# About myself

PhD: KIT 2016
Postdoc: MIT 2016-2021
Right now: CERN (Senior Research Fellow)

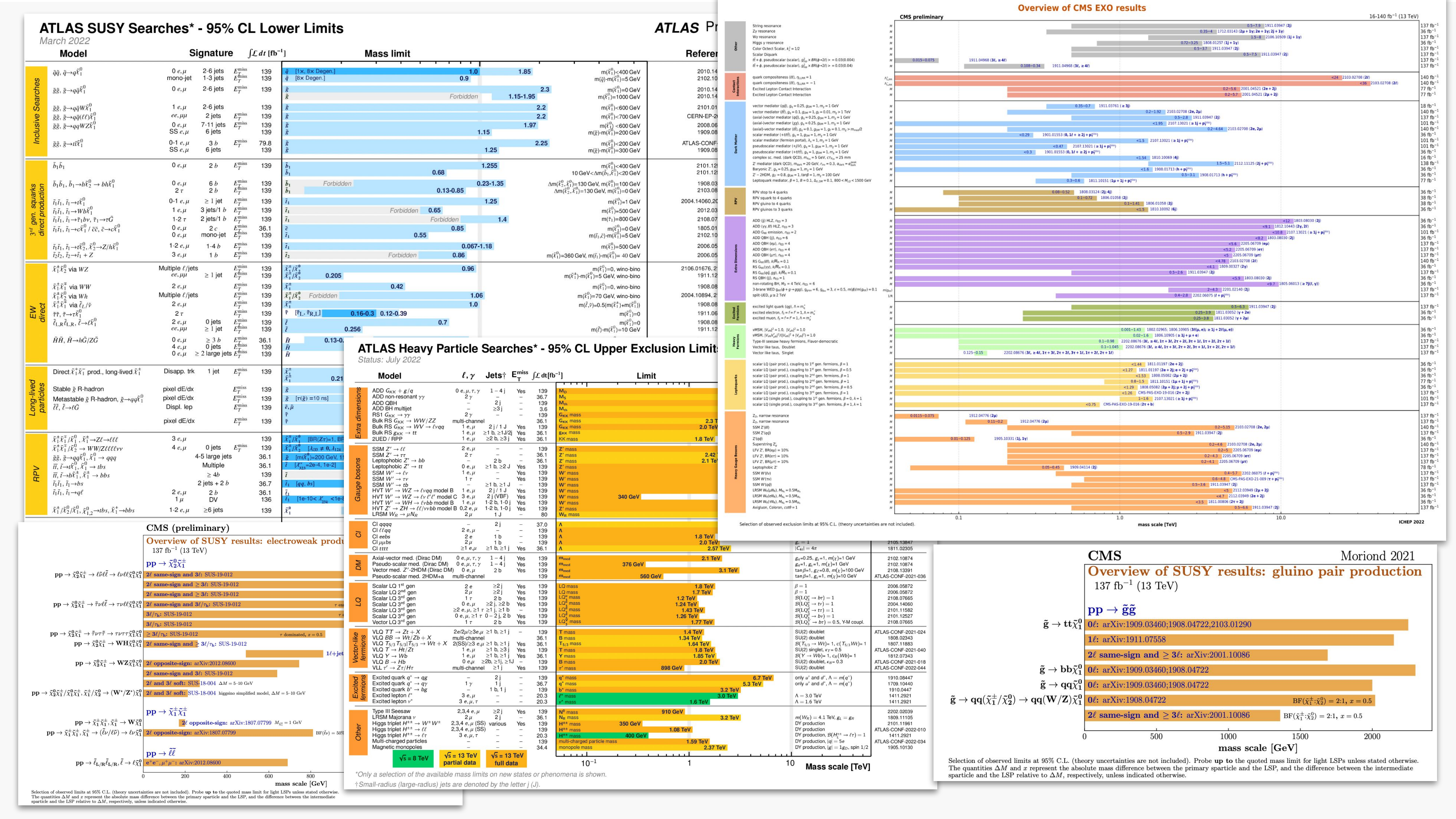
**Principal author of multiple CMS analyses** (mostly focussed on dark matter and dark interactions)

Several **leading roles** in the collaboration: Coordinator of all CMS dark matter searches (2020-2022) and co-leader of CMS data management group (since 2018)

Since September 2022: Co-coordinator of Missing Transverse Energy Object Group

Active contributor to Dark Matter LHC Working Group

Continuous **publication track record outside of CMS**, referee for several journals


ATLAS SUSY Searches* - 95% CL Lower Limits
March 2022
Model
Signature
$\int \mathcal{L}\,dt$ [fb$^{-1}$]
Mass limit
Inclusive Searches
$3^{rd}$ gen. squarks direct production
EW direct
Long-lived particles
RPV
Overview of CMS EXO results
CMS preliminary
16-140 fb$^{-1}$ (13 TeV)
Other
Contact Interactions
Dark Matter
RPV
Extra Dimensions
Excited Fermions
Heavy Fermions
Leptoquarks
Heavy Gauge Bosons
mass scale [TeV]
Selection of observed exclusion limits at 95% C.L. (theory uncertainties are not included).
ICHEP 2022
ATLAS Heavy Particle Searches* - 95% CL Upper Exclusion Limits
Status: July 2022
Model
$\ell, \gamma$
Jets†
$E_T^{miss}$
$\int \mathcal{L}\,dt$[fb$^{-1}$]
Limit
Extra dimensions
Gauge bosons
CI
DM
LQ
Vector-like fermions
Excited fermions
Other
$\sqrt{s}$ = 8 TeV
$\sqrt{s}$ = 13 TeV partial data
$\sqrt{s}$ = 13 TeV full data
Mass scale [TeV]
*Only a selection of the available mass limits on new states or phenomena is shown.
†Small-radius (large-radius) jets are denoted by the letter j (J).
CMS (preliminary)
Overview of SUSY results: electroweak production
137 fb$^{-1}$ (13 TeV)
mass scale [GeV]
Selection of observed limits at 95% C.L. (theory uncertainties are not included). Probe up to the quoted mass limit for light LSPs unless stated otherwise. The quantities $\Delta M$ and $x$ represent the absolute mass difference between the primary sparticle and the LSP, and the difference between the intermediate sparticle and the LSP relative to $\Delta M$, respectively, unless indicated otherwise.
CMS
Moriond 2021
Overview of SUSY results: gluino pair production
137 fb$^{-1}$ (13 TeV)
$pp \rightarrow \tilde{g}\tilde{g}$
$\tilde{g} \rightarrow tt\tilde{\chi}_1^0$
0ℓ: arXiv:1909.03460;1908.04722,2103.01290
1ℓ: arXiv:1911.07558
2ℓ same-sign and ≥ 3ℓ: arXiv:2001.10086
$\tilde{g} \rightarrow bb\tilde{\chi}_1^0$
0ℓ: arXiv:1909.03460;1908.04722
$\tilde{g} \rightarrow qq\tilde{\chi}_1^0$
0ℓ: arXiv:1909.03460;1908.04722
$\tilde{g} \rightarrow qq(\tilde{\chi}_1^\pm/\tilde{\chi}_2^0) \rightarrow qq(W/Z)\tilde{\chi}_1^0$
0ℓ: arXiv:1908.04722
2ℓ same-sign and ≥ 3ℓ: arXiv:2001.10086
BF($\tilde{\chi}_1^\pm$:$\tilde{\chi}_2^0$) = 2:1, $x$ = 0.5
mass scale [GeV]
Selection of observed limits at 95% C.L. (theory uncertainties are not included). Probe up to the quoted mass limit for light LSPs unless stated otherwise. The quantities $\Delta M$ and $x$ represent the absolute mass difference between the primary sparticle and the LSP, and the difference between the intermediate sparticle and the LSP relative to $\Delta M$, respectively, unless indicated otherwise.

# Machine-learning for Maximally Model-Independent Analyses @ LHC

Benedikt Maier

September 12, 2022

# Introduction

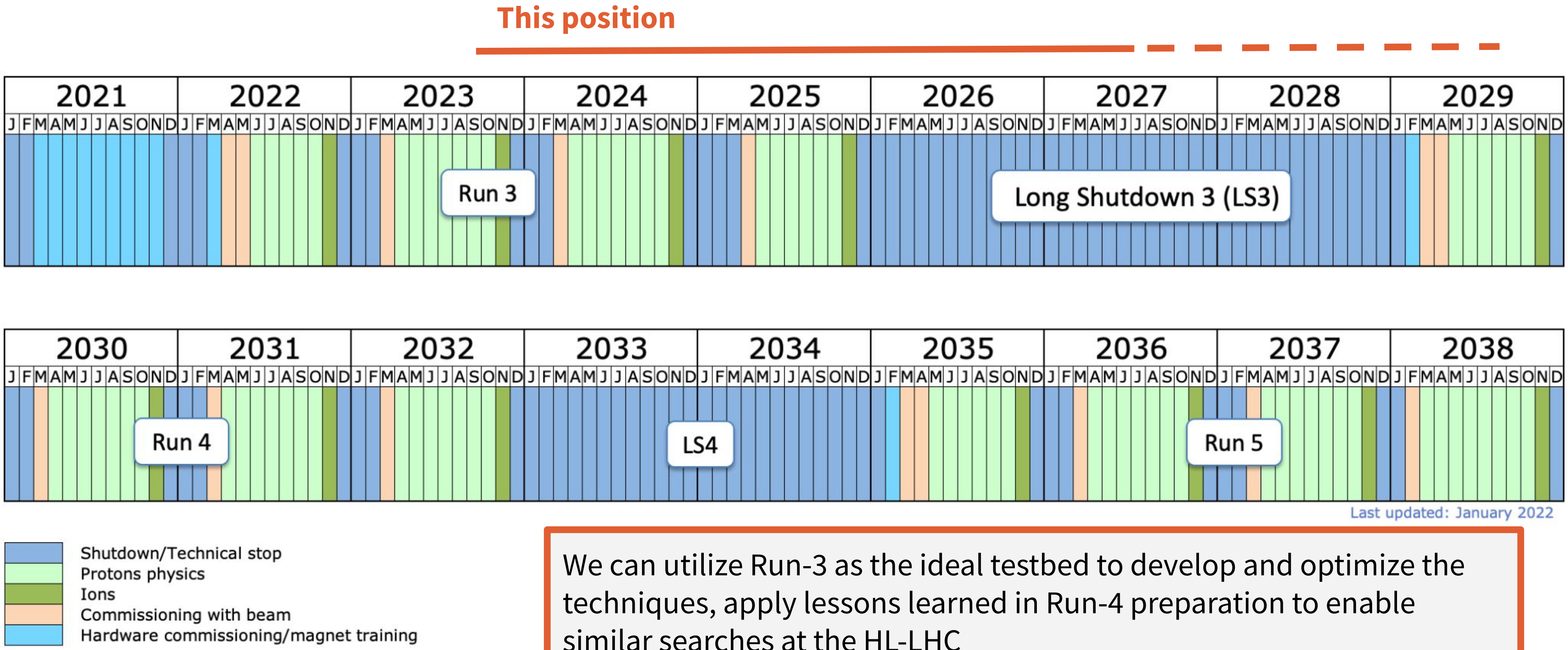
## Where do we stand?

