Symmetry Preserving Attention Networks (SPANets) for Jet-Parton Matching

Presented by Alexander Shmakov

November 11^{th} 2021

Based on

SPANet: Generalized Permutationless Set Assignment for Particle Physics using Symmetry Preserving Attention
Alexander Shmakov, Michael James Fenton, Ta-Wei Ho, Shih-Chieh Hsu, Daniel Whiteson, Pierre Baldi
https://arxiv.org/abs/2106.03898

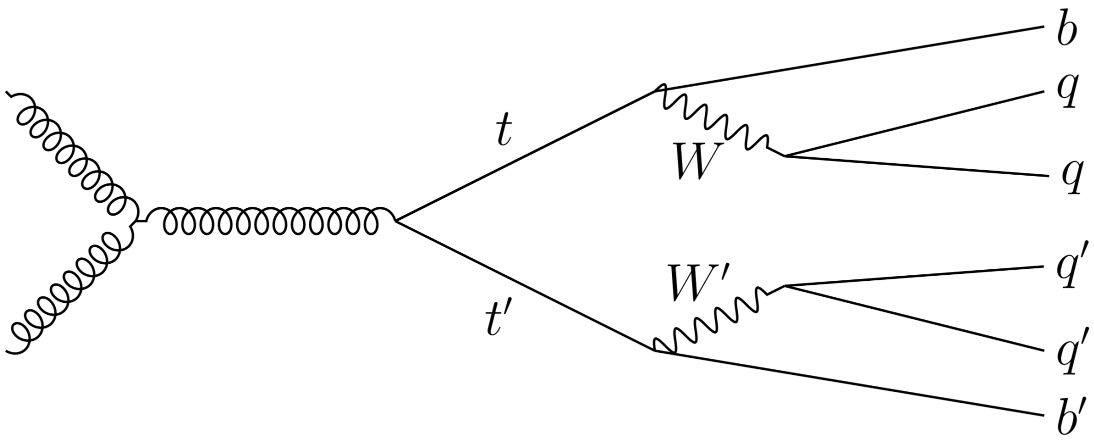
Permutationless Many-Jet Event Reconstruction with Symmetry Preserving Attention Networks
Michael James Fenton, Alexander Shmakov, Ta-Wei Ho, Shih-Chieh Hsu, Daniel Whiteson, Pierre Baldi
https://arxiv.org/abs/2010.09206

and ongoing work by

Michael James Fenton, Alexander Shmakov, Hideki Okawa, Shih-Chieh Hsu, Daniel Whiteson, Pierre Baldi

Overview Jet-Parton Matching $t\bar{t}$ Events

- Primary (all-hadronic) decay channel produces six particles: two qqb triplets with opposite charge.
- After these particles are produced, they are showered and measured as four-momentum *jets*.
- Along with the jets from each of the particles, there may be additional jets in the signal.



Unsorted List of Jets

$$\{j_1, j_2, j_3, j_4, j_5, j_6, j_7, j_8\}$$

Match Jets to Particle Labels

$$\{b, q', \emptyset, q', b', \emptyset, q, q\}$$

Targets

Garbage Jets

Overview Set Assignment

This modeling task may be generalized as a set assignment problem.

Input is a set of size N

$$I = \{j_1, j_2, \ldots, j_N\}$$

Possible Assignments are a set of size $C \leq N$ and a special null assignemnt $\emptyset$

$$T = \{\emptyset, t_1, t_2, \ldots, t_C\}$$

Output is a predicted assignment set of size N

with each $\boldsymbol{p} \in \boldsymbol{T}\ \boldsymbol{s.t.}\ \boldsymbol{p_i} \neq \boldsymbol{p_j}\ \boldsymbol{or}\ \boldsymbol{p_i} = \emptyset$

$$\{p_1, p_2, \ldots, p_N\}$$

Set Assignment Basic Approached

Itemized Approach – Simply train a classification model to predict the assignment.

- How do you prevent two identical targets being predicted? Remove after selection?
- How do you pick which order to go through the targets?
- The network has no signal on the uniqueness of targets.

Permutation Approach – Construct all possible jet permutations and rank each one.

