

DORAEMON T2

Neutrino Event Generator

Surrogate Models – Evaluation Metrics

Roger Huang

9/29/2026

What We Provide

Neutrino Fluxes	Generator
DUNE FD Oscillated Flux	GENIEv3_G18_10a_02_11a
DUNE FD Unoscillated Flux	GENIEv3_G18_10b_02_11a
Monoenergetic 0.6 GeV	NEUT580
Monoenergetic 2.5 GeV	NUWROv25.3.1
Monoenergetic 10 GeV	(+ other model variations)

Considering expanding the number of generators and reducing the number of fluxes

Publicly release:

- 10M (maybe more?) and 1M event files for each combination of (flux, generator) listed above
- Documentation on the file formats
- Notebook/code with a “naive” solution and our baseline solution
- Documentation (and code?) on how we will evaluate solutions

Kept hidden: 1M events for each combination of (flux, generator) as test sets

Evaluation

$$D^2(F, G) = 2 \mathbf{E} \|X - Y\| - \mathbf{E} \|X - X'\| - \mathbf{E} \|Y - Y'\|$$

Energy distance defined as above between two distributions **F** and **G**, with random vectors **X**, **X'** drawn from **F** and random vectors **Y**, **Y'** drawn from **G**

- Provides an unbinned, multidimensional metric between the distributions

We define the distance in **5 dimensions**:

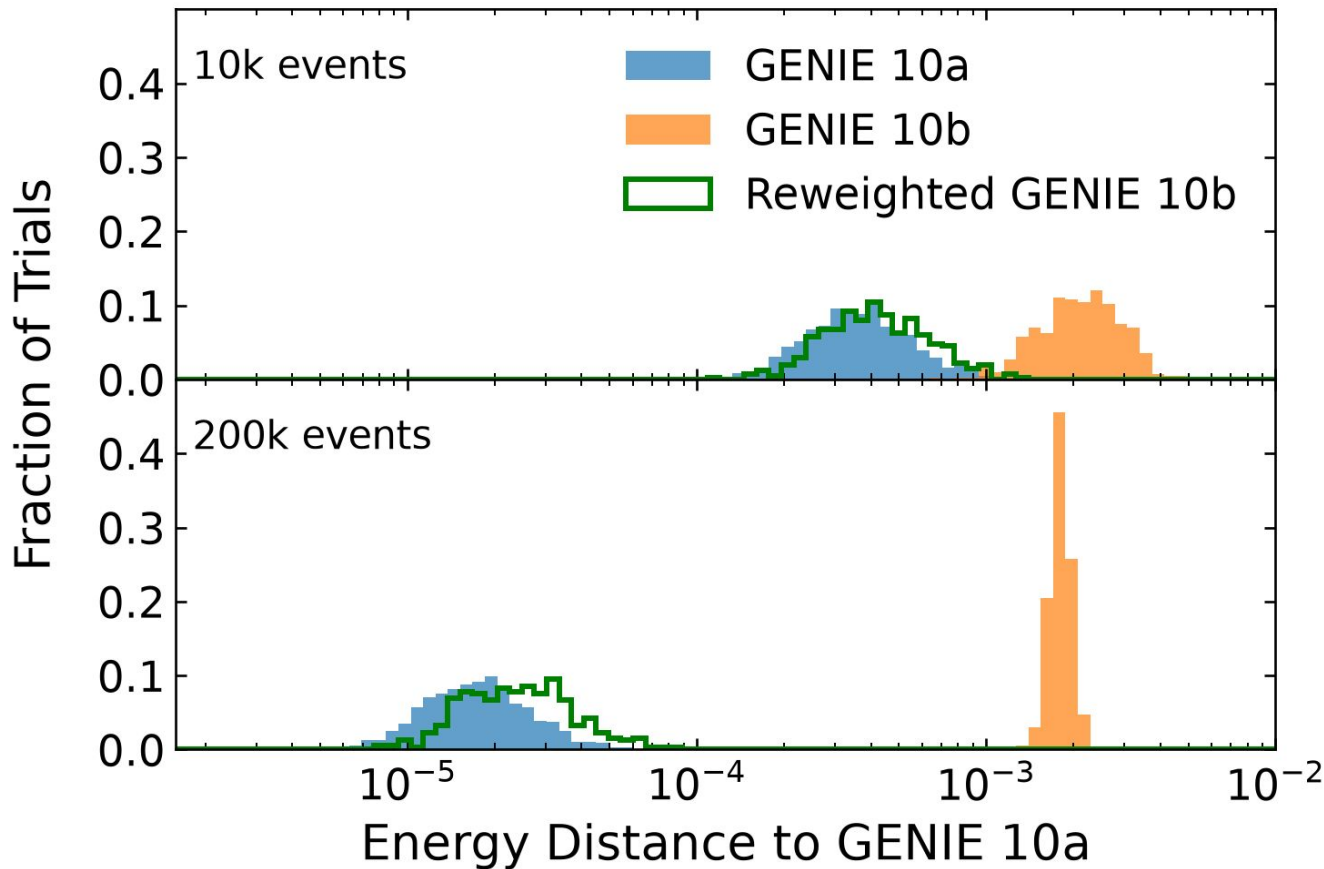
- Incident neutrino energy E_ν
- Reconstructed neutrino energy =
 $E_\mu + \sum E_{\pi^0} + \sum_{\text{Above Threshold}} E_{\pi^{+/-}} + \sum_{\text{Above Threshold}} KE_{\text{Protons}}$
- Number of π^0
- Number of $\pi^{+/-}$ above threshold
- Number of protons above threshold

