Generator to detector object transformation using machine learning

DPF 2019, Boston, July 2019

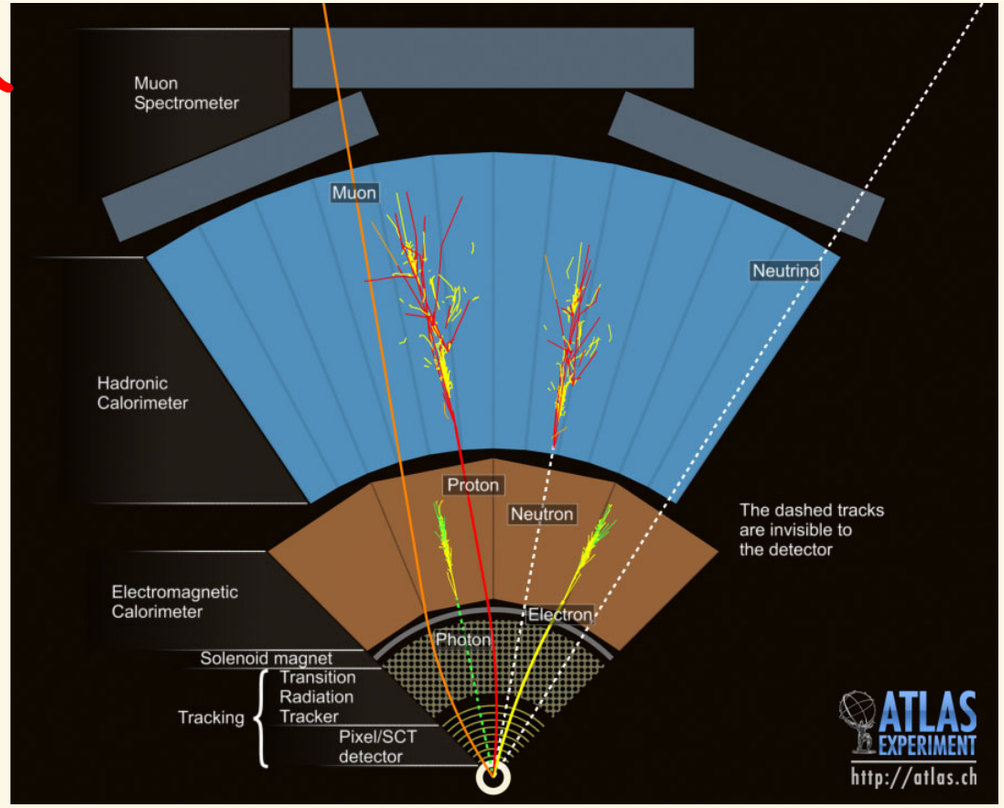
Doug Benjamin, Sergei Chekanov, **Walter Hopkins**, Jeremy Love

Argonne NATIONAL LABORATORY

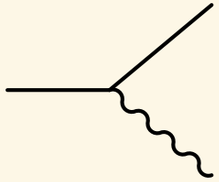
Introduction: simulating particle interactions

Detector transformation: sim+reco

transformed properties/efficiency: $(p_T', \eta', \phi', m'), \epsilon$



generated properties: (p_T, η, ϕ, m)