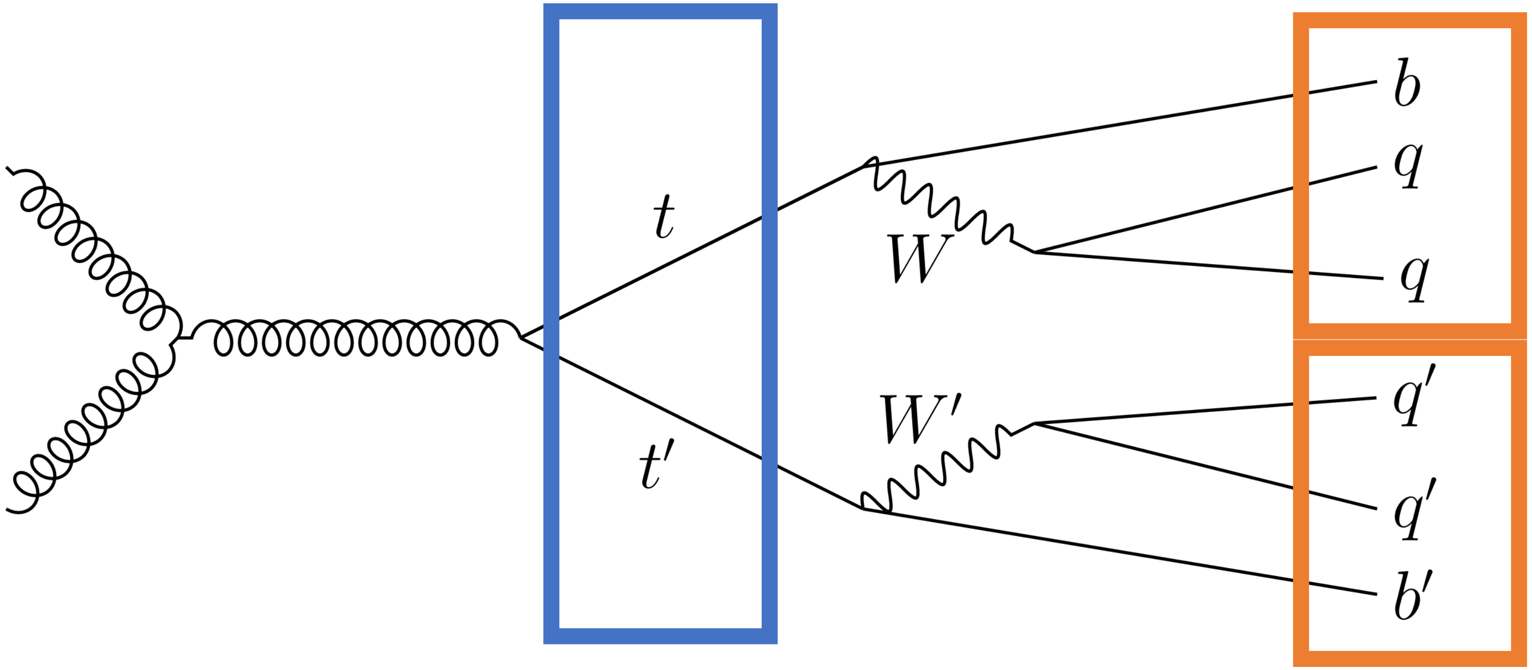
- A good approach for incorporating symmetries and uniqueness.
- Used in existing methods such as χ^2 or BDTs.

- **However,** we need to generate every permutation! Runtime is $O(N^C)$.
- Ranking function may find it difficult to distinguish similar configurations.
- Need to handle variable length inputs in a meaningful way, how do we pick an input order?

Set Assignment **A Combined Approach**

The first improvement in SPANet is to merge these two approaches to get the best of both.

Output **independent sub-permutations scores** for each particle and learn to **differentiate sub-permutations with classification**.



Jet Permutation Stage

Jet Permutation Stage

Particle Classification Stage

Symmetry Target Symmetries

One very interesting property of Feynman Diagram matching is the presence of symmetries.
The following target sets are equivalent due to charge symmetry.

$$\boldsymbol{q_1 q_2} b q_1' q_2' b' \leftrightarrow \boldsymbol{q_2 q_1} b q_1' q_2' b'$$

$$q_1 q_2 b \boldsymbol{q_1' q_2'} b' \leftrightarrow q_1 q_2 b \boldsymbol{q_2' q_1'} b'$$

We call these **jet symmetries** – the light quarks can be freely rearranged. This will be handled with **attention**.

$$\underbrace{q_1 q_2 b}_{\mathcal{T}_1} \underbrace{\boldsymbol{q_1' q_2' b'}}_{\mathcal{T}_2} \leftrightarrow \underbrace{\boldsymbol{q_1' q_2' b'}}_{\mathcal{T}_2} \underbrace{q_1 q_2 b}_{\mathcal{T}_1}$$

We call this **particle symmetry** – the two top quarks cannot be differentiated from each other with kinematic measurements alone. This will be handled with a **special loss function**.

