The Machine Learning Landscape of Top Taggers

G. Kasieczka (ed)[1], T. Plehn (ed)[2], A. Butter[2], K. Cranmer[3], D. Debnath[4], B. M. Dillon[5], M. Fairbairn[6], W. Fedorko[7], D. A. Faroughy[5], C. Gay[7], L. Gouskos[8], J. F. Kamenik[5,9], P. T. Komiske[10], S. Leiss[1], A. Lister[7], S. Macaluso[3,4], E. M. Metodiev[10], L. Moore[11], B. Nachman,[12,13], K. Nordström[14,15], J. Pearkes[7], H. Qu[8], Y. Rath[16], M. Rieger[16], D. Shih[4], J. M. Thompson[2], and S. Varma[6]

arXiv:1902.09914

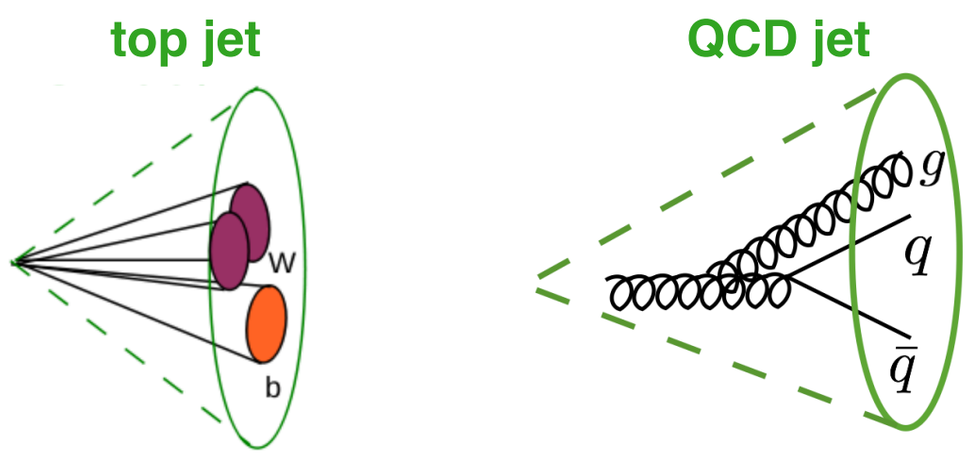
Sebastian Macaluso

New York University

BOOST 2019

July 23, 2019

Jet tagging



- Typical classifiers search for:
  - ✦ Kinematic features induced by the top and W boson mass.
  - ✦ Number of prongs: N-subjettiness.
  - ✦ Flavor: b-tagging

- Goal of this study:

Compare ML based setups to classify top vs QCD jets.

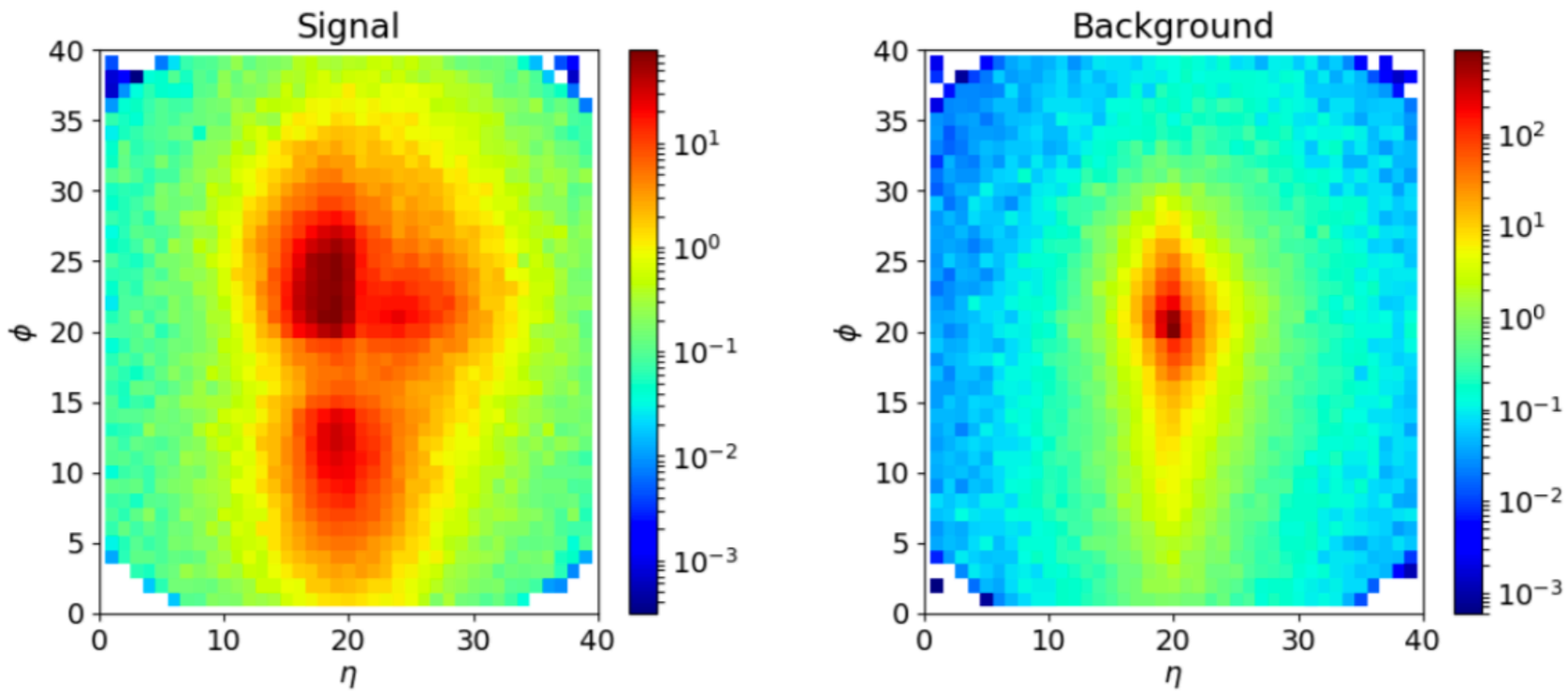
Top Tagging Reference Dataset

- **Top vs QCD jets** produced with Pythia8 @ 14TeV.
- Anti-kt R=0.8 Delphes jets (energy-flow algorithm), pT=[550,650] GeV, match and merge requirements.
- No uncertainties, pile-up.
- Only 4-vectors up to 200 constituents. (No charge or displaced vertex information)
- (1.2M, 400k, 400k) jets as (train, val, test) sets.