Traditional method

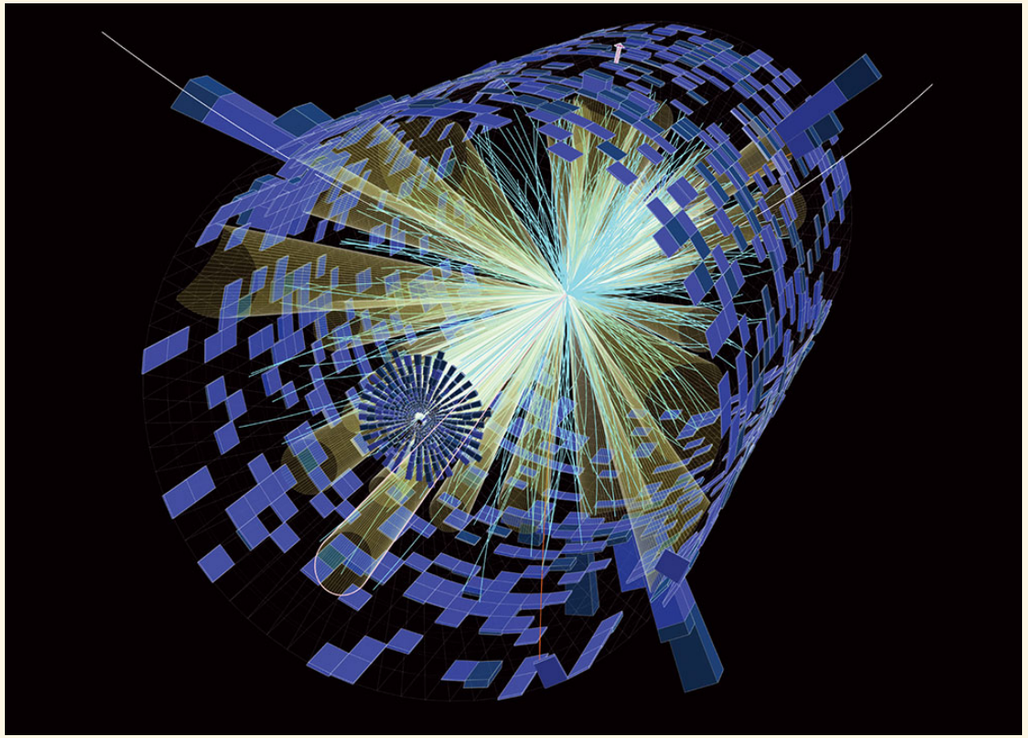
- Geant simulation: 4 → ~100-1000 parameters.
- Reconstruction: reduce info down to 4 parameters.
- **Very compute intensive**

ML method

- Combine simulation and reco step and learn transformation of initial features.
- Learn essential features with good-enough accuracy.
- **Could significantly reduce computational time**

Long term goals

- HL-LHC will produce large amounts of data: 500 PB/year.
- Search for **new physics in new areas of multidimensional space.**
- Reduced systematic uncertainties needed → **find deviations from the SM.**
  - Larger samples and faster turn around → better understanding of systematics.



Dream goal: on-the-fly simulation+reconstruction (hours vs weeks) for full SM background.