Behavior of the Evaluation Metric

For example case of reweighting GENIE 10b to GENIE 10a

Energy distance metric gains more separating power as statistics increase

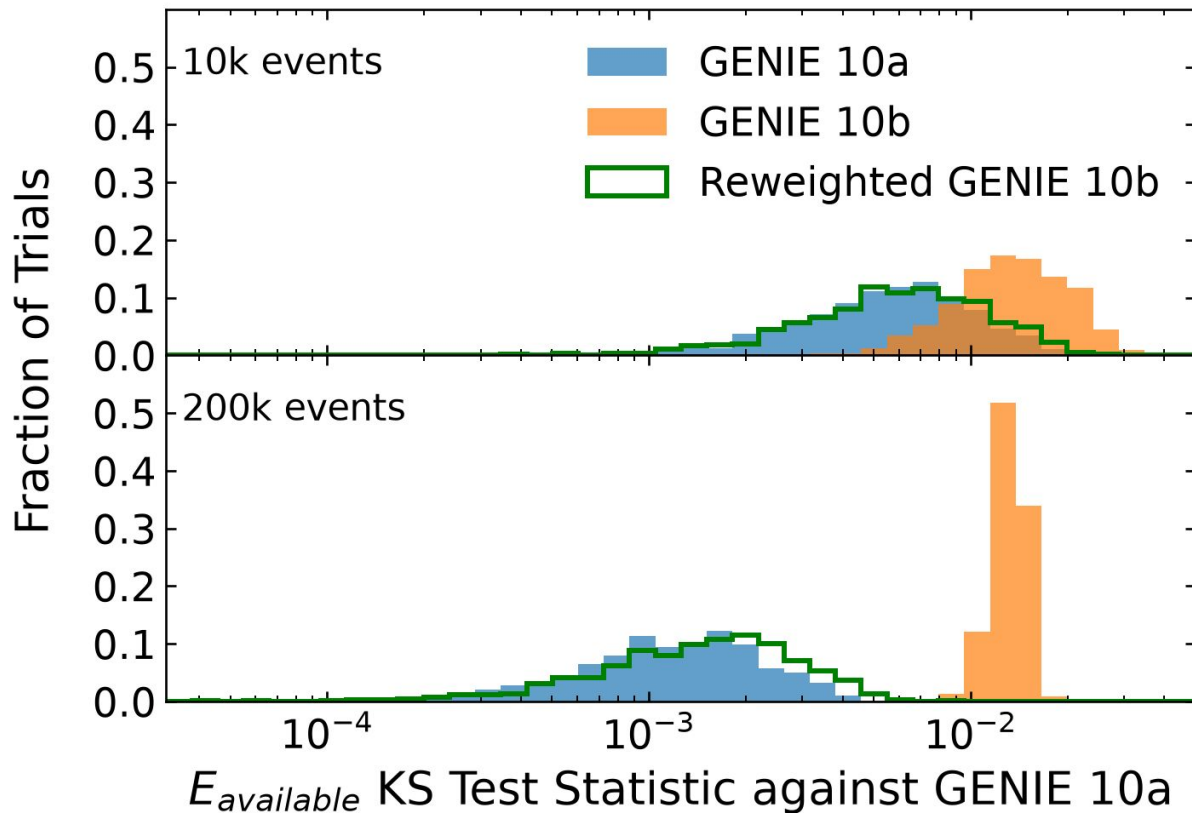
Final evaluation will be a measure of overlap with the reference (self-comparison) distribution with 1M test events