Contact

Gregor Kasieczka (gregor.kasieczka@cern.ch)
Michael Russel (russell@thphys.uni-heidelberg.de)
Tilman Plehn (plehn@uni-heidelberg.de)

Get dataset here



Average of 10k jet images

Taggers

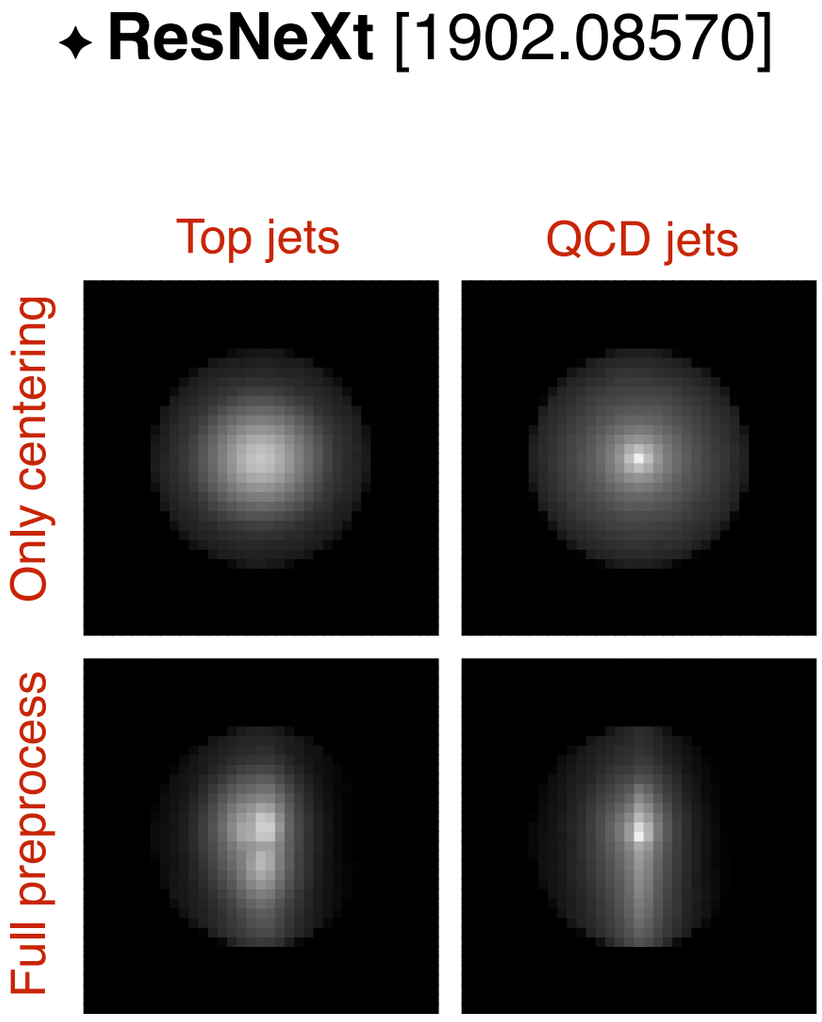
- **Image-based taggers**
  - CNN
  - ResNeXt

- **4-Vector-based taggers**
  - TopoDNN
  - Multi-Body N-Subjettiness
  - TreeNiN
  - P-CNN
  - ParticleNet

- **Theory-inspired taggers**
  - Lorentz Boost Network
  - Lorentz Layer
  - Latent Dirichlet Allocation (LDA)
  - Energy Flow Polynomials
  - Energy Flow Networks
  - Particle Flow Networks

Image-based taggers

- **Image-based taggers**
  - **CNN** [1803.00107]
  - **ResNeXt** [1902.08570]



Average of 100k grayscale jets
Fig. from arXiv:1803.00107

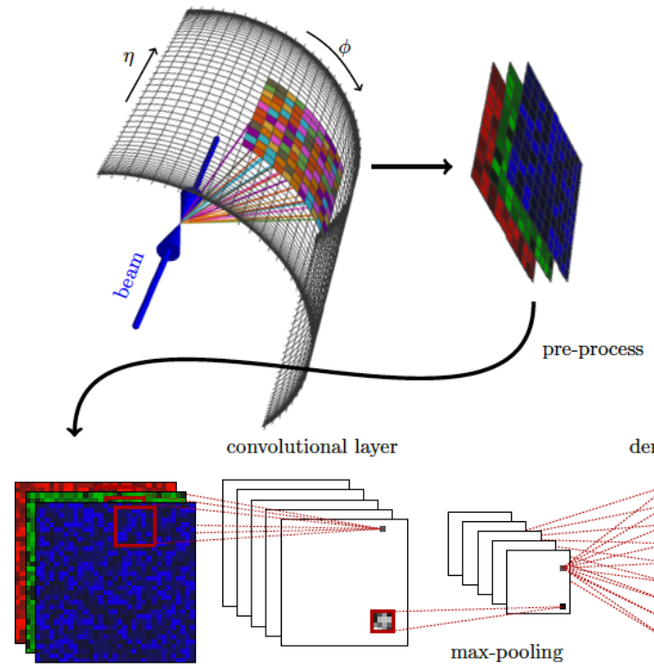
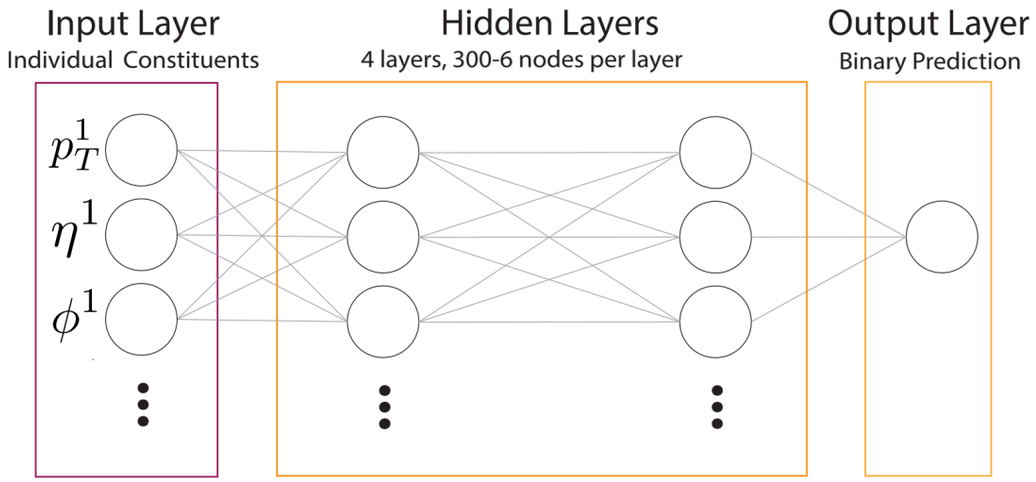


Fig. from [Komiske, Metodiev & Schwartz '16]

- **Features and problems**
  - Implement locality
  - Add multiple colors [arXiv:1612.01551]
  - Sparsity of jet images
  - Pixel resolution
  - Preprocessing (center,rotate,flip,normalize)

4-Vector-based taggers

TopoDNN [1704.02124]



ParticleNet
[1902.08570]
Jets as particle clouds.
Graph for each jet.
(See Huilin Qu talk!)

