteria for Judging Open Data Challenge Proposals

Workshop participants agreed on a few common criteria that can be used to judge the viability of various open data challenge proposals. For each of them, we consider the following questions:

1. Does the challenge target a problem that is of broad interest to neutrino physicists, and possibly even other physicists and non-physicists? Ideal challenges should benefit more than just a single experiment, whether this in terms of the physics impact or in terms of the anticipated technical developments needed for good solutions. This question determines whether a particular data challenge justifies calling for broader community engagement.

2. Do tools already exist to generate the required open dataset, and could the dataset be used for other challenges or purposes as well? This is a practical concern, aimed at maximizing efficiency of effort needed by topical experts to create useful open datasets.
3. Are the metrics that would be used to evaluate solutions clearly defined? Problems where success is still subjective are interesting, but belong more to the realm of exploratory research. For an open data challenge, concrete metrics are needed to allow participants who are not neutrino physicists to meaningfully participate.
4. What computational resources would be needed to support the challenge? This question includes capacity to host and distribute the open dataset, as well as any resources that may be needed to evaluate solutions.

Based on these criteria, each data challenge proposal presented by groups during the NPML workshop was sorted into one of three categories: near-term, mid-term, or long-term proposals. This classification combines questions of resource prioritization, proposal maturity, and expected impact. For instance, challenges where the performance metrics are not clear or where the tools for generating the required dataset are not yet ready get pushed to the mid-term or long-term categories, even if they address an important open problem. We emphasize that in this evaluation, we are not attempting to judge what kinds of research questions are interesting. We are only evaluating which problems seem most well-suited for the format of an open data challenge in the near future. In the following sections, we detail the results of these evaluations.

5 Near-Term Proposals

Near-term proposals are considered feasible to organize in the immediate future, as some of the first open data challenges that the field attempts. They should already have very concrete requirements and goals, such that we have clear ideas of what a good solution should achieve and how it would be applied. Creating an open data challenge centered on one of these proposals should require primarily organizational work, rather than exploratory research. Challenges in this category should also not require significant new infrastructure, which would otherwise need additional development time from the organizers. Lastly, near-term proposals should be in areas where there has already been demonstrated broad interest from the community in applying any novel promising solutions, in order to justify prioritizing them above other possible challenges.

Anomaly Detection in Offline Data This challenge would require the identification of signals from one or more novel processes in an otherwise normal dataset, without specifics of the nature of the signals. These novel signals could be, but would not necessarily be, Beyond the Standard Model (BSM) processes. There is demonstrated broad interest in this topic from the community, and we expect novel technical solutions could be applicable across many types of detectors.

- **Dataset:** Existing event generators and detector simulations would suffice, and the required level of detail is somewhat known from existing BSM studies already. Some exploration would be needed to determine whether low-level detector data or high-level reconstructed quantities are more appropriate for this challenge.
- **Metrics:** Efficiency and purity metrics are well-established for searches for new processes.
- **Other Notes:** A similar data challenge has already been held with the LHC Olympics [10]. Using the resources and experience from that challenge could lower the barrier to setting up this challenge for neutrino physics.

Particle-Level Instance Segmentation This focuses on the low-level task of segmentation of detector-level hits, the results of which are used as input for many types of reconstruction algorithms. Improving performance on this task would improve the downstream performance of many other aspects of reconstruction as well.

- **Dataset:** Labeled simulated datasets suited for this challenge already exist, having been created to train the current generation of reconstruction algorithms. These datasets are generally specific to a particular detector technology, but do not have to be specific to a particular experiment.
- **Metrics:** Purity and Adjusted Rand Index (ARI) metrics are well-established metrics for evaluating current solutions to this task.
- **Other Notes:** This challenge seems particularly well-suited for attracting outside experts, given the problem requires little knowledge of the associated physics and has straightforward metrics.

