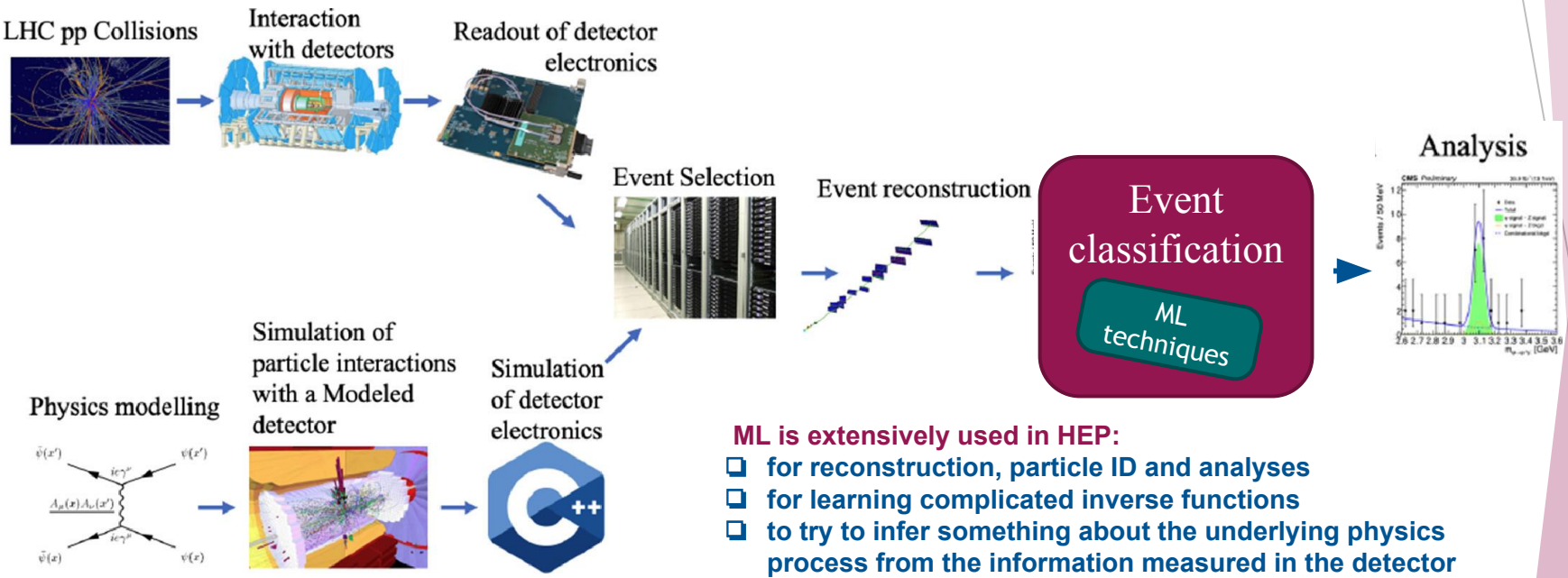
Machine Learning based event classification techniques

Evelin Meoni[1], Elvira Rossi[2], Enrico Tassi[1]
[1]Università della Calabria &INFN Cosenza
[2]Università Federico II di Napoli & INFN Napoli

ML based event classification techniques



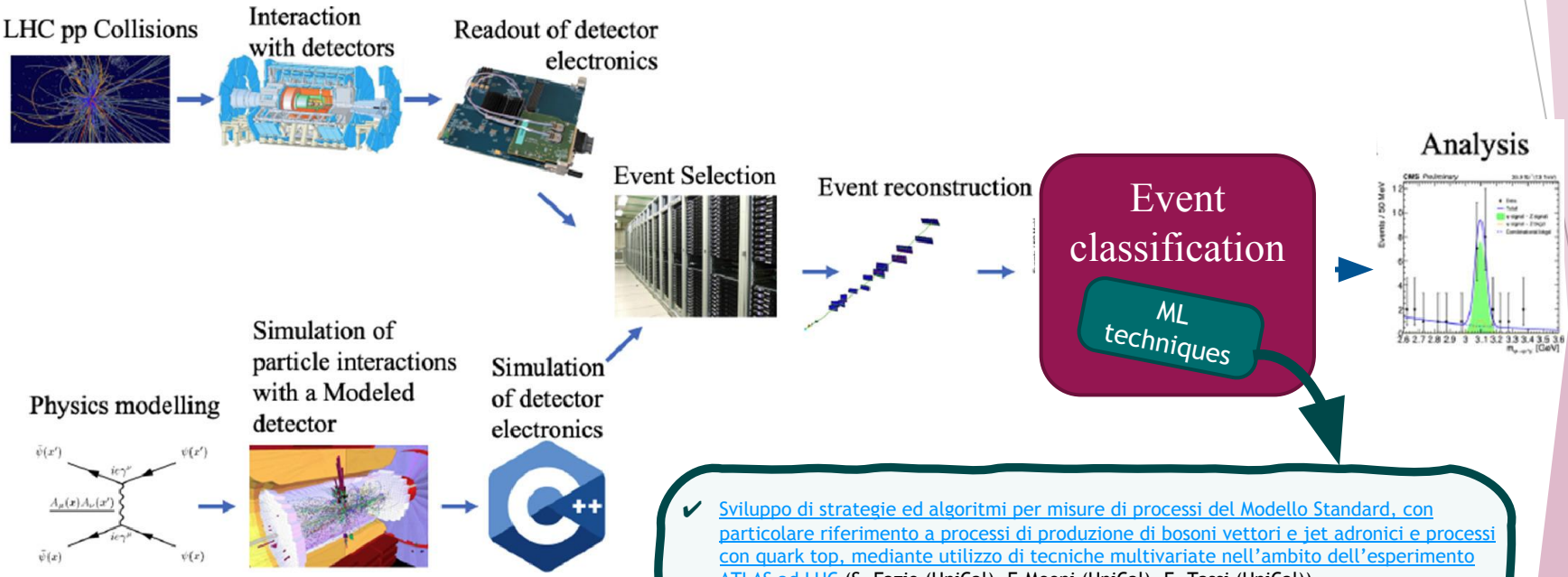
Freely inspired from the original picture from
"Computing models in high energy physics" T. Boccali

ML is extensively used in HEP:

- for reconstruction, particle ID and analyses
- for learning complicated inverse functions
- to try to infer something about the underlying physics process from the information measured in the detector

In this talk we discuss ML applications to data analyses for event classification (signal to background separation)

ML based event classification techniques



Freely inspired from the original picture from
"Computing models in high energy physics" T. Boccali

- ✔ Sviluppo di strategie ed algoritmi per misure di processi del Modello Standard, con particolare riferimento a processi di produzione di bosoni vettori e jet adronici e processi con quark top, mediante utilizzo di tecniche multivariate nell'ambito dell'esperimento ATLAS ad LHC (S. Fazio (UniCal), E.Meoni (UniCal), E. Tassi (UniCal))
- ✔ Sviluppo di modelli per discriminazione di eventi di Vector Boson Scattering (Mariani (Unipg), Tedeschi (unipg, dottorando), Spiga (infn) - daniele.spiga@pg.infn.it)
- ✔ Sviluppo di tecniche di discriminazione con adattamento al dominio (*Lenzi (UniFi), Ciulli (UniF), Viliani (INFN),* B. Camaiani (UniFi,dottoranda), M. Lizzo (UniFi))
- ✔ Sviluppo di modelli per Anomaly Detection per analisi di fisica BSM in eventi di Vector Boson Scattering (Gennai (INFN), Govoni (UniMIB))
- ✔ Anomaly detection e Graph Neural Network (Valerio Ippolito (INFN Roma 1), Elvira Rossi (Federico II))

From “traditional” ML to Deep ML

Event classification

ML techniques

- ✔ Sviluppo di strategie ed algoritmi per misure di processi del Modello Standard, con particolare riferimento a processi di produzione di bosoni vettori e jet adronici e processi con quark top, mediante utilizzo di tecniche multivariate nell’ambito dell’esperimento ATLAS ad LHC (S. Fazio (UniCal), E.Meoni (UniCal), E. Tassi (UniCal))
- ✔ Sviluppo di modelli per discriminazione di eventi di Vector Boson Scattering (Mariani (Unipg), Tedeschi (unipg, dottorando), Spiga (infn) - daniele.spiga@pg.infn.it)
- ✔ Sviluppo di tecniche di discriminazione con adattamento al dominio (*Lenzi (UniFi), Ciulli (UniF), Viliani (INFN)*, B. Camaiani (UniFi,dottoranda), M. Lizzo (UniFi))
- ✔ Sviluppo di modelli per Anomaly Detection per analisi di fisica BSM in eventi di Vector Boson Scattering (Gennai (INFN), Govoni (UniMIB))
- ✔ Anomaly detection e Graph Neural Network (Valerio Ippolito (INFN Roma 1), Elvira Rossi (Federico II))

First collection of proposals spaces from more traditional techniques with supervised approaches up to very new techniques with unsupervised approaches

Multivariate Techniques

Deep Learning

Anomaly Detection (AD)

Traditionally: Boosted Decision Trees (BDTs), Neural Networks (NNs)

