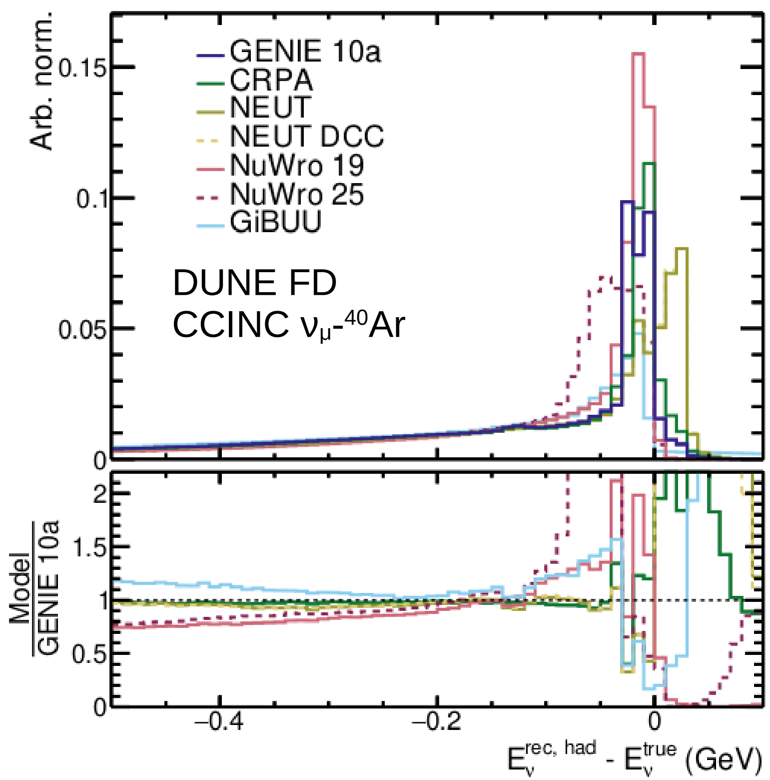
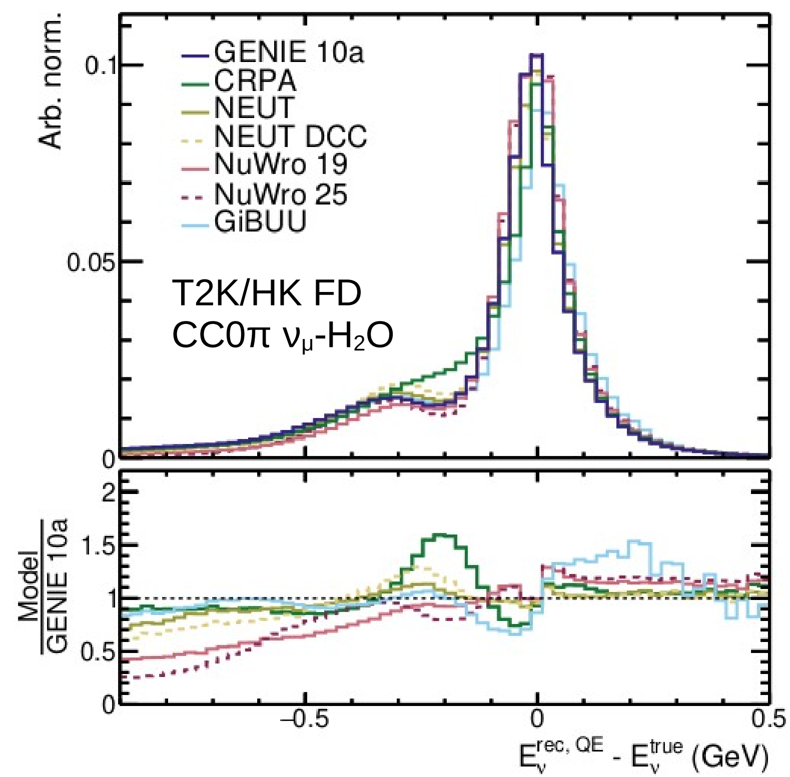
Surrogate event generator as a data challenge

Callum Wilkinson
LBNL, Physics Division

Physics background

- 0.1-10 GeV neutrino interaction physics poorly known
- Experiments rely on event generators predictions: GENIE, NuWro, NEUT, GiBUU, Achilles
- What's the cost of picking a different model? E_ν biases!