- Searches for new physics are generally highly tuned for or target specific signal models
  - In 99% of the cases, the analyses are actually starting with a specific signal in mind
- While this approach allows us to tune the analysis to that specific signal, it increases the **chances of missing a potential excess in data** in other corners of phase space / in other observables
- We need to worry about coverage, and we cannot afford to fill each hole with 10 additional highly tuned analyses
  - Even if theorists had an idea for 10 additional signal models for each hole, we don't have the personpower

## Conclusion

- Run-3 starting now, **right time for a shift in focus**, away from searches targeting specific BSM models to maximally model-independent analysis strategies
- Increasing the coverage dramatically, but the new analysis strategies proposed are very challenging

# LHC Schedule vs. Project Timeline




We can utilize Run-3 as the ideal testbed to develop and optimize the techniques, apply lessons learned in Run-4 preparation to enable similar searches at the HL-LHC

# How to design a model-independent analysis

**Challenge**

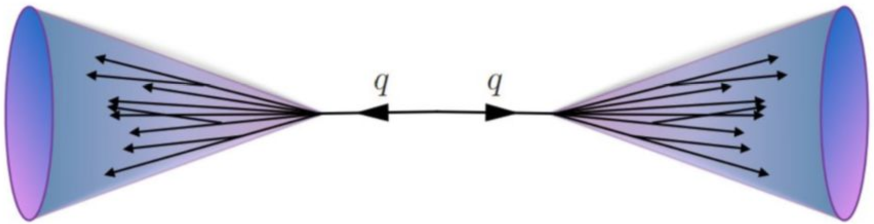
- Find **new physics in final states dominated by hadronic activity**
- Model-agnostic to not be (mis)guided by signal specifics
- Data-driven: limit use of Monte Carlo simulation to the bare minimum (possible even without signal MC)

**Idea**

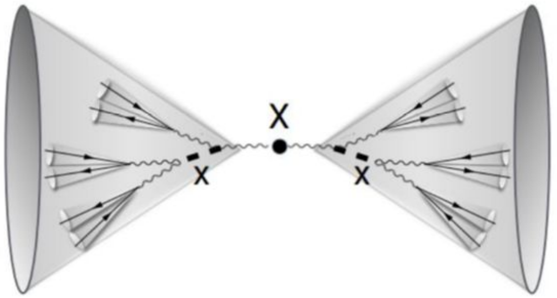
- Selection based on **SM veto**
- Define this veto using a machine-learning algorithm that learns how SM looks like
  - → train on data in control region
- Not targeting a specific BSM scenario … in general anything non-QCD like


# Anomaly Detection

# Incarnation 1: Events with boosted di-jets



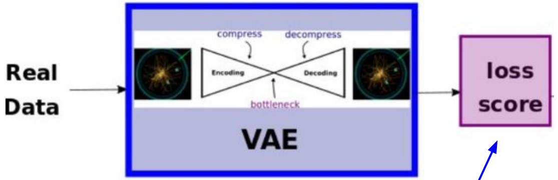

AK8 large-radius jets




**Goal**

- Probe for new physics in di-jet final state in a purely data-driven way
- There is a plethora of good reasons for a new heavy particles that decays into two boosted jets
- Can we design a catch-all analysis workflow based on anomaly detection?
- Typically, the two jets would have exotic **substructure**
  - I.e., non-QCD-like prongginess, color flow, etc
- Ideal setting for ML algorithms feeding on information of jet constituents

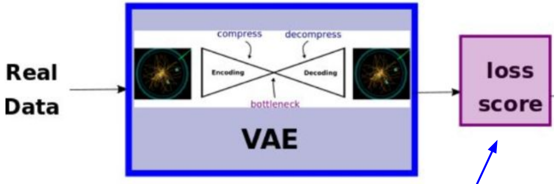
# Analysis workflow

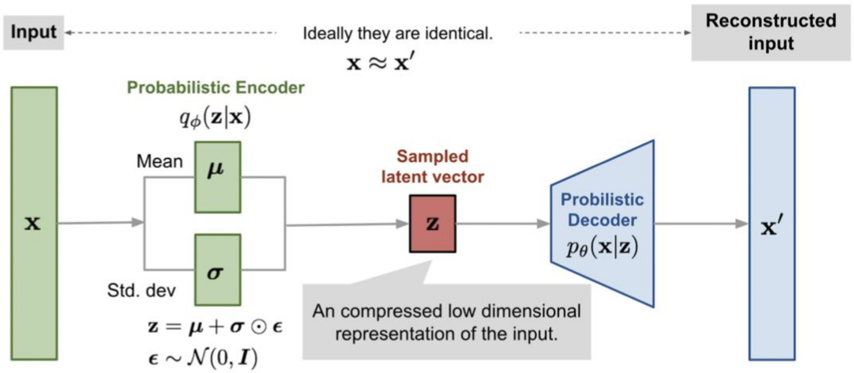



Train variational autoencoder to learn what QCD jets look like

This algorithm returns an *anomaly score* for a given jet

# Analysis workflow



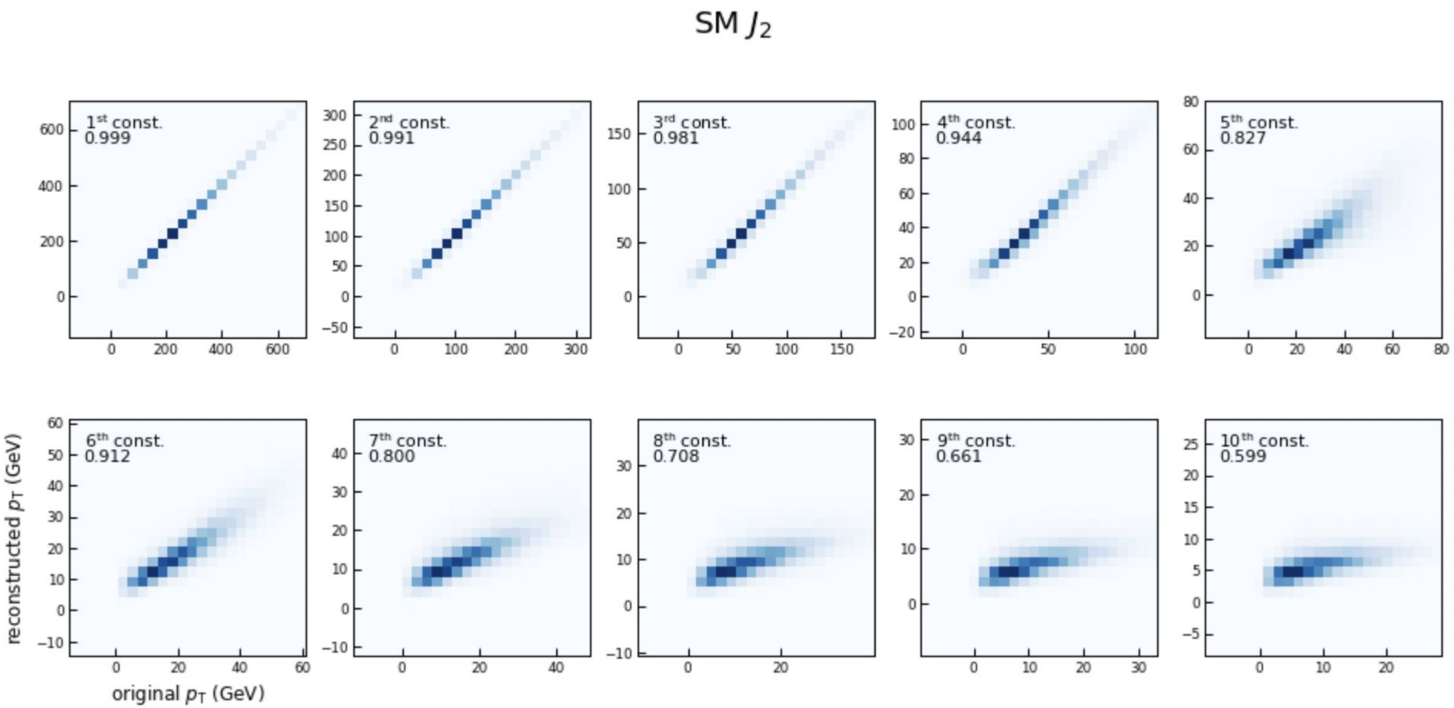

Train variational autoencoder to learn what QCD jets look like

This algorithm returns an *anomaly score* for a given jet

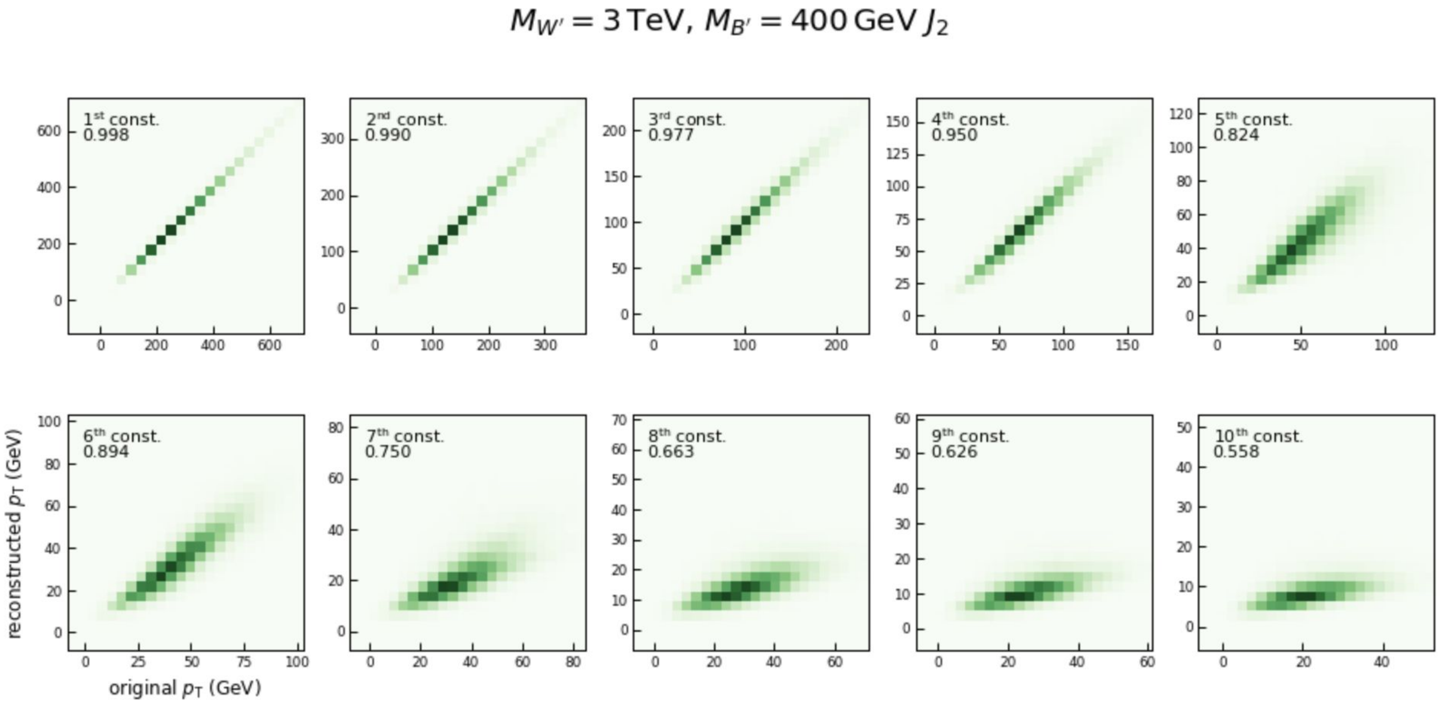
**The training can be done on data, on jets from a $\Delta\eta$(jj) sideband**