TreeNiN

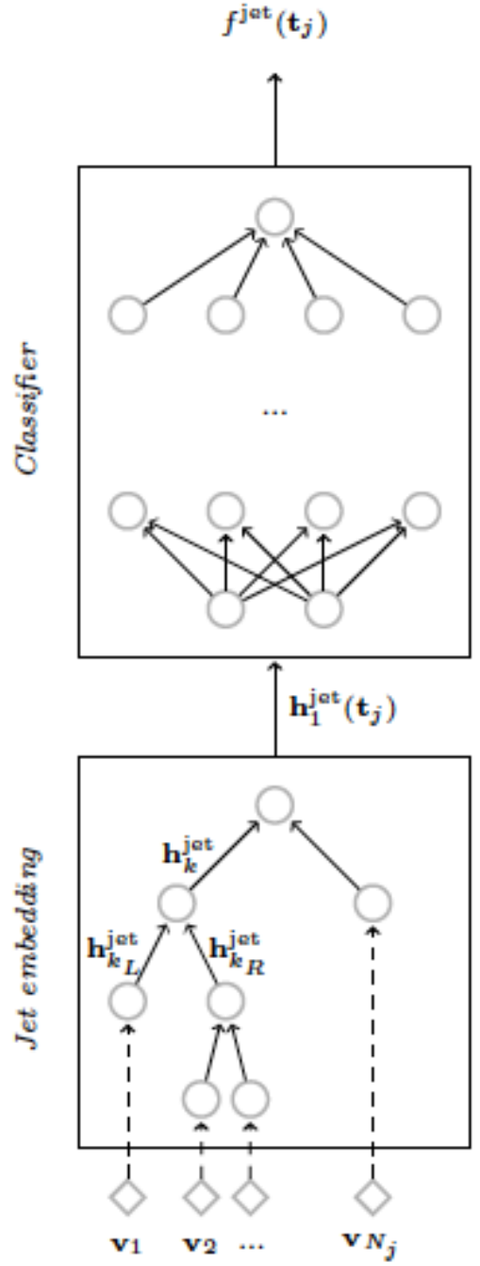


Fig. from arXiv:1702.00748

Multi-Body N-Subjettiness [1807.04769]

$$\left\{\tau_1^{(0.5)}, \tau_1^{(1)}, \tau_1^{(2)}, \tau_2^{(0.5)}, \tau_2^{(1)}, \tau_2^{(2)}, \ldots, \tau_{M-2}^{(0.5)}, \tau_{M-2}^{(1)}, \tau_{M-2}^{(2)}, \tau_{M-1}^{(1)}, \tau_{M-1}^{(2)}\right\}$$

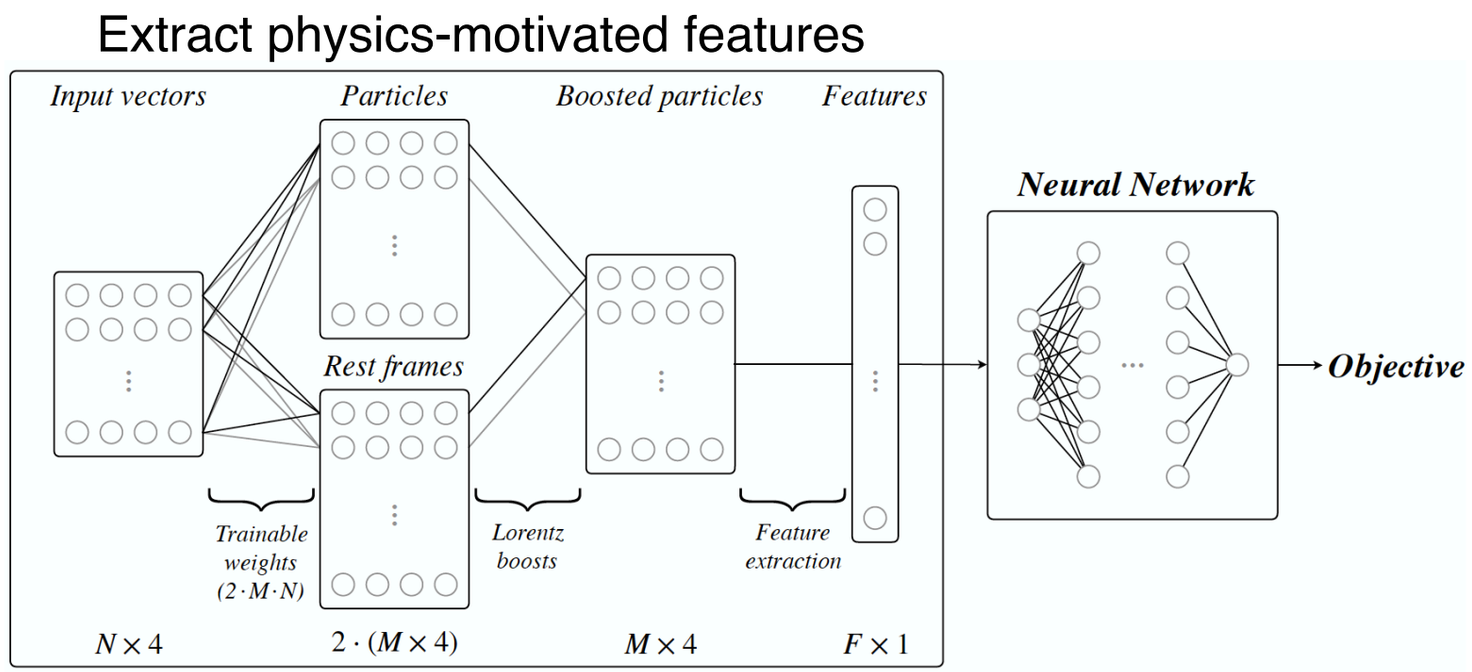
$$\tau_N^{(\beta)} = \frac{1}{p_{T,J}} \sum_{i \in J} p_{T,i} \min\left\{R_{1i}^{\beta}, R_{2i}^{\beta}, \ldots, R_{Ni}^{\beta}\right\}$$

P-CNN: 1D CNN for boosted jet tagging.
7 input features computed from the constituents 4-vectors.
(Similar technology as presented in CMS-DP-2017-049.)