Physics background

- Generators are computationally cheap!
- ...but the cost of pushing many through GEANT4 + det sim + reco + analysis is prohibitely expensive
- Generators reweight the cross section to avoid the same problem. E.g., for an generated with parameters *p*, “reweight” to *p’* with:

$$w(\vec{x}) = \frac{\sigma(\vec{x}, \vec{p}\,')}{\sigma(\vec{x}, \vec{p})}$$

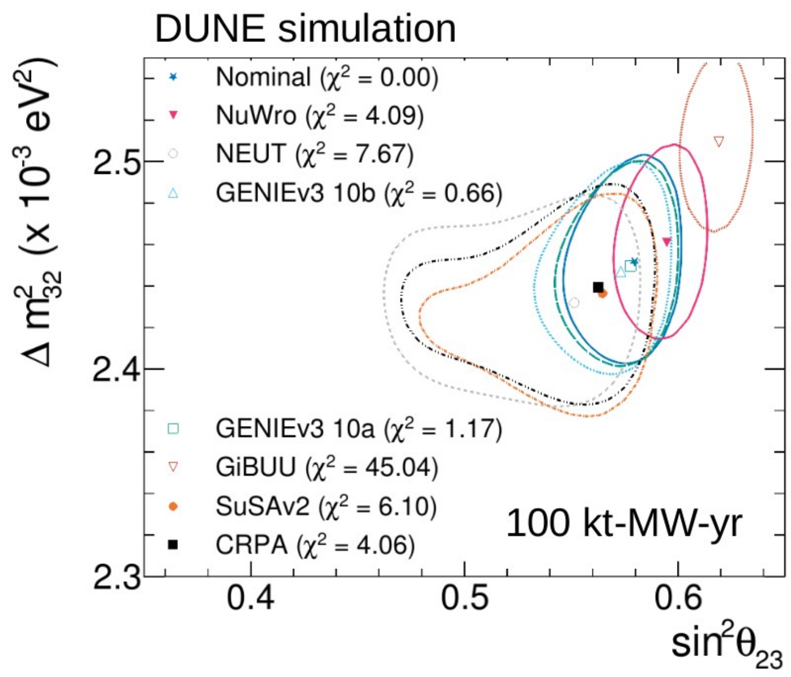
where $\bar{x}$ fully defines the outgoing event state

- Could similarly try to “reweight” from generators A $\rightarrow$ B, if you can estimate the probability density for an event with $\underline{x}$:

$$w(\vec{x}) = \frac{G_{\mathrm{B}}(\vec{x})}{G_{\mathrm{A}}(\vec{x})}$$

Physics background

- Several attempts to do this!
  - **DUNE:** BDT, *Instruments 5 (2021) 4, 31*
  - **MINERvA:** BDT, arXiv:2510.07463
  - **Andrew Cudd:** transformer, see recent NuFact poster



How?

(Stolen wholesale from Andrew's poster:)

Calculating Event Weights:

- Train a model to classify between two datasets **A** & **B** using the weighted cross-entropy loss (L)
- Each event **x** has a true label **p** and receives a prediction **q** from the network

$$L_i = p_i \log q_i + (1 - p_i) \log(1 - q_i)$$

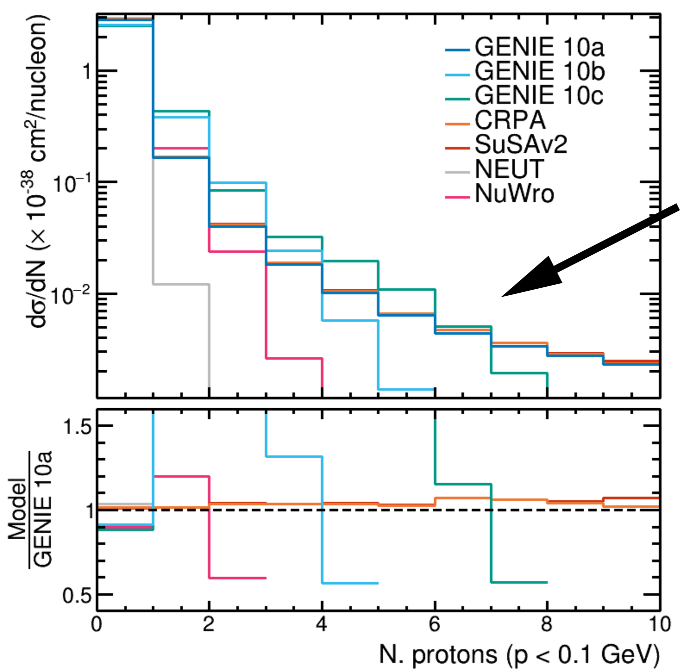
- Model predictions can be used to approximate the ratio of the datasets and used as weights for reweighting

