

An abstract graphic consisting of several thin, black, overlapping lines that form various geometric shapes and polygons, primarily located in the upper left and center of the slide.

AI RESEARCH AT CERN

Sofia Vallecorsa – CERN

May 29th, 2024

ML/DL AT CERN

Machine Learning since LEP years

Limited abstraction

Preliminary feature engineering

Mostly for classification & regression

Then came **Deep Learning...**

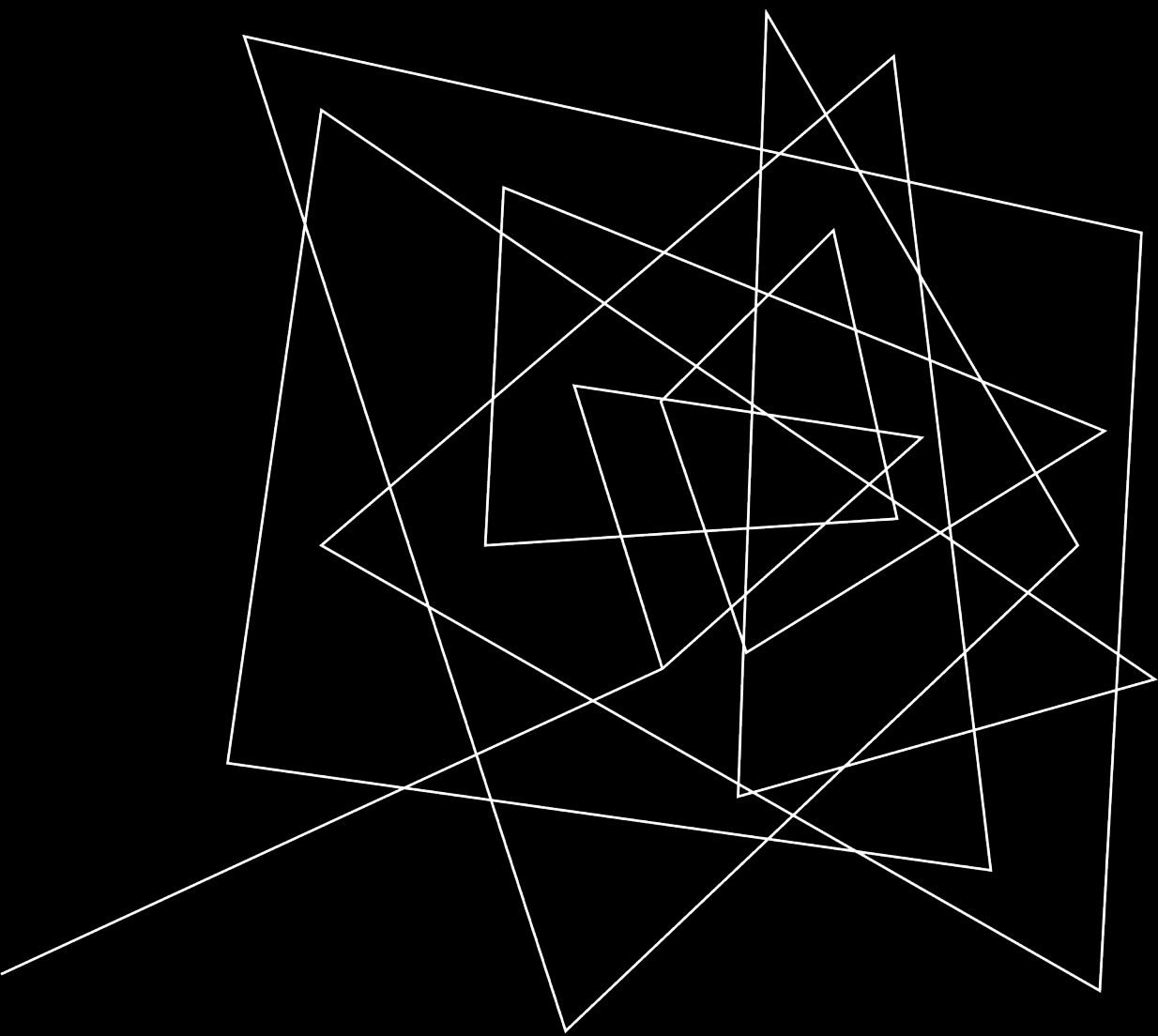
Increasingly abstract representations

Complex computational graphs requiring larger infrastructure

A large range of applications

RESEARCH AREAS

- **Physics**