Note: this is not the same as allowing duplicate targets as the jet **groupings must permute together**.

$$\boldsymbol{q_1 q_2 b q_1'} q_2' b' \neq \boldsymbol{q_1' q_2 b q_1} q_2' b'$$

Symmetry Input Permutation Equivariance

- Another important symmetry permutation invariance on the input.
- We want to ensure that our output matches the order of our input.
- This must work for **any initial ordering** on the input jets and **for any number of jets**.

$$\{j_1, j_2, j_3, j_4, j_5, j_6, j_7, j_8\} \cong \{j_3, j_7, j_1, j_2, j_8, j_4, j_6, j_5\}$$
$$\{b, q', \emptyset, q', b', \emptyset, q, q\} \cong \{\emptyset, q, b, q', q, q', \emptyset, b'\}$$

One common approach is to enforce a consistent ordering, for example sort the jets by p_T.
However, we can avoid fixing an order if we just use a permutation equivariant architecture.

Transformer Attention

Attention Overview

Vectors

Best understood as a *continuous, differentiable* **key-value database**

$$Q = \{q_1, q_2, \ldots, q_m\} \quad \text{QUERIES}$$
$$K = \{k_1, k_2, \ldots, k_n\} \quad \text{KEYS}$$
$$V = \{v_1, v_2, \ldots, v_n\} \quad \text{VALUES}$$

Pick a **SIMILARITY** function. Compute and normalize similarity between all query-key pairs to make **attention weights**.

$$S_{ij} = \text{SIMILARITY}(q_i, k_j)$$

$$W_{ij} = \text{NORMALIZE}(S_{ij}) = \frac{e^{S_{ij}}}{\sum_{l=1}^{n} e^{S_{il}}}$$

Output is the **weighted average** of all values weighted by similarity.

$$O_i = W_{ij} V^j$$

(Abusing Einstein notation a bit)

Attention Self-Attention

$$Q = f_{\theta_q}(X)$$
$$K = f_{\theta_k}(X)$$
$$V = f_{\theta_v}(X)$$

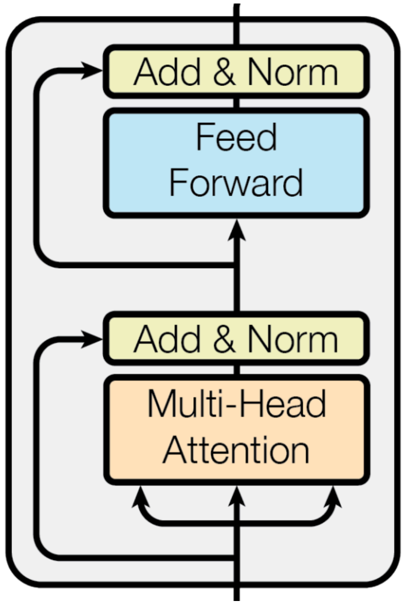
One Interesting case of attention is **Self-Attention**, where the queries, keys, and values are simply **functions of the same set vectors.**

This is leveraged to learn contextual, pair-wise relationships within a set of vectors.

Cheng, Jianpeng, et al. "Long Short-Term Memory-Networks for Machine Reading." Sept. 2016.

Attention Transformers

$$\textsc{Similarity}(q_i, k_j) = \frac{q_i \cdot k_j}{\sqrt{D}}$$



Self-attention[1] with ***scaled dot-product*** as the similarity measure.

The **transformer encoder**[2] combines

- Scaled dot-product attention
- Skip-connections
- Layer Normalization
- Position-independent feed-forward layers.

Transformers are permutation equivariant on their input!

1. The full transformer uses "multi-head" self-attention, but this is conceptually equivalent for our purposes.
2. Vaswani, Ashish, et al. "Attention Is All You Need." Dec. 2017.

Tensor Attention Producing Jet Permutation Rankings

We can also use attention to produce **joint distributions** for over jets.
Generalization of dot-product attention: **Symmetric Tensor Attention**

Suppose $\boldsymbol{X}$ is our list of vectors. This can be viewed as a $(1{,}1)$-tensor with ranks (N, D). We want to create a ranking over K-groups of vectors.