Theory-inspired taggers

Lorentz Boost Network (LBN) [1812.09722]

Get Python Package here



Lorentz Layer (LOLA) [1707.08966]

$$C = \begin{pmatrix} 1 & \cdots & 0 & C_{1,N+1} & \cdots & C_{1,M} \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & C_{N,N+1} & \cdots & C_{N,M} \end{pmatrix}$$

$$k_{\mu,i} \xrightarrow{\text{CoLa}} \tilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij}$$

Combination layer: linear combination of the 4-vectors

$$\tilde{k}_j \xrightarrow{\text{LoLa}} \hat{k}_j = \begin{pmatrix} m^2(\tilde{k}_j) \\ p_T(\tilde{k}_j) \\ w_{jm}^{(E)}\, E(\tilde{k}_m) \\ w_{jm}^{(p_T)}\, p_T(\tilde{k}_m) \\ w_{jm}^{(m^2)}\, m^2(\tilde{k}_m) \\ w_{jm}^{(d)}\, d_{jm}^2 \end{pmatrix}$$

Lorentz layer: extract physics features

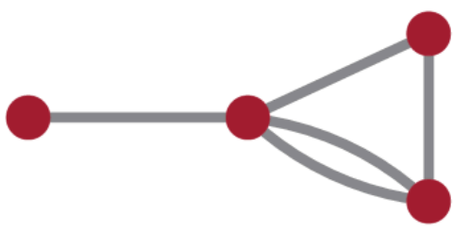
Theory-inspired taggers

Latent Dirichlet Allocation (LDA) [1904.04200]
Used in generative modeling for collections of text documents.
Searches for co-occurrence patterns in high-level substructure features.
(See Barry Dillon's talk!)

Energy Flow Polynomials (EFP) [1712.07124]
Linear basis of IRC safe observables + Fisher's Linear Discriminant

$$z_i = p_{T,i} / \sum_j p_{T,j} \qquad \theta_{ij} = ((y_i - y_j)^2 + (\phi_i - \phi_j)^2)^{\beta/2}$$

$$\bullet_j \iff \sum_{i_j=1}^{M} z_{i_j} \qquad k \text{ —— } \ell \iff \theta_{i_k i_\ell}$$



$$= \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \sum_{i_3=1}^{M} \sum_{i_4=1}^{M} z_{i_1} z_{i_2} z_{i_3} z_{i_4} \theta_{i_1 i_2} \theta_{i_2 i_3} \theta_{i_2 i_4}^2 \theta_{i_3 i_4}$$

Theory-inspired taggers

Energy Flow Networks (EFN) [1810.05165]
IRC-safe observables
+permutation invariance

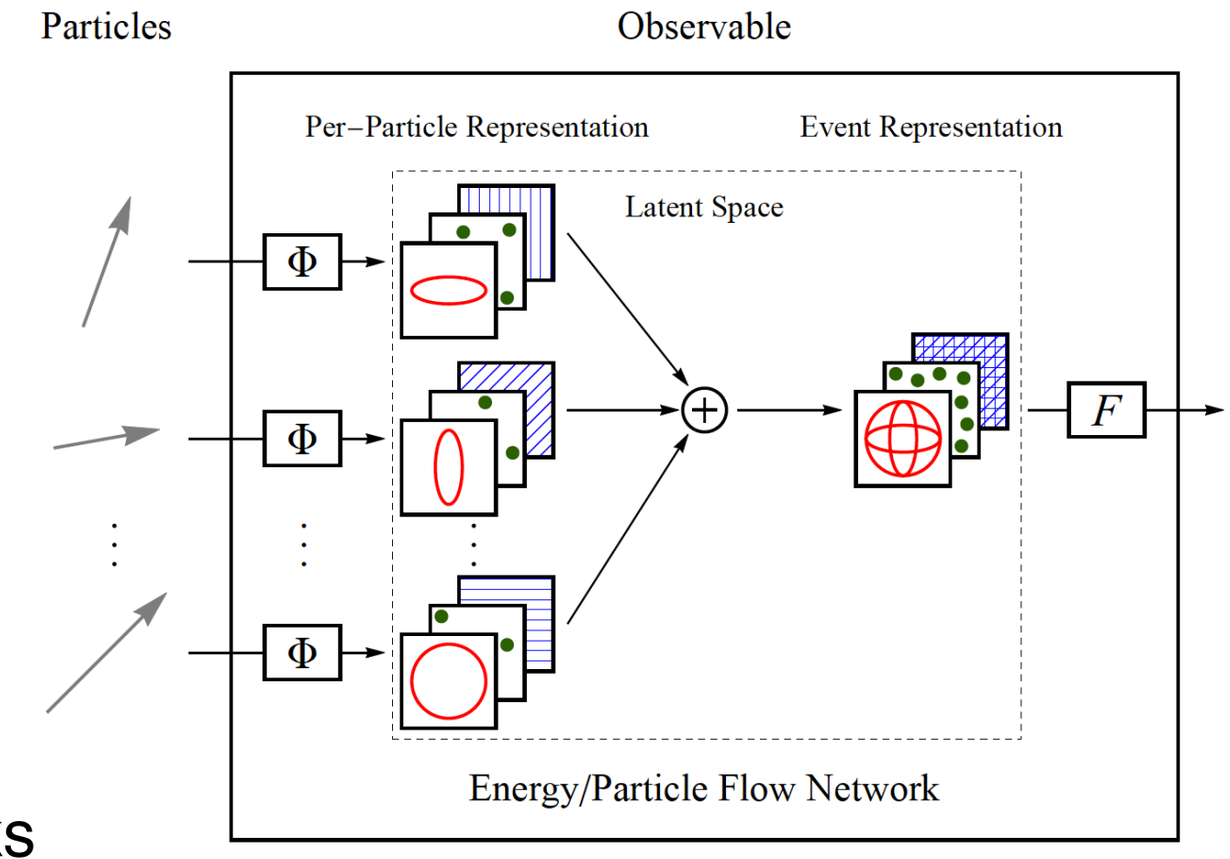
$$\mathrm{EFN} = F\left(\sum_{i=1}^{M} z_i \Phi(y_i, \phi_i)\right)$$

Φ and F parametrized with dense layers.