Interaction Topology and Particle Identification in Water Cherenkov Detectors Water-based detectors that only use Cherenkov light often define “single-ring” and “multi-ring” event topologies. They also often face particular difficulties in distinguishing between $e/\gamma/\pi^0$ interactions in their detectors. The ability to distinguish between these is important for the sensitivity of neutrino oscillation experiments using Water Cherenkov detectors.

- **Dataset:** Labeled datasets containing the detector-level hits and simple truth-level information already exist for studies of these problems with Super-Kamiokande and Hyper-Kamiokande.
- **Metrics:** For particle identification, the classification accuracy as a function of energy, direction, and position in the detector is a commonly used metric already. For ring counting, solutions just have to guess the correct number of rings, but weights for how much to penalize the degree of undercounting or overcounting need to be tuned.
- **Other Notes:** Although this challenge targets a specific detector design, the underlying technical problems do not require unreasonably deep domain expertise to be approachable and seem well-suited for an open data challenge.

Ghost Hit Suppression in Liquid Argon Time Projection Chambers Wire and strip-based liquid argon time projection chambers combine multiple 2D readouts to create a 3D projection of interactions in the detector. However, the ambiguity from combining 2D projections can result in “ghost hits” in 3D space, where no interaction actually occurred. Removing these ghost hits is important to downstream reconstruction tasks in these detectors.

- **Dataset:** Labeled datasets containing the basic detector-level hits and corresponding truth-level information needed for this challenge already exist, as they are used for reconstruction studies in many existing liquid argon time projection chamber experiments.
- **Metrics:** Standard metrics for efficiency and purity in removing ghost hits and not removing true hits already exist for reconstruction studies.
- **Other Notes:** Although this challenge targets a specific detector design, the underlying technical problem does not require domain expertise and seems well-suited for an open data challenge.

Hit/Charge Reconstruction from Waveforms Reconstructing hits, charge, or other high-level quantities from waveforms from detectors' various sensors is the first step of any reconstruction chain. Improvements here can help all other downstream algorithms as well.

- **Dataset:** This requires simulation of raw detector sensor responses, along with the truth-level information that generated those responses. While these simulations already exist, the size of the dataset that would need to be released for this challenge may be large, since uncompressed waveforms require a lot of storage space.
- **Metrics:** The accuracy and resolution of the reconstructed quantities are standard metrics used to evaluate this type of reconstruction. Conventional techniques also already exist for all of these problems and perform quite well, so they can serve as benchmarks.
- **Other Notes:** While the specific types of waveforms and high-level quantities that need to be reconstructed will depend on a choice of detector, the general problem is applicable to many types of detectors, and so is still suitable for an open data challenge.

Fast Simulation with a Surrogate Model This would require participants to develop a model that can closely replicate the results of a detector simulation while having substantially reduced computational requirements. This challenge could also be modified to require a differentiable model in particular.

- **Dataset:** Existing detector simulations would be sufficient. These should include the particle-level deposits and the resulting detector-level quantities. The test dataset would provide only the former and ask submissions to predict the latter.
- **Metrics:** Distance metrics for determining the fidelity of the surrogate models exist. It is more unclear how to incorporate computational efficiency into the evaluations. One possibility is to simply impose a maximum runtime on all submissions, but this requires further exploration.
- **Other Notes:** If computational efficiency is part of the requirement, the organizer must provide computational resources for evaluating submissions, which will also raise questions of optimizing solutions for specific hardware. Experience from the LHC fast calorimeter simulation challenge [11] may be helpful here.

6 Mid-Term Proposals

Mid-term proposals are assessed as requiring more thought or work in one or more of the categories described in section 4, but where the path to fully developing the proposal seems clear and only a moderate amount of additional work may be needed. These proposals would probably be realistic options for one of the later data challenges, after some organizational experience has been accumulated.

Supernova Pointing This would ask participants to use low-level or raw detector data corresponding to detection of a nearby core-collapse supernova to reconstruct the direction of the supernova, within some latency constraint. This is widely acknowledged as an important capability of large neutrino detectors as part of a global Supernova Early Warning System (SNEWS) [27]. Different collaborations have conducted their own studies on their detectors' capabilities, but this

is a topic where people outside of those collaborations would benefit from improving their pointing resolution as well.