Adding in KS Tests

We can do a similar comparison in 1D for each observable we care about

These are cheap to calculate, so could add in $O(10)$ 1D KS comparisons too to make sure solutions don't drastically underperform along one dimension



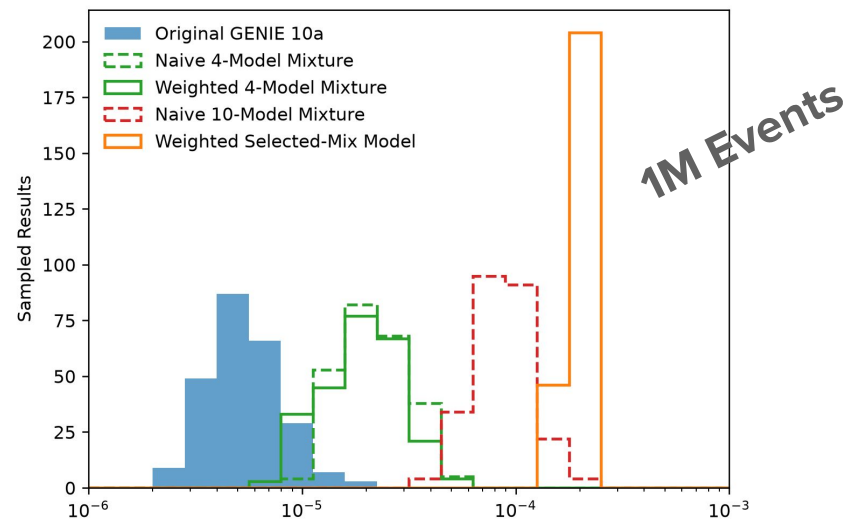
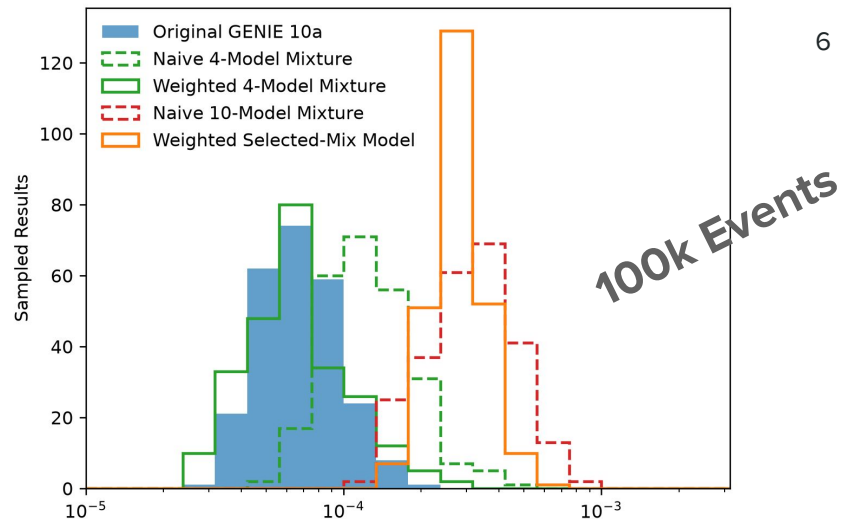
Baseline Solutions

Naive 4-Model Mixture is just an equal mix of provided events from each generator. Events from the target distribution get weights of 1; other events get weights of 0

Weighted 4-Model Mixture uses a DNN that reweights the mixture as a function of the energy distance variables