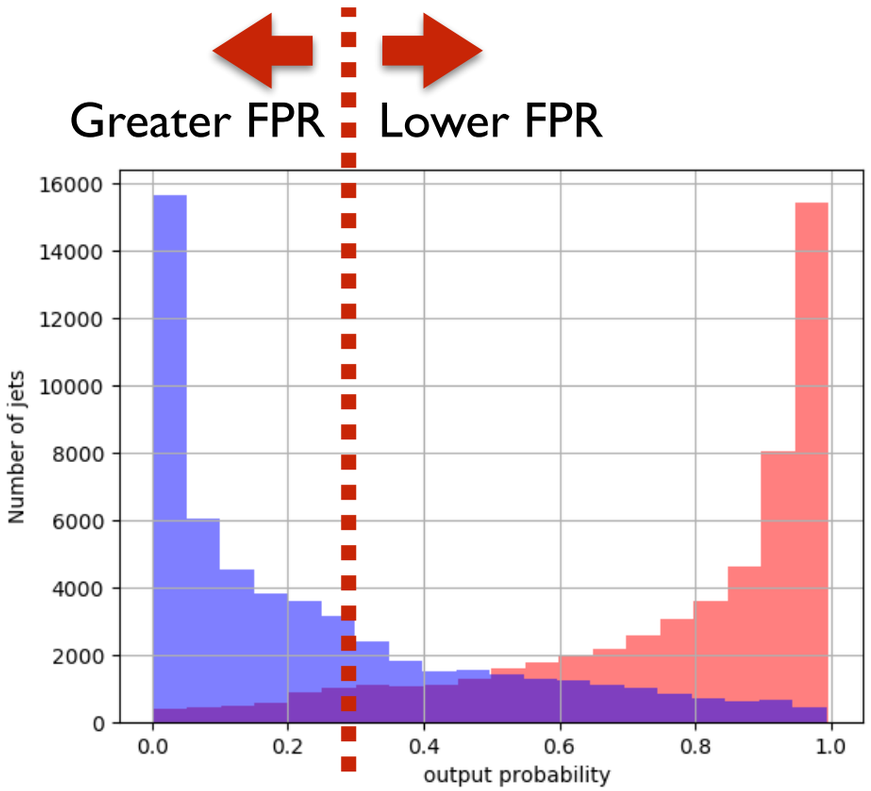
Particle Flow Networks (PFN)
[1810.05165]
Generalization of EFN beyond IRC safety.
Direct contact with Deep Set frameworks

