Ghost Hunter Open Data Challenge

PMT Waveform Analysis and Event Reconstruction

Shengqi Chen[a], Benda Xu[a,b], et al.

[a]Tsinghua University
[b]University of Chinese Academy of Sciences

2026-06-18 NPML @ UC Irvine

The Bottleneck in Neutrino ML

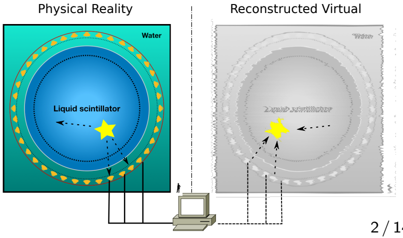
- Machine Learning is reshaping event reconstruction in next-generation neutrino experiments

Gap Between the NP/ML communities

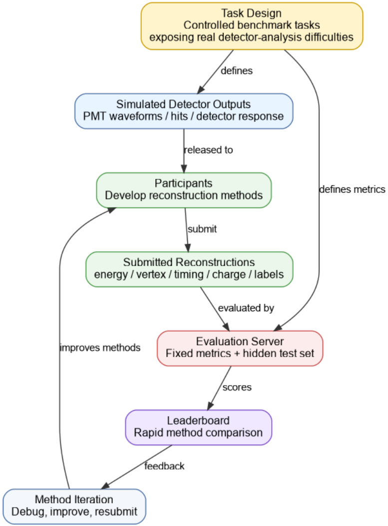
- ML researchers lack access to standardized public datasets with ground-truth labels and standardized metrics.
- Official experiment simulations are bound by collaboration data policies.
- New algorithms often lack reproducibility, overfit to flawed simulator assumptions or struggle to enter the field.

Enter Ghost Hunter as a Standard Candle

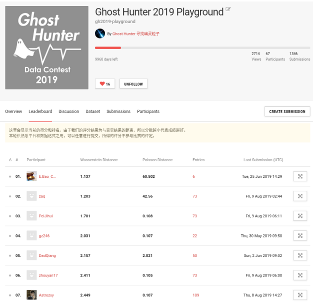
A comprehensive benchmark dataset and challenge framework designed to bridge the gap between AI developers and neutrino physics.



Participation workflow



- A CI-automated leaderboard enables rapid method comparison.
  - instant feedback in algorithm development.
- Competition workshops
  - seminar of reference algorithms.
  - idea sharing.
  - pair programming.



Seven Years of Operational History

Ghost Hunter is a battle-tested infrastructure continuously operating since 2019.
7 Editions Hosted (2019 - 2025/26) with 500+ participants and 180+ teams.

- 1st (2019) 74 participants / 43 teams, PMT waveform analysis
- 2nd (2020) 60 participants / 44 teams, Particle identification
- 3rd (2021) , JUNO-like energy reconstruction
- 4th (2022) ~80 participants / 30 teams, JUNO-like event reconstruction
- 5th (2023) ~90 participants / ~40 teams, Continued JUNO-like reconstruction
- 6th (2024) 106 participants / 36 teams, Optical detector-response studies
- 7th (2025) 75 participants / 24 teams, Continued Optical detector-response

1st 2019 PMT waveform analysis

- An ill-posed blind deconvolution problem with no universally optimal solution.
  - The topic was co-designed with Linyan Wan.
- Summary paper published at JINST 17 P06040

Journal of Instrumentation

PAPER

Towards the ultimate PMT waveform analysis for neutrino and dark matter experiments

D.C. Xu[1,2,3], B.D. Xu[1,2,3,4], E.J. Bao[5,6], Y.Y. Wu[1,2,3], A.Q. Zhang[1,2,3], Y.Y. Wang[1,2,3], G.L. Zhang[7], Y. Xu[8], Z.Y. Guo[1,2,3], J.H. Pei[9], H.Y. Mao[10], J.S. Liu[10], Z. Wang[1,2,3] and S.M. Chen[1,2,3] – Hide full author list

Published 30 June 2022 • © 2022 IOP Publishing Ltd and Sissa Medialab

Journal of Instrumentation, Volume 17, June 2022

Citation D.C. Xu *et al* 2022 *JINST* **17** P06040

Acknowledgments

We are grateful to the participants and organizers of the *Ghost Hunter 2019* online data contest. CNN, Fitting and LucyDDM in this article are developed from the ones submitted to the contest.

