Methodology for Processing Open Data for Machine Learning Models in BSM Searches

Volkov P.*[(1)], Abasov E.[(1)], Dudko L.[(1)], Markina A.[(1)], Vorotnikov G.[(1)], Perfilov M.[(1)], Zaborenko A.[(1)], Savkova N.[(1)], Gorin D.E.[(2)], Vasilevskii O.S.[(2)]

[(1)]SINP MSU [(2)]Physical Faculty MSU

DLCP2026, 7.06.2026

Motivation: Dark Matter searches & Foundation models

- **Dark Matter**: simplified models with scalar/pseudoscalar mediators in **multi-top** final states ($t\bar{t}tW$, $t\bar{t}t\bar{t}$, ...); many SM backgrounds share the same topology
    - Phys. Part. Nuclei 56 (2025) 440
    - Moscow Univ. Phys. Bull. 80 (2025) S816
- **FCNC:** flavor-changing neutral currents (e.g. $tu\gamma$, tug) in single-top production; real CMS analysis target; hand-crafted features vs. raw momenta as classifier inputs
- **Foundation models:** one model pretrained on a broad process pool (0–4 tops) and finetuned for assignment, classification, regression, and unfolding — instead of a new BDT/DNN per analysis
- **LGaTr:** Lorentz Geometric Algebra Transformer; event tokens as multivectors in $\mathrm{Cl}(1,3)$ with **Lorentz equivariance** by construction (arXiv:2605.15910)

Common need: large Monte Carlo simulation (MC) samples + real data → curated, weighted, comparable event tables at scale (Open Data pipeline).

Two paths to ML-ready datasets

Path 1: simulate everything

- generators: **MadGraph**, **CompHEP**, ...
- requires **large computing resources**: background pool, detector effects, systematics
- **reconstruction is hard**: complex final states (multi-top, many jets)
- **no data** $\rightarrow$ must prove modeling is correct

Path 2: use LHC MC and data

- **LHC collision data + MC samples** for training and validation
- best validation of weights, reconstruction, and uncertainties
- but access is mostly **restricted** (Grid, accounts, CMSSW workflows)

CMS Open Data provides a public bridge: real data + MC $\rightarrow$ curated, weighted, comparable event tables at scale.

CMS Open Data: public datasets anyone can use

- **CMS Open Data:** CMS releases selected **real collision data** and **MC samples** to the public
- **License** CC BY 4.0; each dataset has a **DOI** — citable, reproducible
- Read over the network (`xrootd`) from opendata.cern.ch — **no grid credentials**
- Enables independent validation, teaching, and ML on **real collider data** outside the collaboration
- **This work:** build a reproducible pipeline Open Data $\rightarrow$ **flat, weighted tables** ready for neural-network training on a multi-process sample pool

MC/Data (NanoAOD format) → ML model

Reproducible pipeline for DM & cross-process ML on a shared sample pool:

MC/Data (NanoAOD) → event selection → weighted events → features (vars) →

↓ ML model → data analysis

Next stage (outside the scope of this talk).

1. **NanoAOD** — read Open Data (MC + collision data) via `xrootd`; flat input branches per event
2. **Event selection** — skim: trigger, lepton/jet/b-tag cuts, signal region (1–4 lepton, 2–12 jets); merge chunks → selected events
3. **Weighted events** — attach corrections: generator weights, lepton/b-tag SFs, PDF/scale/systematic **weight branches** per event
4. **Features (vars)** — compute kinematic & topology features
5. **ML model** — train on the unified schema; same format for all processes in the pool

NanoAOD format

- **TTree** (ROOT `TTree` objects): one row per collision event; object multiplicity via variable-length arrays
- **Lepton collections:**
  - `Muon`: `Muon_pt`, `Muon_eta`, `Muon_phi`, `Muon_charge`, isolation, ID flags
  - `Electron`: analogous branches (available; channel chosen per analysis)
- **Jet collection:** `Jet_pt`, `Jet_eta`, `Jet_phi`, `Jet_mass`, `Jet_btagDeepFlavB`, `Jet_btagCSVV2`, …
- **Missing transverse energy:** `MET_pt`, `MET_phi` (type-1 corrected in NanoAOD)
- **Vertices:** `PV_npvs` — primary vertex multiplicity
- **Generator-level** (MC only):
  - `GenPart`: generator particles — `GenPart_pt`, `GenPart_eta`, `GenPart_phi`, `GenPart_pdgId`, …
  - weights: `genWeight`, `LHEPdfWeight`, `LHEScaleWeight`, PS weights