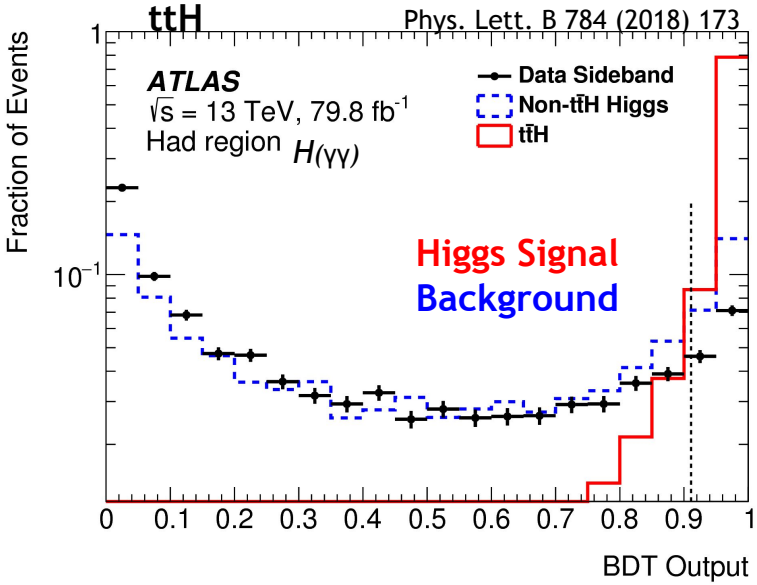
More recently: “**deep learning**” explosion (DNNs, ANNs, RNNs, CNNs, GNNs,VAEs,..)

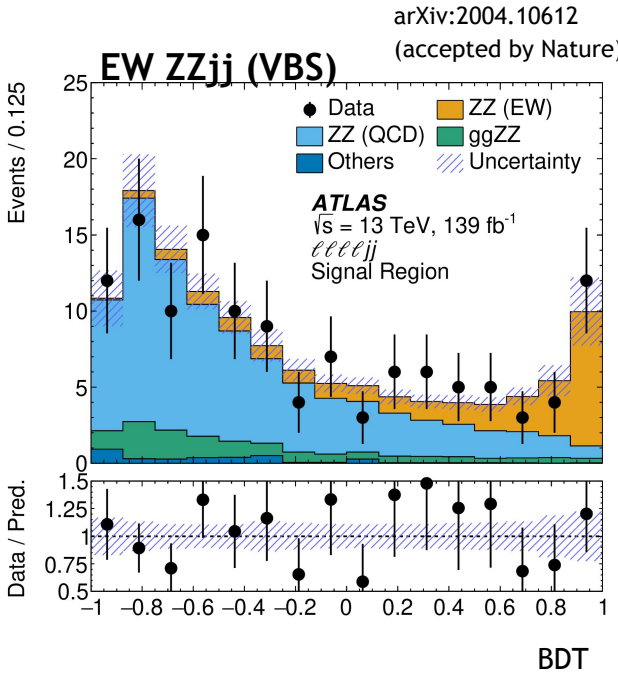
Typically in physics analyses we adopt **supervised** approach (the most used: known features of signal and bkg), recently also proposal for **unsupervised** approach (learn direct from the data)

BDTs and NNs

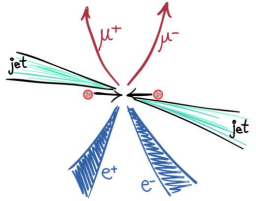
BDTs and NN are used to publish many high-profile HEP results.

Just a few examples at LHC: the observation of ttH at ATLAS and CMS ; observation of Vector Boson Scattering (VBS) (more precisely EW production of dibosons plus 2 jets) at ATLAS and CMS; first evidence for Bs0 → μ+μ− by LHCb.





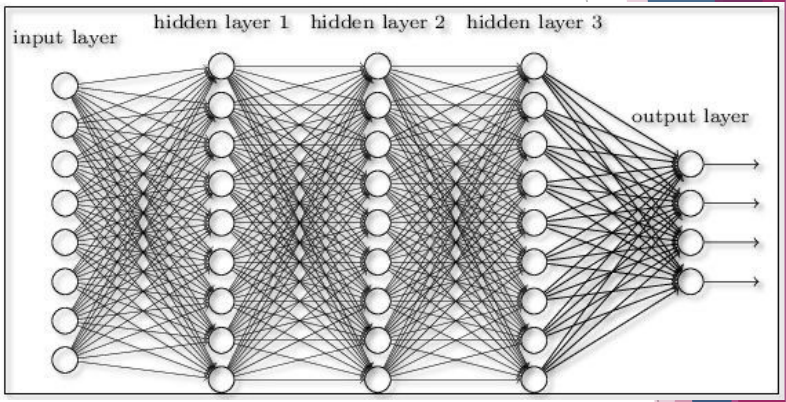
Signal
Irreducible
Background



Inputs: p_T, η, φ, *E of jets, η and φ of each photon,* $p_T/m_{\gamma\gamma}$,..

Input: $\Delta y(jj)$, m_{jj}, pTj1,pTj2, mZZ,pTZZ..

DNNs

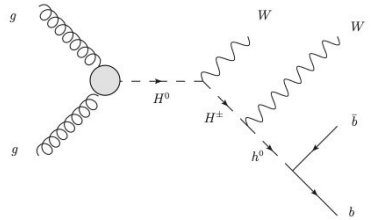


More complex architectures:

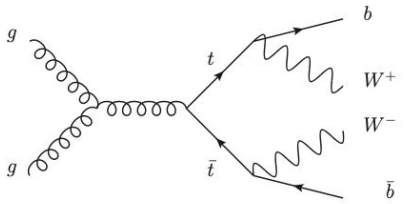
- better performances on NNs
- automated hyperparameters optimization
- should work very well just with low-level inputs

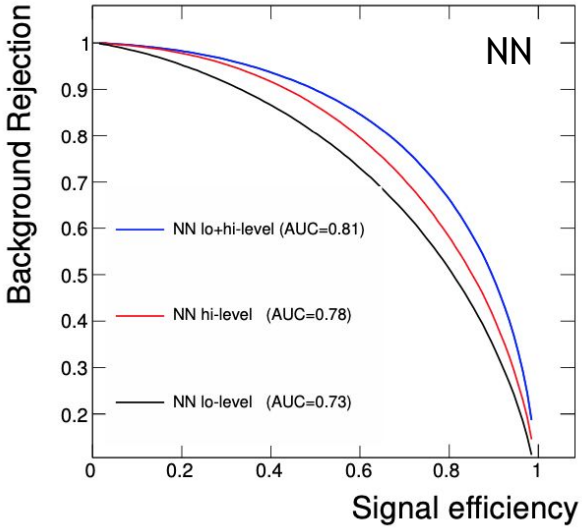
arXiv:1402.4735v2

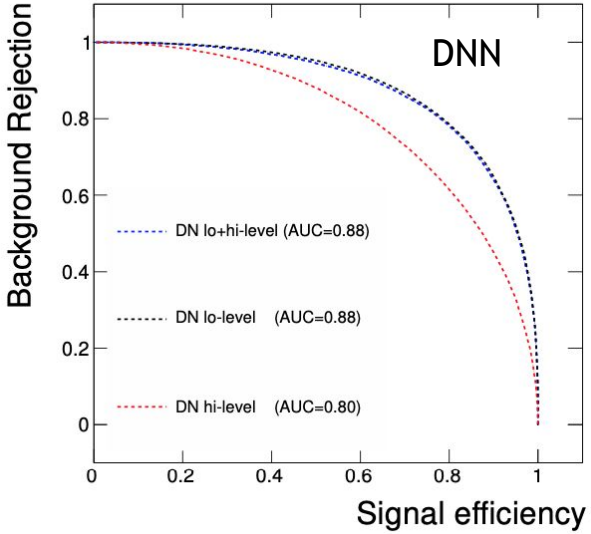
Exotic Higgs (H->WWbb, semileptonic final stal)



SM bkg (ttbar, semileptonic final state)







Low-level inputs= pT and eta jets and lepton
High-level inputs= invariant masses of the various combined objects(bb, Wb,..)