- Pure QCD jets
- VAE learns to compress QCD jets and decompress, i.e., *reconstruct* them
- Will fail at reconstructing non-QCD-like / exotic jets

# Reconstruction quality for QCD jets

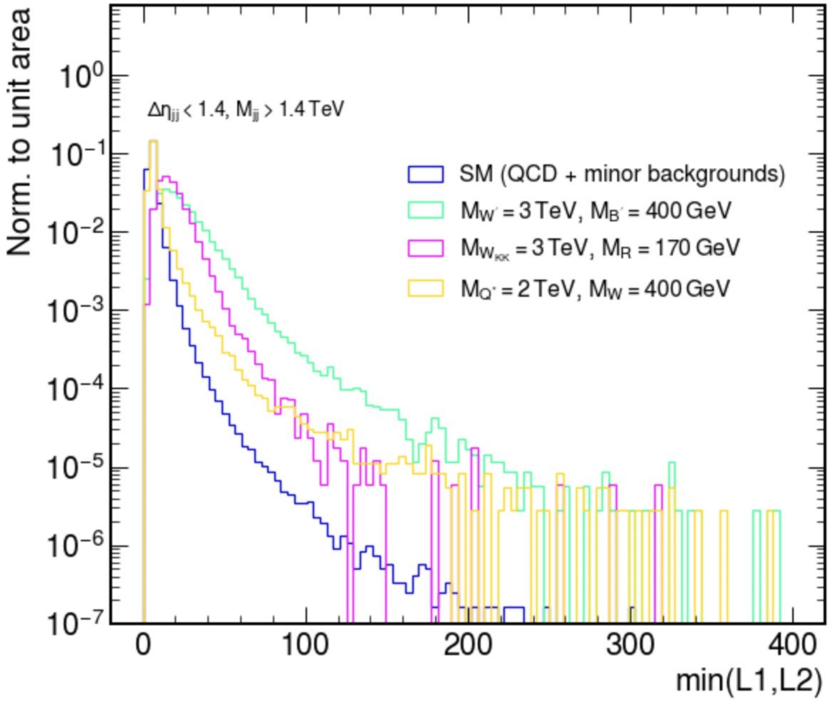


# Reconstruction quality for a BSM signal




Constituents consistently **get reconstructed worse** for this BSM signal

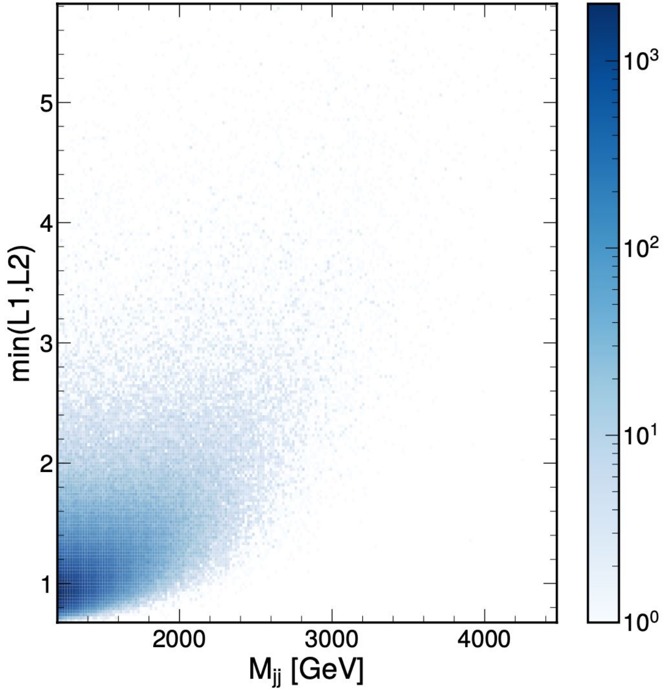
# Discrimination power against various BSM models




**Combining the two jet losses into one powerful event discriminator**

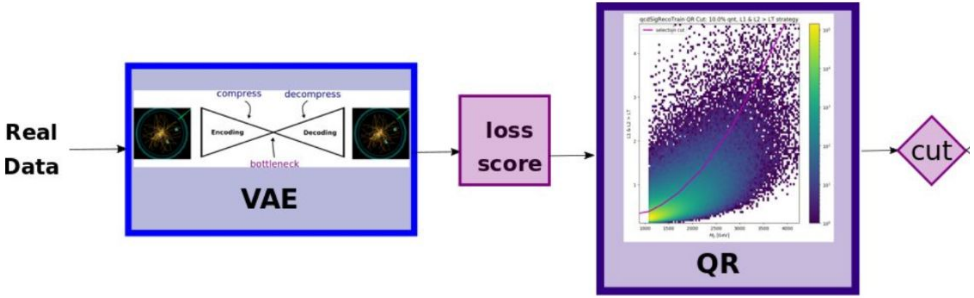
- Classifies well against an entire suite of different signal models predicting non-QCD-like substructure

# What to do with this discriminator?




**Can we somehow bin in this 2D space?**

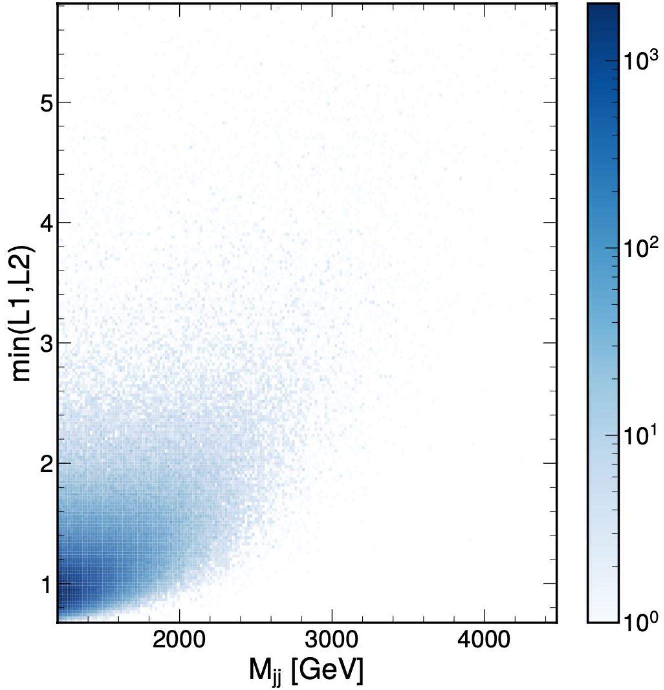
# Analysis workflow




Perform a “quantile regression” to yield a desired target acceptance as a function of observable of choice

Cut leaves background unsculpted

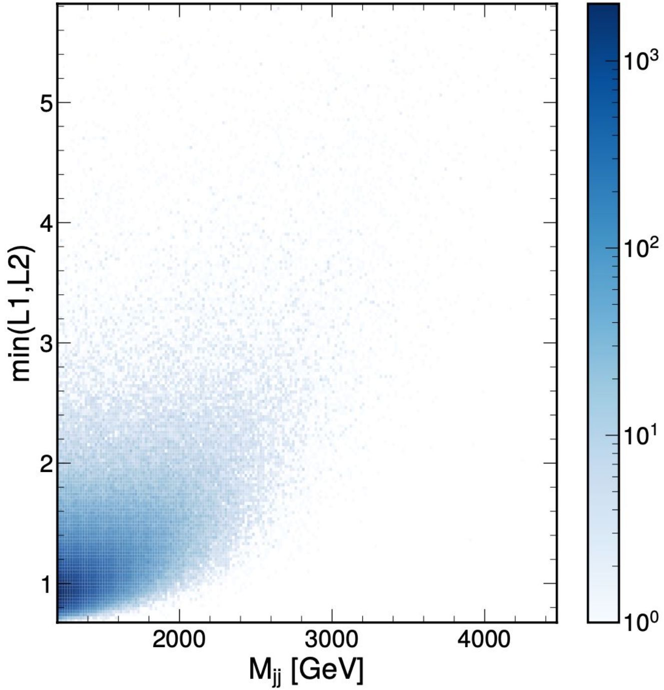
# Leaving $M_{jj}$ unsculpted

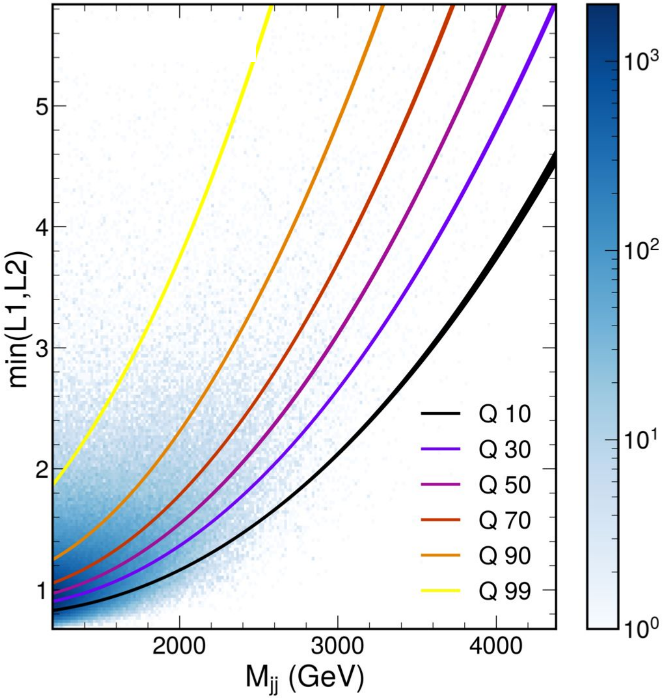



**Train a neural network to find cut as function of dijet mass that gives desired background acceptance → "Quantile regression"**

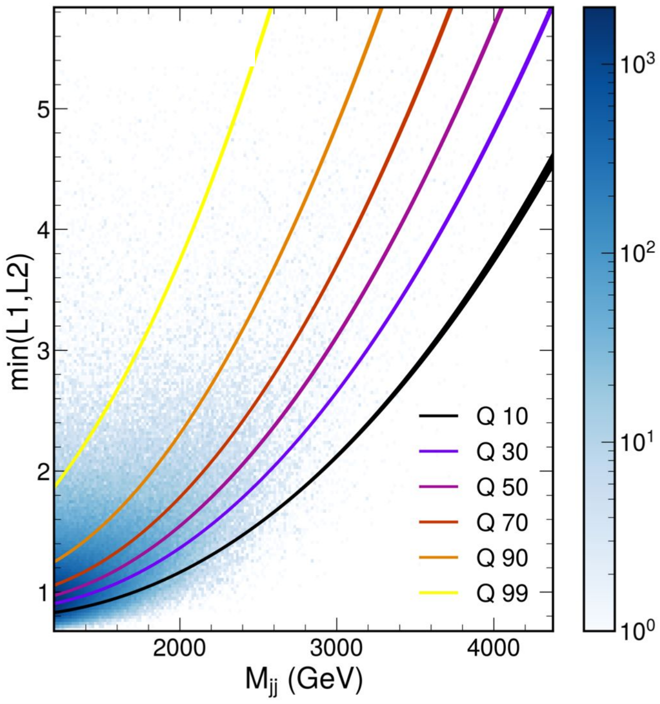
- Allowing to "bin" the space in anomalousness and $M_{JJ}$
- Resulting in unsculpted $M_{JJ}$ spectra that make background estimation straightforward (comes down to ~overall normalization factor)