- **Dataset:** Packages for simulating neutrinos from core-collapse supernovae already exist, but the question of how to structure the open data release requires thought. Raw detector data from a single supernova tends to be extremely large in size, and solutions need to obey certain latency and hardware constraints in order to be practically useful for real experiments. It is not clear whether the open data should simply be many simulated supernova bursts. These details must be ironed out before an open data challenge can be prepared.
- **Metrics:** The resulting supernova pointing resolution is a clear metric of interest for each solution, but their latency and resource consumption should also be taken into consideration. It is not yet clear how to include these latter metrics in the overall evaluation.
- **Other Notes:** Depending on how the structure ends up being structured, it may have nontrivial storage and computational needs from the organizers. In addition to simulated data, it could be interesting to consider making the actual neutrino data from supernova 1987a part of the open data release.

Supernova Data Compression Instead of performing supernova pointing directly from the raw detector data, this challenge would ask participants to compress the raw data while retaining enough information to perform effective pointing. This is of particular interest for underground experiments with heavy computational and bandwidth constraints at their detector site. If supernova data could be efficiently compressed, they could be quickly sent out to remote clusters, where the initial pointing could be done by more computationally intensive advanced algorithms.

- **Dataset:** Raw detector data from a supernova burst would be part of this dataset, like in the previous challenge. However, the storage demands would probably be even higher, since this challenge would not pre-compress any of the raw detector data.
- **Metrics:** Evaluation would be some combination of data compression efficiency and the supernova pointing resolution that is possible with the resulting compressed data. However, it is not obvious how to balance these two in the evaluation.
- **Other Notes:** Since we need to know the pointing resolution that is possible with the compressed data, this challenge would probably also require participants to submit a pointing algorithm intended for use with their compression approach. This pointing algorithm would not be subject to particular hardware or runtime constraints, but the organizers must have a way to verify that its output comes from the compressed data, and not directly from the raw test dataset.

Online Triggering This challenge seeks to improve the ability of live online trigger systems to identify events of interest. The “events of interest” can be defined flexibly for this challenge, but should be targeting something that is difficult for existing conventional triggers. Solutions will have to be subject to some kind of latency and compute-power constraint, in order to simulate realistic experimental conditions.

- **Dataset:** Like the previous challenges, simulated raw detector data are needed as inputs to the trigger algorithm, which brings question of storage capacity. Some effort could be shared with a challenge centered around anomaly detection, if the anomalies are suitable as signal events for testing the trigger system.

- **Metrics:** Standard metrics for evaluating conventional trigger systems already exist, usually considering the true signal efficiency for a given total trigger rate. The latency and compute requirements of solutions do not have to be part of the evaluation metric, so long as they fall under the given constraints.

- **Other Notes:** Since this challenge is only interesting if latency and compute restrictions can be imposed on solutions, the challenge organizers will probably have to run a standard hardware setup. Submissions would have to be configured so that they can all run on this setup.

Multi-task Regression of Reconstructed Quantities From raw detector hits, reconstruction algorithms typically build up higher-level objects and then obtain estimates of various quantities of interest. For instance, for a single identified particle track, this could include estimating its origin vertex, its particle ID, its energy, and its direction. This challenge would ask submissions to simultaneously predict all of these quantities, instead of doing them separately with different models, as is done in most current reconstruction chains.

- **Dataset:** Some enhancement of the simulated detector data currently used for training reconstruction algorithms would be needed, due to a need for more detailed per-particle and per-event labels. The dataset should also be designed to contain events with sufficient complexity for multi-task regression to be an interesting enough challenge over predicting each quantity one at a time.

- **Metrics:** The evaluation would be some combination of accuracy on all the quantities that are being regressed, but it is not obvious how to weight them against each other. A combined metric incorporating all of them will be needed to compare solutions that are not strictly better or worse than each other.