- This is the simplest real solution to the problem

Selected-Mix Model uses importance sampling to build a more efficient mix of the 4 available models



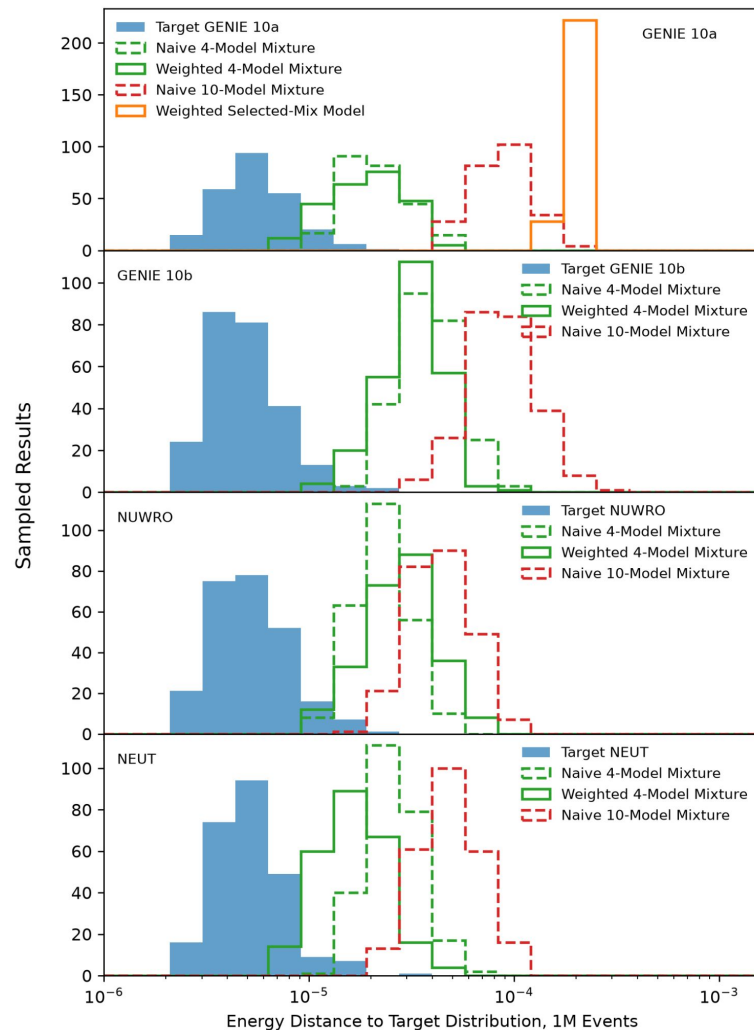
Baseline Solutions – 1M Events

Naive 4-Model Mixture is just an equal mix of provided events from each generator. Events from the target distribution get weights of 1; other events get weights of 0

Weighted 4-Model Mixture uses a DNN that reweights the mixture as a function of the energy distance variables

- This is the simplest real solution to the problem

Selected-Mix Model uses importance sampling to build a more efficient mix of the 4 available models



Evaluation Needs

Participants submit a single file of 1M events in a standardized format, with a separate set of weights for each of the target model+flux combinations (**N total sets of weights, for the N models/fluxes we want to match**).

- This should be lightweight for them to send and for us to receive and store

We need GPU resources to evaluate the energy distance performance of each submission

- With one A100 on Perlmutter, takes **~10 seconds to evaluate 1M events** once
 - Will want to evaluate multiple times on bootstrapped samples to account for statistical uncertainties. These slides used **250 trials**, which seemed like more than enough
 - Note: evaluation scales with (Number of Events)²
- KS tests are fast to evaluate on CPU

Evaluation To-dos

Evaluation tools are all ready for use

Things to decide, which can be adjusted based on baseline solution performance:

- May want to tune the variables used for energy distance calculation depending on how we see different solutions behave
 - In particular to weight sensitivity towards events in the tails of distributions?
- Is there a use for an effective sample size metric $(\sum w)^2 / (\sum w^2)$ on the weights too?
 - Part of this is already covered by the energy distance metric
- Do we want to evaluate at multiple different sample sizes?
 - Expect this will be useful primarily if no solution does very well at the full 1M sample size

Backup - Energy Distance and KS Test Results Example

