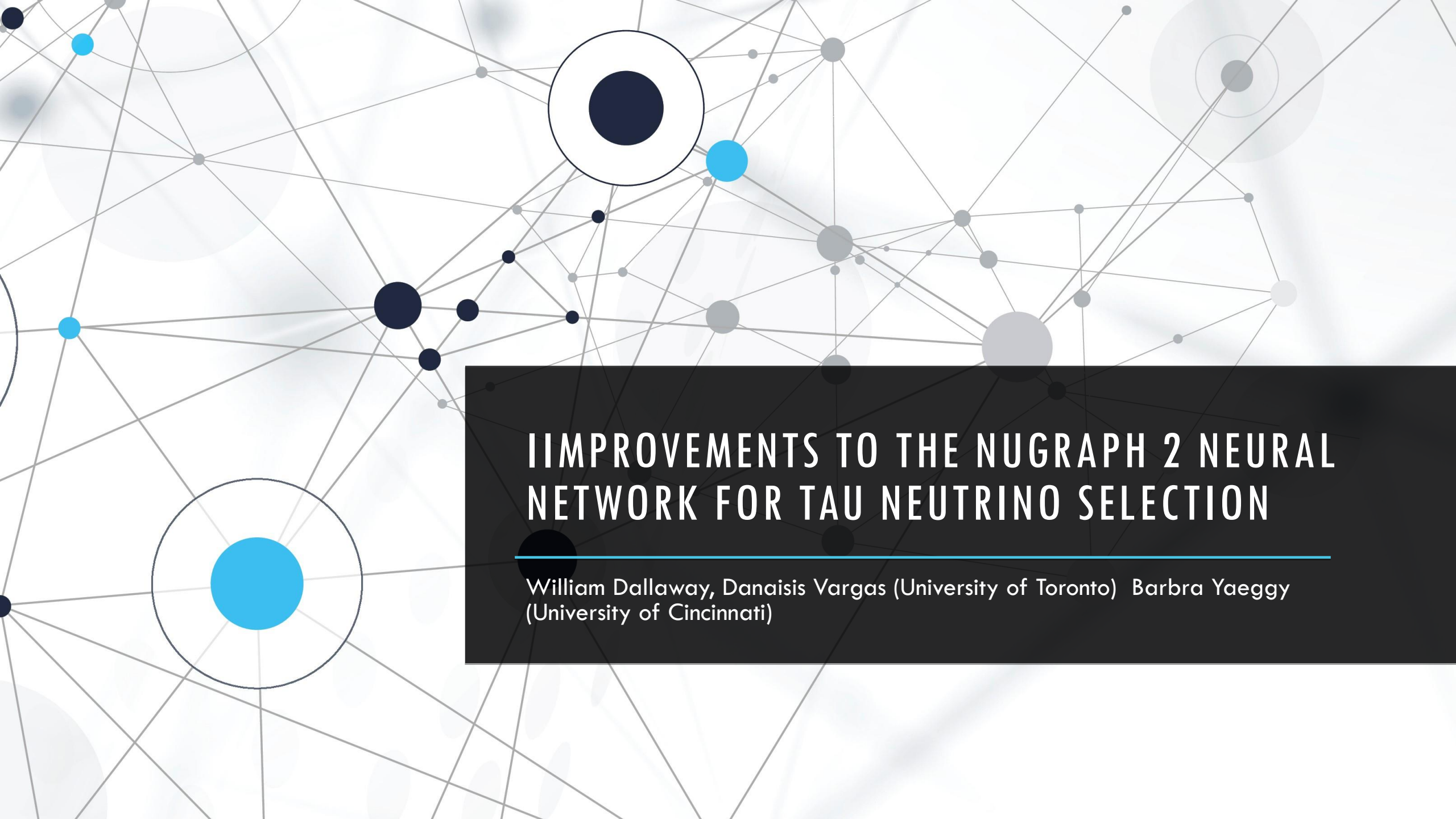


IIMPROVEMENTS TO THE NUGRAPH 2 NEURAL NETWORK FOR TAU NEUTRINO SELECTION

William Dallaway, Danaisis Vargas (University of Toronto) Barbra Yaeggy (University of Cincinnati)

INTRODUCTION

We are attempting to do a simple tau neutrino sensitivity study in the DUNE far detector.

The idea is to use the reconstruction along with the NuGraph2 neural network to select out only tau neutrino events from the DUNE far detector