$$\mathrm{PFN} = F\left(\sum_{i=1}^{M} \Phi(p_i)\right)$$

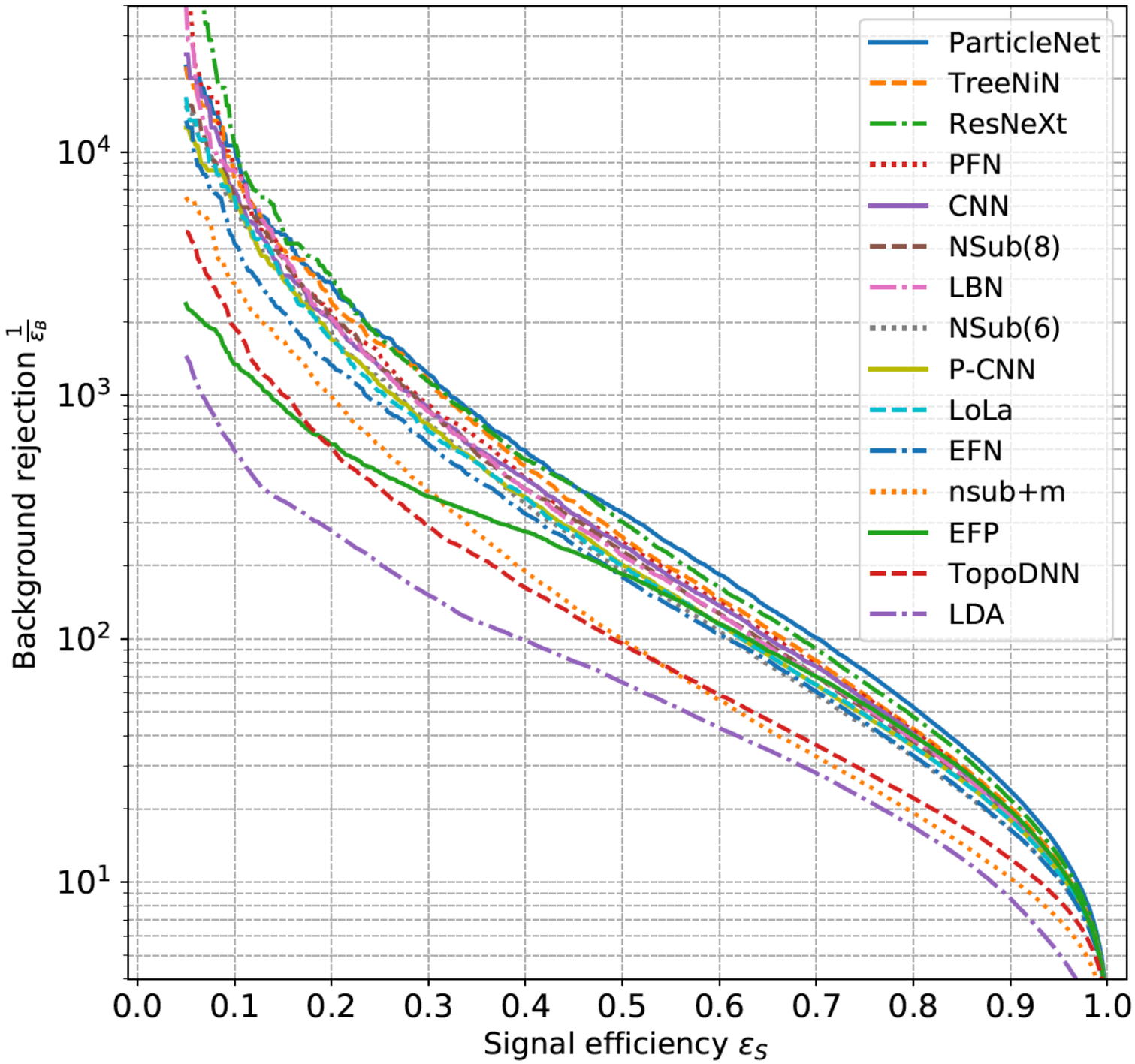
Get Python Package here



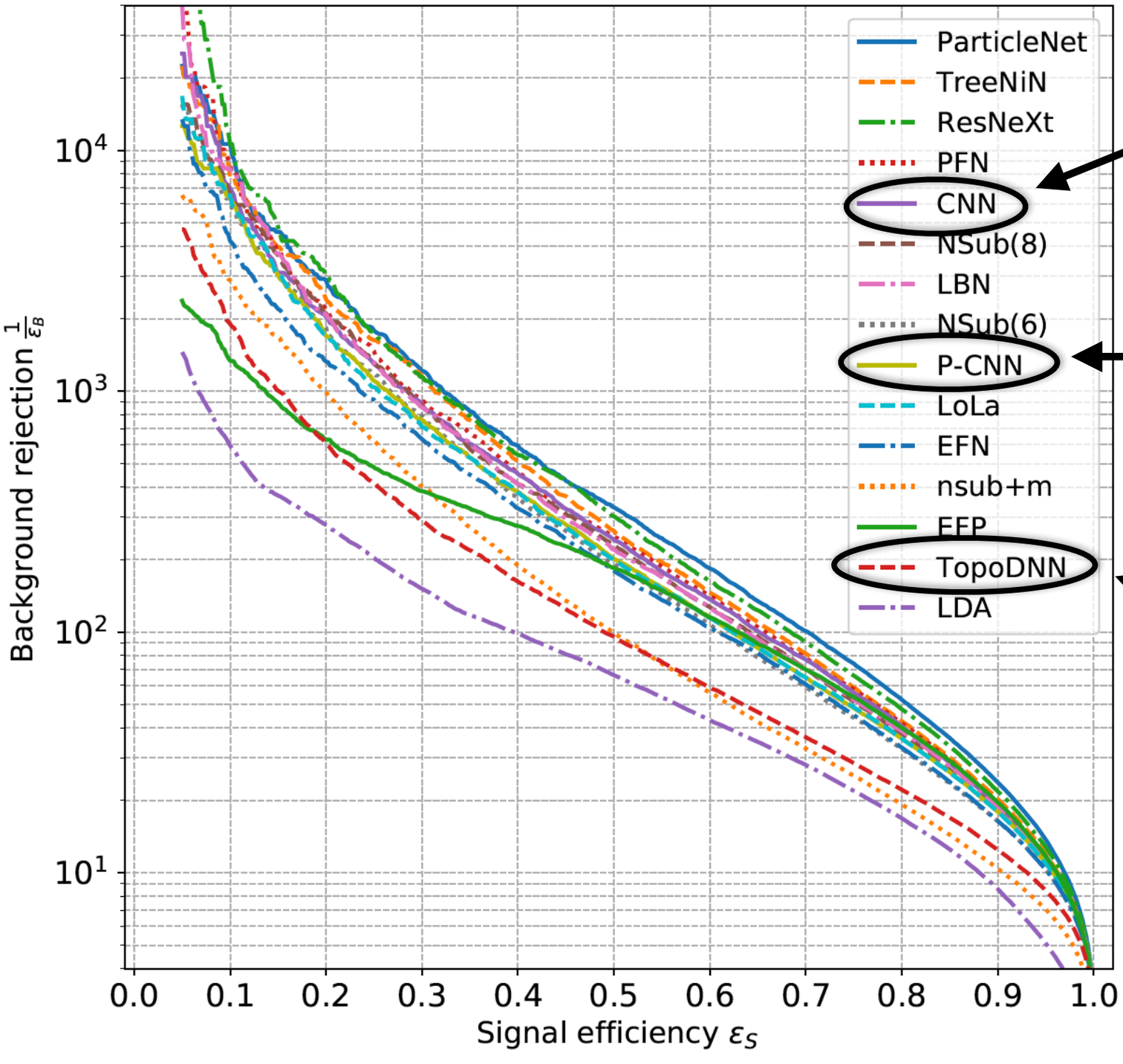
ROC curves



- 9 models trained for each classifier.
- ROC curve for the model that gives the median AUC.



ROC curves



CMS ImageTop architecture initially based on CNN

Similar technology to CMS DeepAK8

Similar technology to ATLAS TopoDNN

Performance Metrics

	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
ResNeXt [31]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
P-CNN	0.980	0.930	732±24	845±13	834±14	348k
ParticleNet [47]	0.985	0.938	1298±46	1412±45	1393±41	498k
LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
LDA [54]	0.955	0.892	151±0.4	151.5±0.5	151.7±0.4	184k
Energy Flow Polynomials [21]	0.980	0.932	384			1k
Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
GoaT	0.985	0.939	1368±140		1549±208	35k

Performance Metrics

News!

	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
ResNeXt [30]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
P-CNN	0.980	0.930	732±24	845±13	834±14	348k
v1 ParticleNet [47]	0.985	0.938	1298±46	1412±45	1393±41	498k
LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
Energy Flow Polynomials [21]	0.980	0.932	384			1k
Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
GoaT	0.985	0.939	1368±140		1549±208	35k
New ***ParticleNet-Lite***	**0.984**	**0.937**	**1262±49**			**26k**
New ***ParticleNet***	**0.986**	**0.940**	**1615±93**			**366k**

Ensemble of all taggers

arXiv:1902.09914

Mean(median)-of-ensemble taggers

- 9 models trained: 84 subsets $\mathcal{A}$ of 6 models each.
- For each subset, get mean-of-ensemble (median-of-ensemble) taggers as a per jet prediction:

Model output probability

$$s^{\mathcal{A}}_{\text{mean}}(j) = \text{mean}_{\{m\in\mathcal{A}\}}\left[s_m(j)\right]$$
$$s^{\mathcal{A}}_{\text{median}}(j) = \text{median}_{\{m\in\mathcal{A}\}}\left[s_m(j)\right]$$

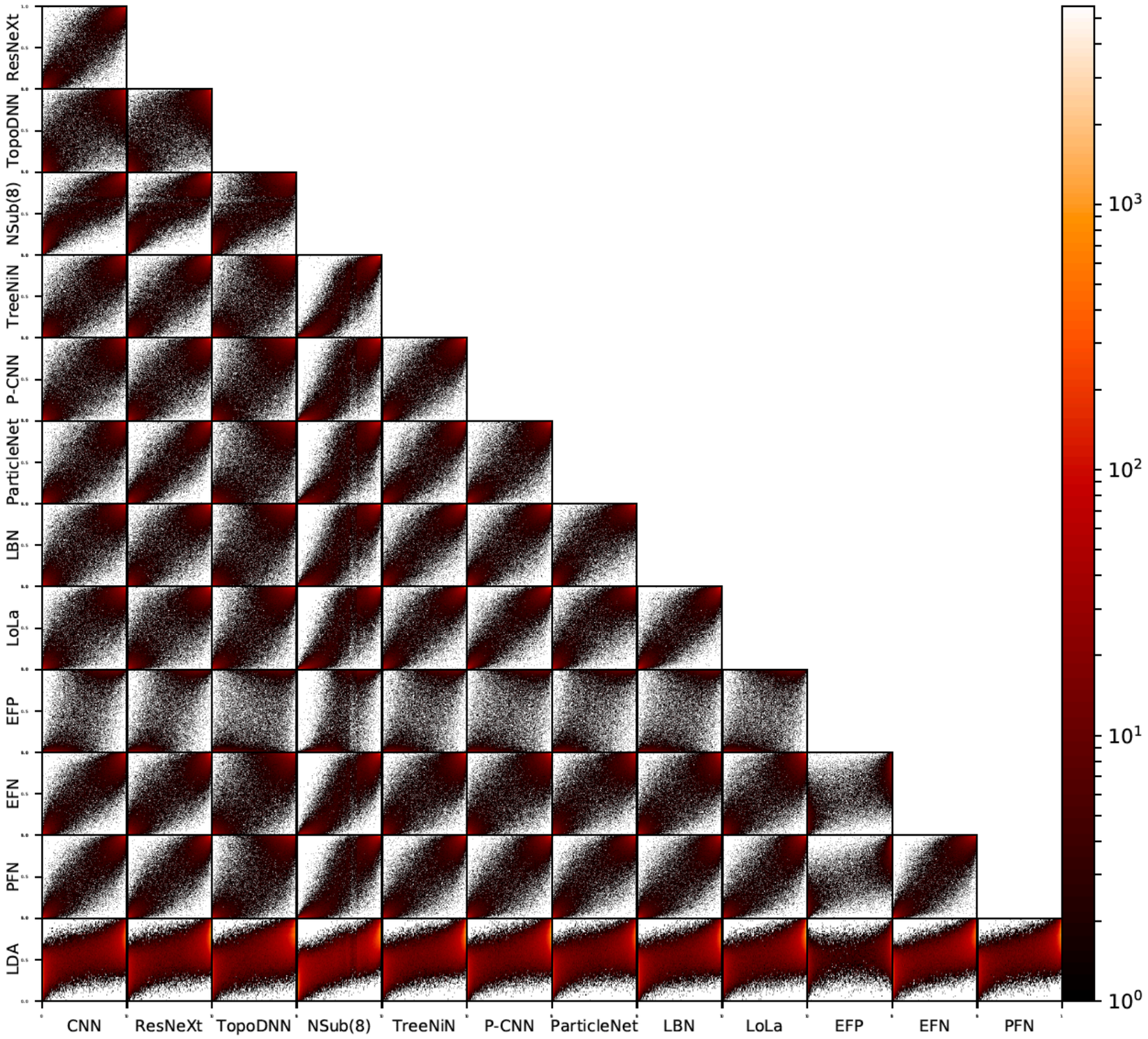
- From 84 such ensembles, choose the one with the median $1/\epsilon_B(\epsilon_S = 0.3)$