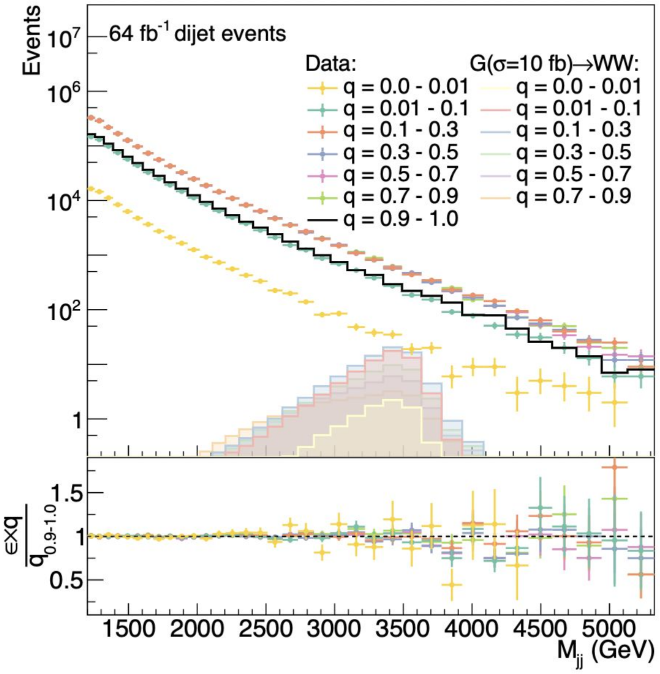
# Leaving $M_{jj}$ unsculpted



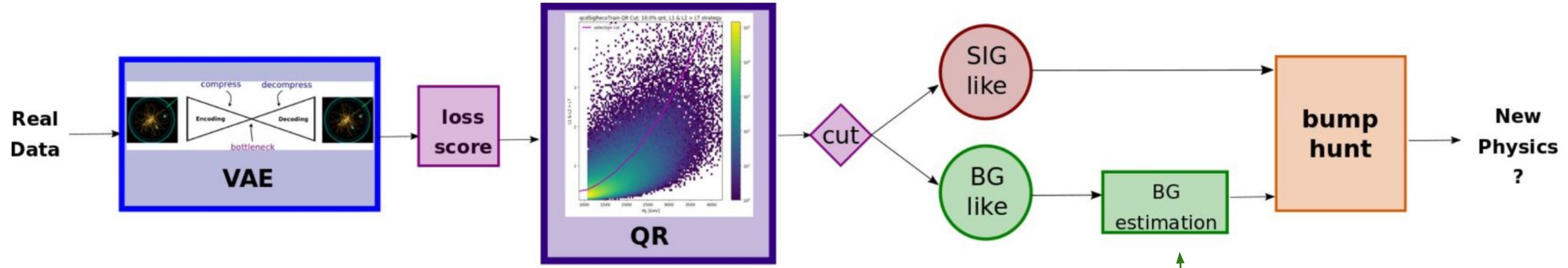



# Leaving $M_{jj}$ unsculpted



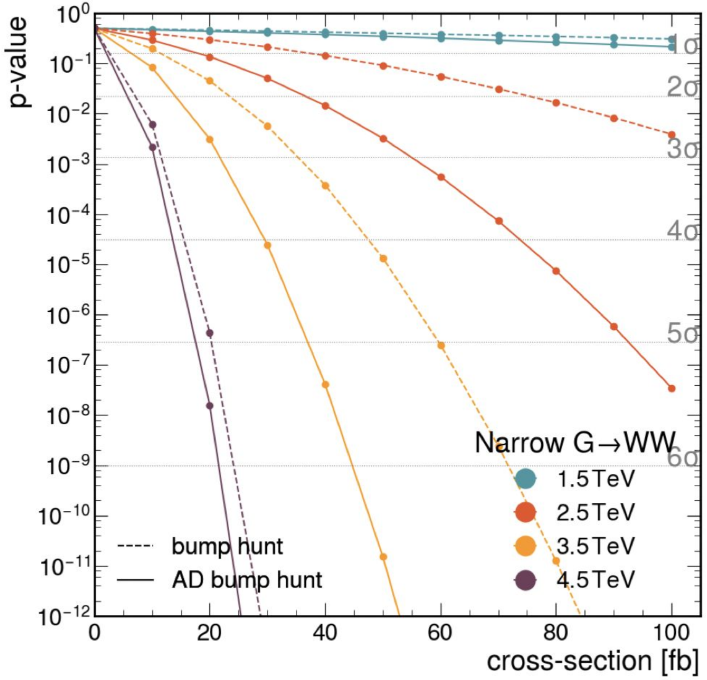
# Analysis workflow




Use most SM-like quantiles to determine background in most anomalous quantiles

*K.A. Wozniak, Maurizio Pierini, BM, et al., paper submission in preparation*

# Large improvement over inclusive fit

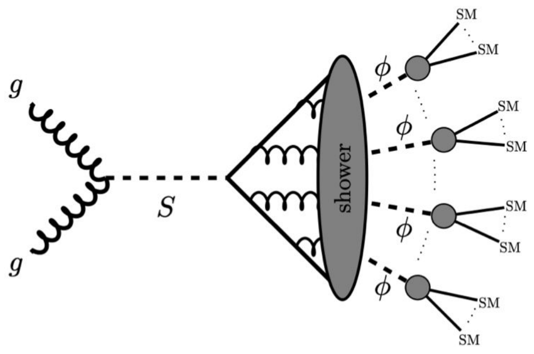



**Large improvement in sensitivity over inclusive bump hunt**

**Turning fully model-independent:**

- The same background shapes across quantiles also allows for a purely model-independent search
- "Goodness-of-fit" test: throw toys around the background prediction and quantify (as a *p* value) how likely the observation is
- Signal MC used **nowhere**

# Incarnation 2: Unclustered hadronic activity



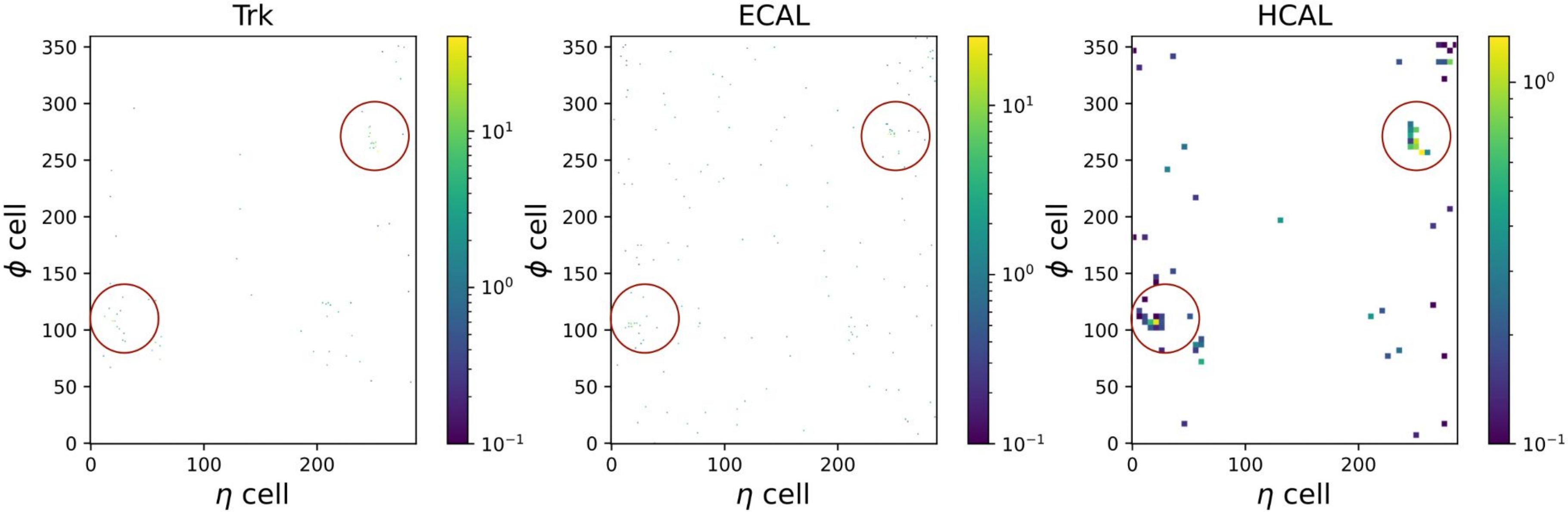

Example: “soft bombs”, “SUEPs”
(Soft Unclustered Energy Patterns) – Dark Showers

**Goal**

- I just explained to you how to look for exotic hadronic activity clustered into highly Lorentz-boosted jets
- **Complementary to that**, we can apply the *same* analysis approach to look for **exotic unclustered hadronic activity**
- Instead of operating on jets, we can encode the entire event in a (graph) autoencoder
- Several exotic processes could lead to significant unclustered energy: *dark showers, SM instantons, …*