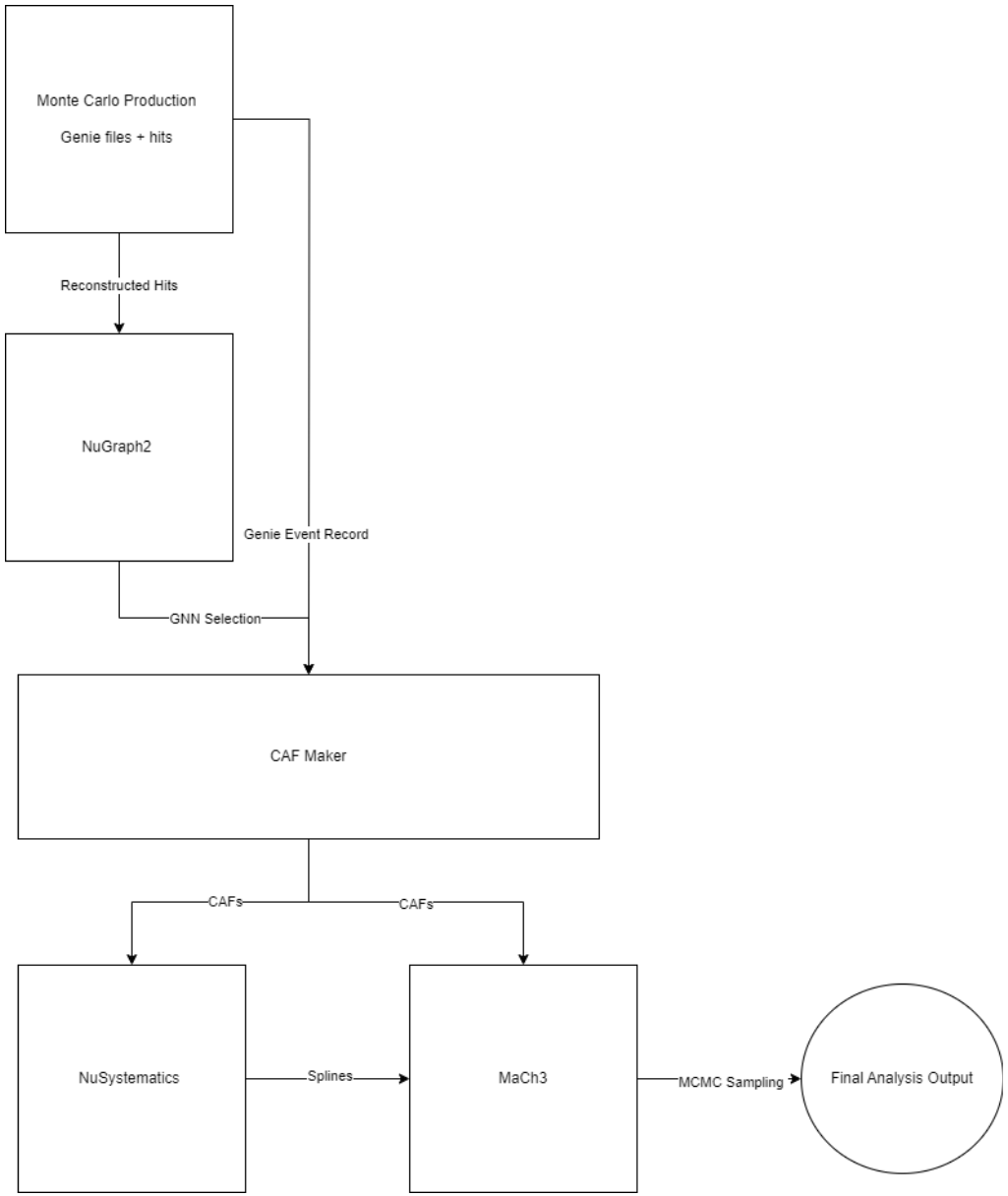
We then use the tools that were recommended to us to get the sensitivity to tau neutrinos specifically we use NuSystematics and MaCh3 for to convert our selections to a sensitivity.

Our major challenge is that the processing of large amount of events into graphs and training are very computationally intensive for processing with 32 mpi processes 600k DUNE FD events requires ~600 GB of RAM and several hours of wall time and for training we have to use small batch sizes (<32) for more than 100k events due to limited memory on our GPUs

PROPOSED WORKFLOW

Please do give suggestions for changes we should make or if I am misunderstanding the functions of the components!