Short term goals

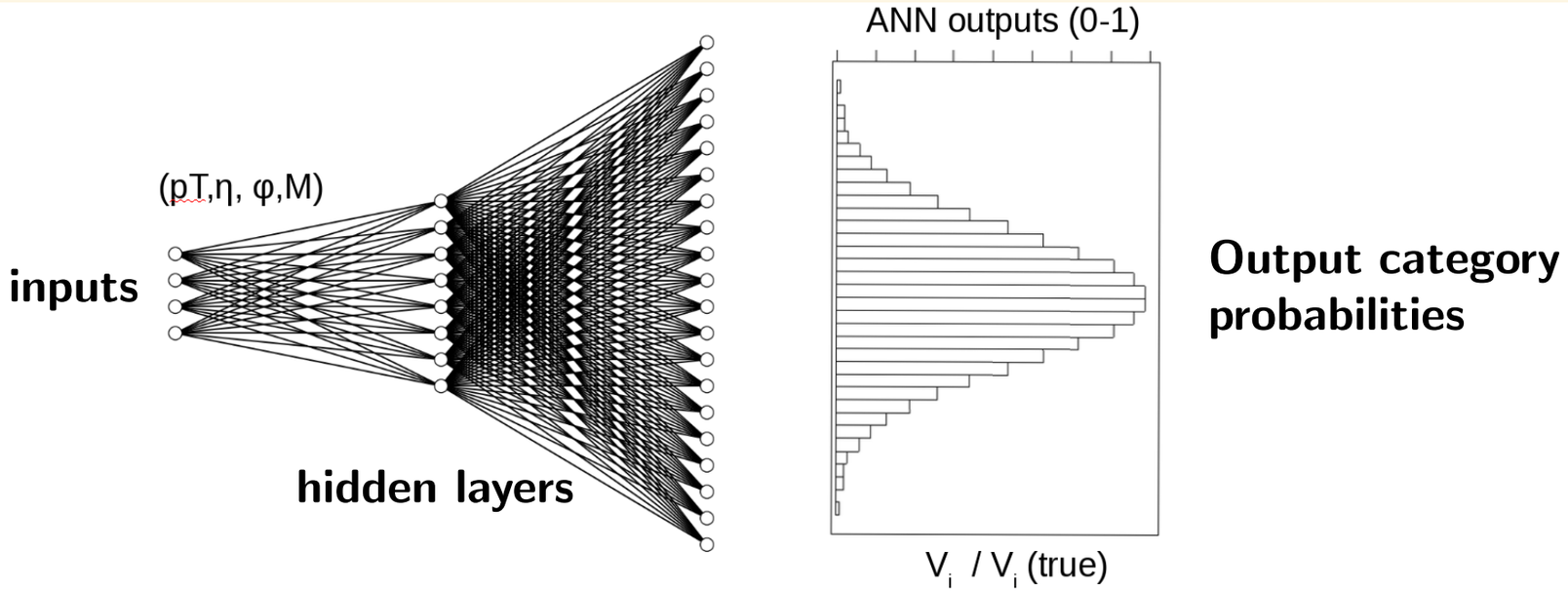
- Replace "truth smearing" with ML detector smearing.
- Previous method used binned ($\mathrm{p_T}$, η, etc), parametrized resolution histograms.
- NN can skip binning: correlations between binned variables can be learned.

Truth smearing uses

- Useful for signal region optimization: produce large amounts of data to train a discriminating NN.
- MC filtering: e.g. $E_{\mathrm{T}}^{\mathrm{miss}} > 250$ GeV for trigger threshold.