- **Computing infrastructure optimization**
- **AI for sustainability**
- **Sustainable AI: workload optimisation**
wrt hw and computing models



PHYSICS APPLICATIONS

SIMULATION EXAMPLE: DETECTOR RESPONSE AS IMAGES

Monte Carlo simulation of detector response is extremely **demanding in terms of computing resources**

→ 50 % of LHC Computing Grid resources today

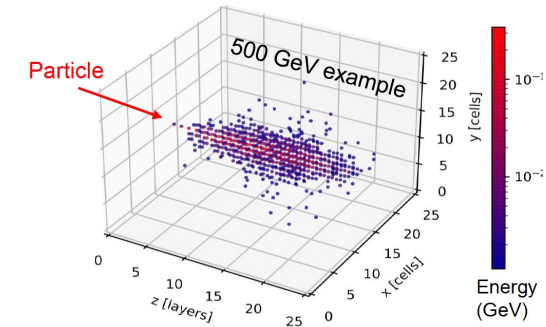
Interpret **detector output as images**

Sensors outputs become pixels in a image

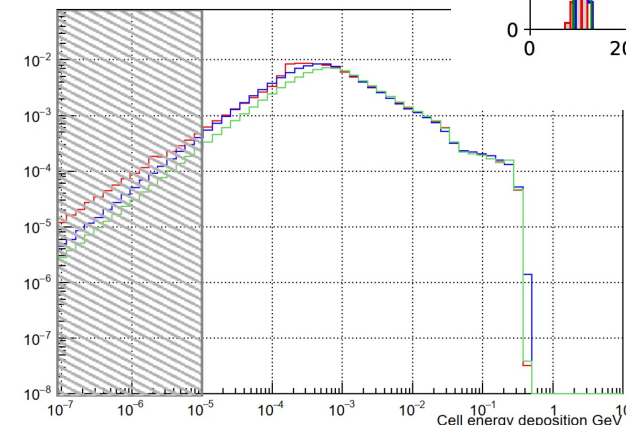
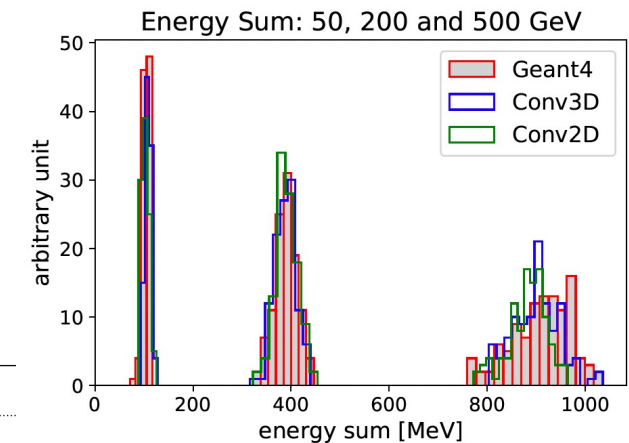
Use **computer vision techniques** to interpret results