We will see a 5-15% improvement in performance!

Performance Metrics

	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
ResNeXt [31]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
P-CNN	0.980	0.930	732±24	845±13	834±14	348k
ParticleNet [47]	0.985	0.938	1298±46	1412±45	1393±41	498k
LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
LDA [54]	0.955	0.892	151±0.4	151.5±0.5	151.7±0.4	184k
Energy Flow Polynomials [21]	0.980	0.932	384			1k
Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
GoaT	0.985	0.939	1368±140		1549±208	35k

Pair-wise correlations



GoaT (Greatest of all Taggers)

	AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
ResNeXt [31]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
P-CNN	0.980	0.930	732±24	845±13	834±14	348k
ParticleNet [47]	0.985	0.938	1298±46	1412±45	1393±41	498k
LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
LDA [54]	0.955	0.892	151±0.4	151.5±0.5	151.7±0.4	184k
Energy Flow Polynomials [21]	0.980	0.932	384			1k
Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
GoaT	0.985	0.939	1368±140		1549±208	35k

Reproducible Open Benchmarks for Data Analysis Platform (ROB)

Work in progress [Heiko Mueller, Irina Espejo, Kyle Cranmer & SM]

Exploratory work for enabling such community benchmarks.

Outline of ROB usage

1. Benchmark workflow defined by **coordinator** along with input data.
2. Users provide docker containers that satisfy workflow stages
3. Back-end processes entries, evaluates metrics (powered by REANA).
4. Front-end displays results.

Benchmark Workflow Example:

1. Common input data set used for benchmarks (defined by benchmark coordinator).
2. Preprocessing stage (code provided by user).
3. Prediction stage (code provided by user).
4. Evaluate metrics, update tables & plots (defined by benchmark coordinator).

