Machine Learning Methods in high jet multiplicities

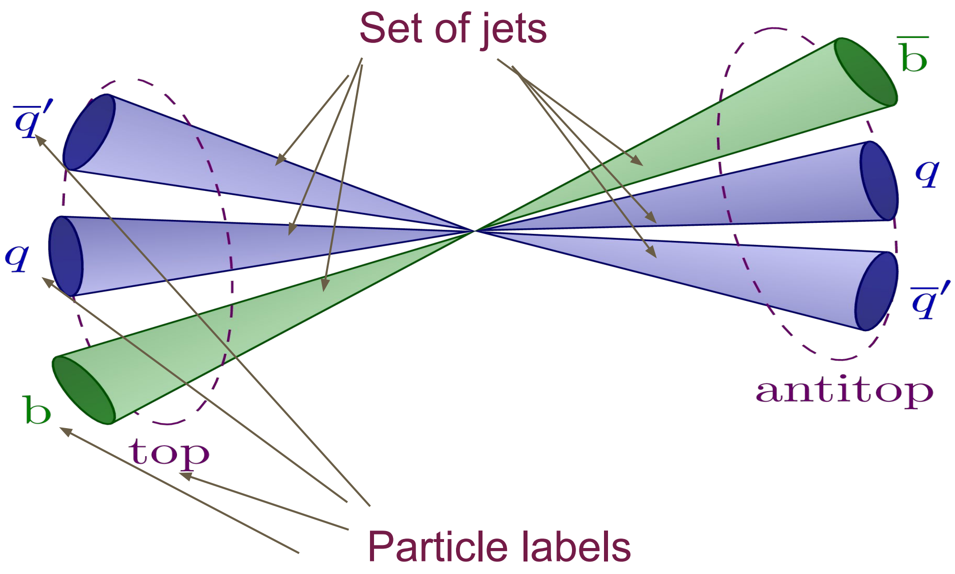
Denys Timoshyn

Supervisor at CU: Peter Berta

16 Sept 2024

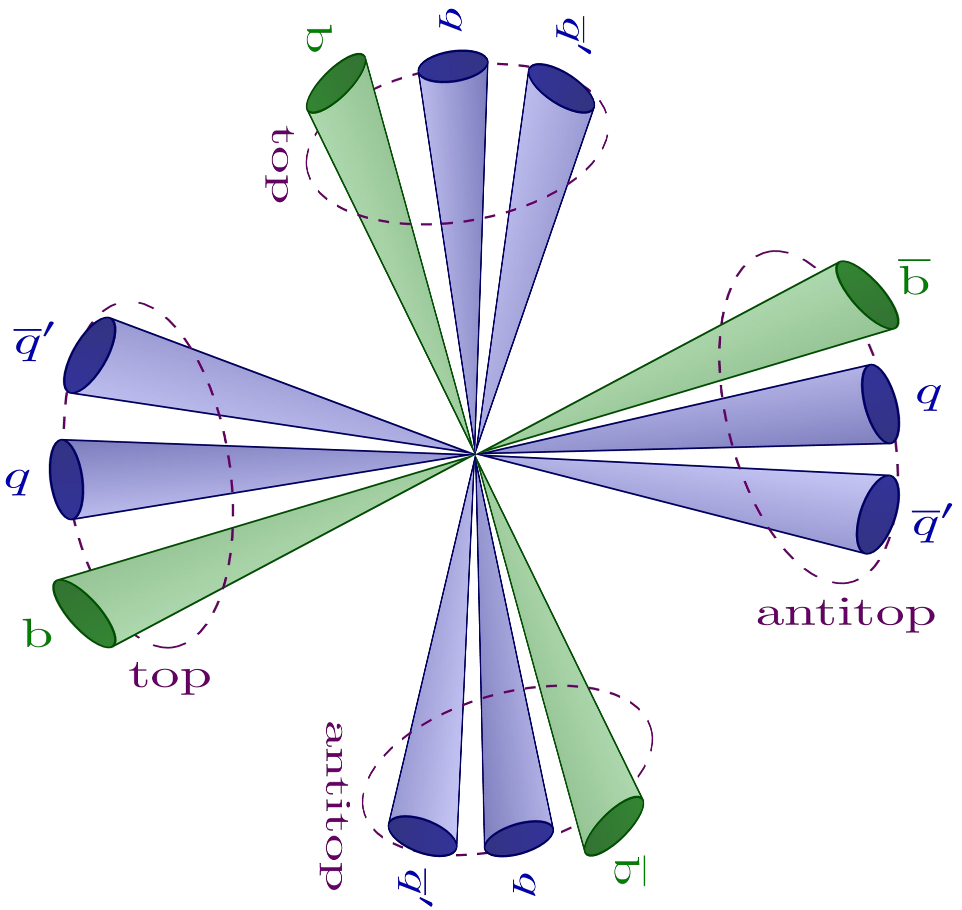
Particle physics objects

- Jets and leptons are the key detector signatures we observe at the end of particle collisions, helping us to identify what happened during the process
- A particle assignment to a jet set - i.e. a reconstruction of event - plays a significant role in analyses
- For low-jet multiplicity processes (pp->tt), signature is 2-6 jets, depending on decay channel
  - 6 jets lead to 90 combinations
- Pile-up, final state radiation and other underlying processes can increase final number of jets
- The methods: KLFitter, χ^2-minimisation, BDT can deal with them



High jet multiplicity process

- A process like all-hadronic four top quark production has a jet signature of 12, which leads to 1247400 combinations to check using χ^2-minimisation
  - There is a need for new reconstruction techniques!



SPANet

❖ **S**ymmetry **P**reserving **A**ttention **Net**work (arxiv:2106.03898) is a transformer based NN

➢ SPANet improves run-time performance over baseline permutation methods by avoiding having to construct all valid assignment permutations via special attention head

Training and prediction files format structures from SPANet github

Training input

Source:
[{E_0, η_0, sin ϕ_0, cos ϕ_0, b-tag$_0$}, {E_1, η_1, sin ϕ_1, cos ϕ_1, b-tag$_1$},... …],
[{E^{miss}_T, sin ϕ^{miss}, cos ϕ^{miss}}]
Targets:
t1:
q1: 0 <- Matched using
q2: -1 <- True level parton
b: 4 <- kinematics
t2:……..
t3:……..
t4:……..
Symmetries:
[t1,t2,t3,t4], [t1/q1, t1/q2] ,...

Prediction results

Targets:
t1:
q1: 0 <- predictions
q2: 3 <- predictions
b: 7 <- predictions
detection Probability
assignment Probability
marginal Probability
t2:……..
t3:……..
t4:……..

SPANet model - short summary of SPANet article arxiv:2106.03898

- Number of particles - m
- Each particle - k_p partons\products
- For each particle (p) model searches for best parton - jet assignment
- G_p is a permutation group of partons (b, [q1,q2] our case). σ are its elements
- Encoded jet tensor X_p has dimensions: Xp $\in R^N \otimes R^N \otimes ..(\text{D-times}) = R^{N \otimes D}$
  - N - total number of jets
  - D - hidden dimensions
- A Symmetric Tensor Attention a rank k_p $\Theta \in R^{D \otimes kp}$
- Due to summation over G_p permutations we have jet-symmetric outputs
- There is an implemented symmetry of particles via symmetrized loss function, where T_i are true assignments

$$\mathcal{L}_{min} = \min_{\sigma \in G_E} \sum_{i=1}^{m} CE(\mathcal{P}_i, \mathcal{T}_{\sigma(i)})$$

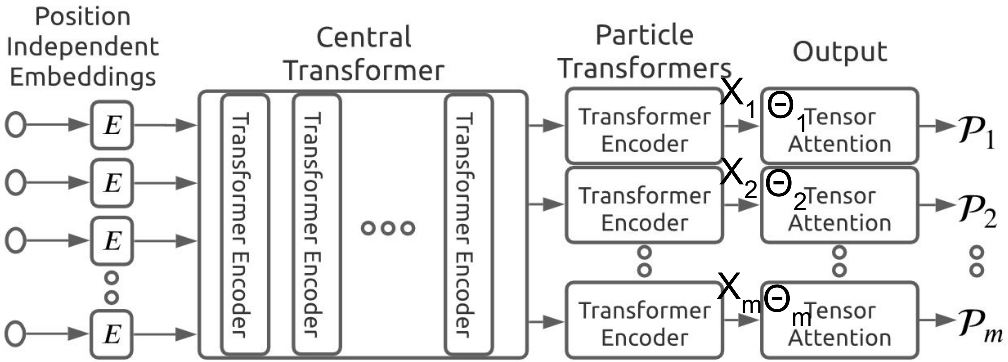


Figure 2: A visualization of the high level structure of SPA-NET.