Store $\boldsymbol{\Theta}$: a $(0, K)$-tensor of **learnable weights** with rank $(D, D, \dots, D)$.

$$O^{j_1 j_2 \cdots j_N} = X^{j_1}_{n_1} X^{j_2}_{n_2} \dots X^{j_N}_{n_N} \Theta^{n_1 n_2 \dots n_N}$$

1. Perform generalized dot-product self-attention on $\boldsymbol{X}$ with the mixing weights $\boldsymbol{\Theta}$ to produce attention weights $\boldsymbol{O}$.
2. Normalize $\boldsymbol{O}$ to create a valid joint distribution $\boldsymbol{P}$ over K-groups of vectors.

$$\mathcal{P}^{j_1 j_2 \cdots j_N} = \frac{\exp O^{j_1 j_2 \cdots j_N}}{\sum \exp O}$$

Tensor Attention Incorporating Group Symmetries

Suppose our vector scores obey additional **symmetries**. We encode this as a permutation group on the indices of $\boldsymbol{\Theta}$

Suppose $\boldsymbol{G_P} \subseteq \boldsymbol{S_K}$ is a permutation group acting on the jet assignments $\{J_1, J_2, \ldots, J_K\}$ associated with particle $\boldsymbol{P}$.

1. Create an augmented symmetric weights tensor $\boldsymbol{S}$ by summing over the symmetric indices of $\boldsymbol{\theta}$ according to $\boldsymbol{G_P}$.
2. Perform tensor attention as before with this new symmetric parameter tensor $\boldsymbol{S}$.

$$S^{i_1 i_2 \cdots i_K} = \sum_{\sigma \in G_P} \Theta^{i_{\sigma(1)} i_{\sigma(2)} \cdots i_{\sigma(K)}}$$

$$O^{j_1 j_2 \cdots j_K} = X_{i_1}^{j_1} X_{i_2}^{j_2} \ldots X_{i_K}^{j_K} S^{i_1 i_2 \cdots i_K}$$

$$\mathcal{P}^{j_1 j_2 \cdots j_K} = \frac{\exp\left(O^{j_1 j_2 \cdots j_K}\right)}{\sum_{j_1, j_2, \ldots, j_K} \exp\left(O^{j_1 j_2 \cdots j_K}\right)}$$

Tensor Attention $t\bar{t}$ Example

For full hadronic $t\bar{t}$ events, we want to score possible qqb triplets ($K = 3$) associated with each top quark.

Suppose $\boldsymbol{X}$ stores our jets after the transformer phases. We interpret these encoded jets as a $(1,1)$-tensor with rank (N, D).

Suppose $\boldsymbol{\Theta}$ is a $(0,3)$-tensor of learnable weights with rank (D, D, D). D may be chosen arbitrarily, and we use $D = 128$ in our experiments.

$$O^{ijk} = X_n^i X_m^j X_l^k \Theta^{nml}$$

1. Perform generalized dot-product self-attention on $\boldsymbol{X}$ with the mixing weights $\boldsymbol{\Theta}$ to produce attention weights $\boldsymbol{O}$.
2. Normalize $\boldsymbol{O}$ to create a valid joint distribution $\boldsymbol{P}$ over **triplets** of vectors.

$$\mathcal{P}^{ijk} = \frac{\exp O^{ijk}}{\sum \exp O}$$

Tensor Attention Symmetric $t\bar{t}$ Example

Each $q_1 q_2 b$ triplet obeys a charge symmetry on the light quarks. Such a permutation group may be generated by the transposition $(q_1 q_2)$

Suppose $\boldsymbol{G_t} = \langle (q_1 q_2) \rangle \subseteq \boldsymbol{S_3}$ is a permutation group acting on the jet classes $\{q_1, q_2, b\}$ associated with particle $\boldsymbol{t}$.

1. Create an augmented symmetric weights tensor $\boldsymbol{S}$ by summing over the symmetric indices of $\boldsymbol{\theta}$ according to $\boldsymbol{G_P}$. In this case, simply ensure that parameter $\boldsymbol{S}$ is commutative in the first two axes.
2. Perform tensor attention as before with this new symmetric parameter tensor $\boldsymbol{S}$.