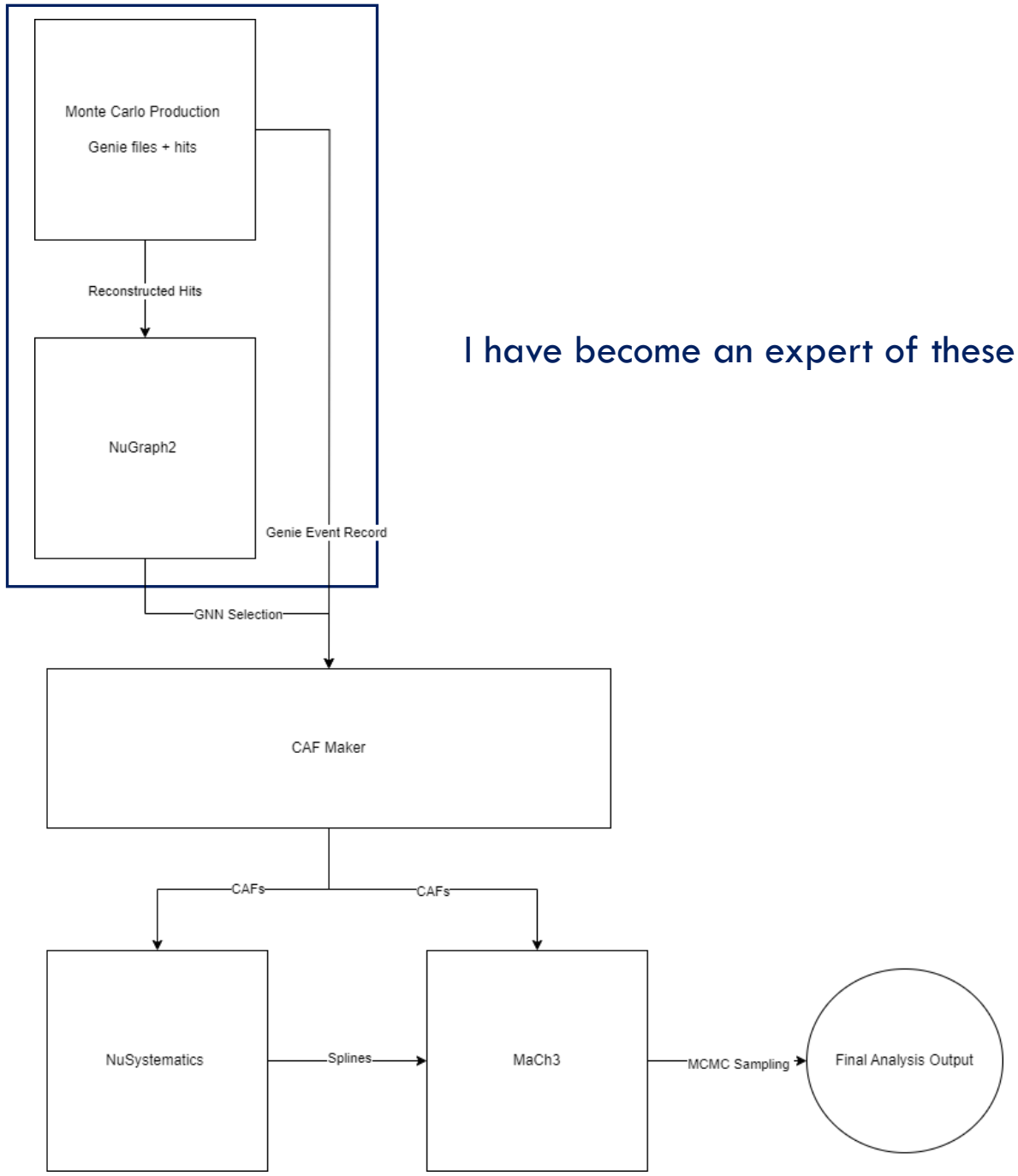
In the next couple of slides I will describe our MC production and the NuGraph neural network since I understand these the most and they are most relevant for the Sim/Reco group



PROPOSED WORKFLOW

Please do give suggestions for changes we should make or if I am misunderstanding the functions of the components!

In the next couple of slides I will describe our MC production and the NuGraph neural network since I understand these the most and they are most relevant for the Sim/Reco group