Replace Monte Carlo approach with **Generative Models**

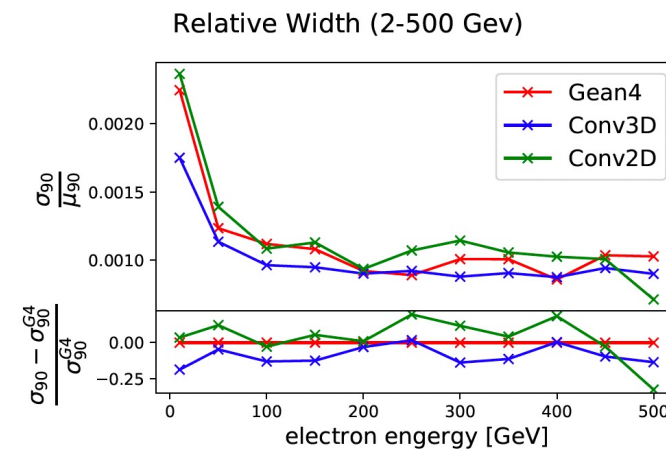
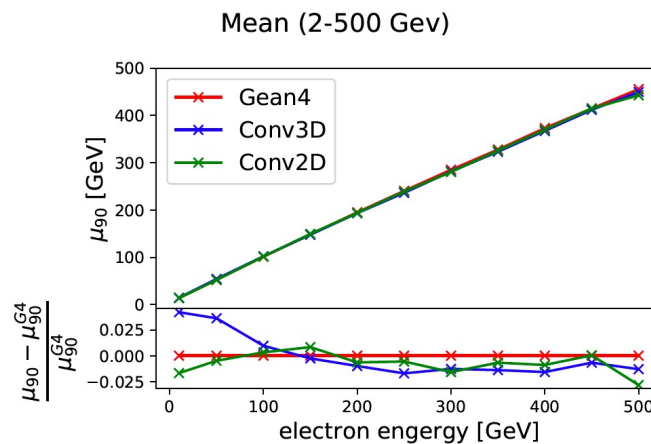
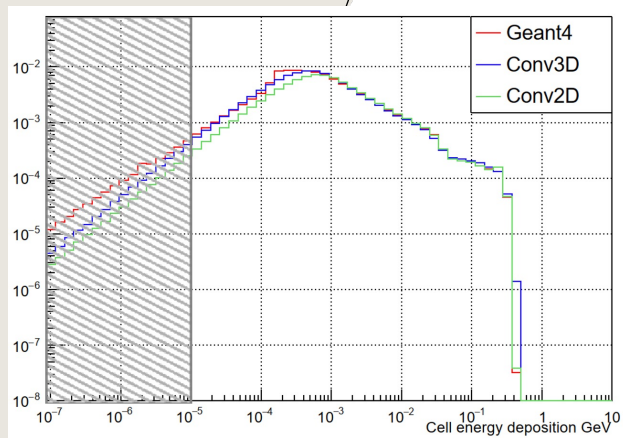