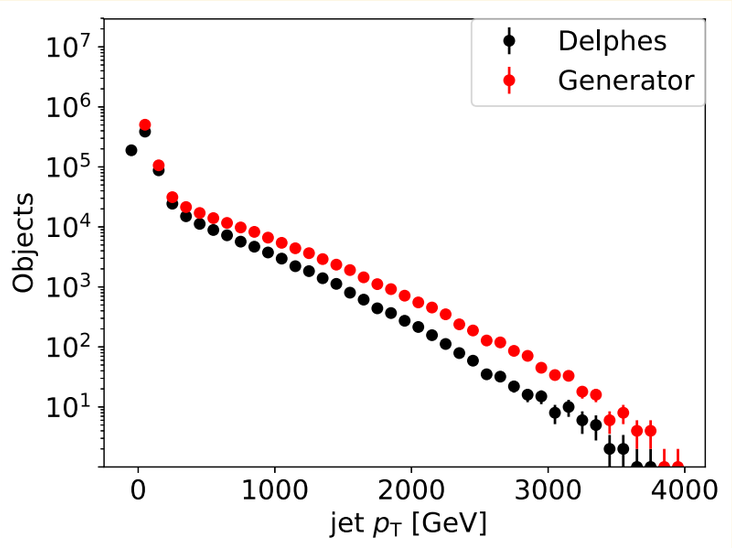
Resolution neural network

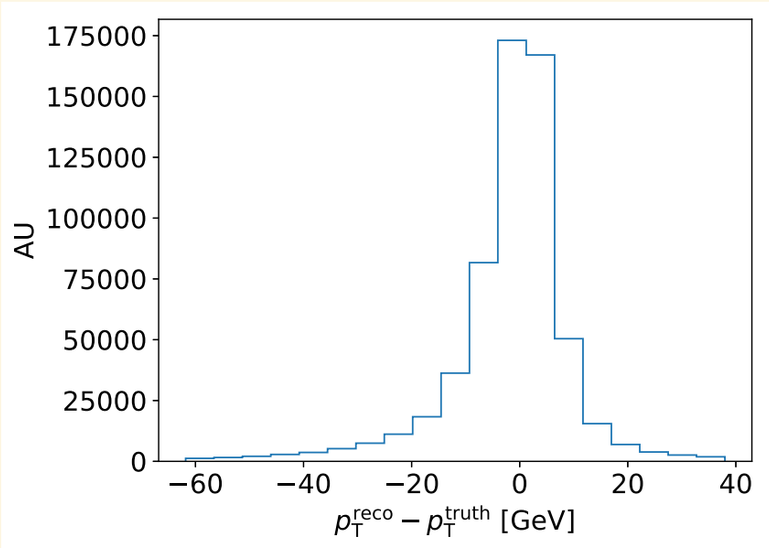
- NN inputs: features that affect resolutions, e.g. four-vectors of objects.
- NN output: 100 bins representing resolution of quantity.
  - $p_T{}^{reco} - p_T{}^{truth}, \eta^{reco} - \eta^{truth}, \phi^{reco} - \phi^{truth}, m^{reco} - m^{truth}$.
  - 1 NN per output quantity: p_T, η, ϕ, m.
- NN structure: multiclassifying fully connected network → each resolution bin=category.



Data preprocessing

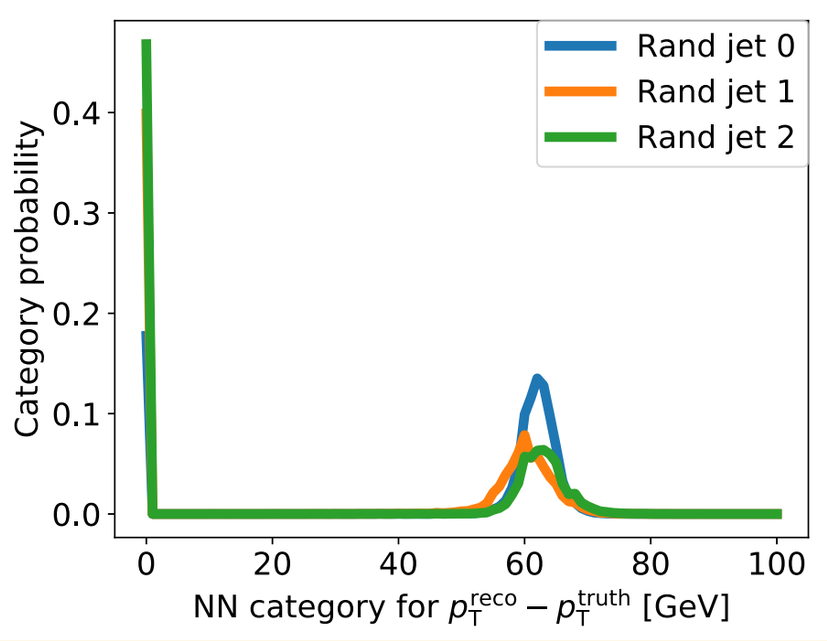
- Delphes ATLAS-like samples used: mix of $t\bar{t}$ and γ+jets.
  - First study was on jets, other objects tested but not shown.
- Inputs are individually scaled to be in 0-1.
- Outputs: $p_T^{reco} - p_T^{truth}$, scaled to be from 0-1.
  - Only objects within 1-99% of the distribution are selected.
  - Inefficiencies are represented by lowest bin: objects generated but not in Delphes reconstruction.