2nd 2020 particle identification

- α/β particle identification in liquid scintillators
  - Organized completely online due to COVID-19.

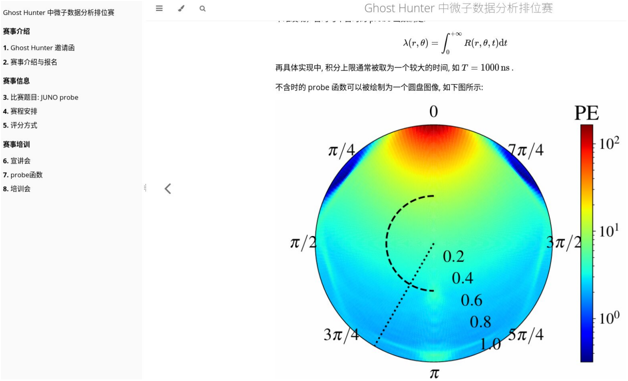
3rd 2021 JUNO energy resolution

4th 2022 JUNO energy resolution refocused

1. drop waveforms and simplify
2. drop events at the acrylic boundaries
3. most realistic detector modeling, imperfect acrylic sphere

2023–2025 JUNO-like detector response

- Portal website `https://ghosthunter.thudep.com/2025/` (Use your favorite browser translation plugin)

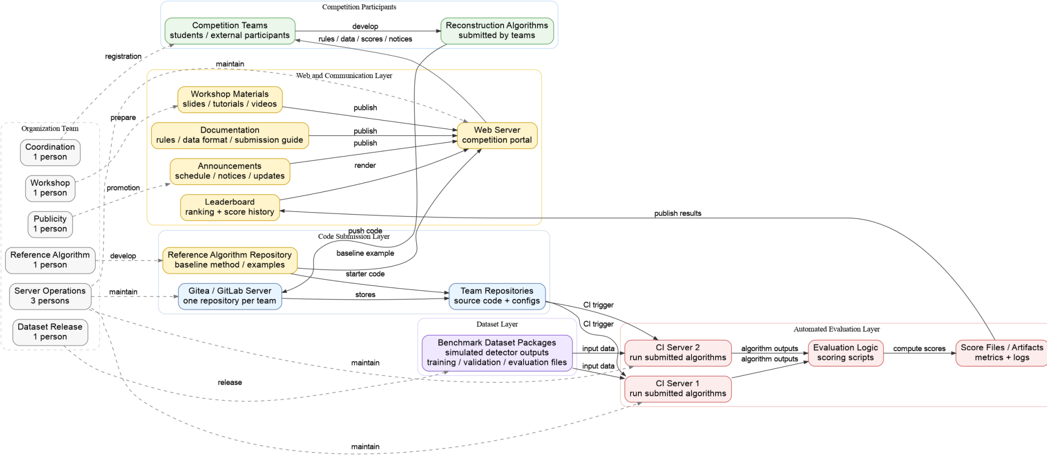


- Public repository `https://github.com/thudep/Ghost-Hunter-JUNO-probe-2025`

Ghost Hunter Open Data Challenge

Shengqi Chen[a], Benda Xu[a,b], et al.

Infrastructure and Organization Readiness



- A team of 8 people hosts the contests with reproducible infrastructure.

Physics-Motivated Metrics and Dataset Anatomy

Physics-motivated metrics enforce meaningful loss functions and evaluation criteria, bridging CS algorithms with detector reality.
We standardize model inputs, outputs, and reproducible command-line interfaces.

Example Metrics

Optimal Transport Wasserstein Distance for discrete time-profile comparison.

Likelihoods Unbinned time-dependent Poisson likelihoods for photo-electron time series.

A history of producing useful reconstruction ideas

- Ghost Hunter crowd-sources diverse methods and AI-agent workflows.

Tiered data governance model

Public challenges can precede collaboration-approved datasets.