# Incarnation 2: Unclustered hadronic activity

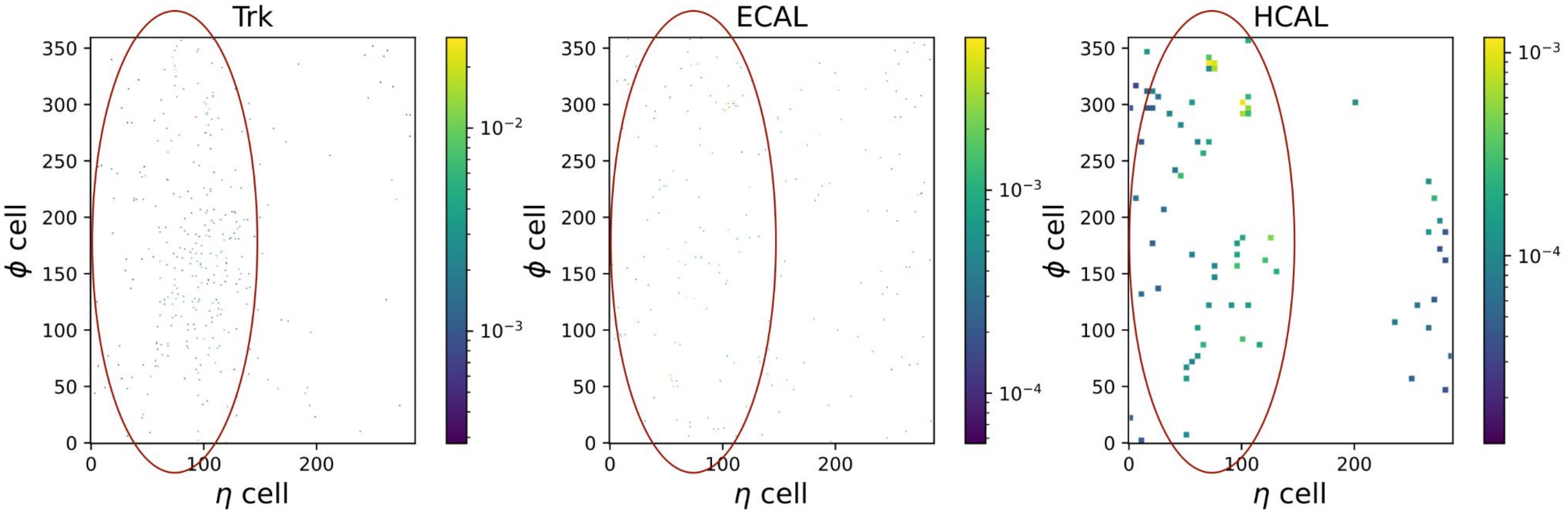
**QCD event**



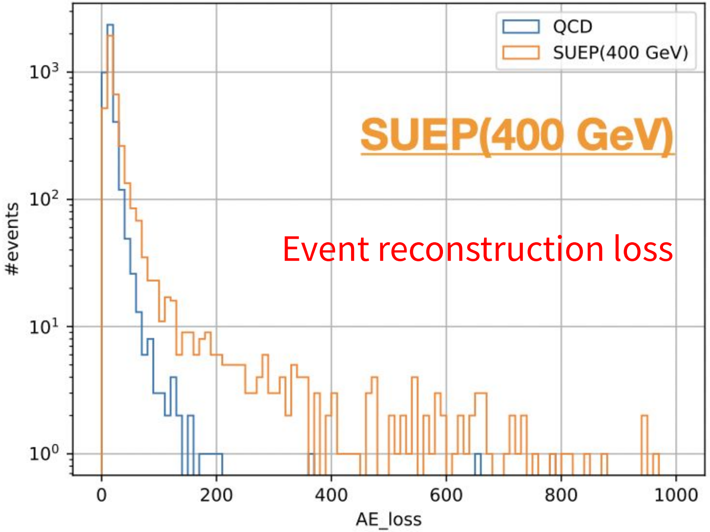
# Incarnation 2: Unclustered hadronic activity

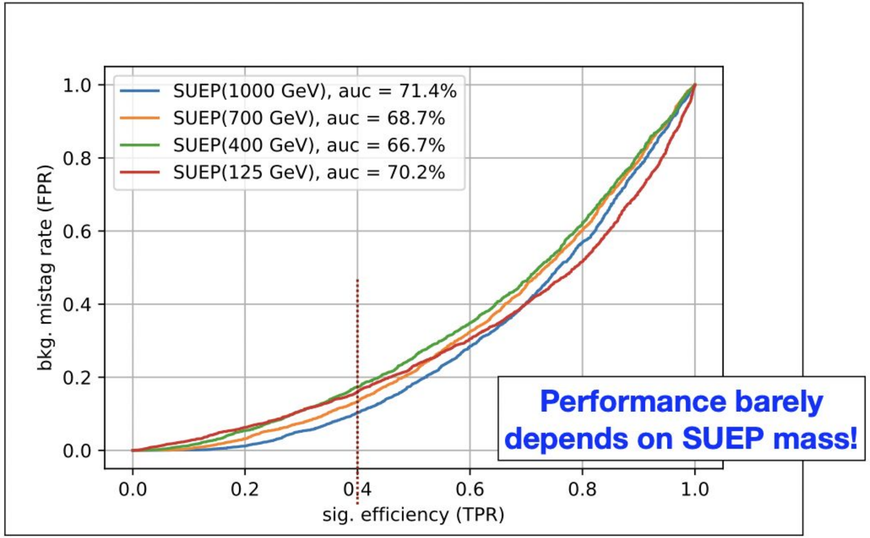
**SUEP event**

Banded structure



# Incarnation 2: Unclustered hadronic activity




Performing quantile regression as a function of particle multiplicity or sphericity

Nadezda Chernyavskaya, Simranjit S. Chhibra, Syed Hasan, Benedikt Maier, Maurizio Pierini, *paper submission in preparation*

# Run-3 analyses: summary

- Find new physics *in final states dominated by hadronic activity*
  - Do it for highly collimated signatures (particles clustered into fat jets)
  - Do it for unclustered signatures (soft sprays of particles)
- Purely data-driven techniques with machine learning
- **NB**: **ML not a facilitator, but an enabler!**
- Model-agnostic analysis strategy

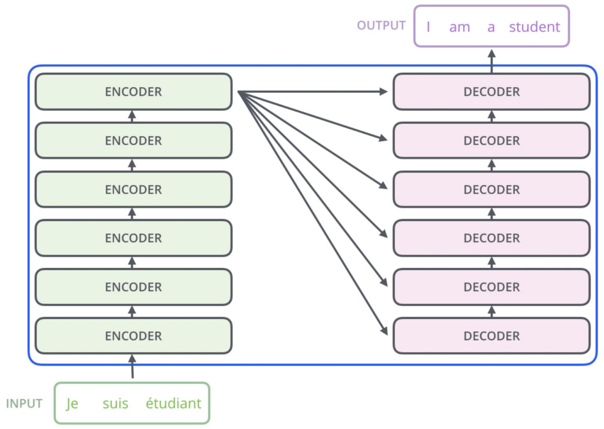
**Can we perform the same analyses at the HL-LHC beginning in 2029?**

Some work needs to be invested before!

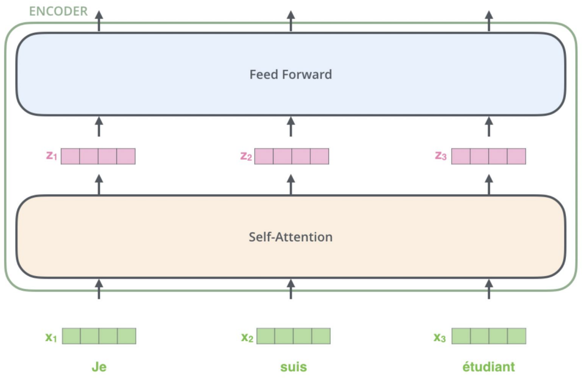
# Enabling searches at the HL-LHC

# Pile-up mitigation

# HL-LHC Pile-Up



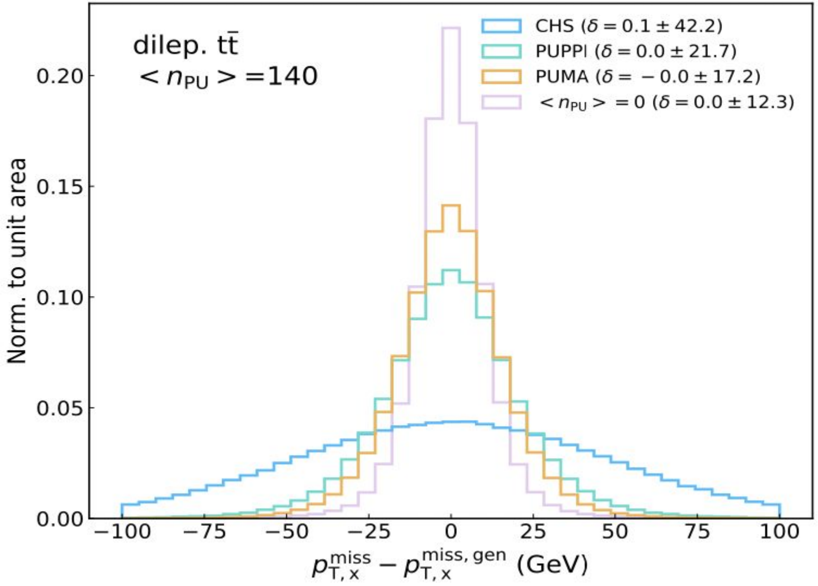

- With **up to 200 simultaneous collisions at the HL-LHC**, prospects for anomaly detection with hadronic activity without taking appropriate action are poor
  - Anomaly detection and pile-up rejection **joined at the hip**

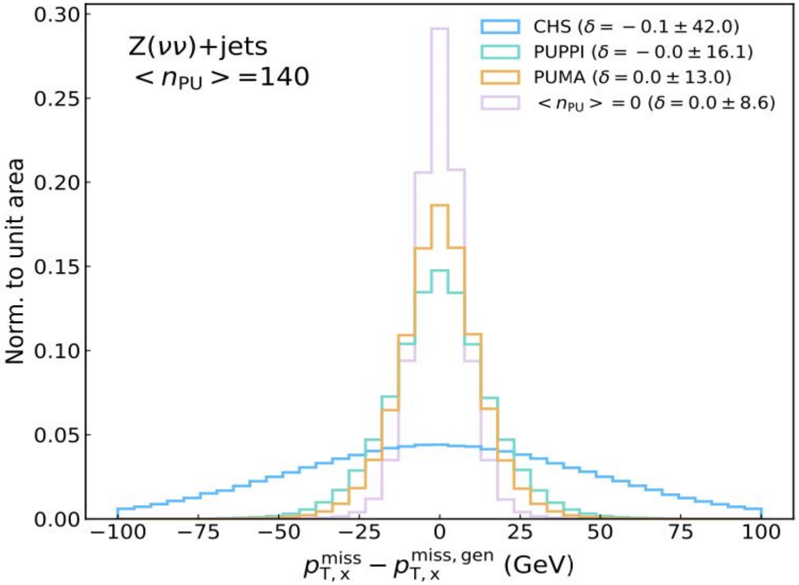



- Adapt techniques from natural language processing to reject pile-up

- *Transformers* utilize **self-attention** to enrich elements of a sequence with information of neighboring elements

# Missing transverse energy resolution




Network (PUMA) **outperforms** state-of-the art traditional algorithm (PUPPI)

B Maier *et al* 2022 *Mach. Learn.: Sci. Technol.* 3 025012

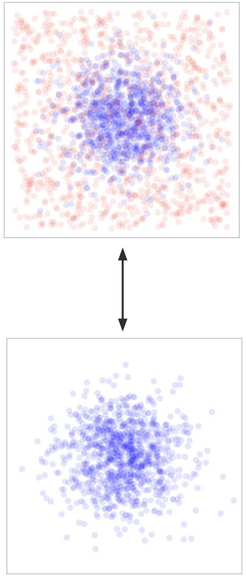
Study on toy (DELPHES) data

# The problem with the ground truth

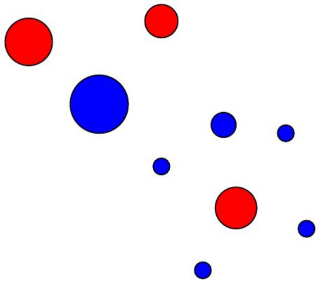
The truth label "In neutral particle from Primary Vertex?" is practically **impossible** to obtain in a GEANT4-based detector simulation with highly complex reconstruction as employed by CMS