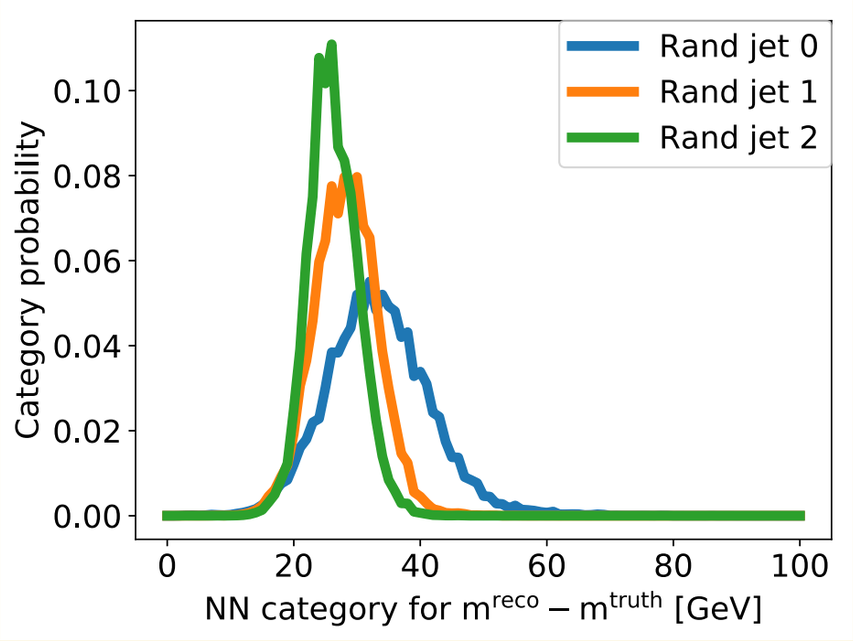




Sampling

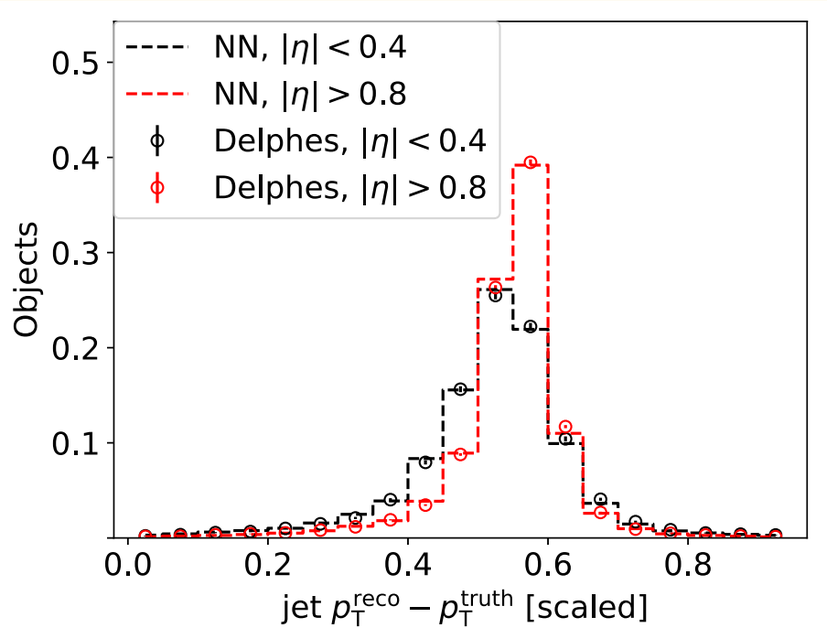
- NN produces category probabilities for each object.
- Detector response is stochastic: randomly sampling category probabilities.
  - Identical input can result in different outputs.