Tier	Scope	Public?	Purpose
0	Standalone toy benchmark	Yes	outreach, ML prototyping, leaderboard
1	Physics-inspired stress tests	Semi	domain shift and calibration robustness
2	Internal collaboration benchmark	No	official MC / internal detector validation
3	Collaboration-approved dataset	Future	sanitized public benchmark if approved

- Support collaboration-internal benchmarks for selecting reconstruction and data-production algorithms.

Tiered data governance model

Public challenges can precede collaboration-approved datasets.

Tier	Scope	Public?	Purpose
0	Standalone toy benchmark	Yes	outreach, ML prototyping, leaderboard
1	Physics-inspired stress tests	Semi	domain shift and calibration robustness
2	Internal collaboration benchmark	No	official MC / internal detector validation
3	Collaboration-approved dataset	Future	sanitized public benchmark if approved

- Support collaboration-internal benchmarks for selecting reconstruction and data-production algorithms.

Detector-agnostic expansion

Ghost Hunter started from optical scintillator-motivated problems, but the infrastructure is detector-agnostic. Possible modules from current brainstorming

- Water-Cherenkov reconstruction,
- hybrid Cherenkov/scintillation separation,
- LArTPC surrogate-response benchmarks,
- calibration-robustness challenges,

Positioning in the NPML Open-Data Ecosystem

NPML is forming an open-data benchmark ecosystem across detector technologies.

- Ghost Hunter fills a critical niche by providing mature infrastructure for recurring data challenges.

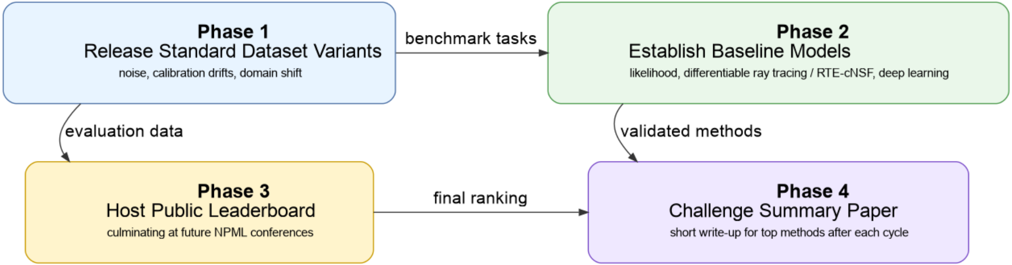
Advanced baselines under discussion

- PDE motivated models.
- Differentiable simulation models.
- Likelihood-based waveform reconstruction.
- Neural-network density estimation.
- Transformer models.
- Surrogate detector-response models.
- ...

A benchmark becomes valuable only when strong methods are evaluated on the same field.

Proposal for a Recurring NPML Open Data Challenge

We propose to develop Ghost Hunter into an annually recurring NPML Open Data Challenge.



For NPML

- recurring open-data activity,
- visible outreach and education output,
- continuity across workshops.

Summary

1. Ghost Hunter has run seven editions since 2019.
2. It has existing ranking and server infrastructure.
3. Originally built for education, it is now being upgraded into an open benchmark infrastructure for neutrino ML.

Ghost Hunter Proposal

As a recurring NPML detector-agnostic open-data challenge framework.

Scientific output beyond outreach

- Example outputs:
  - waveform-analysis algorithms,
  - deconvolution methods,
  - likelihood-based reconstruction ideas,
  - FSMP-related developments,
  - student-driven ML and AI-agent pipelines,
  - follow-up waveform-analysis publication.

Example: waveform-analysis task

- Input:
  - digitized PMT waveform,
  - noisy baseline,
  - possible pile-up,
  - single-PE response fluctuation.
- Hidden truth:
  - number of photoelectrons,
  - true PE times,
  - charge contribution.
- Output:
  - reconstructed hit times,
  - charge estimate,
  - uncertainty estimate.

A controlled version of a real detector bottleneck.

Dataset contents

Level	Example labels
Event-level	true vertex, true energy, event type
PMT-level	PMT position, channel response, waveform
Hit-level	true hit time, photon arrival time, PE count
Detector-level	geometry, noise configuration, response parameters
Stress-test level	drift, mismatch, background, domain shift

Full truth labels are the reason toy benchmarks are useful.