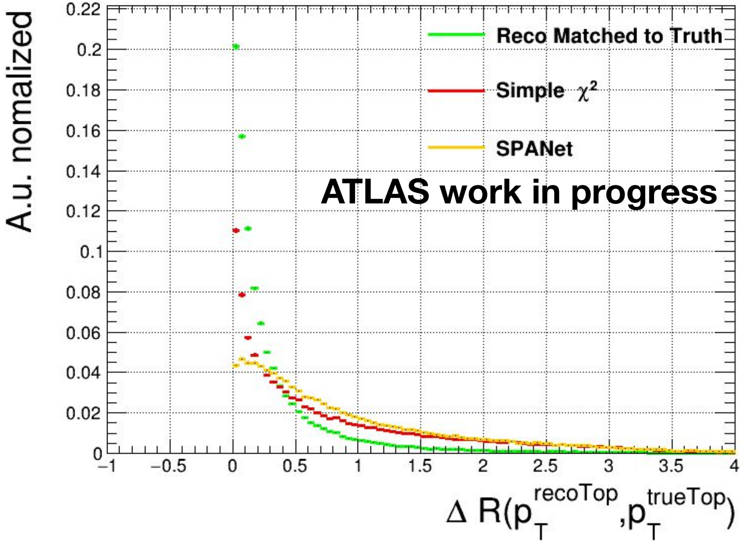
$$\mathcal{S}^{i_1 i_2 \dots i_{k_p}} = \sum_{\sigma \in G_p} \Theta^{i_{\sigma(1)} i_{\sigma(2)} \dots i_{\sigma(k_p)}},$$

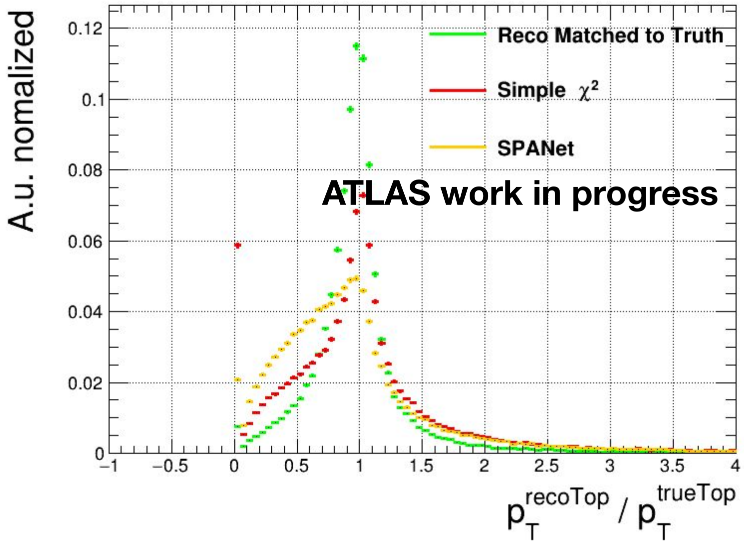
$$\mathcal{O}^{j_1 j_2 \dots j_{k_p}} = X_{i_1}^{j_1} X_{i_2}^{j_2} \dots X_{i_{kp}}^{j_{kp}} \mathcal{S}^{i_1 i_2 \dots i_{k_p}},$$

$$\mathcal{P}_p^{j_1 j_2 \dots j_{k_p}} = \frac{\exp(\mathcal{O}^{j_1 j_2 \dots j_{k_p}})}{\sum_{j_1, j_2, \dots, j_{kp}} \exp(\mathcal{O}^{j_1 j_2 \dots j_{k_p}})}$$

SPANet results for tttt sample

- SPANet was trained using tttt information
- Trained model can be used on process





- For each true top reconstructed top was found via smallest Delta R between. Pairwise Delta-R (angular distance) and transverse momenta ratios are shows

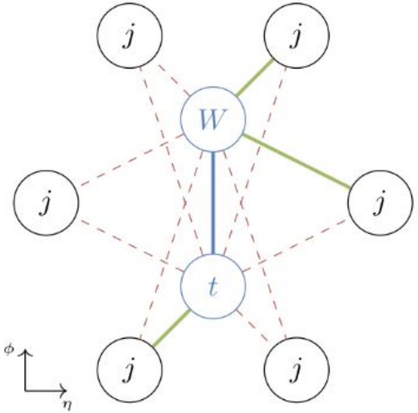
Summary

- High multiplicity processes can help getting extra precision on fundamental interactions
- Need appropriate tools to deal with combinatorial complexity
- SPANet provides a symmetry preserving approach
- SPANet models is also integrated into TopCPToolkit, a framework for producing Monte-Carlo simulations for studies