**Now what?** What if we could make a network learn *relative* information instead of *absolute*?

Could use a **sample simulated with 0 PU and the same events with PU added!**



# Optimal transport problems are similar



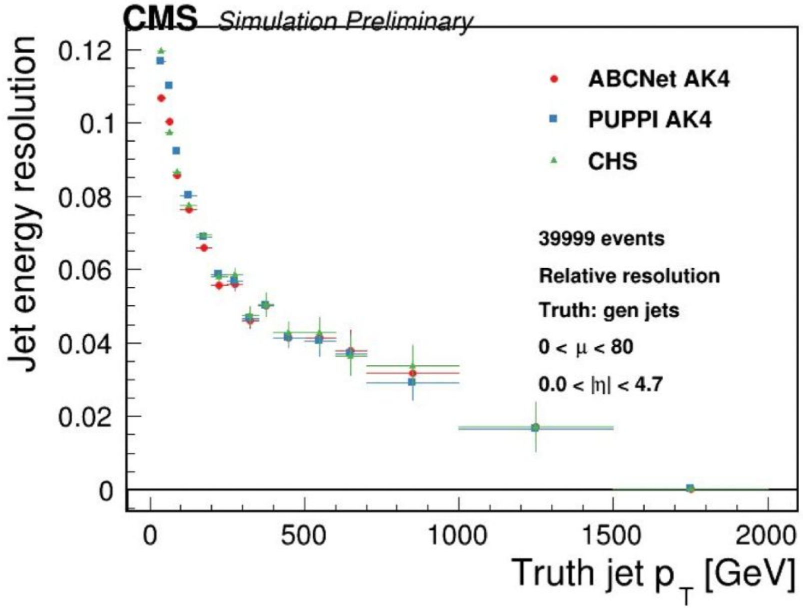
Earth Mover's Distance (EMD) is the **minimum work required** to move earth into to fill some holes

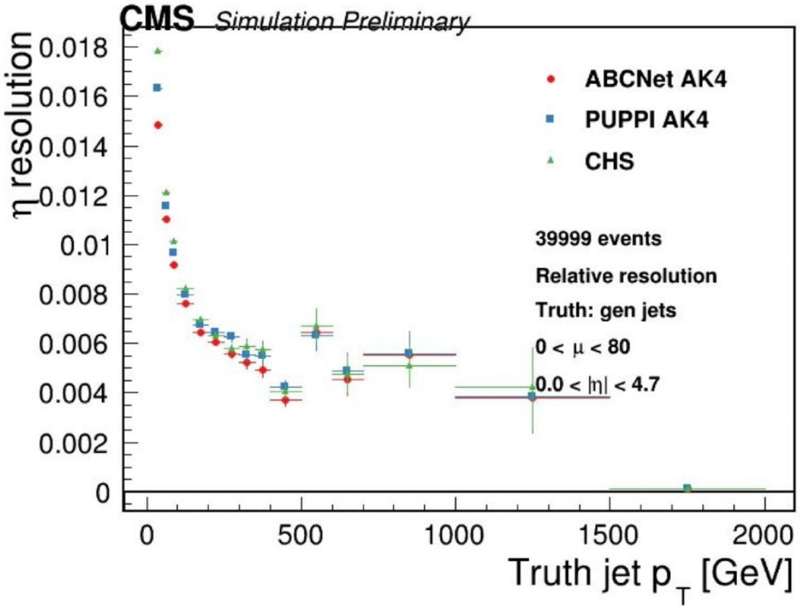
$$EMD(\vec{x}, \vec{y}) = \min_f W(f, \vec{x}, \vec{y})$$

Should leave particles from PV unchanged and destroy PU particles

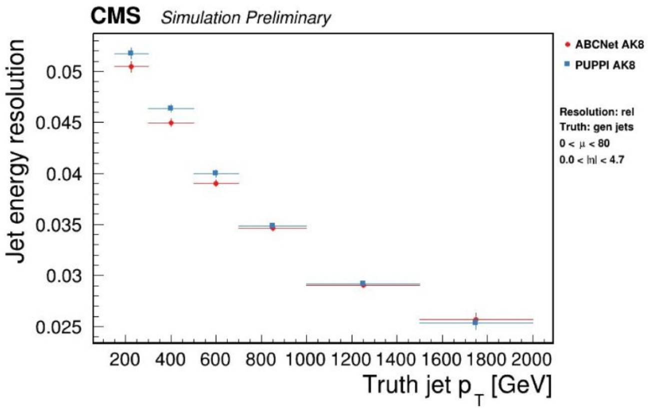
Employ EMD as loss function in a *graph neural network*. Training for a per-particle weight [0,1] and use it **to scale particle 4-momenta**

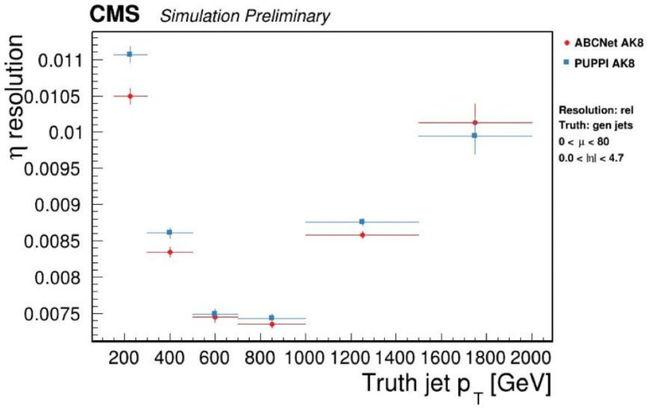
# Improved resolution in AK4 jets for CMS!

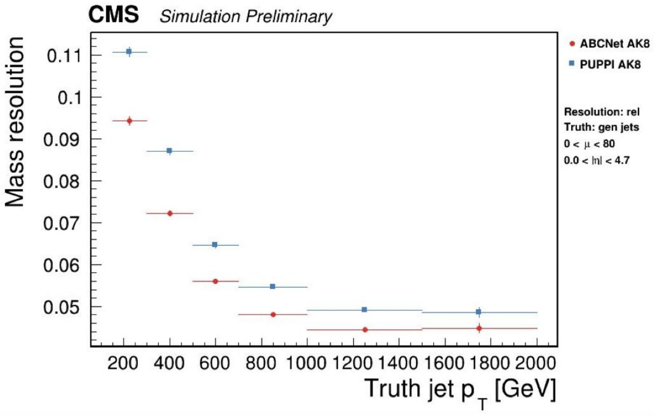


# Top quark AK8 jets: Z’ → tt




Improvements up to 20% in **large-radius jets** consistent with improvement in narrow-cone jets

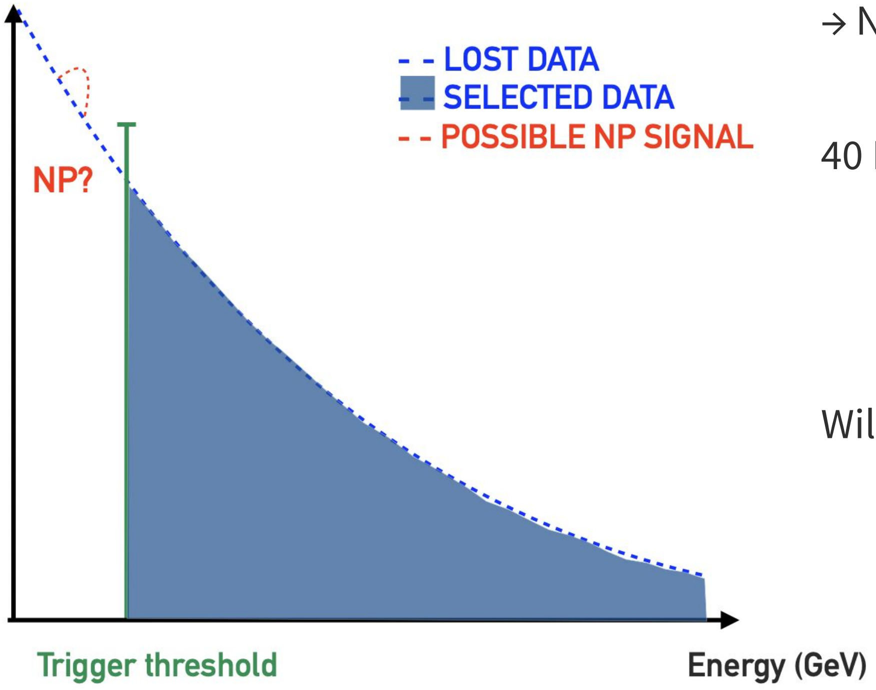
Loukas Gouskos, Fabio Iemmi, Sascha Liechti, BM, Vinicius Mikuni, Huilin Qu, *paper submission in preparation, talk at ML4Jets in Nov.*

# Enabling searches at the HL-LHC

# Machine-learning for the Level-1 trigger

# Deploying algorithms in the L1 trigger

Sketch by Thea Årrestad (ETH)




At 1 billion pp collisions per second, we'd have to save 1 PB/s to disk → impossible

→ Need to discard events below certain E **forever**

40 MHz → 750 kHz → 7.5 kHz

L1 HLT

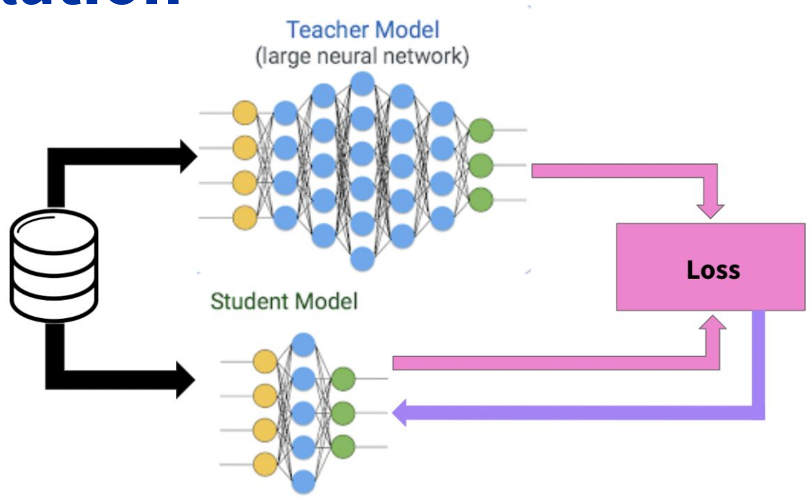
FPGAs GPUs

Will have tracking information at L1 for Run-4

Porting anomaly detection* and pile-up mitigation algorithms to L1

* Can also be used for Data Quality Monitoring/ experiment protection

# Knowledge Distillation




The **teacher** is a complex, pre-trained graph neural network for anomaly detection or PU mitigation

The **student** is a simple feed-forward neural network with fewer inputs that tries to regress the output of the complex teacher; can easily be deployed on FPGA

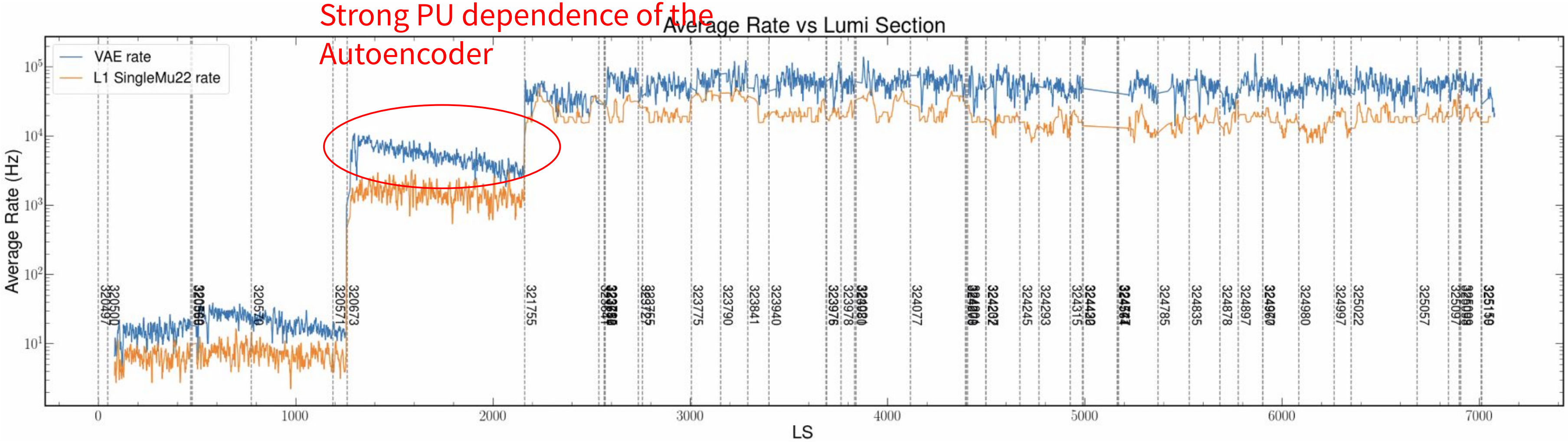
# The link between Anomaly Detection and Pile-Up Mitigation

By definition, anomaly detection relies on a **clean** event content

Without a proper pile-up mitigation, anomaly detection cannot be performed properly

… in particular at the L1

Strong PU dependence of the Autoencoder



Thanks for your attention!