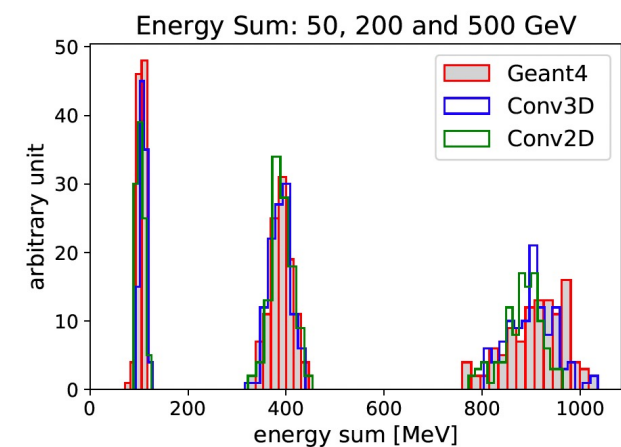
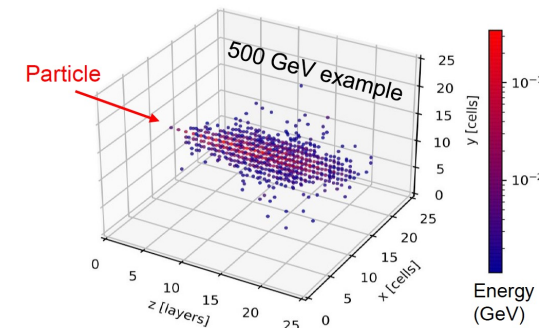
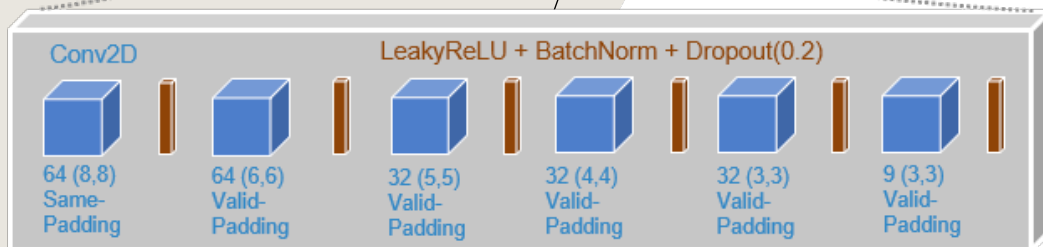
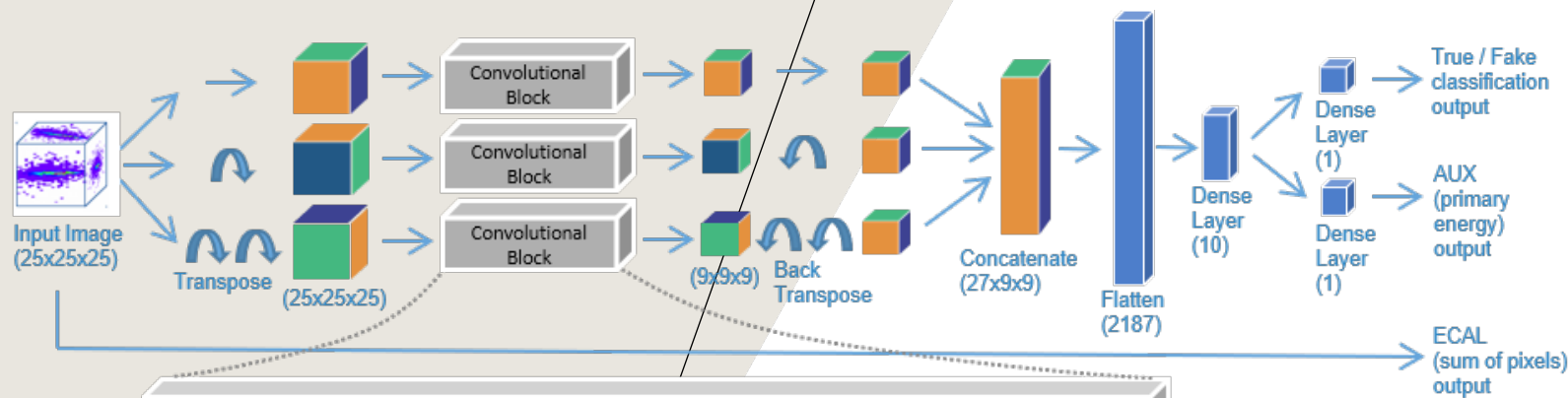
Interfacing DL to standard software is not trivial!