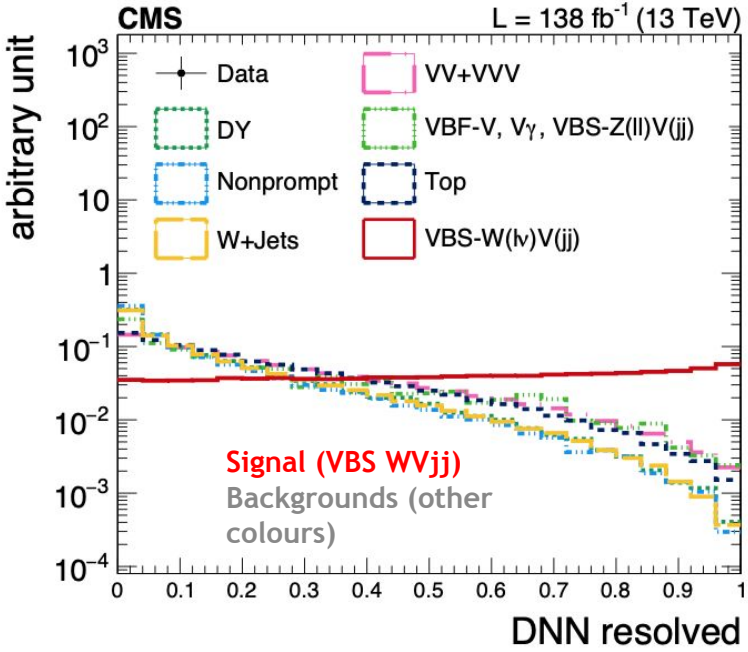
DNNs

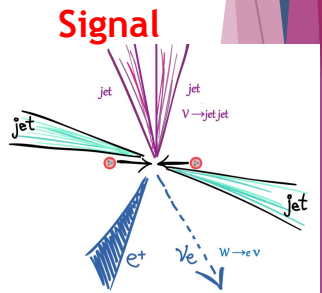
Already LHC physics paper using DNN-based analysis: various BSM searches, also VBS measurements and di-boson polarisation measurements

Variable	Resolved
Lepton pseudorapidity	✓
Lepton transverse momentum	✓
Zeppenfeld variable for the lepton	✓
Number of jets with $p_T > 30$ GeV	✓
Leading VBS tag jet p_T	-
Trailing VBS tag jet p_T	✓
Pseudorapidity interval $\Delta\eta_{jj}^{VBS}$ between tag jets	✓
Quark/gluon discriminator of leading VBS tag jet	✓
Azimuthal angle distance between VBS tag jets	✓
Invariant mass of the VBS tag jets pair	✓
p_T of the leading V_{had} jet	✓
p_T of the trailing V_{had} jet	✓
Pseudorapidity difference between V_{had} jets	✓
Quark/gluon discriminator of the leading V_{had} jet	✓
Quark/gluon discriminator of the trailing V_{had} jet	✓
p_T of the AK8 V_{had} jet candidate	-
Invariant mass of V_{had}	✓
Zeppenfeld variable for V_{had}	-
Centrality	-

Inputs: not only low-level inputs (i.e. single objects observables)

Phys. Lett. B 834 (2022) 137438





Proposed physics cases

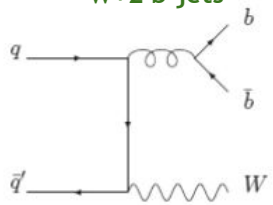
Contatti: S. Fazio, E. Meoni, E. Tassi **UniCal (Cosenza)**

Descrizione: Lo studio di processi previsti dal Modello Standard (MS) è centrale per la validazione del Modello alle energie finora inesplorate di LHC e risulta essere un potente tool di ricerca indiretta di fenomenologia di Nuova Fisica (NF) mediante osservazione di deviazioni delle predizioni del MS. L'utilizzo di tecniche multivariate per la reiezione del fondo determina un sostanziale miglioramento del rapporto segnale-fondo rispetto a tecniche tradizionali cut-based, permettendo la rivelazione di processi rari e la misura di processi in regioni complesse dello spazio delle fasi nelle quali i fondi dominano sul segnale. Inoltre poiché i processi del MS, in particolar modo processi con bosoni vettori e jet adronici e processi con quark top, costituiscono fondo per la fisica dell'Higgs ed i segnali di NF, tali misure sono fondamentali per il tuning delle simulazioni Monte Carlo (MC) dei fondi utilizzate appunto per studi di Higgs e ricerche dirette di NF.

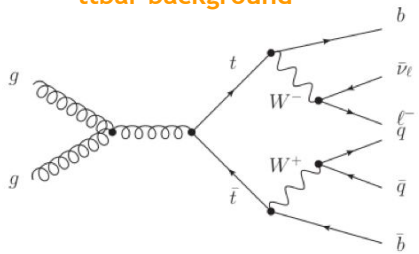
Esperimento: ATLAS; **Stato:** Attività non iniziata

W+b-jets

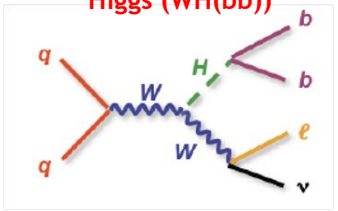
W+2 b-jets



ttbar background



Higgs (WH(bb))



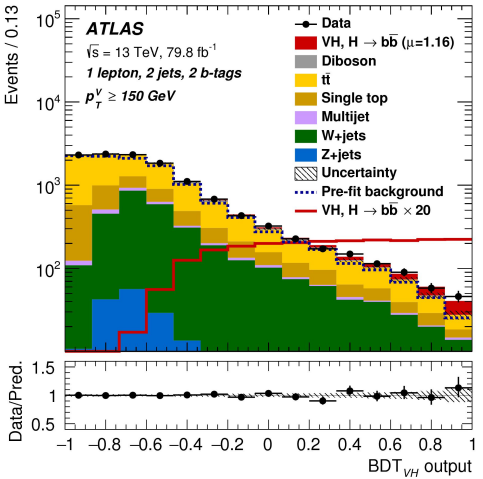
Frequent background for Higgs (WH(bb)) and BSM Searches

Precise modeling important to reduce the associated systematics

Missing a W+2 b-jets differential measurement at LHC due to the huge ttbar background

→ Develop a DNN for S/B separation

Phys. Lett. B 786 (2018) 59



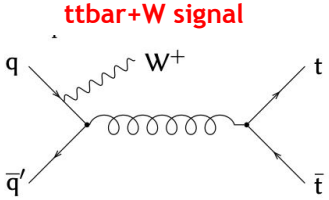
Proposed physics cases

Contatti: S. Fazio, E. Meoni, E. Tassi **UniCal (Cosenza)**

Descrizione: Lo studio di processi previsti dal Modello Standard (MS) è centrale per la validazione del Modello alle energie finora inesplorate di LHC e risulta essere un potente tool di ricerca indiretta di fenomenologia di Nuova Fisica (NF) mediante osservazione di deviazioni delle predizioni del MS. L'utilizzo di tecniche multivariate per la reiezione del fondo determina un sostanziale miglioramento del rapporto segnale-fondo rispetto a tecniche tradizionali cut-based, permettendo la rivelazione di processi rari e la misura di processi in regioni complesse dello spazio delle fasi nelle quali i fondi dominano sul segnale. Inoltre poiché i processi del MS, in particolar modo processi con bosoni vettori e jet adronici e processi con quark top, costituiscono fondo per la fisica dell'Higgs ed i segnali di NF, tali misure sono fondamentali per il tuning delle simulazioni Monte Carlo (MC) dei fondi utilizzate appunto per studi di Higgs e ricerche dirette di NF.

Esperimento:ATLAS; **Stato:** Attività non iniziata

ttbar+W

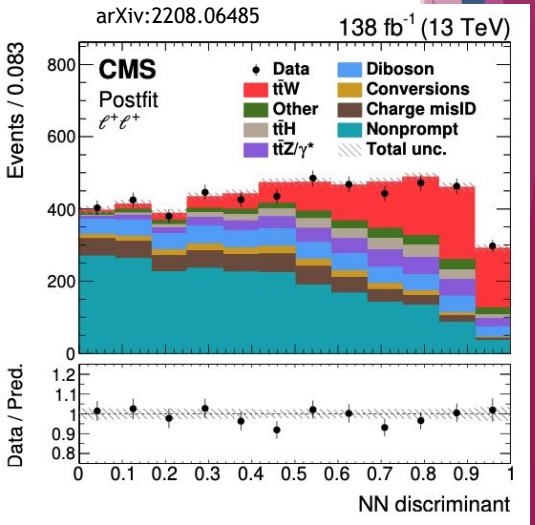


Highly sensitive to BSM physics

Important background for BSM final states with leptons and b-jets

Missing a differential measurement at LHC →Develop a DNN approach for S/B sep.

Expertise: a bachelor thesis on ttbar+Z with a DNN



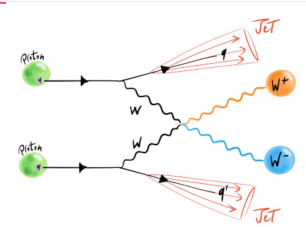
Proposed physics cases

Contatti: Mariani, Tedeschi, Spiga **Perugia/INFN**

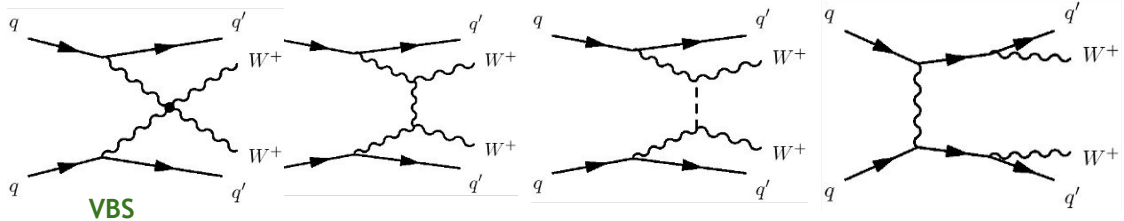
Descrizione: Sviluppo di modelli DNN per la discriminazione di eventi di Vector Boson Scattering, sia a livello di caratterizzazione globale di evento (usando le proprietà cinematiche degli oggetti ricostruiti) che di jet tagging. E' possibile l'esplorazione di soluzioni basate su Graph Neural Network.