$$S^{i_1 i_2 i_3} = \Theta^{i_1 i_2 i_3} + \Theta^{i_2 i_1 i_3}$$

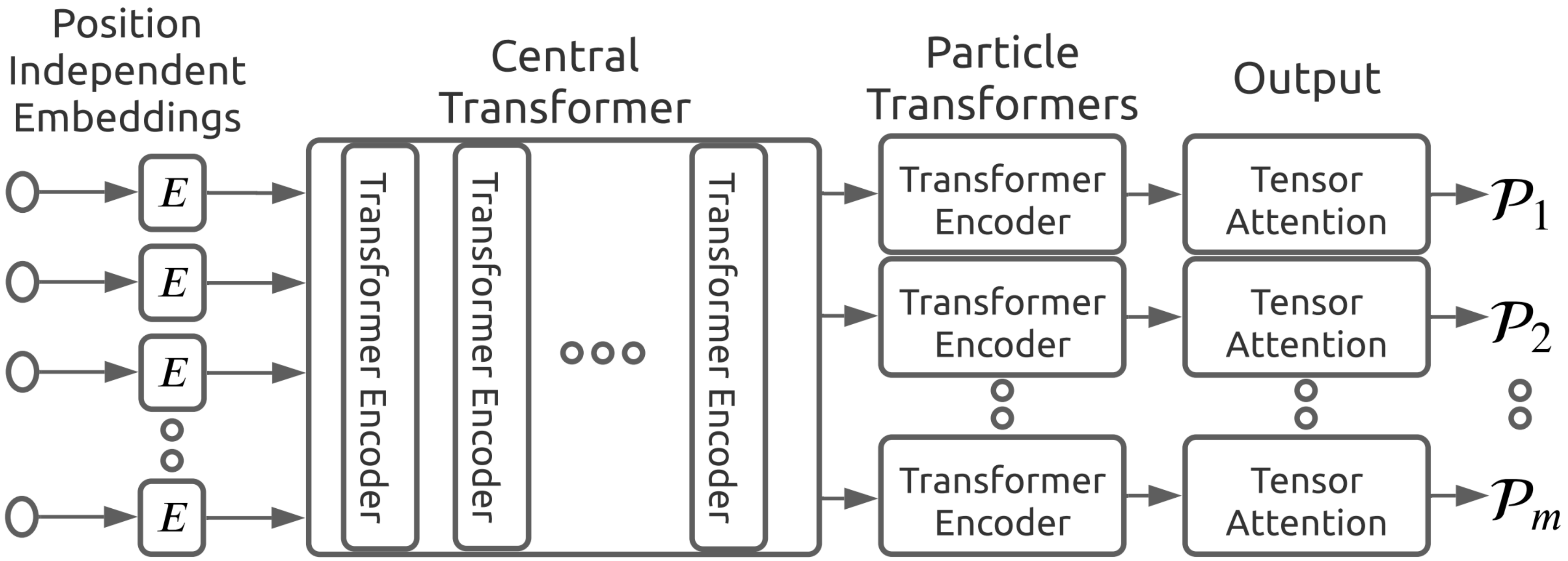
$$O^{j_1 j_2 j_3} = X_{i_1}^{j_1} X_{i_2}^{j_2} X_{i_3}^{j_3} S^{i_1 i_2 i_3}$$

$$\mathcal{P}^{j_1 j_2 j_3} = \frac{\exp\left(O^{j_1 j_2 j_3}\right)}{\sum_{j_1, j_2, j_3} \exp\left(O^{j_1 j_2 j_3}\right)}$$

SPANet Architecture

Split the information stream into a finite collection of particles.

Input a set of 4-momentum vectors.



Output distributions predicting the most likely permutation of jets associated with each particle.

Event-level context-aware encoding

Particle-level encoding

Symmetric jet matching

SPANet Training the Permutation Ranker

We output a joint distribution matrix summing to 1, representing the networks belief (score) that the given permutation is the correct assignment.

Each particle may be trained using regular **Cross-Entropy Loss.**

$$\mathcal{L}_P(\mathcal{P}, \mathcal{T}) = \sum_{j_1, j_2, \cdots, j_N} -\mathcal{T}^{j_1 j_2 \cdots j_N} \log \mathcal{P}^{j_1 j_2 \cdots j_N}$$

SPANet Combining Particle Loss with Symmetries

If we had two independent particles, we could simply sum the two particle losses and train both outputs simultaneously.

However, we have additional **particle symmetries**, which we encode as a **particle-level symmetry group** $G_E \subseteq \boldsymbol{S}_m$ which acts on our resonance particles $\{P_1, P_2, \dots, P_m\}$ in a similar way to the jet symmetry groups.