Testing correlations

- Various input parameters (η, $\mathrm{p_T}$) impact resolutions.
- Traditional method:
  - Build resolutions in bins of η, $\mathrm{p_T}$, occupancy, etc.
- ML algorithms are good at finding correlations!
  - No binning: **1 NN learns correlations between input and resolution.**

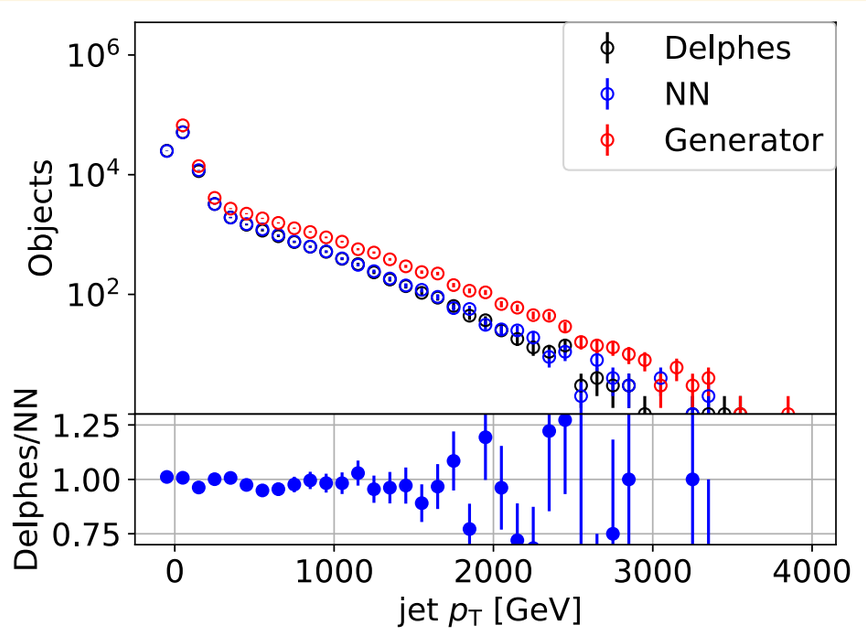


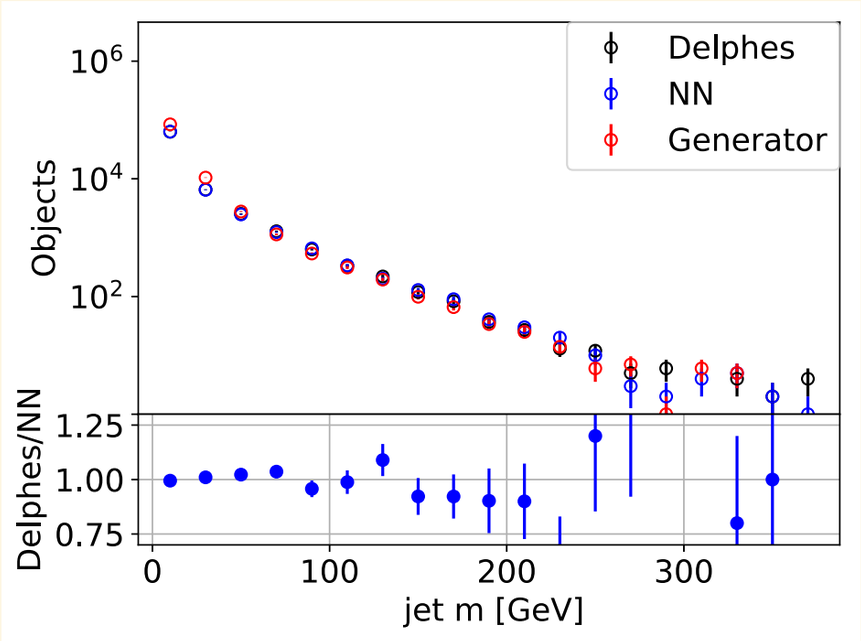
NN correlation test

- Delphes changes the $\mathrm{p_T}$ resolution based on η.
- Compare $\mathrm{p_T}$ resolutions for $|\eta| < 0.4$ and $|\eta| > 0.8$.
  - In Delphes and in NN trained without η binning.