Esperimento:CMS; **Stato:** Attività in corso; **Tecnologie:** CPU, GPU

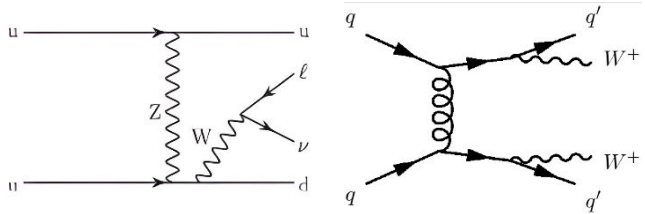
VBS: Proposed final state $q\bar{q} \to W^{\pm}W^{\pm}q\bar{q} \to \tau_h^{\pm}\nu_\tau \ell^{\pm}\nu_\ell q\bar{q}$



EW ssWWjj (signal)



SM Backgrounds (examples)



Main: W+Jets, ttbar, ZZ, Opposite Sign

Important test of SM and of Higgs mechanism

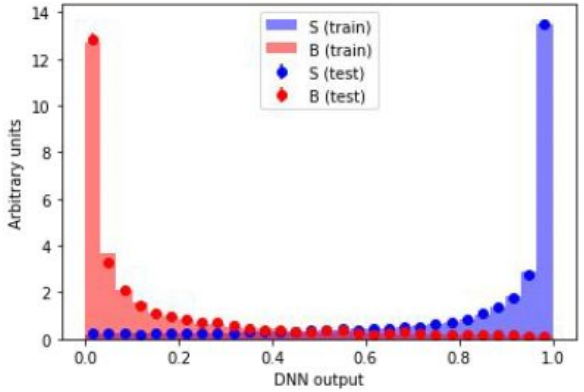
Missing measurement with tau at LHC, important to probe BSM processes

Proposed physics cases

Perugia/INFN

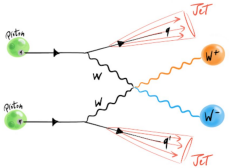
- 2 ML approaches for S/B separation: BDT and DNN
- Classifiers trained using event and objects information
  (invariant masses of various objects combinations, transverse mass of lep-MET, number of jets, angular separations, lepton pT)
- The analysis looks for both SM and BSM effect (EFT approach): different classifiers are trained for each different signal (SM, BSM_dim6, BSM_dim8)

Best DNN for BSM (dim6)

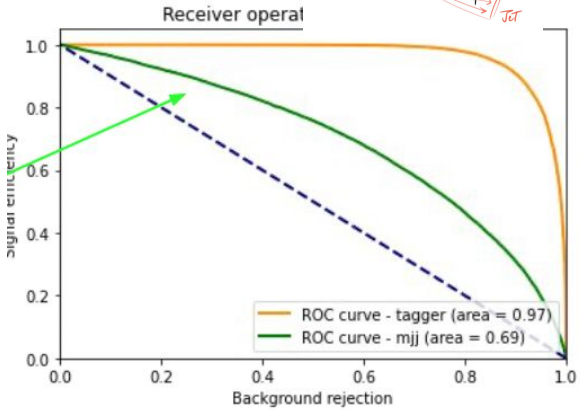


Further developments:

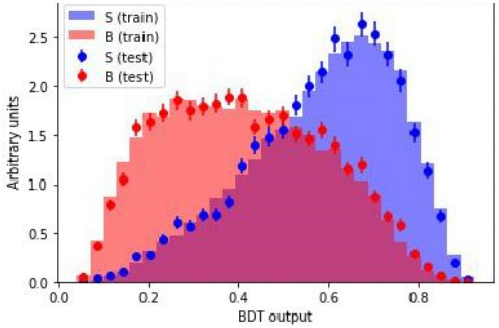
- ML jet tagging algorithm (BDT)



Select the 2 VBS jets on the basis of the jet tagger score, in place of the criterium of higher invariant mass



- BDT to separate $W_L W_L$ (signal) from other WW polarisation states (background)



Classification with domain adaptation

Contatti: *Lenzi (UniFi), Ciulli (UniFi), Viliani (INFN)* **Firenze/INFN**

Descrizione: Discriminazione di segnale rispetto al fondo in modo agnostico rispetto al modello di segnale, tramite un addestramento competitivo di due sotto-reti, l'una dedicata alla soluzione del problema di discriminazione e l'altra al riconoscimento del dominio di training.

Esperimento:non specifico (genesi in CMS) ; **Stato:** Attività in corso; **Tecnologie:** CPU, GPU

Model used in the train can introduce bias (i.e. training done on a sets of MC events, therefore the NN shape depends on the underlying physics of the MC)

→ **Goal: perform signal to background discrimination in a way that is agnostic with respect to the signal model**

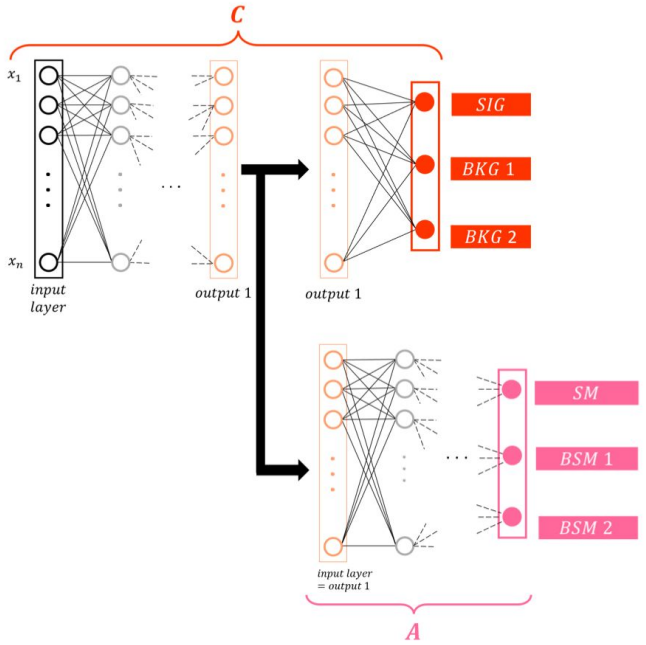
The **ADNNs** is a system of two networks:

- a classifier NN trained for the S/B separation using different S models (domains: S1,S2,S3..)

- an Adversary NN trained in a competitive way to identify the signal model (S1 or S2 or S3)

- S/B discriminator is penalized if an adversary network can correctly discriminate among S1, S2, S3…

arXiv:2207.09293v1 [hep-ex] (Accepted by EPJC)



A Physics Case

VBF H(WW)

XS measurements can be used for reinterpretation in term of BSM → Minimising the bias from signal model used for XS extraction important

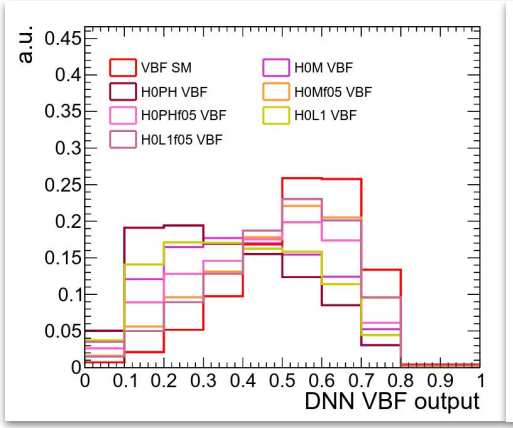
Signal = VBF H(WW) - several models (SM, S_BSM1, S_BSM2,..)
Background (B) = H (WW) ggF and WW and ttbar

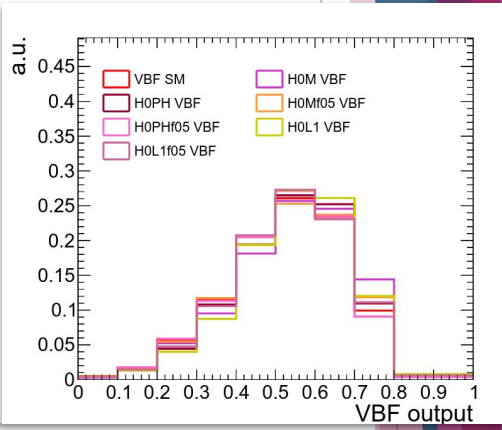
Built a DNN and ADNN (with jet, lepton, jj and ll variables) and extract XS

Built pseudo-data (S_BSM+B) → Bias much smaller for ADNN than DNN trained on S_SM only

DNN
(trained on SM only)

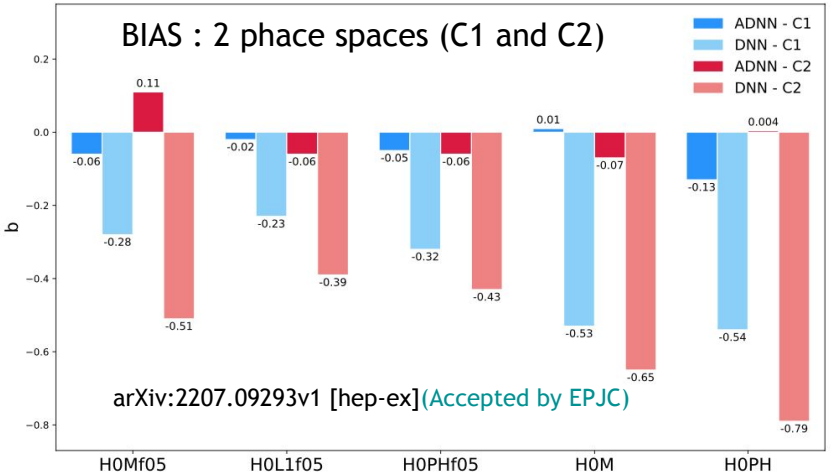
ADNN
(W/Domain Adaption)