Rehm, Florian, et al.
arXiv:2105.08960 (2021).



GAN-BASED SIMULATION



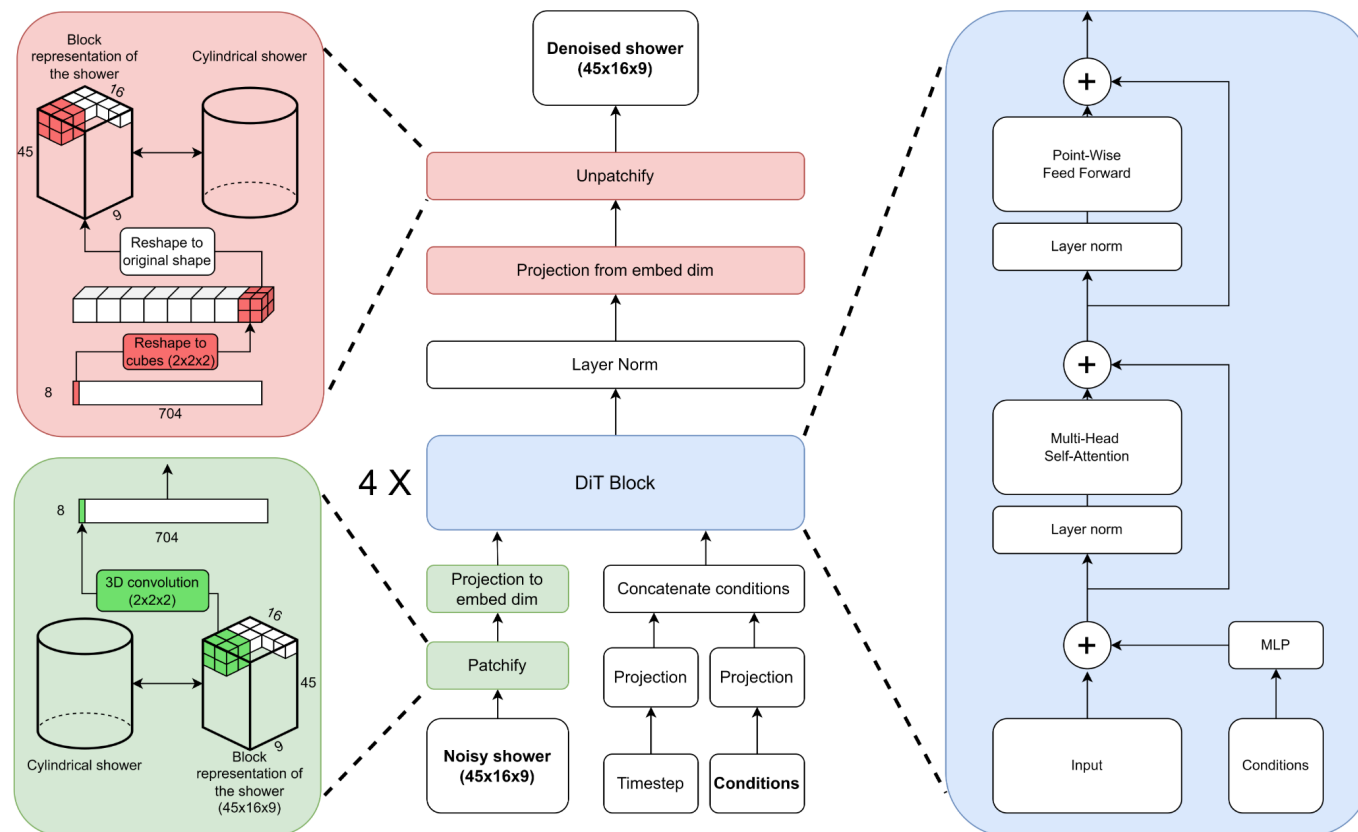
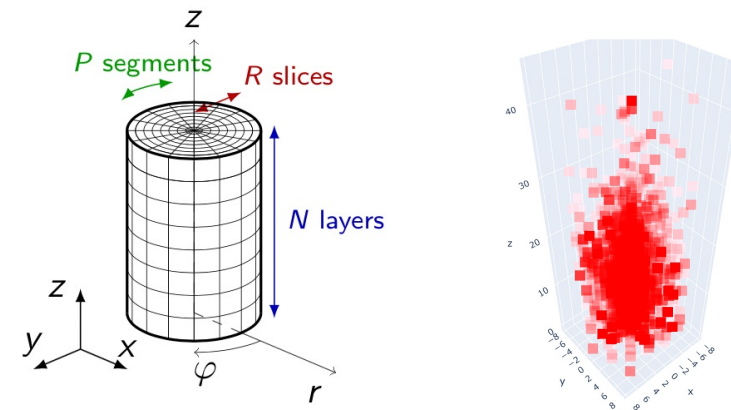
DIFFUSION + TRANSFORMERS