Results

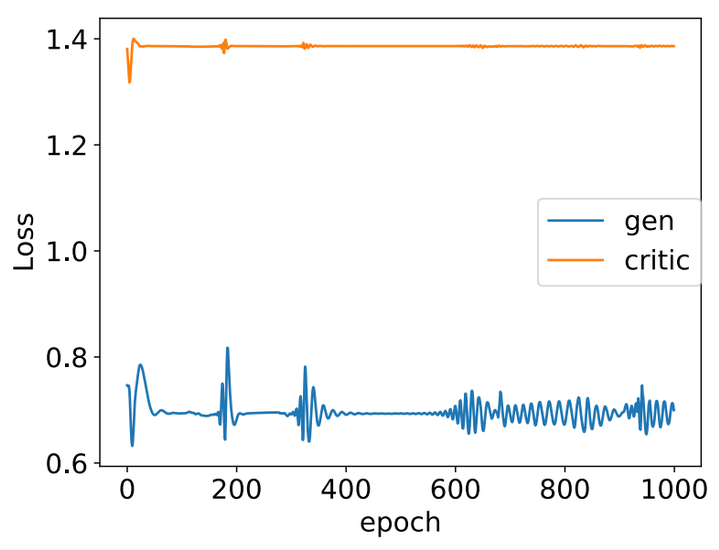
- Produce reco objects from generated objects+NN+random sampling.
- First negative bin are Delphes objects that couldn't be reconstructed.
- NN results within 15% of Delphes objects (when stat uncert. <10%).
  - Simple 5-layer, 100-node, 101-category NNs.
  - Flattening of input distributions could improve performance.

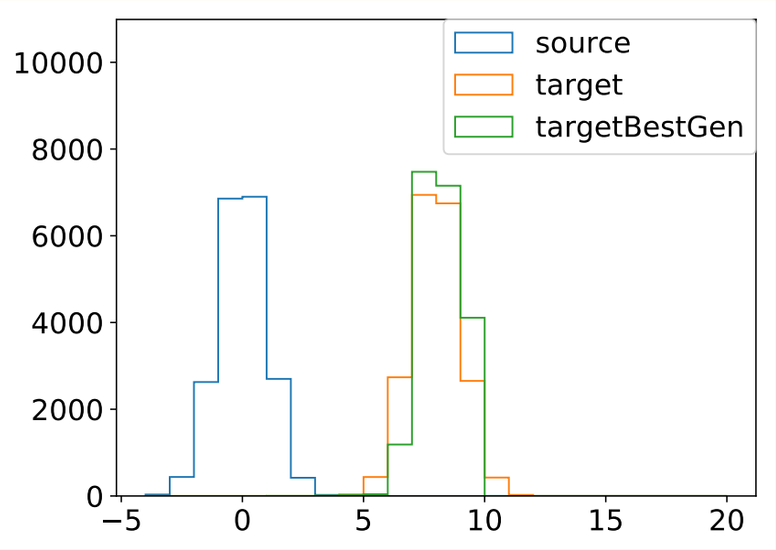




Future plans: generative adversarial models

- Train generative models adversarial models (GANs) to learn generator to detector transformation.
- Optimize the generative models with a hyperparameter scan on HPCs.
- Will require careful selection of figure of merit.
  - Loss function can look stable but tails of generated distributions are badly modelled.
  - Common problems of GANs: training and interpretability.