To ensure that the network may jointly swap any two symmetric particles, set the loss to the **minimum** achievable loss on all possible permutations of our particle targets.

$$\mathcal{L} = \min_{\sigma \in G_E} \sum_{i=1}^{m} \mathcal{L}_P \left(\mathcal{P}_{\sigma(i)}, \mathcal{T}_{\sigma(i)} \right)$$

SPANet Training on Partial Events

By constructing such a separatable loss function, we can also better use existing data by training on **partial events** where one or more of the jets are unable to be reconstructed.

As long as **at least one complete resonance particle** is present, we can use the loss from that particle, along with symmetric alternatives, to still recover a training signal. We mark reconstructable particles with a masking tensor $\boldsymbol{M}$

$$\mathcal{L}^{masked} = \min_{\sigma \in G_E} \left(\sum_{i=1}^{m} \mathcal{M}_{\sigma(i)} \mathcal{L}_P(\mathcal{P}_i, \mathcal{T}_{\sigma(i)}) \right)$$

Very important for larger events, where complete events are rare!

SPANet $t\bar{t}$ Example

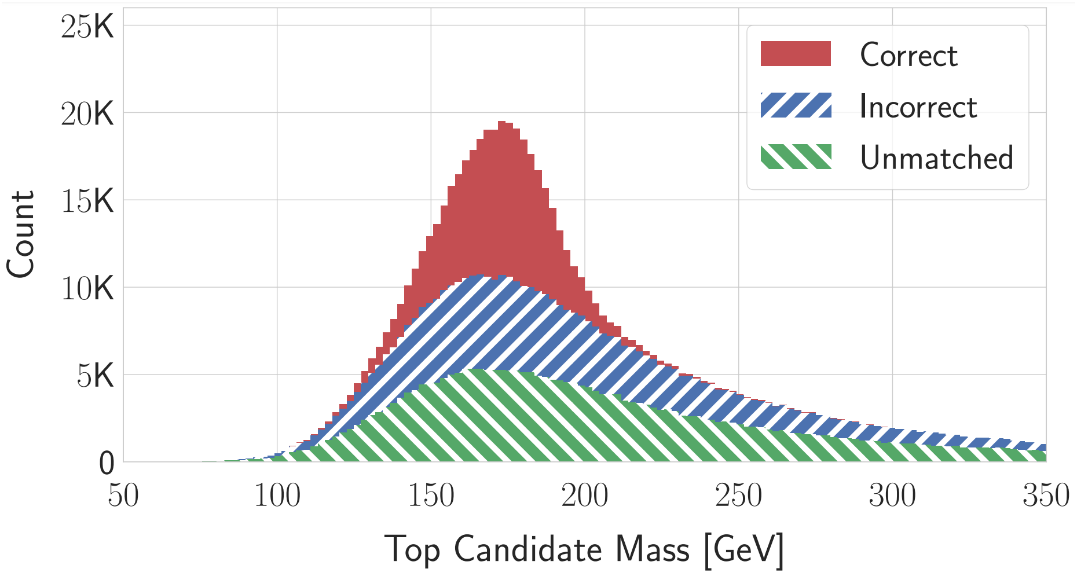
Our particle symmetry group for $t\bar{t}$ is the entire $S_2 = \langle (t\ \bar{t}) \rangle = G_E$.
Our symmetric loss term becomes.

$$\mathcal{L} = \min\left\{\mathcal{L}_P\left(\mathcal{P}_t, \mathcal{T}_t\right) + \mathcal{L}_P\left(\mathcal{P}_{\bar{t}}, \mathcal{T}_{\bar{t}}\right), \mathcal{L}_P\left(\mathcal{P}_t, \mathcal{T}_{\bar{t}}\right) + \mathcal{L}_P\left(\mathcal{P}_{\bar{t}}, \mathcal{T}_t\right)\right\}$$