Change to cylindrical geometry (more realistic)

Match SOA diffusion models to transformers to ensure:

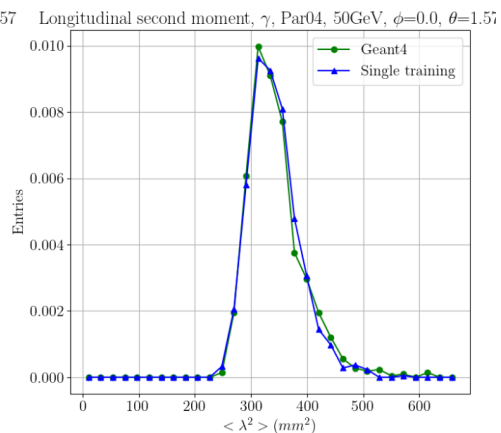
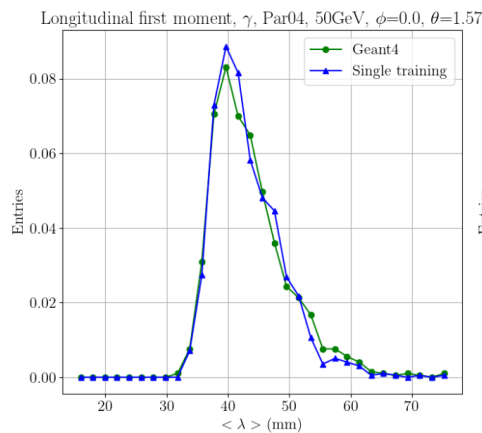
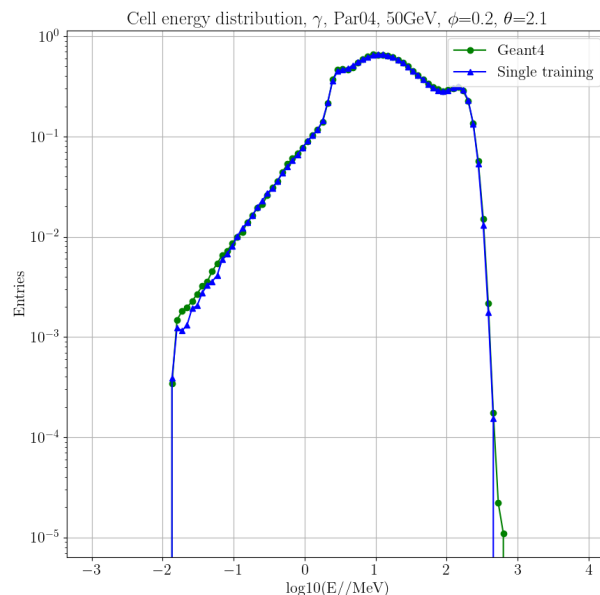
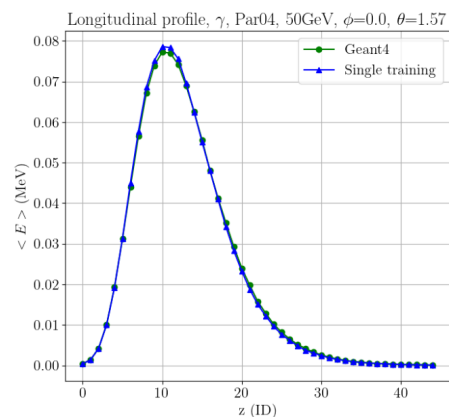
High quality images