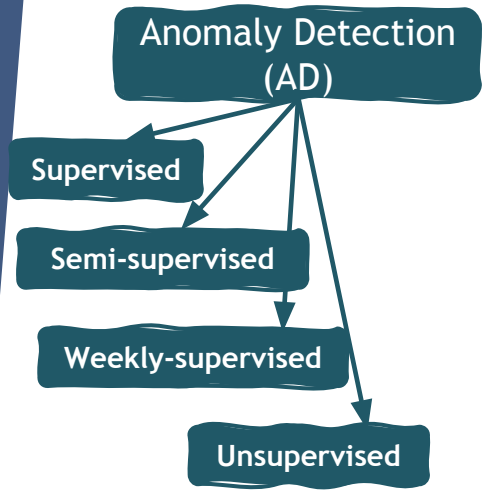
Strategy applicable in many cases whenever designing a model that is insensitive to a given quantity is desirable

Collaboration ongoing with Baker & Hughes Oil & Gas (Florence) to regress temperature/pressure in turbines operated in different conditions with respect to those of a testbed



Anomaly Detection

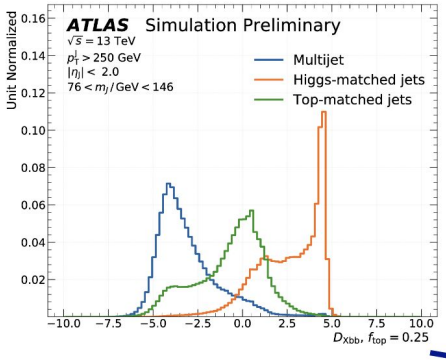
Identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well defined notion of normal behaviour
- In HEP language: identify features of the data that are inconsistent with a background-only model



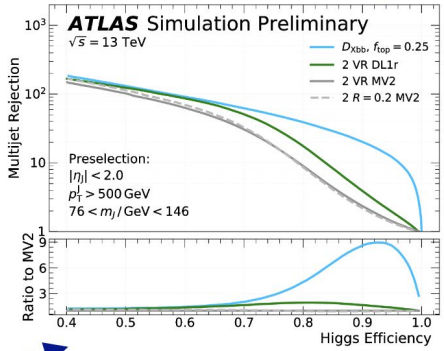
Supervised learning (labeled training datasets) has a long & effective history in HEP:
- High statistics, complex dataset with many correlated features
- Examples: neural net to classify boosted Higgs bosons from QCD/top backgrounds (ATLAS)

H→bb

NN Output Scores



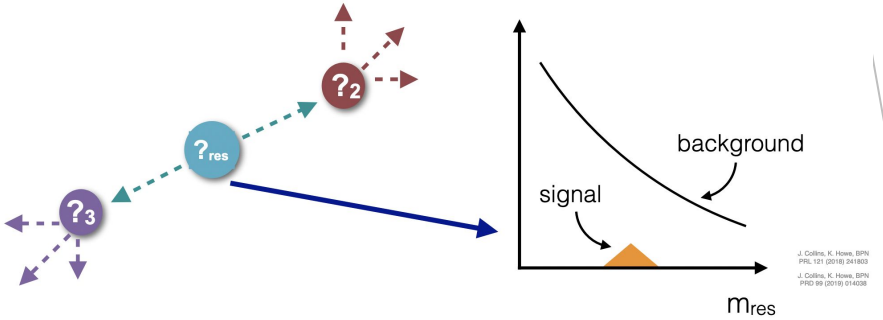
Background Rejection vs. Signal Efficiency



Anomaly Detection

But, what if we don't know the characteristics of our signal?

A broad search for "new physics" means we want to be sensitive to anything not predicted by the Standard Model (and perhaps, not even predicted by us)



- **Semi-supervised approach:** labels for some events
- **Weakly supervised approach:** noisy labels ("signal-enriched" instead of pure)
- **Unsupervised approach**: train over unlabeled events

Unsupervised approach

Autoencoders

I know how to predict all collisions

Are there any collisions that I cannot predict?

Weakly supervised approach

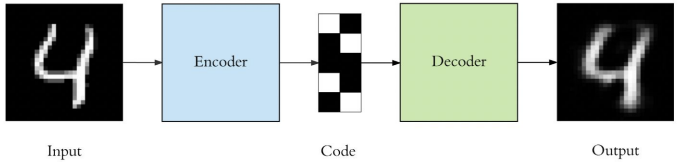
Weakly-Supervised

I know regions where new physics does not exist

I want to leverage those regions against other parts of the data to find differences

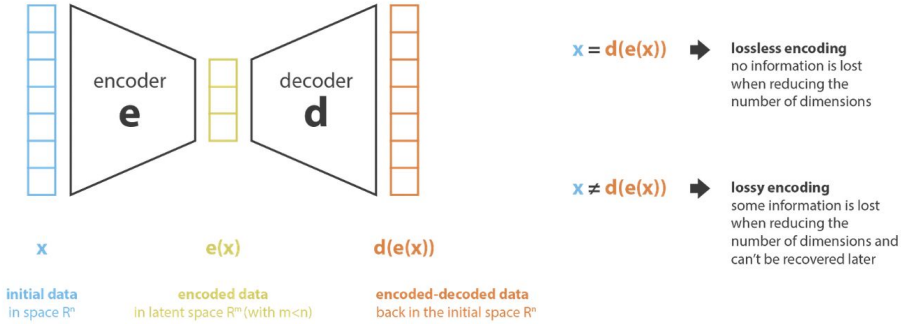
Anomalies in Open Data

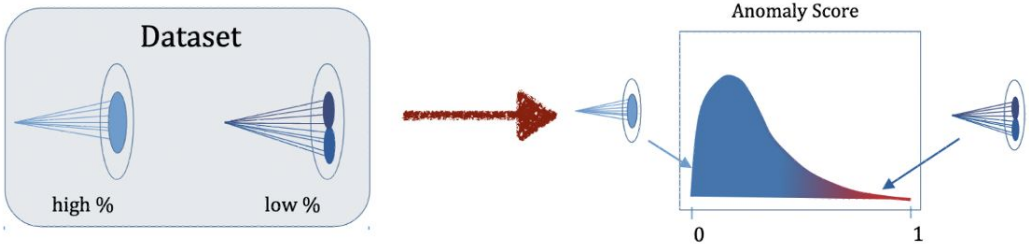
Anomaly Detection: Autoencoders & Variational autoencoders



An autoencoder (AE) is a type of NN used to learn efficient codings of unlabeled data. An autoencoder consists of 3 components: **encoder**, **code** and **decoder**. The encoder compresses the input and produces the code, the decoder then reconstructs the input only using this code.

A variational autoencoder (VAE) is a type of NN, a deep learning based generative model, a generalization of an AE to a generative model. Basically, an AE will learn the latent structure of a feature set, whereas a VAE will learn to *generate* examples in this latent space.





AE e VAE act on each event, try to reconstruct the event, assign an anomaly score. From the anomaly score, identifying the anomalous regions and then conduct the analysis

Anomaly Detection

Identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well defined notion of normal behaviour
- In HEP language: identify features of the data that are inconsistent with a background-only model

ML for anomaly detection: a physics cases

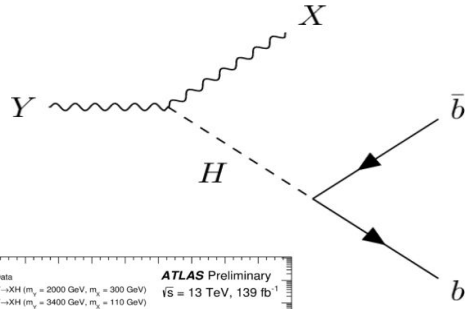
Search for heavy resonance Y decaying into SM Higgs and new particle X in a fully hadronic final state

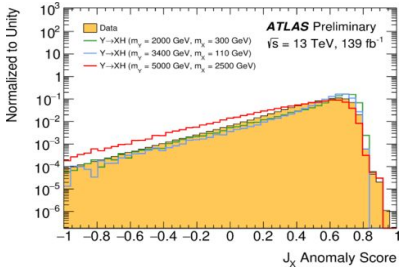
- Targeting high Y mass regime
- Boosted H and X with collimated decay products

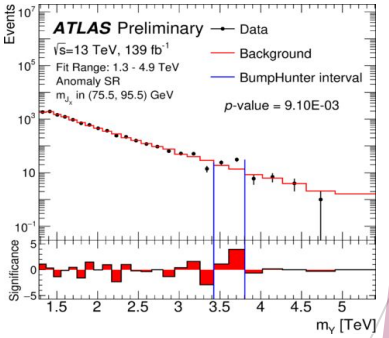
Variational recurrent neural network (VRNN - extension of VAE in the recurrent framework for the purpose of modelling high dimensional sequences) trained on jets in data

- Selection of X based on incompatibility of jet sub-structure with background jets

First application of ML technique fully usupervised for anomaly detection to ATLAS analysis