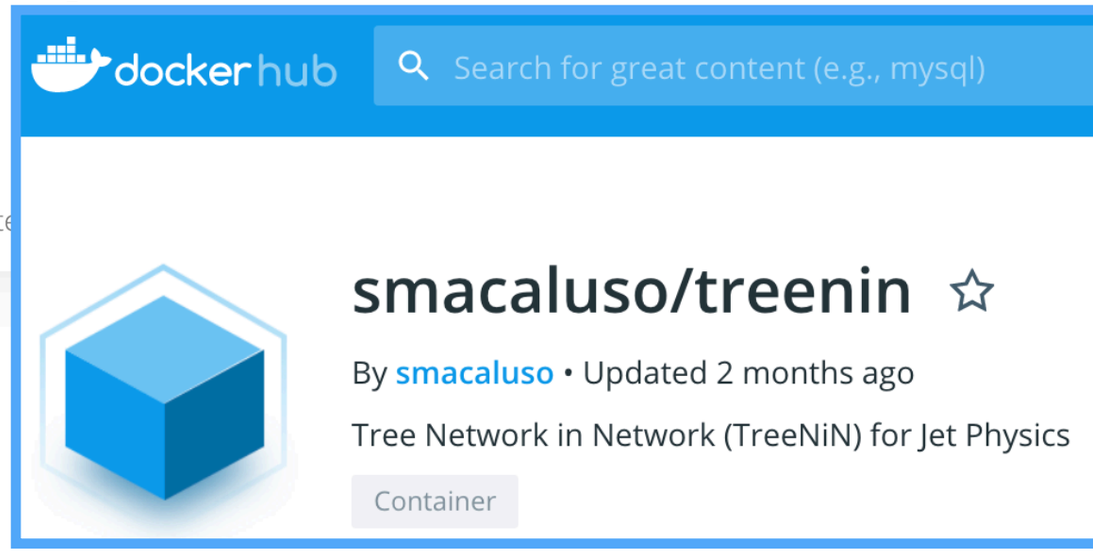
ROB elements

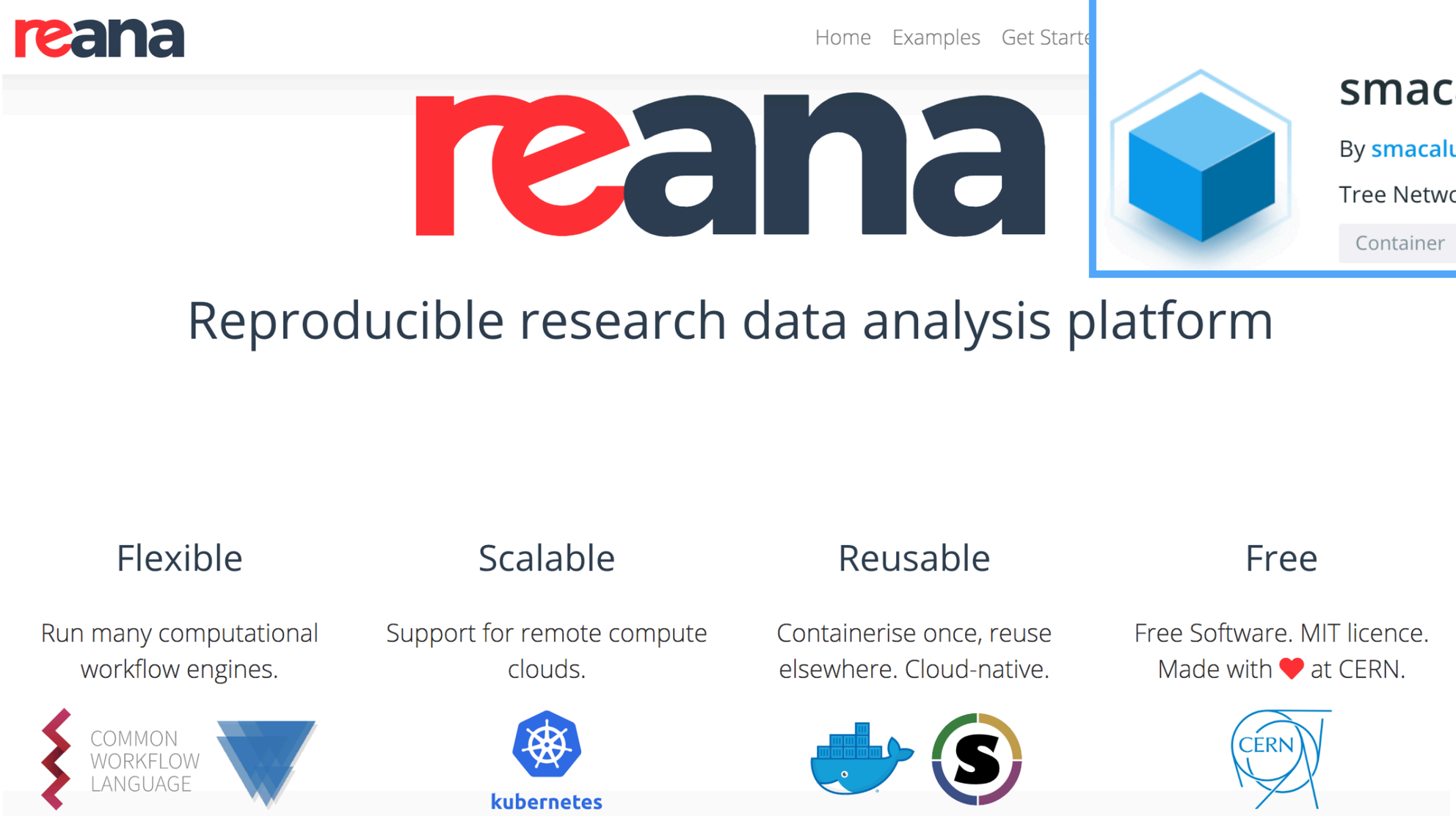
Workflow templates powered by reana

http://www.reana.io/

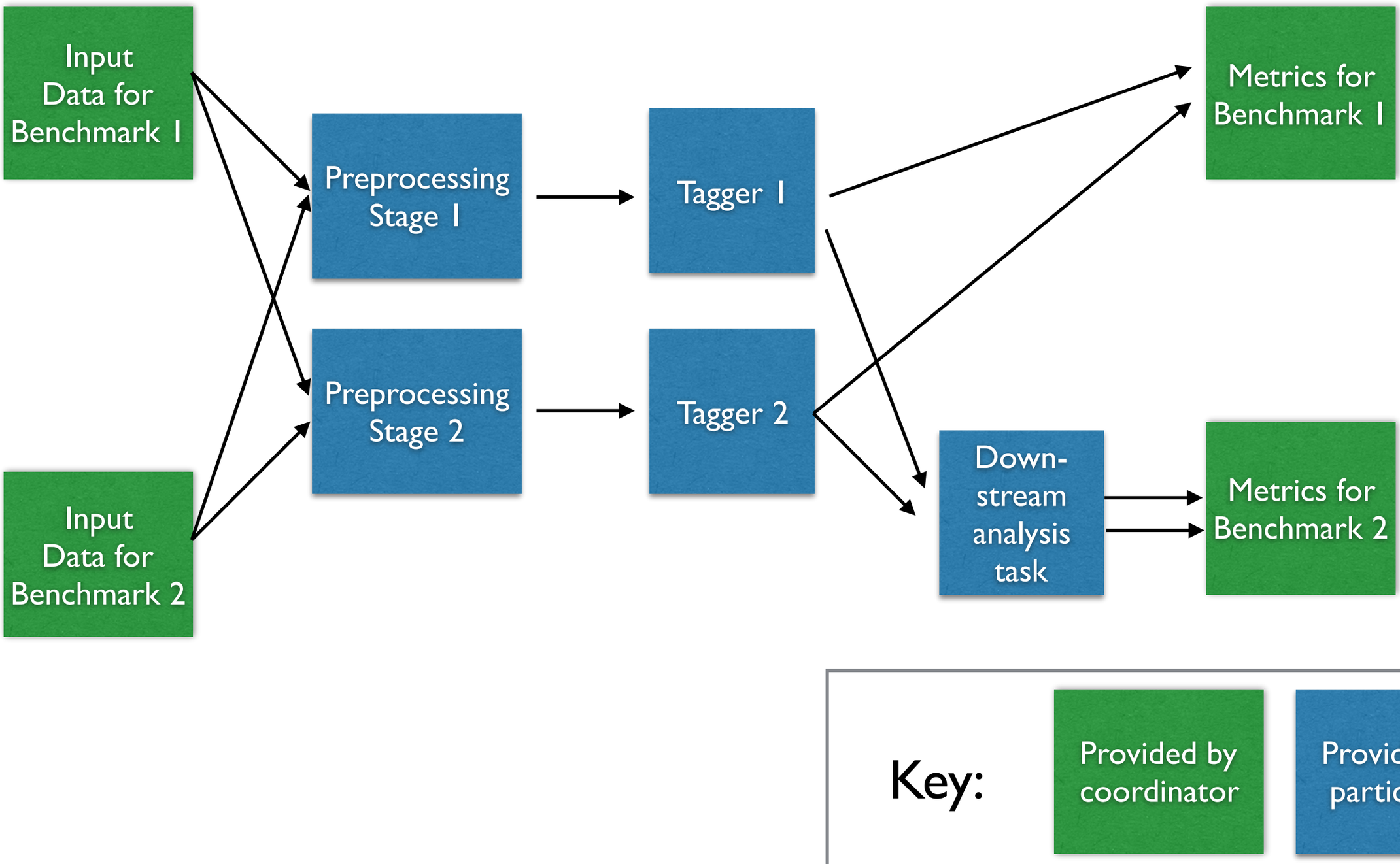
Users provide a Docker container, e.g. TreeNiN docker image

https://hub.docker.com/r/smacaluso/treenin





Idea: Compossible workflows support contributions used for multiple benchmarks



Final remarks

- Many competitive taggers with comparable performance!
- Interesting to see ATLAS and CMS implementations of other jets representations, e.g. point clouds/sets, trees, etc.
- Algorithms include very different levels of physics-inspired structures.
- Explore performance with softer jets, higher pT jets (detector resolution) and stability for wider pT windows.
- Study samples with effects from underlying events, secondary vertex information.
- Consider stability tests for pile-up and systematic uncertainties (see Jennifer Thompson's talk! [1904.10004]).
- Compare evaluation time and number of training events to saturate performance.

Thanks for your attention!