# Group establishment

# Establishing myself and my group at HEPHY

Applying for a START FWF Grant

Using my network to **attract extra personpower** by encouraging young students/postdocs to apply for Marie Curie, Humboldt fellowships

Large **visibility for all group members** by assuming coordinating roles in the collaboration

Applying for additional funding for computing resources: GPU servers, etc to build an analysis facility or integrate it into existing computing clusters

# Backup

# Variational autoencoder

### We are training on jets, not on events

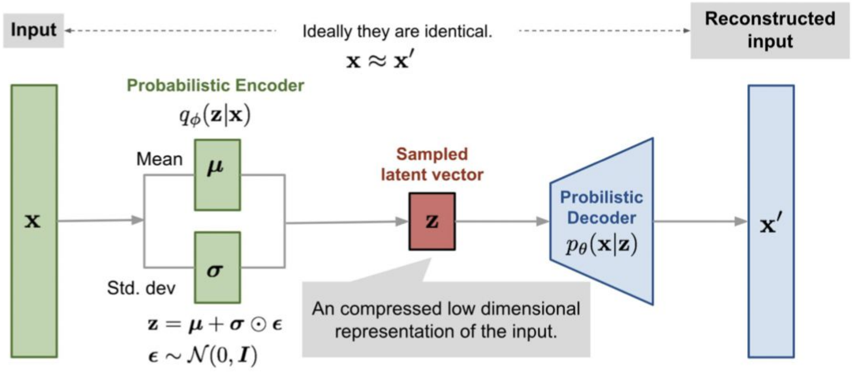
- Roughly 3M jets from dEta(jj) sideband

### Using a 100 x 3 input matrix

- Truncate at/pad to 100 jet constituents
- px, py, pz
- Inputs get standardizes (mean 0, std 1)

### Architecture

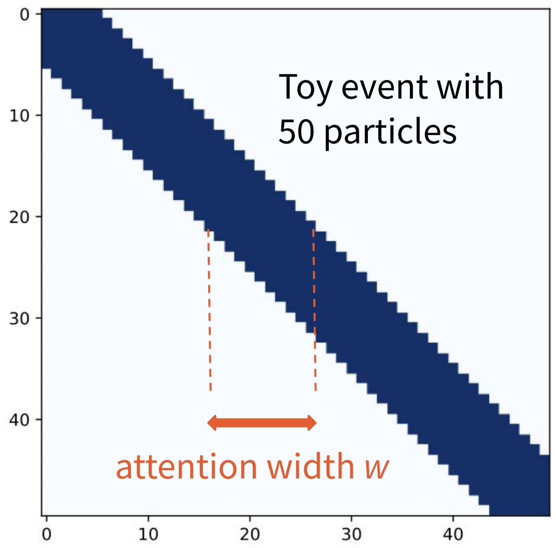
- Encoder: series of 2D + 1D convolutions, flattened in last layer
- Latent space: dense networks compressing into 12 Gaussians: 12 $\mu$'s and 12 sigmas
- Decoder: 1D and 2D transposed convolutions




$$\mathcal{L}_{\text{tot}} = \underbrace{\mathcal{L}_{\text{reco}}}_{\text{point-wise dissimilarity}} + \underbrace{\beta}_{0.5} \cdot \underbrace{\mathcal{L}_{\text{KL}}}_{\text{latent distribution soundness}}$$

# Sparse transformers

- **Sparse transformers** can compute attention scores in a *sliding window* without first having to do N x N computation
- Complexity scales as O(N*$w$) rather than O($N^2$)
- Longer-range information can be exchanged by *stacking transformer layers*
- Clearly, the **sequence order** is important because for some particle pairs attention scores would never be considered
- Embedding **a sense of locality** in the sequence by kMEANS-clustering all particles, then sorting by cluster_pT and particle_pT → Particles close in detector space will be close in the sequence

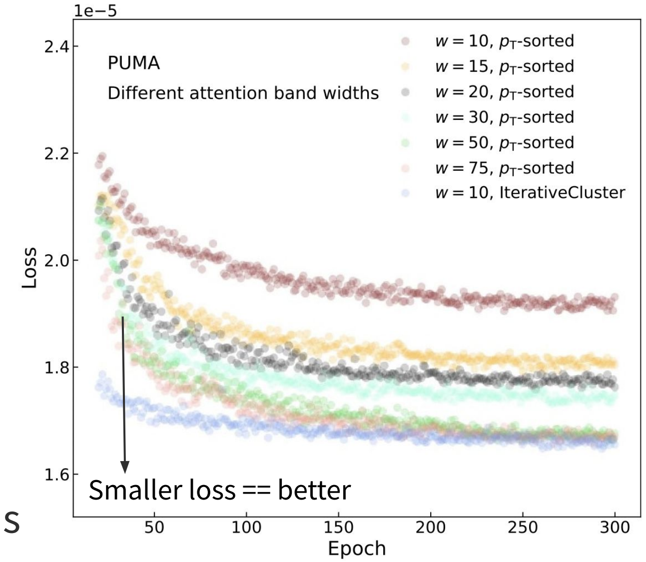



40 in reality is up to 9000, and *w* up to 200

# Where does the useful information come from?

- Loss function is the MSE of y_true (1 for PV particles, 0 for PU particles, something between [0,1] for some merged CaloTowers) and y_pred

- Comparing training on kMEANS-ordered sequence with training on sequence ordered by particle pT
  - Clearly the most information is contained in the **local vicinity** of the particle.
  - Same performance can only be recovered with a **very large receptive field** for pT-ordered sequence
  - (But even large receptive fields would be possible with this implementation!)

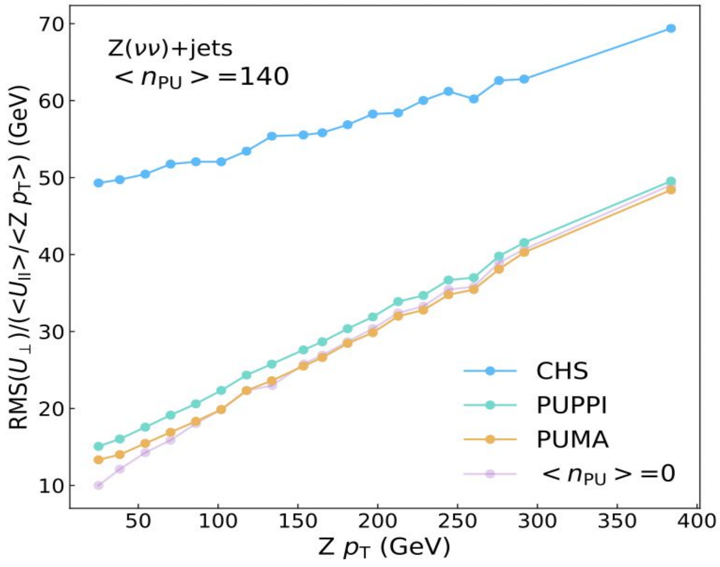


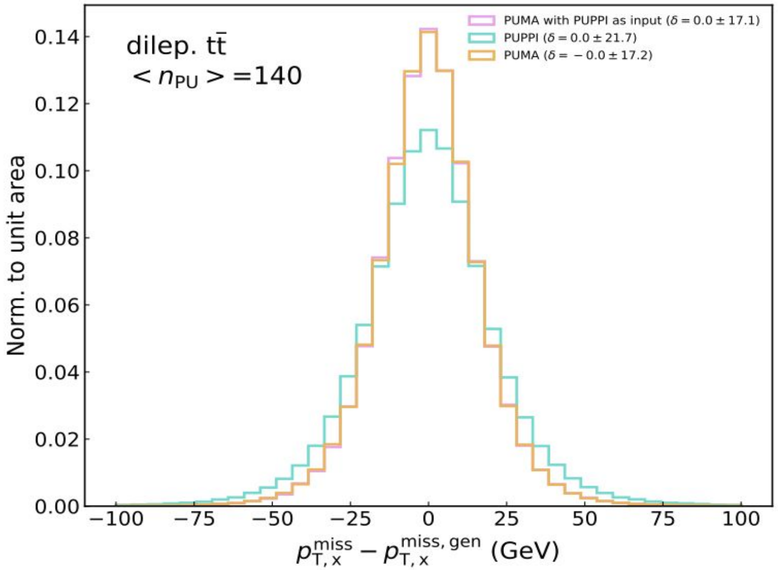

B Maier *et al* 2022 *Mach. Learn.: Sci. Technol.* 3 025012

- Let's look at some observables …

(DELPHES3 data)

# Hadronic recoil resolution, PUPPI as additional input




Sparse transformers can efficiently remove pile-up. Adding PUPPI as input feature doesn't further improve performance.

# What have we learned

**Attention mechanisms** are extremely well suited for surfacing information on particle provenance

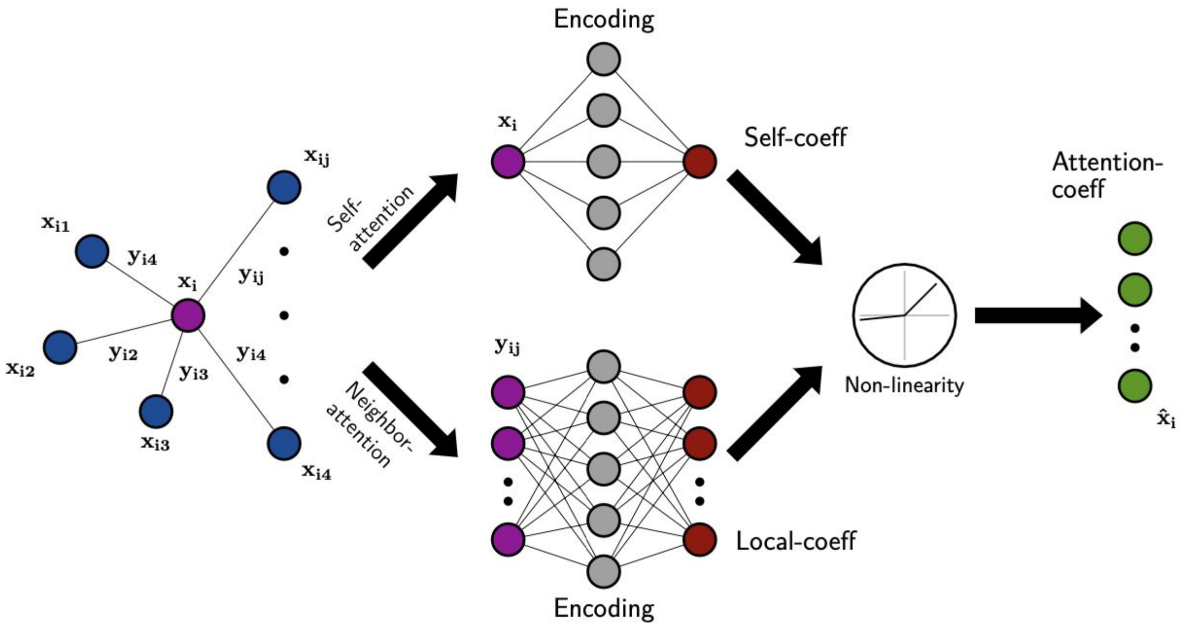
The most information is **contained in** a close **neighborhood** (in $\eta$ - $\phi$) of the query particle.