ATLAS-CONF-2022-45

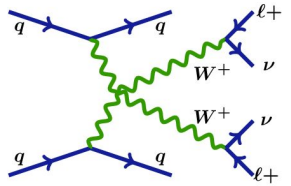
Proposed physics cases: Sviluppo di modelli per Anomaly Detection per analisi di fisica BSM in eventi di Vector Boson Scattering

Milano/INFN

Contatti: *Gennai (INFN), Govoni (UniMIB)*
Descrizione: Sviluppo di un modello VAE per selezionare fisica BSM generata tramite operatori di EFT in processi di VBS in CMS.
Stato: Attività in corso
Esperimento: CMS
Tecnologie: CPU (per il momento attuale, stiamo progettando un nuovo modello che potrebbe richiedere GPU)

Same sign W vector bosons scattering (SSWW) with fully leptonic final state:

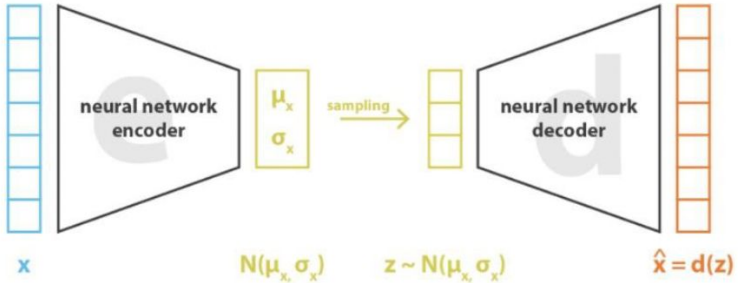
$$W^{\pm}W^{\pm}jj \rightarrow \ell^{\pm}\nu_{\ell}\ell^{\pm}\nu_{\ell}jj$$



Variational Autoencoders (VAEs) for Anomaly Detection (AD) in VBS events:

Using unsupervised learning methods for anomaly detection to search for BSM physics in Vector Boson Scattering event **The aim is to reduce the dependence of the analysis on the BSM theory assumed**

- Variational Auto Encoders (**VAEs**) VAE→VAE+DNN
- Effective Field Theories (**EFTs**)



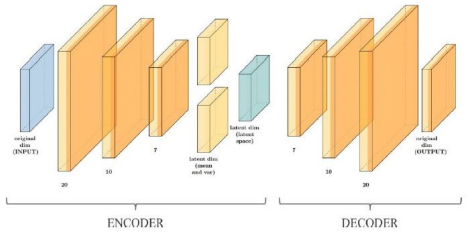
Proposed physics cases: Sviluppo di modelli per Anomaly Detection per analisi di fisica BSM in eventi di Vector Boson Scattering

The VAE+DNN Model

Milano/INFN

VAE:
Trained to reconstruct SM and EFT events

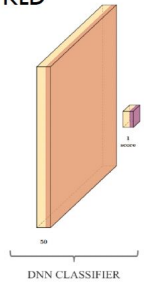
Minimized losses:
- MSE (reconstruction loss)
- KLD (regularization)

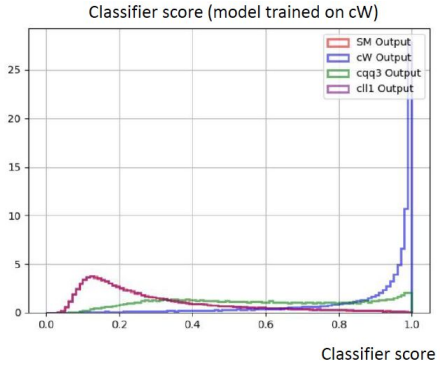


Classifier:
Can take as inputs:
- The input data (SM and EFT events)
- MSE loss computed between input and output of the VAE
- 2D loss comprised of MSE and KLD

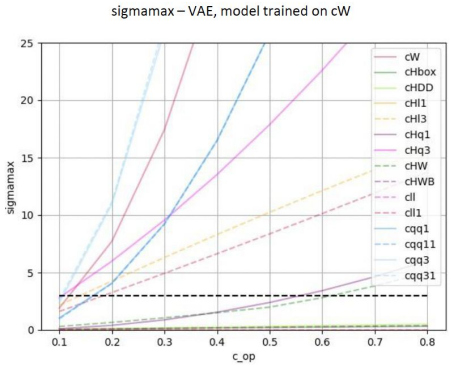
Trained to discriminate between SM and BSM events

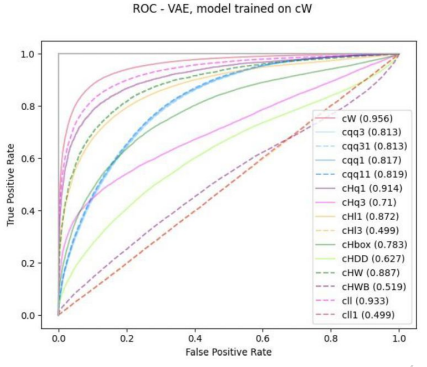
Minimized loss:
- Binary crossentropy





- ✔ It is possible to use a VAE model to single out EFT contributions due to different operators
- ✔ Though reducing the model independence of the analysis, the addition of a pure DNN classifier to the model improves its performance in terms of SM–BSM discrimination.

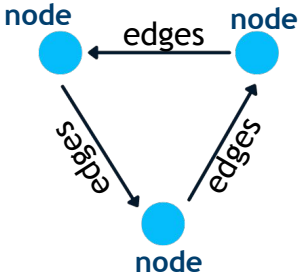


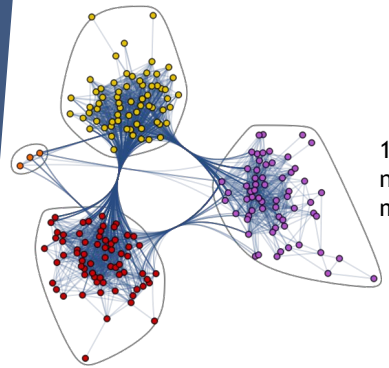


Graph Neural Network

Graph Neural Network (GNN) is a deep learning model that handles a graph as input data.

A **Graph** is the type of data structure that contains nodes and edges. A node can be a person, place, or thing, and the edges define the relationship between nodes. The edges can be directed and undirected based on directional dependencies.





Community Graph Plot by dataset

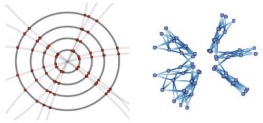
Jazz Musicians Network

198 nodes and 2742 edges: different colors of nodes represent various communities of Jazz musicians and the edges connecting them

Graphs are excellent in dealing with complex problems with relationships and interactions.

They are used in pattern recognition, social networks analysis, recommendation systems, and semantic analysis. Creating graph-based solutions is a whole new field that offers rich insights into complex and interlinked datasets.

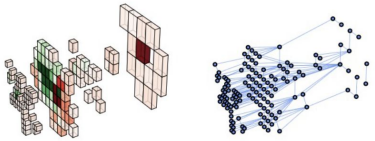
Data structure in HEP



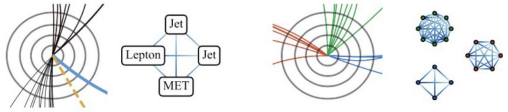
2007.13681

(a) **Tracking**

2203.12852



(b) **Calorimeter clustering**



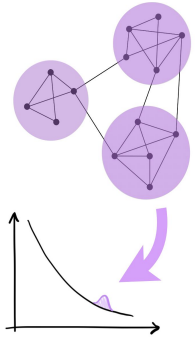
(c) **Event classification** (d) **Jet classification**

Task Definition: the first step is to decide what function one wants to learn with the GNN. In some applications this is trivial - for example jet, event or particle classification. In those cases a GNN is used to learn some representation of the node or the entire graph/set and a standard classifier is trained on that representation. For tasks such as segmentation or clustering this definition is less trivial.

Proposed physics cases: Anomaly detection e Graph Neural Network

Roma/Napoli/INFN

Contatti: Valerio Ippolito (INFN Roma 1), Elvira Rossi (Federico II),
Descrizione: Sviluppo di algoritmi basati su graph neural networks, con applicazioni per anomaly detection a livello di oggetto ricostruito e di classificazione dell'evento. Implementazione a livello di trigger e di analisi, e possibile implementazione su acceleratori tipo FPGA.
Stato: Fase prototipale iniziata basata su OpenData (LHC Olimpics)
Esperimento: ATLAS
Tecnologie: GPU, FPGA, DataLake



Anomaly Detection to be more sensitive to New Physics:
Offline algorithms to perform analysis, **e.g. heavy diboson resonances**

The idea for analysis with AD:

- Create **graphs** from energy deposits and track hits
- Build **unsupervised graph-based autoencoder** and **Anomaly Score**
- **Train only on background (Standard Model events)**
- Test on anomalous events
- Compare Standard Model distribution in Anomaly Score

Graph NN code - Up to now:

Input data: events with >2 large radius jets for an anomaly detection task (LHC Olympics dataset)

Net: GIN (graph isomorphism network) and we are evaluating the performance approaches against more conventional (**DNN without graph**): a graph neural network is implemented for the classification of large-radius jets. The network starts from the **topological clustering in the calorimeters** (topoclusters) used to build each jet and **uses the spatial information to connect N neighbors** (**nodes**) **with weights on their distance in ΔR** (**edge**); **the code run on Rome server with 2 GPU RTX 3090 24GB RAM**