- **Other Notes:** Evaluation of solutions is very cheap, as we only have to compare lists of predicted versus true quantities in the test dataset.

Multi-modal Detector Reconstruction Many detectors consist of multiple components or types of sensors specialized for different purposes (for instance, having separate charge and light readout systems in a liquid argon time projection chamber). Reconstruction techniques can treat each detector mode individually and then combine them afterwards, but treating them all together from the beginning can yield better results.

- **Dataset:** This requires a combination of simulated data from each component of the detector, along with information on how the data from the different components should be aligned or relate to each other. These simulations and information all already exist in principle, but some work would be required to release them together in a format that is intelligible to non-experts in the detector design.

- **Metrics:** Overall accuracy and resolution on an event could be evaluated, but some investigation is needed to determine how well this gauges solutions' abilities to combine information from multiple detector modes.

- **Other Notes:** This challenge would likely initially pick some specific existing detector design, but afterwards, the dataset design and metrics could likely be generalized to be usable for other detectors as well.

Reconstruction Robustness against Detector Variations Most detector simulations rely on several tunable parameters in order to match the responses seen in measured data. These tend to represent various physical conditions that are not known *a priori*, such that they must be extracted from the actual detector data. This tuning is often not perfect, and so reconstruction algorithms should ideally be robust against small changes in these parameters.

- **Dataset:** This challenge would reuse datasets from typical reconstruction-focused challenges, but there would need to be multiple additional copies of the datasets with different types of detector variations applied. The test datasets would then use other unseen detector variations, but which should not be entirely out of distribution compared to the provided training sets.
- **Metrics:** The typical reconstruction performance metrics can still be used, but we may wish to evaluate them for multiple datasets with different degrees and types of detector variation.
- **Other Notes:** The main question for this challenge will be determining the types of detector variations that are reasonable to include in the training and test datasets.

7 Long-Term Proposals

Long-term proposals are those that require significant additional thought in multiple areas before we can realistically consider setting them up as data challenges. For some of these proposals, it is not clear they are well-suited for the format of an open data challenge, even if they are attempting to address an important research question. Nonetheless, we include everything discussed during the workshop for completeness.

Neutrino Oscillation Analysis with Uncertainty Estimates This would consist of performing a typical neutrino oscillation analysis given certain experimental inputs, with proper incorporation and propagation of all relevant uncertainties. While this is clearly of interest as one of the primary goals of neutrino oscillation experiments, it is not yet known exactly what the uncertainties will look like.

- **Dataset:** The tools to simulate the neutrino oscillation data that would be used for this challenge do not exist yet, but we expect they will be developed in due course by upcoming neutrino oscillation experiments.
- **Metrics:** Typical statistical metrics of bias and coverage will be sufficient.
- **Other Notes:** While having an analysis method that satisfies our bias and coverage demands is clearly useful, it is not yet clear that the analysis method itself will be a limiting factor in next-generation neutrino oscillation analyses. As such, this challenge would be more worth revisiting after those experiments have more mature analyses.

Learning Interpretable Maximally Separating Observables For some particular sets of signal and background processes, some choices of observables do better jobs of separating them than others. Classical analyses hand-choose observables to optimize an analysis. ML techniques combine high-dimensional information to improve this separation, but typically at the cost of interpretability. A technique that maximizes signal/background separation with a physically interpretable output combines the best of both worlds and could yield more generalizable insight to the problem.

- **Dataset:** Typical detector-level simulations used for regular analyses would probably suffice. Some thought would have to be given to whether low-level quantities should be given in addition to high-level observables, to allow for more flexibility in types of solutions.
- **Metrics:** Standard signal-to-background metrics are enough for the quality of the learned observables. However, the key requirement is determining their interpretability, and there is no clear metric for this. Without it, the output of a standard ML classifier would already be a solution.
- **Other Notes:** This could be tied into a bigger challenge around optimizing analysis chains in general, which would likely face similar problems in evaluating solutions.