Flat branches → selected physics objects → one analysis row per accepted event

Reconstructed objects for ML feature extraction

Leptons

- p_{T}, η, ϕ, charge
- relative isolation, muon/electron ID
- trigger matching (data)

Jets

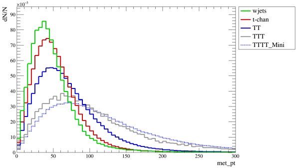
- p_{T}, η, ϕ, mass
- JER smearing applied nominally in engine

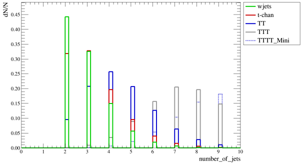
***b*-tagging**

- DeepJet score (UL2016 WP)

High level particles

- $m_{\mathrm{T}}(\ell, \mathrm{MET})$, dilepton/dijet masses
- analytic neutrino p_z solver (Orso algorithm)
- reconstructed top/W kinematics for topology variables





Embedded corrections & systematic weights

Systematic uncertainties stored as **per-event weight columns** — no full reco re-run per variation.

Source	Branches	Effect
Generator PDF	`weight_pdfUp/Down`	shape (+ norm envelope)
Ren./fac. scales	`weight_Ren/Fac/RenFac Up/Down`	norm (9-point `LHEScaleWeight`)
b-tag efficiency	`weight_btag_*`, `weight_btag_*_jes`	shape (incl. JES proxy)
JEC/JER corrections	jet energy scale/resolution variations	shape
Lepton SFs	`weight_LepId`, `weight_LepIso`, `weight_LepTrig`	norm
Pileup, prefire	`weight_pu`, `weight_pref*`	norm
ISR / FSR	`weight_Isr*`, `weight_Fsr*`	shape

Central weight: `weight` = `genWeight` $\times$ SFs.

Part II

Toy example: validate weights & shapes
against a published CMS analysis

Toy example: why and how we validate

- Shapes on MC are necessary but **not sufficient** — weights, SFs, and systematics must be checked too
- Only way to validate the weight machinery: compare with a **published CMS measurement** on Open Data
- **Benchmark:** CMS single-top t-channel, μ+jets, 2 jets + 1 b-tag (CMS-TOP-17-011)
- Compare **MC/data ratios** and **distribution shapes** (16.8 vs 35.9 $\mathrm{fb}^{-1} \rightarrow$ shapes only)
- MC: t-channel, $t\bar{t}$, tW, W+jets, DY, QCD; data SingleMuon G+H

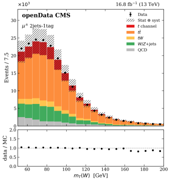
Toy example: event selection & regions

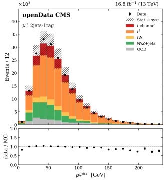
Channel: single muon ($p_{\mathrm{T}} > 26$ GeV, $|\eta| < 2.4$, relIso< 0.06)

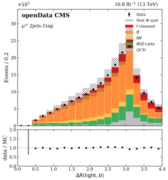
Cutflow (example MC signal): all events $\rightarrow$ trigger+PV $\rightarrow 1\ell \rightarrow$ 2 jets ($p_{\mathrm{T}} > 40$ GeV) $\rightarrow$ 1 b-tag

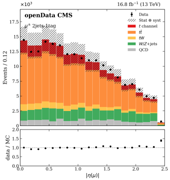
Region	Definition
SR (base)	2 jets + 1 b-tag — variable tables
SR (plots / DNN)	+ $m_{\mathrm{T}}(W) > 50$ GeV

108 output columns: 12 reference vars + 45 extended topology vars + weight branches.