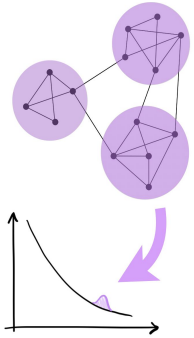
Computing time:

- bottleneck: building the graph itself as it starts from energy deposit clusters ~50 events a second
- training ~hours/ 1 day but it strongly depends on the size of the dataset

Proposed physics cases: Anomaly detection e Graph Neural Network

Roma/Napoli/INFN

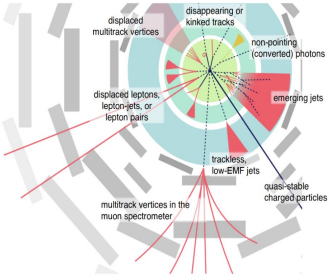
Contatti: Valerio Ippolito (INFN Roma 1), Elvira Rossi (Federico II),
Descrizione: Sviluppo di algoritmi basati su graph neural networks, con applicazioni per anomaly detection a livello di oggetto ricostruito e di classificazione dell'evento. Implementazione a livello di trigger e di analisi, e possibile implementazione su acceleratori tipo FPGA.
Stato: Fase prototipale iniziata basata su OpenData (LHC Olimpics)
Esperimento: ATLAS
Tecnologie: GPU, FPGA, DataLake



Anomaly Detection to be more sensitive to New Physics:

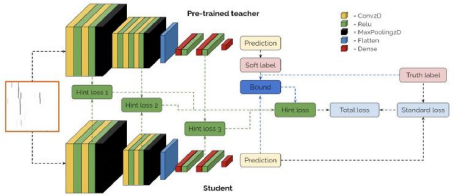
Online algorithms to select (and for not discarding) interesting event @ trigger level

Low level trigger: Ready to implement the method on FPGA with more flexibility than traditional algorithms (like pathfinder) → Ultra-small CNN models (700 parameters) trained and performance loss is recovered thanks to compression techniques; Synthesis on XCV13P FPGA performed through *HLS4ML* library but only with a serial input (for now)



HLT-Physics motivations: What if we are not sensitive to select beyond SM events @ trigger level and, instead, we discard these events?
New Physics could have unconventional anomalous signature

High level trigger: Test with Long Living Particles (LLP) decay tracks in the MS using CNN; work on a full chain from ML model to FPGA implementation is ongoing (Implementation on Xilinx Alveo U50 FPGA with *Vitis AI)*



Data, Computing infrastructure, test-beds etc…

Data policy: which data we will use (i.e. open data, embargoed, others…)?

Short term

- Survey of the computing infrastructures available in the various institutes (at WG2 level or better at spoke2 level)
- CINECA (ISCRA)/INFN (Grid), INFN-Cloud (what are the available resources? resources from related projects?)
- Can Cern resources be used (CondorGPU, lxplusGPU w/o software stack (-> cvmfs, containers,...)?
- Colaboratory, Swan GPU Server, Kubeflow, etc…
- Tools/Services: what tools/services are available and what need to be created ad hoc?

Medium term

- Develop our own test-beds customized on the basis of our use cases
- As of now difficult to be quantitative

Backup

ML applications to VBS analysis with a hadronic tau

Perugia/INFN

Workflow:

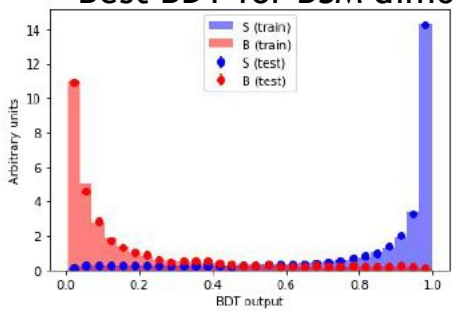
Algorithms tested:

- *BDT (XGBoost)*
- *DNN (Keras on top of Tensorflow), minmaxscaling*

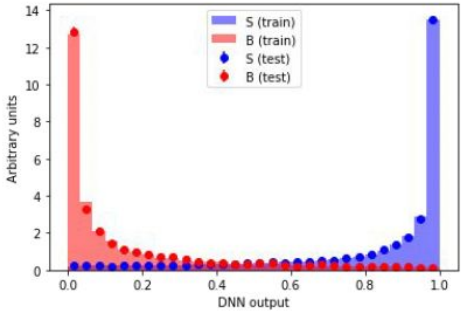
For each classifier :

- ***Feature selection***
- ***Hyperparameter optimization*** *(objective: ROC-AUC-val - |ROC-AUC-val - ROC-AUC-train|)*
  - *GridSearchCV (2fold) for BDT*
  - *Bayesian optimization (gaussian process with EI) for DNN*
- ***50-25-25 training-validation-test split***
- *Use the output of optimized algorithm as the variable for the final statistical analysis*
- *Test the impact of the algorithm in the analysis*

Best BDT for BSM dim6



Best DNN for BSM dim6



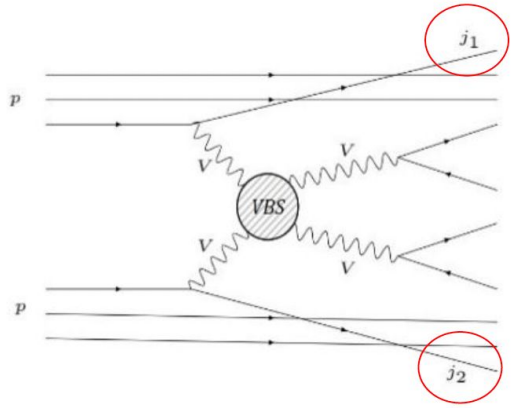
This is an ongoing activity with a lot of room for improvements and extensions

ML applications to VBS analysis with a hadronic tau

Perugia/INFN

ML applications to VBS analysis with a hadronic tau:

- ***Main application*** □ *Signal vs background discriminator*
- ***Further possible improvements:*** *VBS Jet Tagger* and *Polarization discriminator*



VBS jet tagger

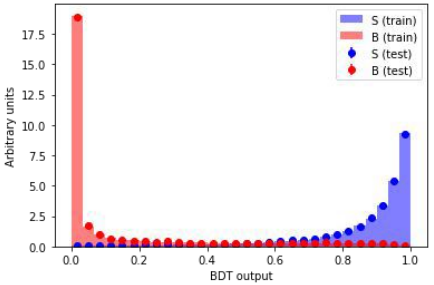
Problem: ***identification of VBS jets***

Current strategy: select jet pairs satisfying tight criteria with highest invariant mass

New strategy: select jet pairs satisfying tight criteria with highest ML algorithm score

Aim: to recognise true VBS jets pairs against fakes - The trained algorithm is then implemented in the analysis: for each event, the algorithm is applied to each pair of tight jets, and the pair with highest score is selected as VBS jet

Algorithm: XGBoost Classifier, optimized with GridSearch (objective: ROC-AUC-val - |ROC-AUC-val -ROC-AUC-train|)



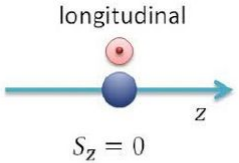
ML applications to VBS analysis with a hadronic tau

Perugia/INFN

ML applications to VBS analysis with a hadronic tau:

- ***Main application*** □ *Signal vs background discriminator*
- ***Further possible improvements:*** *VBS Jet Tagger* and *Polarization discriminator*

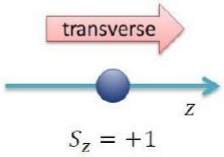
Polarization discriminator



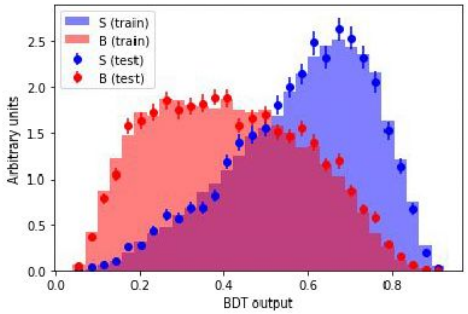
ML study performed in order to discriminate among different W-boson polarization states:

- *W boson polarization is related to its decay products kinematics*
- *Challenges: polarization studies in a analysis with hadronic tau decays*

The idea is to discriminate events with longitudinal (LL) polarization from transverse (XT) polarization using info related to leptons.