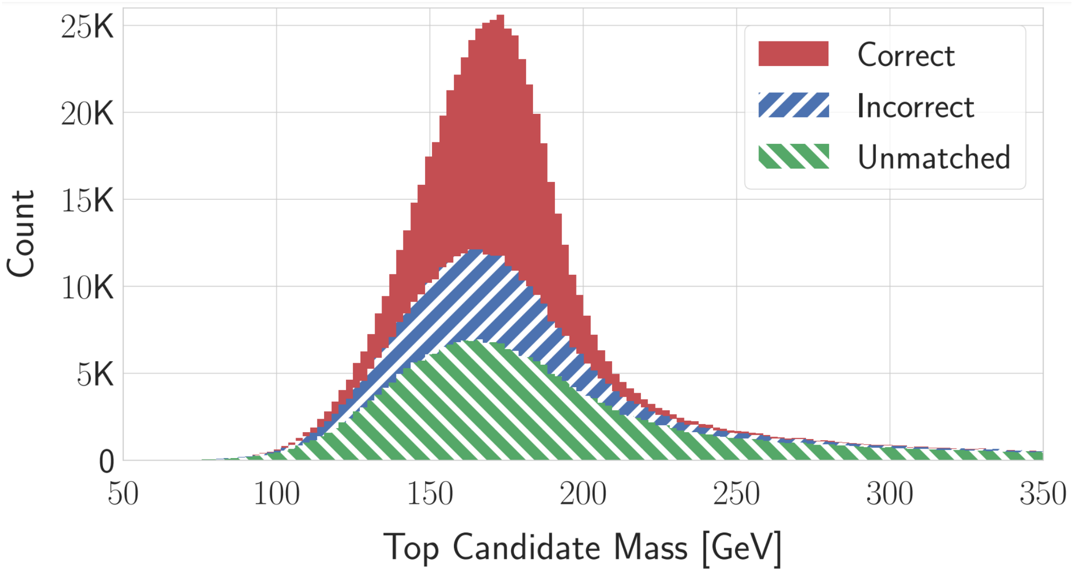
This way, our network will not be penalized for a correct jet assignment with an incorrect charge.

SPANet Results – ATLAS $t\bar{t}$

	N_{jets}	Event Fraction	SPA-NET Efficiency Event	SPA-NET Efficiency Top Quark	χ^2 Efficiency Event	χ^2 Efficiency Top Quark
All Events	== 6	0.245	0.643	0.696	0.461	0.523
	== 7	0.282	0.601	0.667	0.408	0.476
	≥ 8	0.320	0.528	0.613	0.313	0.395
	Inclusive	**0.848**	**0.586**	**0.653**	**0.387**	**0.457**
Complete Events	== 6	0.074	0.803	0.837	0.664	0.696
	== 7	0.105	0.667	0.754	0.457	0.556
	≥ 8	0.145	0.521	0.662	0.281	0.429
	Inclusive	**0.325**	**0.633**	**0.732**	**0.426**	**0.532**



χ^2



SPANet

SPANet Results – ATLAS Challenge Events ttH & $tttt$

ttH

Testing different particle types with more complex symmetries

	N_{jets}	Event Fraction	SPA-NET Efficiency Event	Higgs	Top	χ^2 Efficiency Event	Higgs	Top
All Events	== 8	0.261	0.370	0.497	0.540	0.056	0.193	0.092
	== 9	0.313	0.343	0.492	0.514	0.053	0.160	0.102
	≥ 10	0.313	0.294	0.472	0.473	0.031	0.150	0.056
	Inclusive	**0.972**	**0.330**	**0.485**	**0.502**	**0.045**	**0.164**	**0.081**
Complete Events	== 8	0.042	0.532	0.657	0.663	0.040	0.220	0.135
	== 9	0.070	0.422	0.601	0.596	0.019	0.152	0.079
	≥ 10	0.115	0.306	0.545	0.523	0.004	0.126	0.073
	Inclusive	**0.228**	**0.383**	**0.583**	**0.572**	**0.016**	**0.153**	**0.087**