Online Data Compression Many detectors save waveforms from their sensors that contain significant periods of time where not much interesting is happening. However, these samples do usually carry some information about baselines and noise, which are needed to properly analyze signal hits when they arrive. Finding a way to compress these waveforms with minimal loss of information would free up storage and reduce data transfer latency.

- **Dataset:** Both simulated and actual measured waveforms would be good to include in this challenge, since in this case the truth label is simply the original waveform. No particular cleaning or preparation would be needed for them either, if the goal is to be able to handle arbitrary waveforms.
- **Metrics:** There exist general metrics for compression efficiency, but using those would just reduce this to a computer science problem. Some metric that measures the loss from using the compressed waveforms instead of the full waveforms in an actual analysis would be ideal, but this requires further work to develop.
- **Other Notes:** Most experiments already have methods of zero suppression or otherwise being efficient in how they store information from their detectors. There should be some exploration of what kind of improvement is realistic, and how much this would help those experiments.

Real-time Detector Monitoring Given data from all available sensors in a detector, search for anomalies that indicate degraded detector conditions, events of interest, or anything else that is worth human intervention. This would reduce reliance on human monitoring for ensuring detectors are operating well.

- **Dataset:** The dataset would ideally be a near-full set of monitoring data from an actual detector, which would include any unusual conditions that human detector shifters already flagged. However, there are likely significant non-technical barriers to making this kind of data available.
- **Metrics:** Being able to reproduce human-flagged periods of detector instability would be a requirement, but it is unclear how to evaluate anything novel that submissions flag in the dataset.
- **Other Notes:** The type of data requested for this challenge is generally not made publicly available, as it includes run conditions databases, logs, and data from monitoring sensors.

Detector Design Optimization When developing new techniques, the detector design is often treated as immutable, simply because it has already been built or is already under construction. When looking toward future experiments, there is the potential to use ML-based techniques to explore the space of possible detector designs and better optimize the detector’s sensitivity to particular processes.

- **Dataset:** The primary content of the dataset would be a set of truth-level interactions in some near-infinite medium of interest. Solutions would propose a detector design to construct around this volume. The dataset may need to include sets of allowed detector components and sensors, along with defined responses for them, so that designs cannot simply propose unrealistically high-performing components.
- **Metrics:** Many possible metrics exist, but is unclear how to choose and combine them in an appropriate way. Detector performance metrics, such as energy resolution, spatial resolution, and particle identification capabilities, would be relevant, but many of these depend on downstream analysis tools as well. Metrics could also be aimed at specific physics processes instead of low-level detector metrics, but this has similar difficulties.
- **Other Notes:** The amount of information needed to constrain detector designs to be realistic and the number of metrics that need to be optimized make this difficult to define as an open data challenge.

AI-Assisted Literature Search Most researchers would benefit from a reliable tool for more efficiently conducting literature searches. There is a much bigger market for this kind of tool already, but tuning them for maximum effectiveness within some subfield could require more specific expertise.

- **Dataset:** This just requires the current body of available literature, which likely does not require explicit construction into a designated dataset.
- **Metrics:** There are no obvious metrics for judging the usefulness of any particular output, besides simply having humans read and evaluate them.
- **Other Notes:** It is not clear that this is well-suited for a challenge organized by neutrino physicists, given that such a tool would have very general utility.

8 Conclusions

Bottom line still to be written

References

- [1] M. Wilkinson et al., *The FAIR Guiding Principles for scientific data management and stewardship*, *Scientific Data* **3** (2016) .
- [2] X. Chen et al., *Open is not enough*, *Nature Phys* **15** (2019) 113.
- [3] A. Zuiderwijk and H. Spiers, *Sharing and re-using open data: A case study of motivations in astrophysics*, *International Journal of Information Management* **49** (2019) 228.