Algorithm: XGBoost Classifier, optimized with GridSearch (objective: ROC-AUC-val - |ROC-AUC-val -ROC-AUC-train|)



ML applications to VBS analysis with a hadronic tau

Perugia/INFN

Tentative forcecast of resources:

- Use of nanoAOD (a new compact event data format in CMS ▫ Storage: ~2TB
- First tests will require: ~10 core e and 20 GB ▫
- In full operation: O(100 core) and O(200GB ram)
- Software requirements: standard python libraries like tensorflow, xgboost et al
- To train DNN:
  - using “derived-feature” ▫ CPUs
  - using “raw feature” ▫ GPUs

Classification with domain adaptation

B. Camaiani, V. Ciulli, P. Lenzi, M. Lizzo, L. Viliani (Florence)

Goal: **perform signal to background discrimination in a way that is agnostic with respect to the signal model** → reduce/remove bias towards the model used in the training

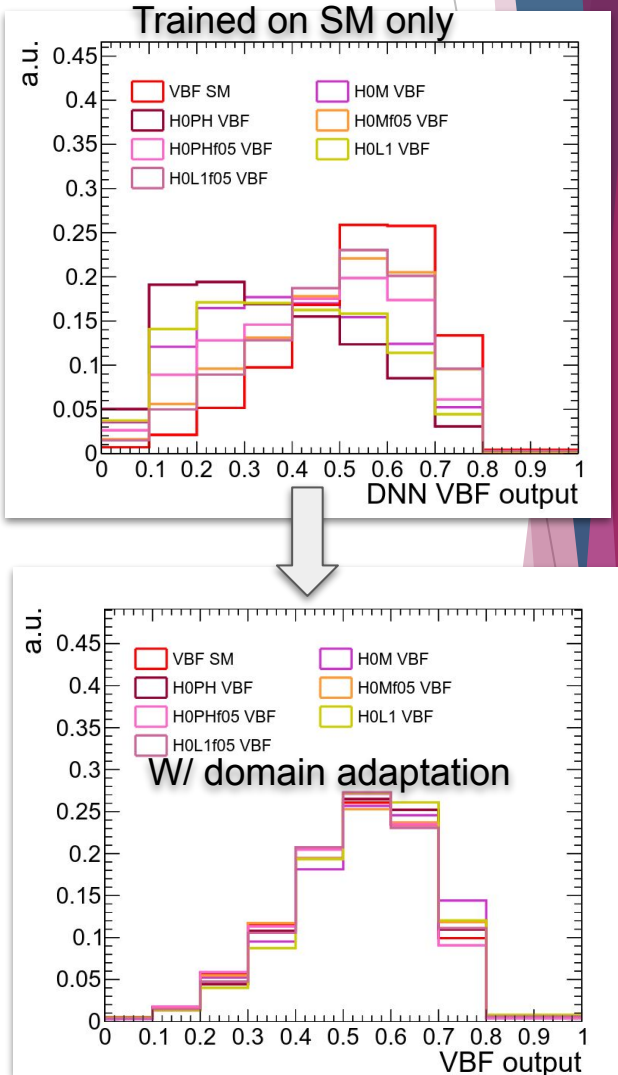
Means: Achieved with an **adversarial approach**, in which the **S/B discriminator is penalized if an adversary network can correctly discriminate between SM and BSM**

Case studied in the context of Higgs cross section measurement https://arxiv.org/abs/2207.09293 (Accepted by EPJC)

Similar to a case studied in ATLAS (F. Peri PhD thesis and github) to design **"systematic-insensitive" observables**

Strategy applicable in many cases whenever designing a model that is insensitive to a given quantity is desirable

Collaboration ongoing with Baker & Hughes Oil & Gas (Florence) to regress temperature/pressure in turbines operated in different conditions with respect to those of a testbed



Resources & Plans

B. Camaiani, V. Ciulli, P. Lenzi, M. Lizzo, L. Viliani (Florence)

The training of this network (~100k events) is ~1h on a CPU, going down to minutes on Tesla T4 GPU instantiated via INFN cloud.

Tools used are **TensorFlow** (with a bit of Keras, but the network is ad-hoc) and **Optuna** for optimization

Optimization of the network is demanding, as it requires to the very least 100 sequential trials to explore 6 hyperparameters

Improvements to explore in the context of CN1 (**in close relationship with WP5**):

- Optimization would benefit from **parallelization strategies** (project ongoing with Fondazione Cassa di Risparmio di Firenze, PI Lorenzo Viliani, currently recruiting).
- Explore **distributed training strategies**.

Both could be useful for many applications, and the domain adaptation via adversarial training itself is applicable to many different use cases.

A testbed with ~10 GPU (guesstimate) to test optimization and training parallelization would be desirable.

Difficult to estimate a production phase for this specific use case, but a GPU farm would benefit many cases.

The activity is currently ongoing and we are exploring other fields of application (discrimination with limited systematic dependency and anomaly detection, see the talk from S. Gennai).

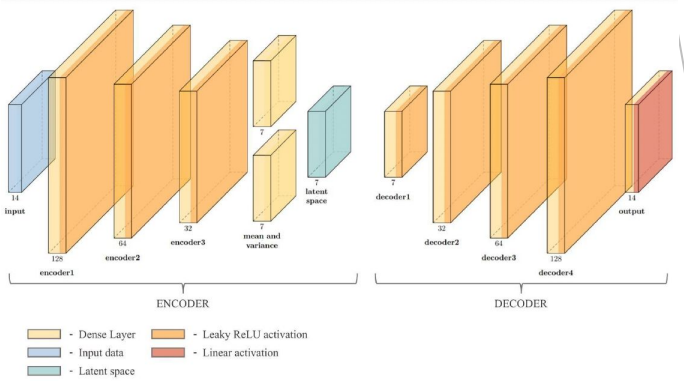
Sviluppo di modelli per Anomaly Detection per analisi di fisica BSM in eventi di Vector Boson Scattering

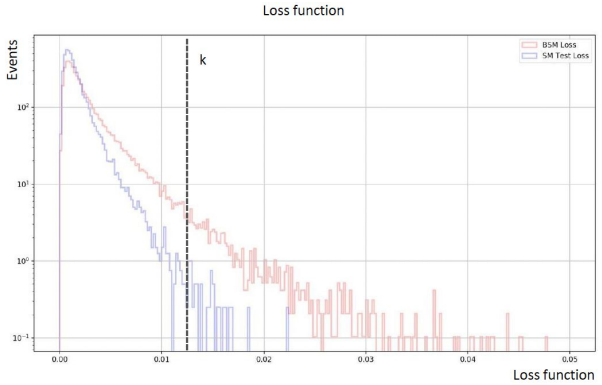
Milano/INFN

Encoding phase: the VAE maps an input as a distribution (mean and variance) in the so-called latent space
Sampling: a point is then sampled from the distribution in the latent space (generative model)
Decoding: the sampled point is decoded

The **model** is built via sub classing on Tensor Flow and Keras libraries
Training on SM: The training of the VAE is performed on a SM sample
Test on BSM=SM+EFT:
The BSM sample comprises EFT events,
which are badly reconstructed, resulting
in higher values of the loss function

→Anomalies are expected to lie in the tail of the loss function



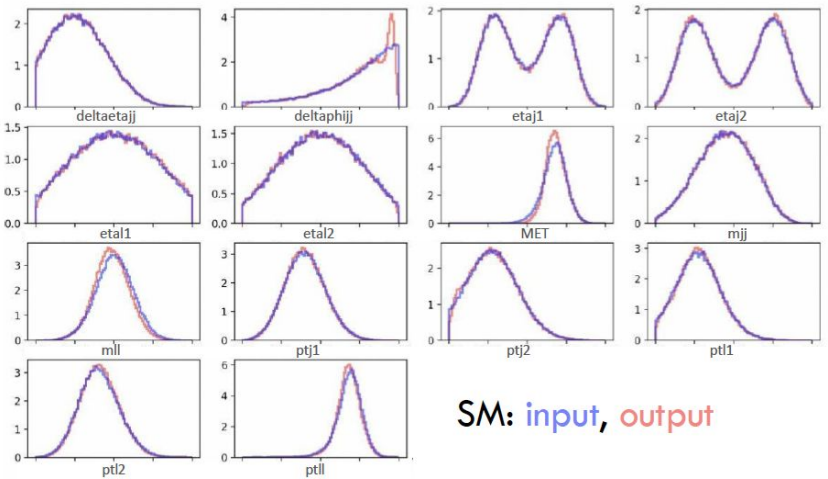


Significance: $\sigma = \frac{|BSM - SM|}{\sqrt{SM}} = \frac{|LIN + QUAD|}{\sqrt{SM}}$

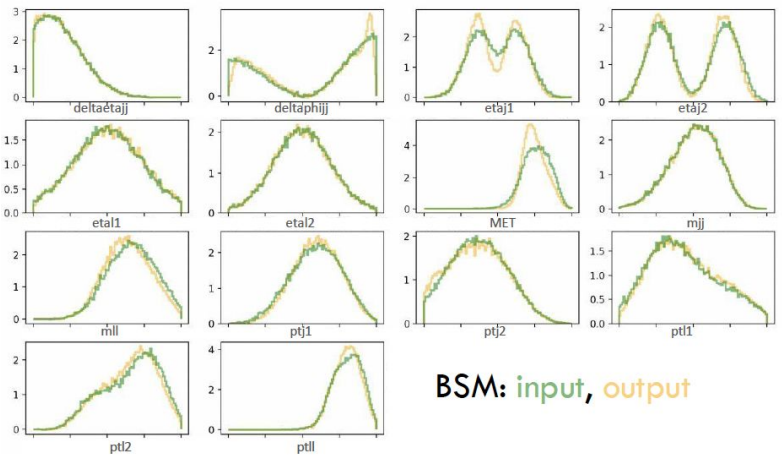
Sviluppo di modelli per Anomaly Detection per analisi di fisica BSM in eventi di Vector Boson Scattering

Milano/INFN

RECONSTRUCTION OF THE SAMPLES



SM: input, output



BSM: input, output

The SM sample is reconstructed better than that comprising both SM and EFT contributions