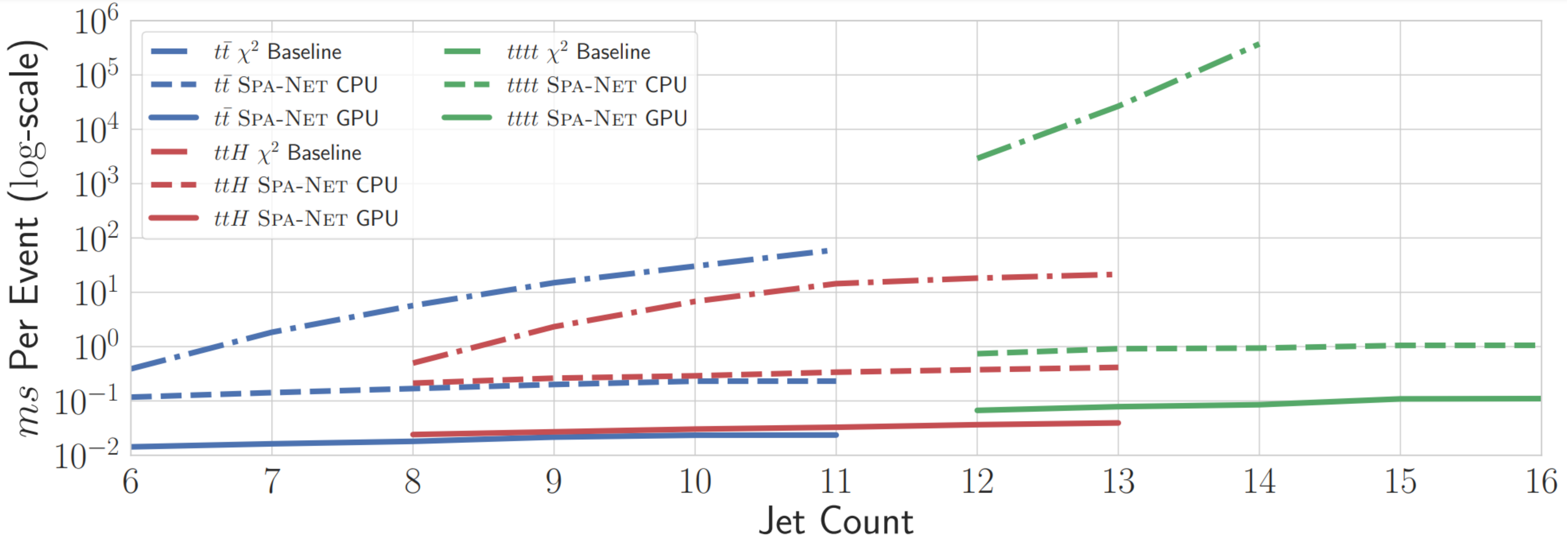
$tttt$

Testing extreme event sizes with at least 13 jets in most events.

	N_{jets}	Event Fraction	SPA-NET Efficiency Event	Top Quark
All Events	== 12	0.219	0.276	0.484
	== 13	0.304	0.247	0.474
	≥ 14	0.450	0.198	0.450
	Inclusive	**0.974**	**0.231**	**0.464**
Complete Events	== 12	0.005	0.350	0.617
	== 13	0.016	0.249	0.567
	≥ 14	0.044	0.149	0.504
	Inclusive	**0.066**	**0.191**	**0.529**

SPANet Results – Performance

SPANet is able the achieve these reconstruction efficiencies while being **up to 3 orders of magnitude faster** than baseline methods, with further potential to accelerate computation on GPUs.



SPANet Results – Transfer Learning $\mathrm{e}^-\mathrm{e}^+ \rightarrow t\bar{t}$

Preliminary results show that knowledge may be quickly transferred between different detectors & processes so long as the event topology remains the same

We take the pre-trained model from before, trained on 5 Million ATLAS $t\bar{t}$ events.

- Simply applying the ATLAS model directly works but not very efficient.
- We can significantly reduce required event count by transfer learning.
- More events allowed better results but require much fewer events than training from scratch.
- Transfer learning is very fast, only 10 minutes on a single GPU.

	ϵ^{event}	ϵ_2^{top}	ϵ_1^{top}
Direct ATLAS Model	31.7%	36.4%	15.4%
Transfer Learning on **10K** $\mathrm{e}^-\mathrm{e}^+$ events	69.0%	71.1%	42.2%
Transfer Learning on **100K** $\mathrm{e}^-\mathrm{e}^+$ events	77.2%	78.8%	49.3%

We thank Gang Li and Qiang Li for their feedback on CEPC Delphes to generate e^-e^+ events.

SPANet Library

We’ve open sources all our techniques so that you can start applying to your events and experiments!

https://github.com/Alexanders101/SPANet

Included is a full guide on how to run SPANets on $t\bar{t}$ events and a general configuration guide for any event topology.

https://github.com/Alexanders101/SPANet/blob/master/docs/TTBar.md

SPANet Event Configuration

```
[SOURCE]
mass = log_normalize
pt = log_normalize
eta = normalize
phi = normalize
btag = none

[EVENT]
particles = (t1, t2)
permutations = [(t1, t2)]

[t1]
jets = (q1, q2, b)
permutations = [(q1, q2)]

[t2]
jets = (q1, q2, b)
permutations = [(q1, q2)]
```

Configure your event by specifying the event permutation groups mentioned before and the network construction will be automated according to the specification.

$\boldsymbol{G_E}$ Particle Symmetry Group

$\boldsymbol{G_{t_1}}$

$\boldsymbol{G_{t_2}}$

Jet Symmetry Groups

SPANet Summary