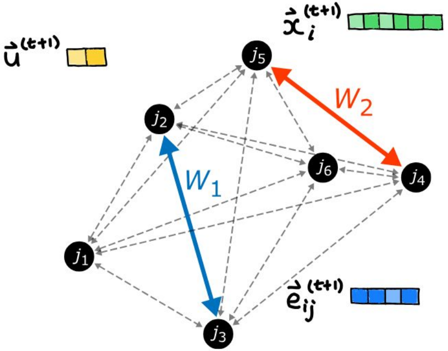
Backup

Other state of the art approaches

Topograph - a GNN
arxiv:2303.13937



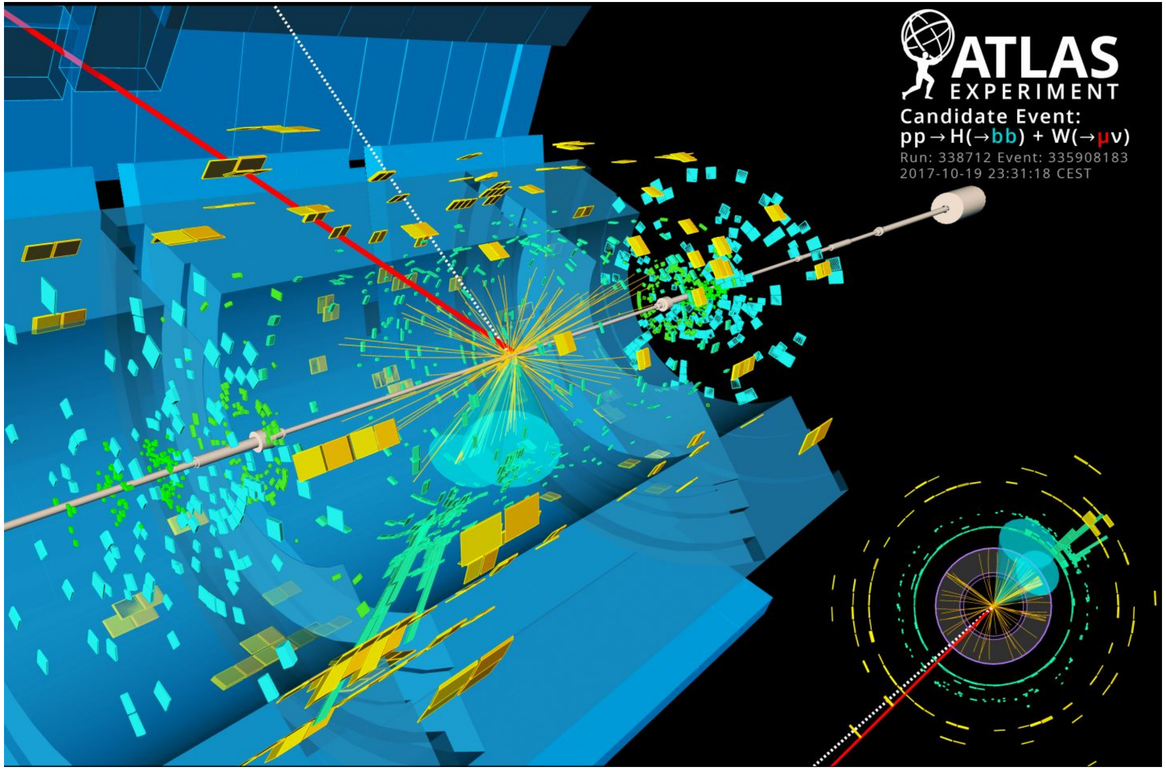
HyperGraph - a digraph NN structure
arxiv:2402.10149



HyperGraph presentation

Detector view of jets and leptons

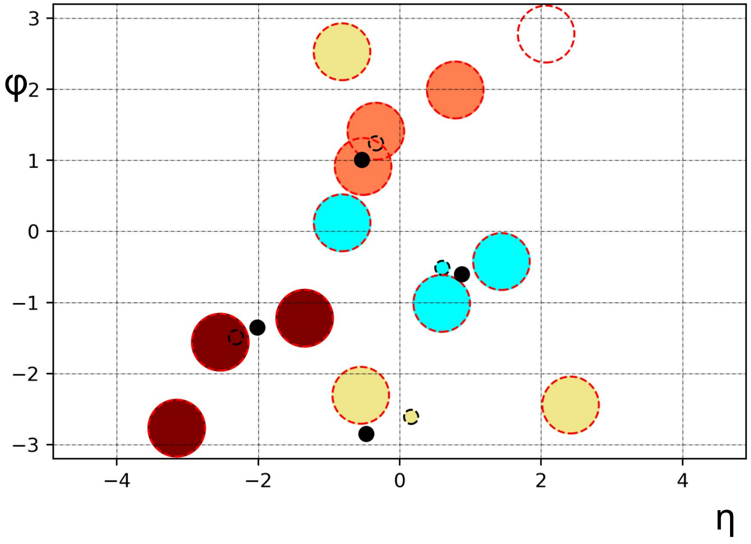
[ATLAS Event Displays]