Residual long-range dependencies can be exploited with a **sparsification of the adjacency matrix** and sliding window attention + stacked transformer layers

This study indicates that a *graph network* enhanced with attention mechanisms that performs convolutions on close-by particles should be suited for the task as well as a transformer.

Such a graph network exists and has already been used for analysis in CMS: *ABCNet* → Let’s study its applicability for the pile-up problem with CMS data …

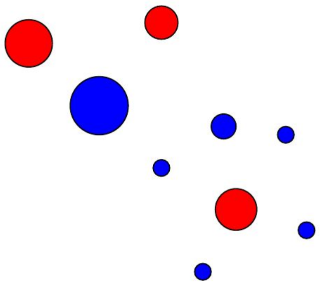
# ABCNet




A graph neural network with attention mechanisms
Can perform per-particle regression tasks

Vinicius Mikuni, Florencia Canelli, EPJ Plus 135 (2020) 463

# Optimal transport in 1D

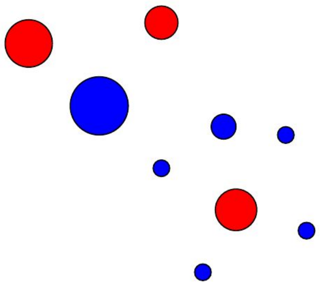


For Run-3: ~ 4000 particles expected

High computational cost to consider 4000 x 4000 feasible flows

Can we move to a 1D space? There, solving for the EMD becomes a **sorting problem**

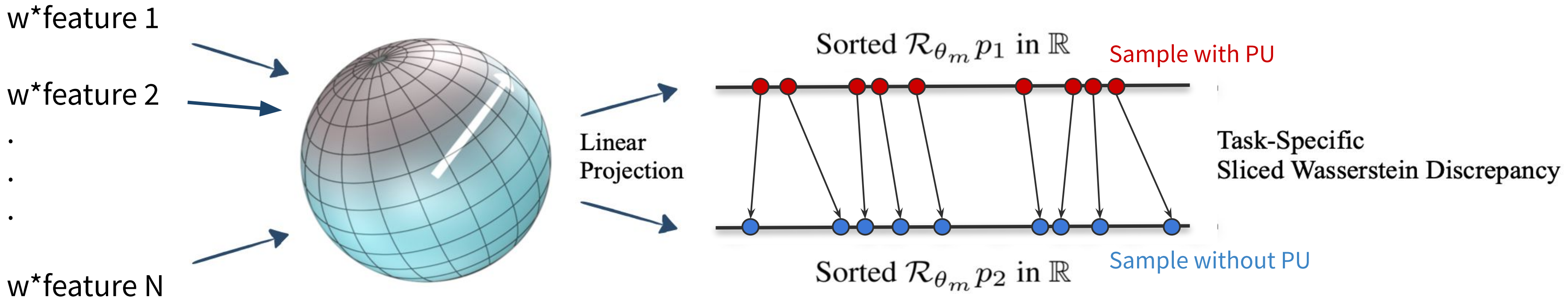
# Sliced Wasserstein Distance (SWD)



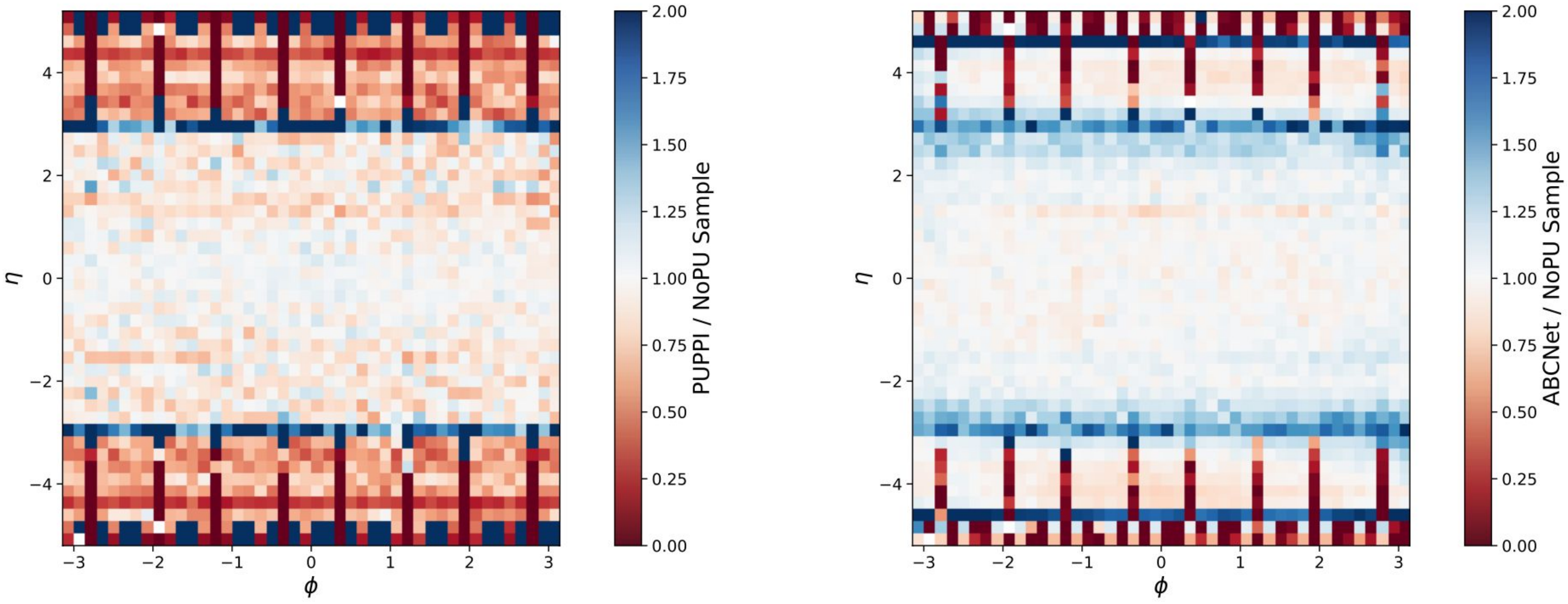
Each particle is characterized by its **N features** (pt, eta, …)

We can throw a unit vector in this N-dim. space and **project the feature vector multiplied with a learned weight** *w* ϵ [0,1] onto this dimension

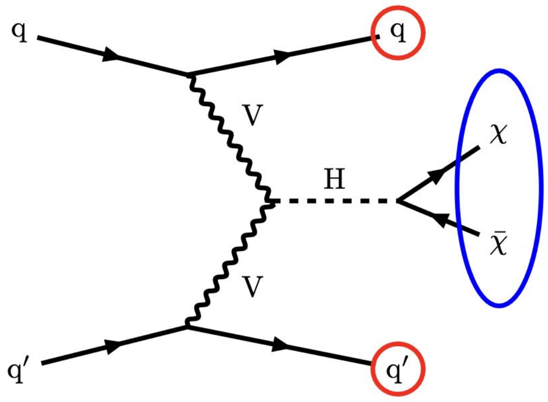
SWD is the sum of pairwise 1D distances after sorting both distributions
→ Use **SWD as loss function**, use **w to scale particle 4-momenta**

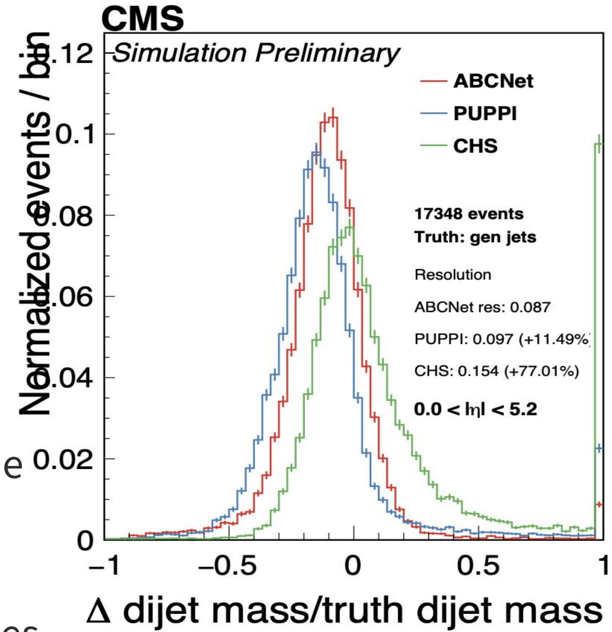


# Energy flow in the detector: PUPPI vs ABCNet



# Improved resolution in AK4 jets for CMS!



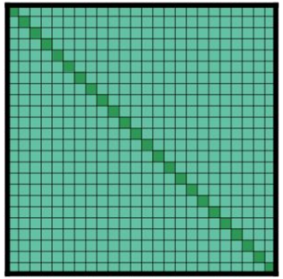

**Energy and pointing resolution** improved → let's look at some invariant mass plots

VBF **Higgs → invisible** one of the most important BSM processes

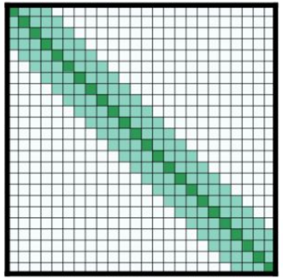
Improvement of >10% in $M_{jj}$

# Attention patterns

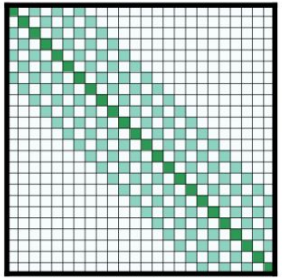
https://arxiv.org/pdf/2004.05150.pdf



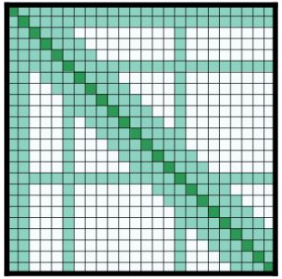
(a) Full $n^2$ attention



(b) Sliding window attention



(c) Dilated sliding window



(d) Global+sliding window

# How a nice latent space looks like in a VAE