- [4] C. Krause, R. Winterhalder, M. Feickert, B. Nachman, J. Raine and HEP ML Community, *A living review of machine learning for particle physics*, July, 2026. 10.5281/zenodo.21626667.
- [5] C. Krause, R. Winterhalder, M. Feickert and B. Nachman, *The Living Guide of Machine Learning for Particle Physics*, 2608.09531.
- [6] D. Rousseau and A. Ustyuzhanin, *Machine Learning scientific competitions and datasets*, 2012.08520.
- [7] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kégl and D. Rousseau, *The Higgs boson machine learning challenge*, in *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, G. Cowan, C. Germain, I. Guyon, B. Kégl and D. Rousseau, eds., vol. 42 of *Proceedings of Machine Learning Research*, (Montreal, Canada), pp. 19–55, PMLR, 13 Dec, 2015, <https://proceedings.mlr.press/v42/cowa14.html>.
- [8] S. Amrouche et al., *The tracking machine learning challenge: Accuracy phase*, in *The NeurIPS '18 Competition*, S. Escalera and R. Herbrich, eds., (Cham), pp. 231–264, Springer International Publishing, 2020.
- [9] S. Amrouche et al., *The tracking machine learning challenge: Throughput phase*, *Comput. Softw. Big Sci.* **7** (2023) 1.
- [10] G. Kasieczka, B. Nachman, D. Shih, O. Amram, A. Andreassen, K. Benkendorfer et al., *The lhcb olympics 2020: A community challenge for anomaly detection in high energy physics*, *Reports on Progress in Physics* **84** (2021) 124201.
- [11] C. Krause, M. Fucci Giannelli, G. Kasieczka, B. Nachman, D. Salamani, D. Shih et al., *Calochallenge 2022: a community challenge for fast calorimeter simulation*, *Reports on Progress in Physics* **88** (2025) 116201.
- [12] “FAIR Universe.” <https://fair-universe.lbl.gov/>.
- [13] R. Chakkappai et al., *Fair Universe Higgs Uncertainty Challenge*, 2509.22247.
- [14] L. Benato et al., *Fair Universe Higgs Uncertainty Challenge*, 2410.02867.
- [15] B. Dai et al., *FAIR Universe Weak Lensing ML Uncertainty Challenge: Handling Uncertainties and Distribution Shifts for Precision Cosmology*, 2604.14451.
- [16] G. Cerati, *Microboone public data sets: A collaborative tool for lartpc software development*, *EPJ Web of Conf.* **295** (2024) 08012.
- [17] MINERvA collaboration, *MINERvA Experiment Open Data Release*, .
- [18] CUORE Collaboration, *CUORE Data Release for ML Applications: Pulse Shape Analysis Dataset*, 2607.02548.
- [19] MAJORANA Collaboration, *Majorana Demonstrator Data Release for AI/ML Applications*, 2308.10856.
- [20] G. Parker, J. Brodsky and I. Chakraborty, *Single-phase xenon tpc charge readout simulation (neutrinoless double beta decay ai summer school)*, June, 2026. 10.5281/zenodo.20776761.
- [21] K. Mistry, *NEXT Simulation Dataset for AI Summer School UC Irvine 2026*, 2605.07566.

- 552 [22] M. Macko, *SuperNEMO Dataset for Ovbb AI Summer School*, June, 2026.
553 10.5281/zenodo.20698789.
- 554 [23] IceCube Collaboration, *Public Kaggle Competition "IceCube – Neutrinos in Deep Ice"*,
555 2307.15289.
- 556 [24] "Ghost Hunter Open Data Challenge." <https://ghosthunter.thudep.com/>.
- 557 [25] *NPML 2025* <https://indico.ipmu.jp/event/462>, .
- 558 [26] National Energy Research Scientific Computing Center, "NERSC User Documentation."
559 <https://docs.nersc.gov/>, 2026.
- 560 [27] S. Al Kharusi, S.Y. BenZvi, J.S. Bobowski, W. Bonivento, V. Brdar, T. Brunner et al.,
561 *Snews 2.0: a next-generation supernova early warning system for multi-messenger*
562 *astronomy*, *New Journal of Physics* **23** (2021) 031201.