Simple χ^2 tttt reconstruction

- Is a very combinatorially complex task due to high jet multiplicity
- Following objects become top candidates
  - All Large-R jet in event, tagged by Contained Top 50%, become top candidates
  - Each top candidate should have 3 small-R jets
    - Can have less as majority of events have less than 12 jets
- Candidates must satisfy following conditions
  - Reconstructed top mass should be in [130, 220] GeV mass range
  - If applicable:
    - 2 out of 3 jets should yield mass close to 80 GeV and the third is b-tagged with DL1 WP 77&
- From candidates, the final reconstructed tttt set is obtained via finding the set with minimum of metric:

$$\chi^{2,sum}_{reco} = \sum_{i=1}^{4} \frac{\chi^2_{t,i}}{30} + \frac{\chi^2_{W,i}}{20} + \frac{\chi^2_{b,i}}{10}$$



- true tops momenta vector position in (η, φ) plane

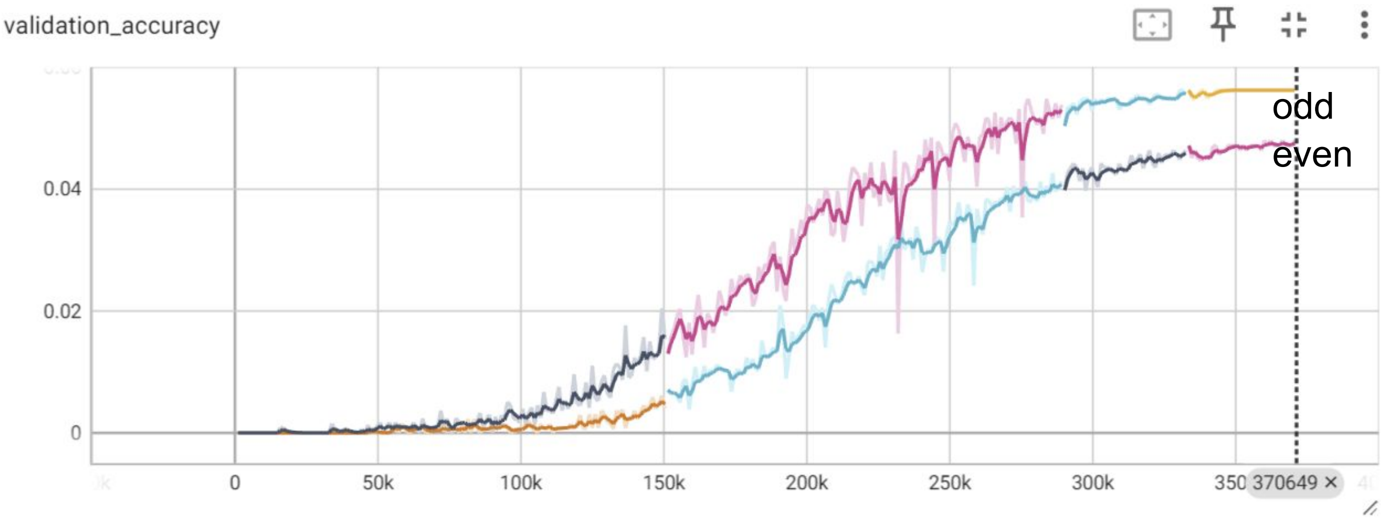
- detector level jets the make up one group - a reconstructed top

- reconstructed tops momenta vector

- detector level jets that are left out of reconstruction

SPANet training

- 28 hours on 1 GPU L40 NVIDIA for an even and an odd
- Batch size 4096
- Model has 832 K (Input embedder 38.2 K; Jet Encoder 63.3 K; Decoder 730 K) parameters with 4.5 entries per training (1000 step per epoch), 350 epochs
- Can state that training reached maximum performance for even
- Learning rate **0.003** for odd **0.0015** for even



SPANet results

- We can also see some separation in assignment probabilities for signal and background processes