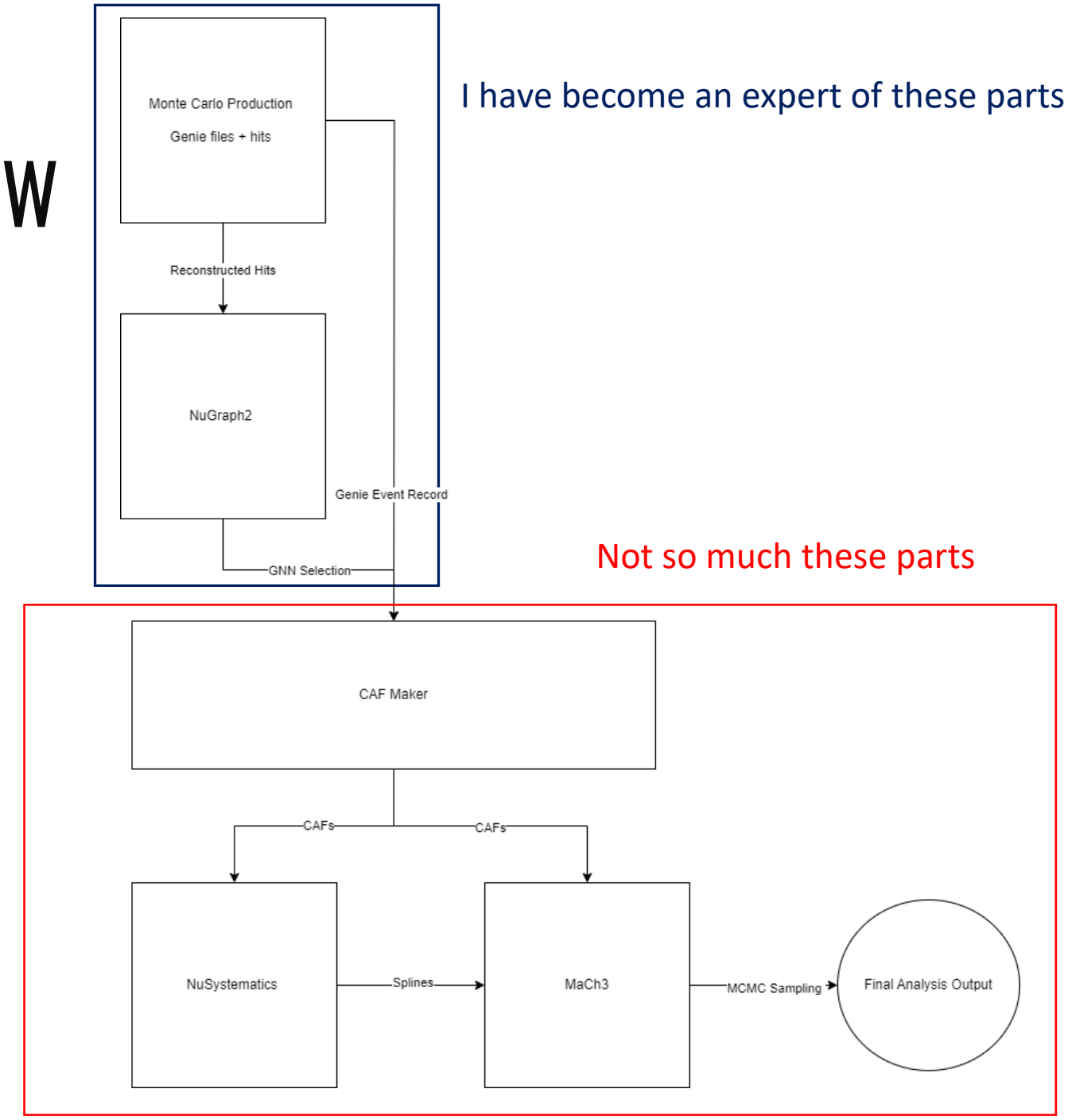


I have become an expert of these parts

PROPOSED WORKFLOW

Please do give suggestions for changes we should make or if I am misunderstanding the functions of the components!

In the next couple of slides I will describe our MC production and the NuGraph neural network since I understand these the most and they are most relevant for the Sim/Reco group



OUR MC PRODUCTION

We have generated 600k MC far detector beam events 200k of each generated of each flavour.

We have put these through the full reconstruction chain and are storing them on our Canadian computing cluster.

We have used these so far for our training BUT we will need to do some things differently (more on this later)

We are also going to need to generate more for our analysis.

We are open to sharing these/integrating them to DUNE analysis if people are interested and there's space for them

MORE ON OUR MC SAMPLE

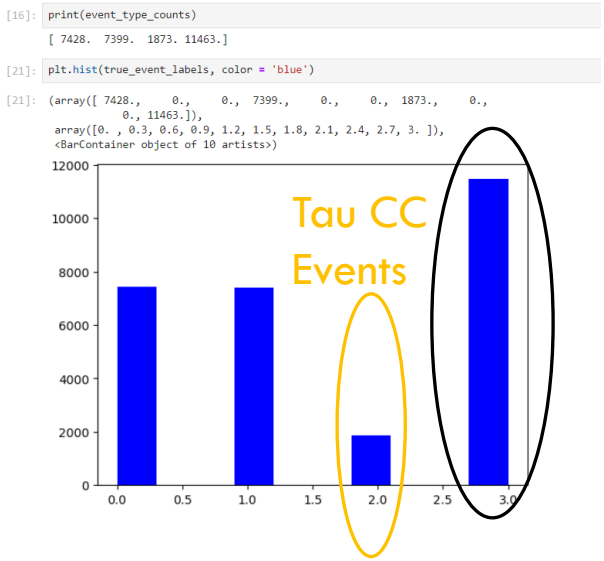
For all the results I'll show of the training the samples we used are just all the 600k events concatenated together.

This is not ideal because there is a large phase space suppression of the tau charged current cross section due to phase space effects from the large tau mass

MORE ON OUR MC SAMPLE

For all the results I'll show of the training the samples we used are just all the 600k events concatenated together.

This is not ideal because there is a large phase space suppression of the tau charged current cross section due to phase space effects from the large tau mass



Neutral Current Events

MORE ON OUR MC SAMPLE

This (as you will see) causes issues since this is a systematic bias in the training set. Based on this there are two potential solutions. We are implementing them both and soon I should have the data to empirically compare the solutions

MORE ON OUR MC SAMPLE

This (as you will see) causes issues since this is a systematic bias in the training set. Based on this there are two potential solutions. We are implementing them both and soon I should have the data to empirically compare the solutions

Solution 1:

Simply don't include the neutral current events from the tau samples (since this is where most of them are

Pros:
Very simple to implement
Less resource intensive for processing and training.

Cons:
This reduces the data available for training meaning that the predictive power is worsened.

MORE ON OUR MC SAMPLE

This (as you will see) causes issues since this is a systematic bias in the training set. Based on this there are two potential solutions. We are implementing them both and soon I should have the data to empirically compare the solutions

Solution 1:

Simply don't include the neutral current events from the tau samples (since this is where most of them are

Pros:
Very simple to implement
Less resource intensive for processing and training.

Cons:
This reduces the data available for training meaning that the predictive power is worsened.

Solution 2:

Change the loss function to accommodate the bias (i.e punish the network more for misclassifying tau cc as neutral current. Specifically, we want to use weighted categorical cross entropy.

Pros:
This allows us to use the full dataset for training and improve our results.

Cons:
Slightly harder to implement (but can still be done simply)
Will increase processing time/resources (since we need to calculate the weights for each category)

ATMOSPHERIC SAMPLES

We have also generated atmospheric muon and tau events (again 200k) these have the same issue but worse.

Why? I'm not sure I thought they should have energies larger than the beam and the so they should have less tau cc suppression, but it seems just as bad if not worse.

I will need look into the flux files that were used to understand why this is happening. Ideally we could also do a sensitivity study with these.

NEXT RUN OF SAMPLES

To train or do inference with NuGraph's GNN we need the full geant4 truth record but the standard fcls used roll up these showers.