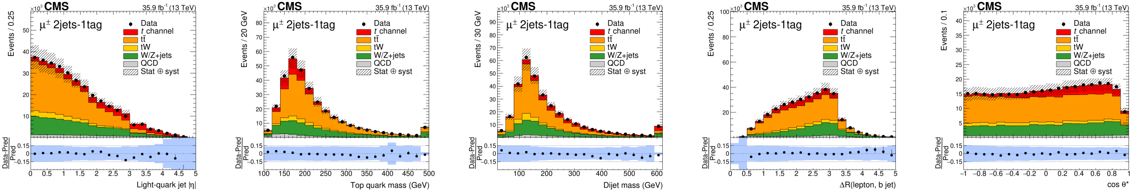




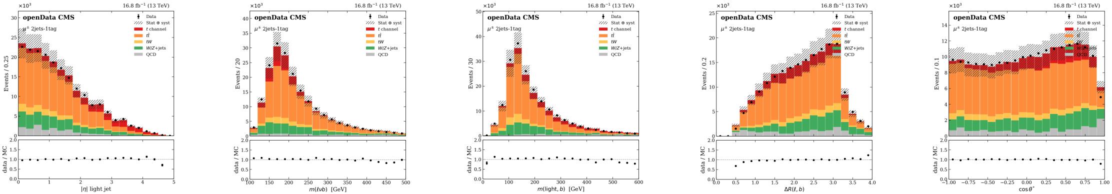


Toy example: input variables (CMS vs Open Data)

CMS

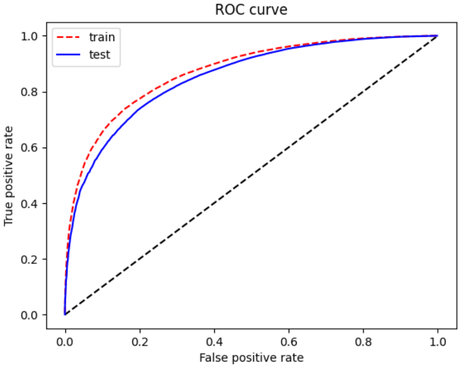


Open Data

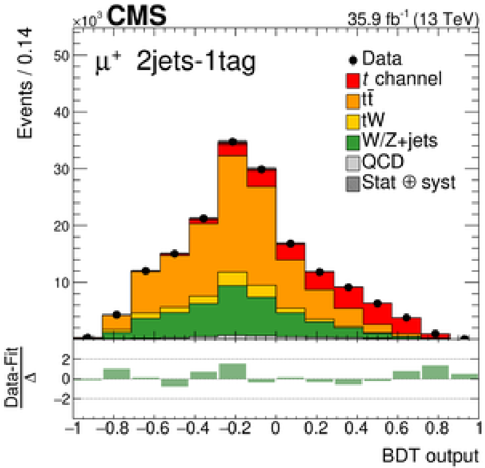


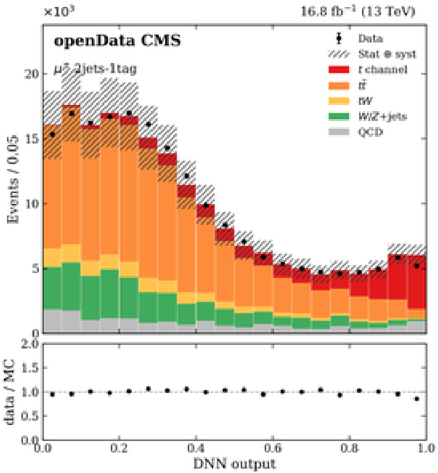
Toy example: DNN performance summary

- Architecture: 2 hidden layers $\times$ 100 neurons, dropout 0.2
- Train / exam: 399,999 / 399,771 events; weight `weight_scl`$\times(m_{\mathrm{T}}(W)>50)$
- ROC-AUC: train 0.8687 ± 0.0017; exam 0.8492 ± 0.0018
- Toolchain **bnn-hep** $\rightarrow$ PyTorch on Open Data derivatives



Toy example: BDT (paper) vs DNN 2 × 100 (ours)





Conclusion

- **Pipeline is ready:** Open Data $\rightarrow$ flat weighted tables with corrections and systematic variations
- **Toy example:** good MC/data agreement in a published CMS analysis (shapes), with systematics treated via weight columns
- **Next step:** full background processing and validation/control plots at scale — **already in progress**
- **Extend:** check additional decay channels / event topologies beyond the μ+jets benchmark
- **Framework availability:** the processing framework will be available ASAP at github.com/higen-sinp

Product: a reusable Open Data $\rightarrow$ ML factory for DM, FCNC, and cross-process foundation-model studies.