Candidate ML tasks

- Waveform denoising.
- Charge reconstruction.
- First-hit timing extraction.
- Multi-PE deconvolution.
- Vertex reconstruction.
- Energy reconstruction.
- Uncertainty estimation.
- Calibration-robust reconstruction.
- Fast likelihood approximation.

The same infrastructure can host multiple difficulty tiers.

Evaluation metrics

- Timing resolution.
- Charge bias and resolution.
- Vertex resolution.
- Energy resolution.
- Calibration robustness.
- Domain-shift robustness.
- Inference speed.
- Memory footprint.
- Interpretability / physics consistency.
- Likelihood quality where applicable.

Leaderboard metrics must reward generalization.

From toy problems to real detector problems

- Real detector problems cannot always be exposed through public data.
- But their structure can be abstracted into benchmark stress tests:
  - PMT gain drift,
  - time-offset drift,
  - dark-noise variation,
  - afterpulsing,
  - scintillation time-profile mismatch,
  - wavelength-dependent optical response,
  - radial nonuniformity,
  - background contamination,
  - data/MC domain shift.

The goal is to stress-test reconstruction logic.

Shengqi Chen[a], Benda Xu[a,b], et al.

Example advanced benchmark question

How should we compare physics-informed and ML-based photon-timing models?

- Does the model reproduce the full timing PDF?
- Does it preserve likelihood normalization?
- Does it extrapolate under optical-parameter shifts?
- Does it remain stable near detector boundaries?
- Does it scale to large PMT arrays?
- Does it generalize beyond nominal simulation?

This is where RTE–cNSF, LUCiD-style models, and neural surrogates can be compared.

What we need from the community

- Benchmark-design feedback.
- Metric definitions.
- Baseline model contributions.
- Documentation help.
- Server / leaderboard advice.
- Detector-technology module proposals.
- Stress-test scenarios.
- Governance input.
- Volunteers for an NPML challenge working group.

The immediate goal is a sustainable benchmark protocol.

Proposed 2026–2027 roadmap

- NPML 2026:
  - collect community feedback,
  - define scope of first NPML Ghost Hunter module,
  - identify baseline contributors.
- Late 2026:
  - release prototype toy-detector dataset,
  - publish baseline models and metrics,
  - test server-side submission.
- 2027:
  - run first NPML-associated challenge cycle,
  - produce summary paper,
  - evaluate expansion to additional detector modules.

Start small. Make it recurring. Then scale.

2019 waveform-analysis benchmark

- Task:
  - reconstruct photoelectron timing from digitized PMT waveforms.
- Challenge:
  - additive noise,
  - pile-up,
  - fluctuating single-PE response,
  - unknown number of pulses,
  - timing ambiguity.
- Evaluation:
  - compare reconstructed PE-time distribution with truth.
  - distance-based metrics such as Wasserstein distance are natural for discrete timing distributions.

This was an early example of waveform-level detector ML benchmarking.

Why Wasserstein-style metrics appeared naturally

- Waveform timing reconstruction does not always produce the same number of hits as truth.
- Simple pointwise residuals are ill-defined.
- Discrete truth distributions and continuous inferred distributions need a distribution-level distance.
- Optimal transport / Wasserstein distance provides one principled comparison.

Benchmark metrics must match the mathematical structure of the output.

avoiding simulator overfitting

A public benchmark should not reward models that only overfit a nominal simulator.

Mitigations:

- hidden detector-response shifts,
- multiple noise regimes,
- calibration-drift scenarios,
- train/test mismatch,
- uncertainty-aware scoring,
- physics-consistency checks,
- stress-test leaderboards.

The hard problem is not fitting toy MC. The hard problem is surviving detector-response mismatch.

relationship to official experimental data

- Official public datasets are essential but expensive to release and maintain.
- Toy benchmarks are faster, safer, and truth-complete.
- The two should coexist:
  - official data for realism,
  - toy benchmarks for controlled algorithm development,
  - derived datasets for intermediate validation.

Ghost Hunter is the controlled sandbox layer.

Possible working-group structure

Role	Responsibility
Physics scope leads	define tasks and detector-response stress tests
ML baseline leads	provide reference models
Infrastructure leads	server, leaderboard, submission workflow
Documentation leads	tutorials, dataset schema, examples
Governance leads	versioning, policy, publication rules