$$\mathcal{L} = p_A(x_i)/p_B(x_i) \approx q_i/(1 - q_i)$$

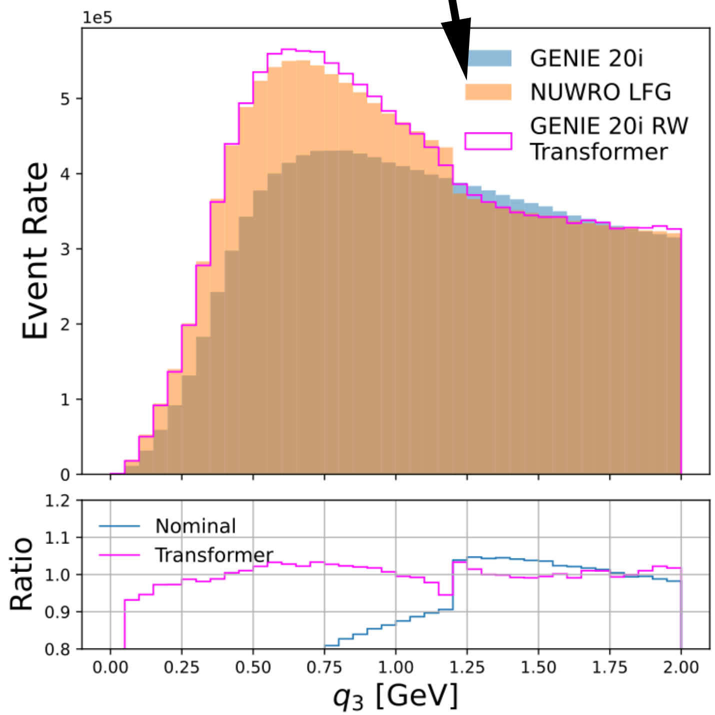
What’s the problem?

- If there are regions of non-overlapping phase space, this approach breaks
- Also a problem for general generator reweighting in some cases

Unphysical cut off in model component causes issues



Major phase space differences, even between generator configurations!



From Andrew Cudds’s poster!

What's the solution?

- Generate in a wider phase space and then deweight regions which are not there in your target model

$$w(\vec{x}) = \frac{G_{\mathrm{B}}(\vec{x})}{G_{\mathrm{A}}(\vec{x})}$$

This can be 0

This cannot be 0

(This approach currently used in some GENIE tunes to make nuclear binding energy reweightable)

- **Challenges:**
  - Very difficult to generalize analytically
  - Deweighting reduces effective MC statistics

Data challenge?

- This seems like an impactful problem with an ML solution
- **Challenge:** develop a surrogate model for fast neutrino event generation with phase-space covering an ensemble of generators
- **Dataset:** release a set of events for each generator, for a given E_ν range and set of target materials
- **Target audience:** natural appeal for neutrino people, +collider folk who have done similar work. Not sure about non-HEP appeal

Datasets

- Easy to generate large samples of events from a variety of models and put in the same format

- For each generator, can provide complete information ($\bar{x}$): E_ν, target, and four vector + PDG code of outgoing particles
  - Simple to document
  - In ROOT this would be ~200 MB/million events
  - Easy to put in hdf5(?), simple file layout
  - Generators are public, no permission required to release datasets

- Could also provide containers with simple event generation wrappers included if there’s a benefit (maybe overkill?)

Metrics

Here's where I fall down a bit and would invite some discussion

1) A clear metric is: $$w(\vec{x}) = \frac{G_{\mathrm{target}}(\vec{x})}{G_{\mathrm{surrogate}}(\vec{x})} \neq 0$$

For all $\bar{x}$ generated by the target generators of interest

2) Minimal efficiency loss when applying the deweighting factors $w(\bar{x})$:

- A particle bomb can fulfill (1), but is a bad solution
- Minimizing MC stat. uncertainties when applied to a test set?
- Maximize the average value of $w(\bar{x})$ for a test set?

+Maybe
3) The model should be as computationally inexpensive as possible to train, as it will need to be retrained to add new generators

Barriers to physics impacts?

Solves a key problem for all few-GeV neutrino experiments, but some technical barriers:

- Can’t be backported, and it would be a breaking change
- If the *efficiency* is low, the cost might seem prohibitive
- Adds a technical barrier as reweighting required downstream
- Integration with flux drivers which place vertices in a complex geometry from a complex flux would require some effort
- Retraining required to include new physics models