Since we need this truth information, we have used modified fcls in our generation of the samples which keeps this information

It's looking likely that we will need to generate another run to do our inference and analysis since this is not likely to be included in the official samples

Based on discussion with the MaCh3 developers they recommend 10*(#Number of events for the exposure you're considering)

This will be much larger than what we needed for the training and we don't use DUNE computing resources only my PI's RAC allocation from Compute Canada so knowing that we only need the hits and the genie record suggestions for streamlining are welcome!

NUGRAPH NEURAL NETWORK

The NuGraph neural network is a GNN that is based on the ExaTrkx collaboration which was originally used for hadron physics but has been repurposed for doing neutrino interactions

NuGraph core convolution network is a self attention message passing-network network which uses a categorical embedding.

The nodes of the graphs are hits, and the features are the time and wire of the hit along with the integral and rms energy deposit in the wire and I also added the coordinates of the TPC module where the wire is and the drift direction.

The vertices have edges drawn between them using Delaunay triangulation.

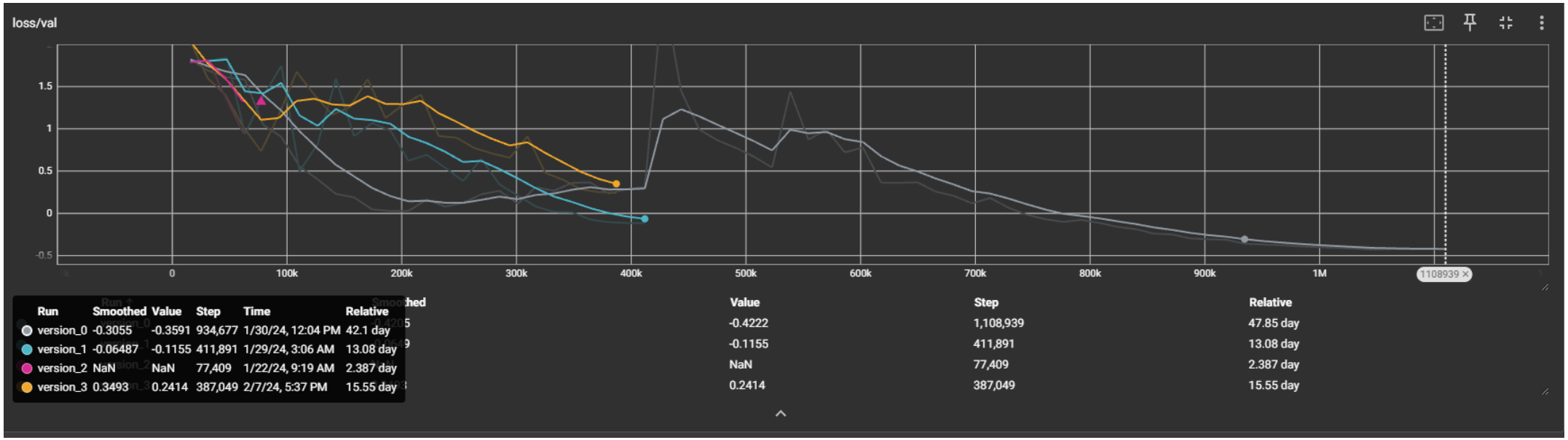
MORE ON NUGRAPH

The graphs are then pooled (in the original network this is using max pooling but we changed it to LSTM) and fed through an encoder layer then a nexus layer is used for the message passing then there are decoder layers.

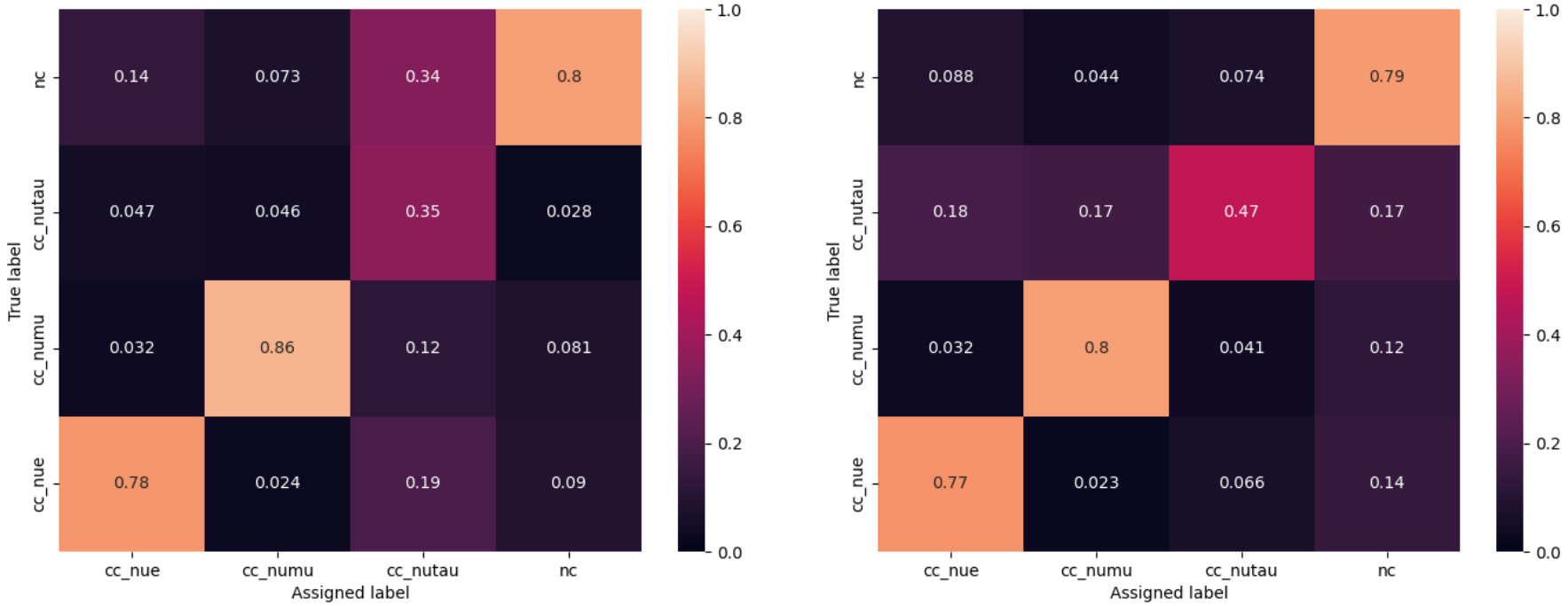
The decoder layer has several different heads which we can use. The two we have been using for training are the semantic head which predicts the type of particle associated with the energy deposit as well as the event head decoder which predicts the neutrino event type.