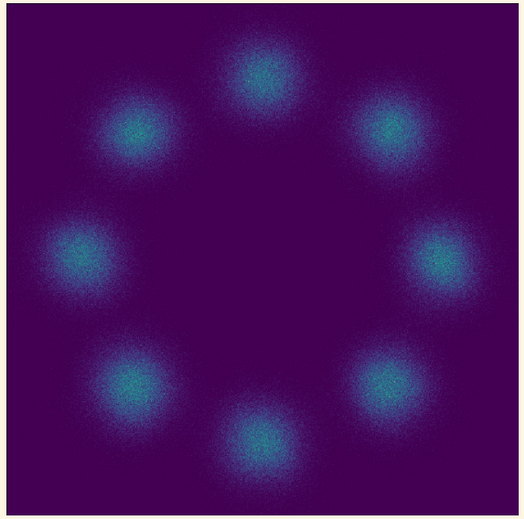




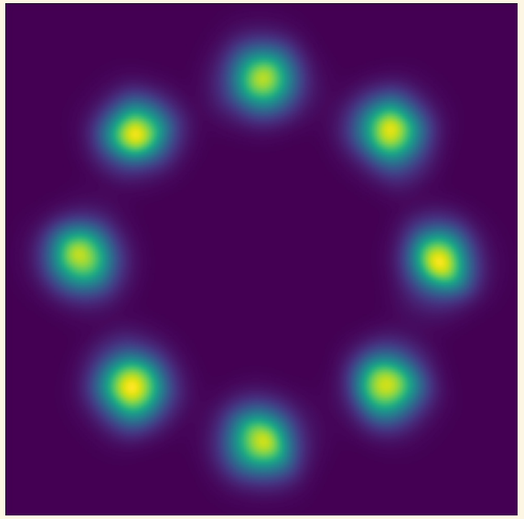
Future plans: autoregressive models

- More interpretable and reversable models developed: GLOW, NICE, FFJORD.
- These methods estimate the log-density of a target distribution.
- Methods are reversable: detector smearing and truth unfolding.

True distribution



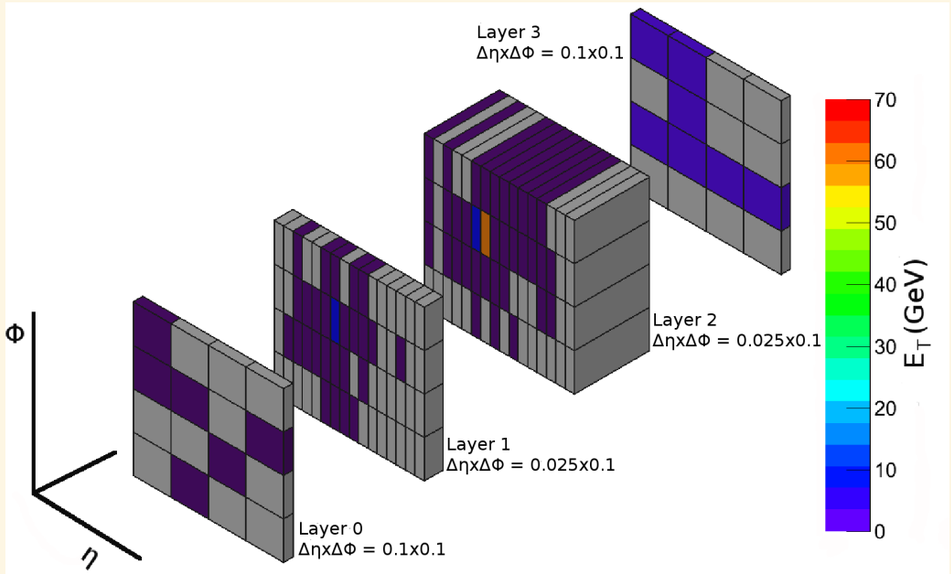
Predicted distribution



W. Grathwohl, et al, arXiv:1810.01367

Future plans: calorimeter simulation

- Use what we learned in generator to detector transformation on calorimeter simulation.
- Calorimeter simulation takes significant fraction of compute time.



- **Hyperparameter scans will be essential in reducing the NN parameters.**
  - Shift compute from evaluation to training.
  - Smaller network should result in faster evaluation.

Conclusion

- Generator to post-detector response transformation can be modeled by NNs.
- Started studying GAN implementation of the same transformation.
- Working on autoregression model implementation.
- **Ultimate goal: fast modelling of calorimeter deposits.**
  - Explore various techniques: GANs, autoregressive models.
  - Minimize network size: hyperparameter scans on HPCs, reinforcement learning.

Know your network

J. Zhu et al, arXiv:1703.10593