We haven't made any changes to these parts of the network

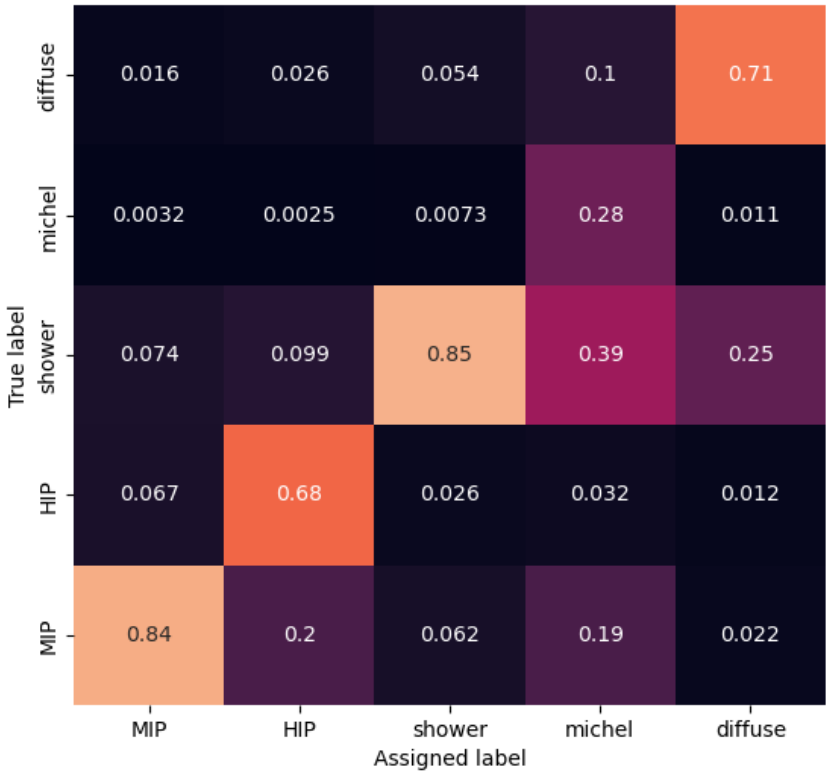
OUR RESULTS WITH NUGRAPH

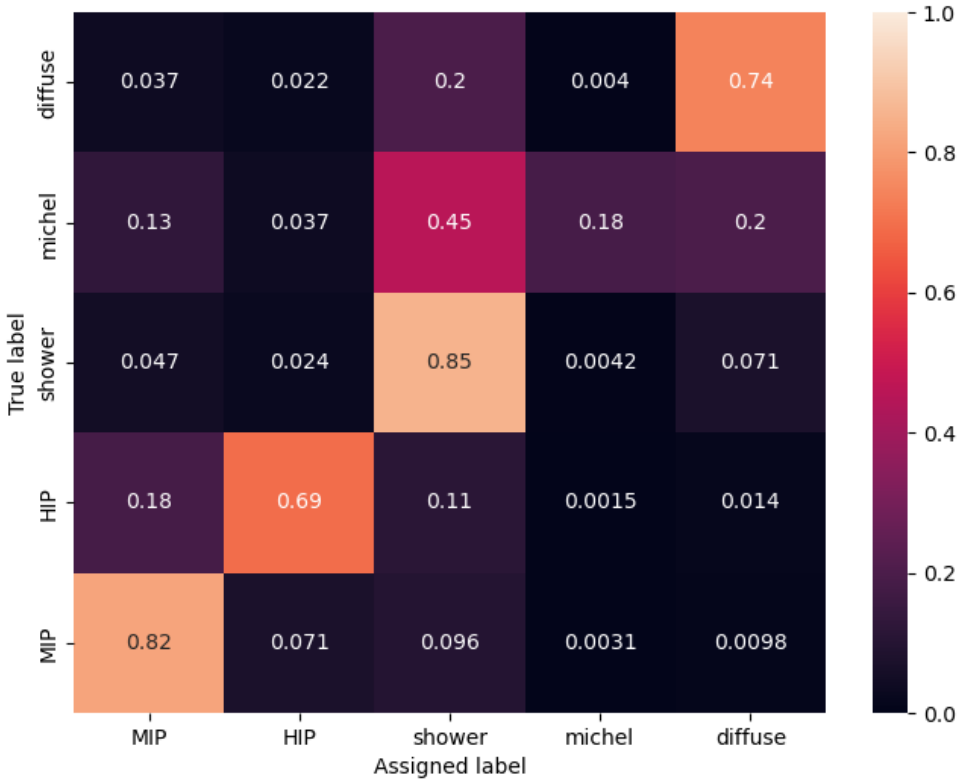


OUR RESULTS WITH NUGRAPH

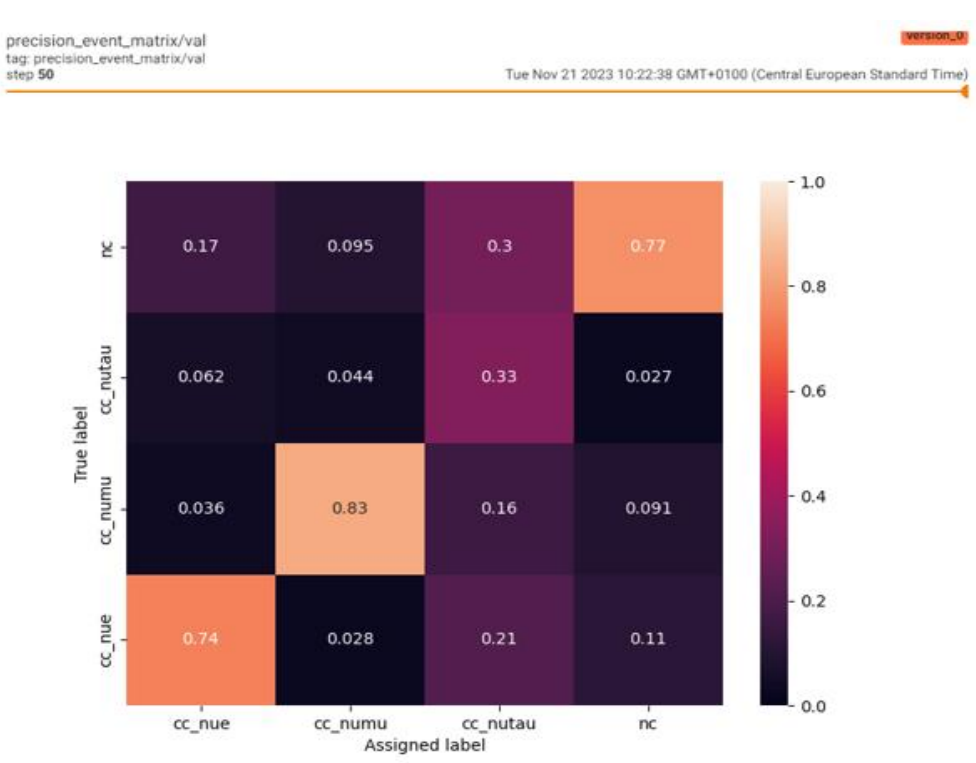


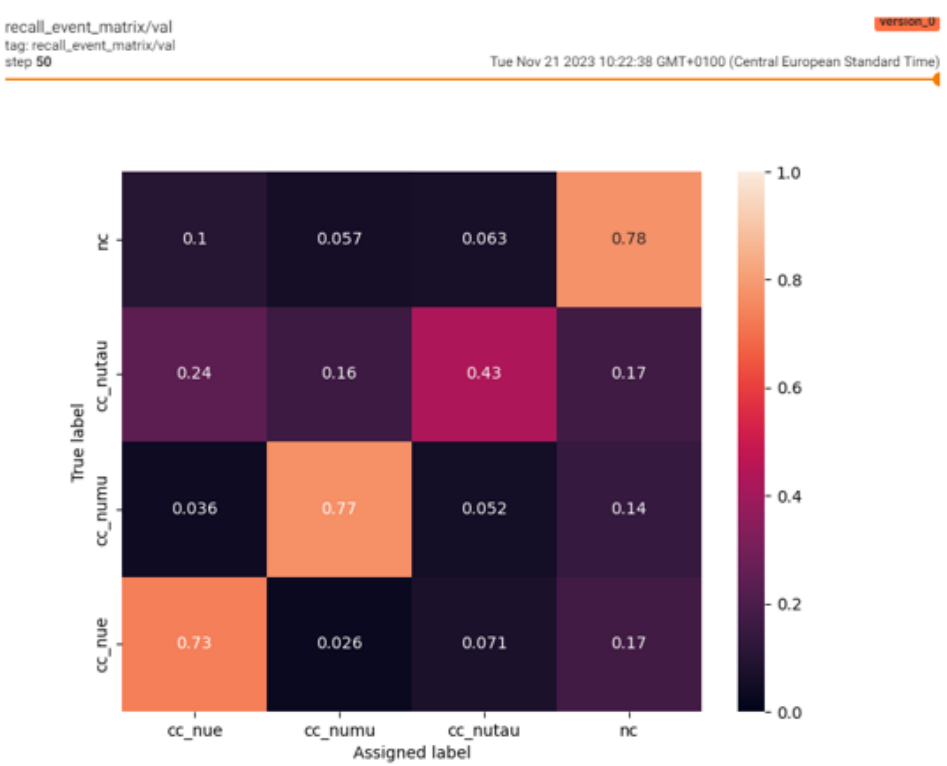
OUR RESULTS FROM NUGRAPH



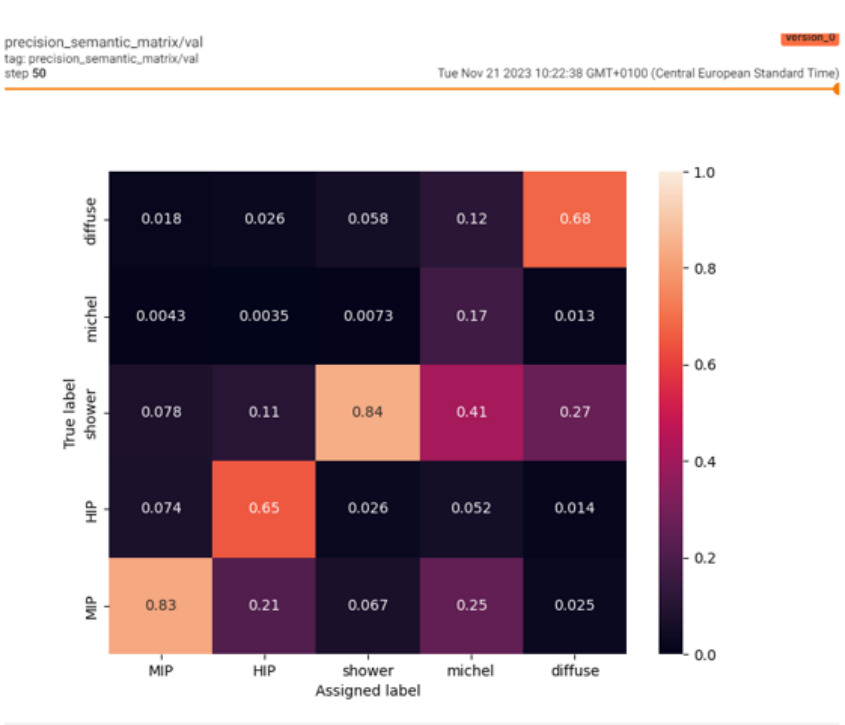


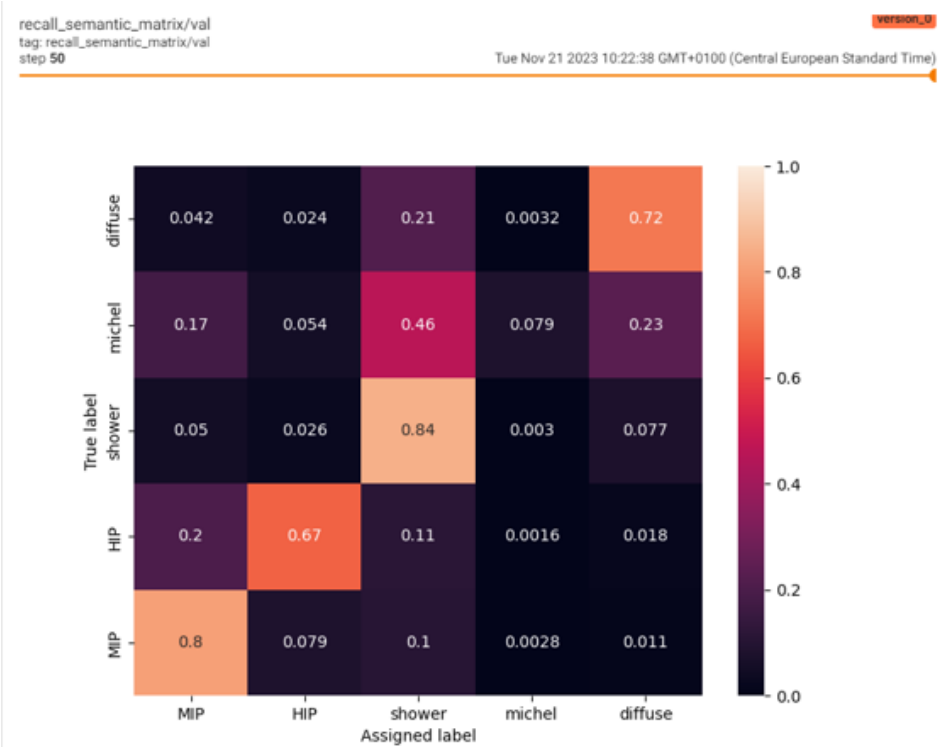
IMPROVEMENTS WITH THE NEW FEATURES/POOLING





IMPROVEMENTS WITH THE NEW FEATURES/POOLING





NEXT STEPS

In order to continue with our analysis we need to:

1. Retrain the model with the weighted loss function or smaller (but balanced) dataset whichever performs better.
2. Plot ROC curves and choose a threshold for tau selection.
3. Make new files and convert them to CAFs.
4. Run through NuSystematics and then MaCh3 for sensitivity.

HOW THIS ALL FITS IN

We are using the event head decoder for doing the tau selection study

FD sim/reco might also be interested in the semantic decoder to do proper tau reconstruction

Also are developing a vertex decoder which might be helpful to work with pandora ect.

POINTS FOR DISCUSSION

How to use the duneana CAFmaker module? I get errors when I try to run from the fcls in the duneana repository so I wonder who maintains it?

Does the proposed workflow makes sense?

Any thoughts on the systematics to include which may not be there?

THANKS FOR YOUR ATTENTION