Generalizable results



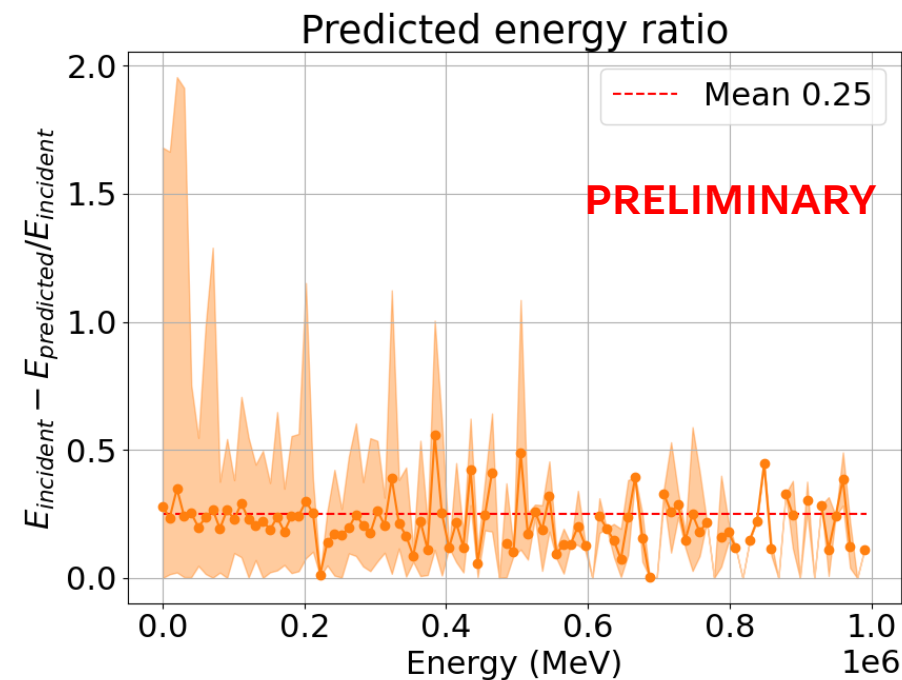
RESULTS

Pixel distributions are correct over a range spanning 5 orders of magnitude



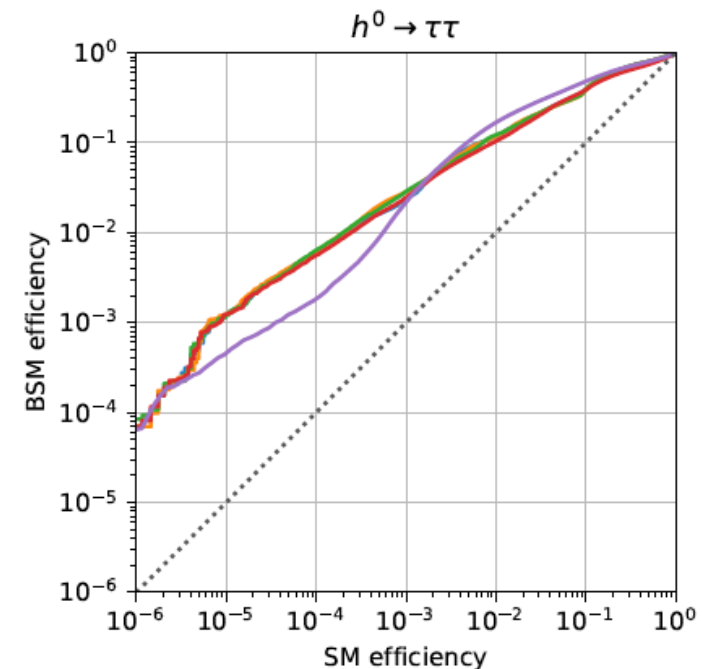
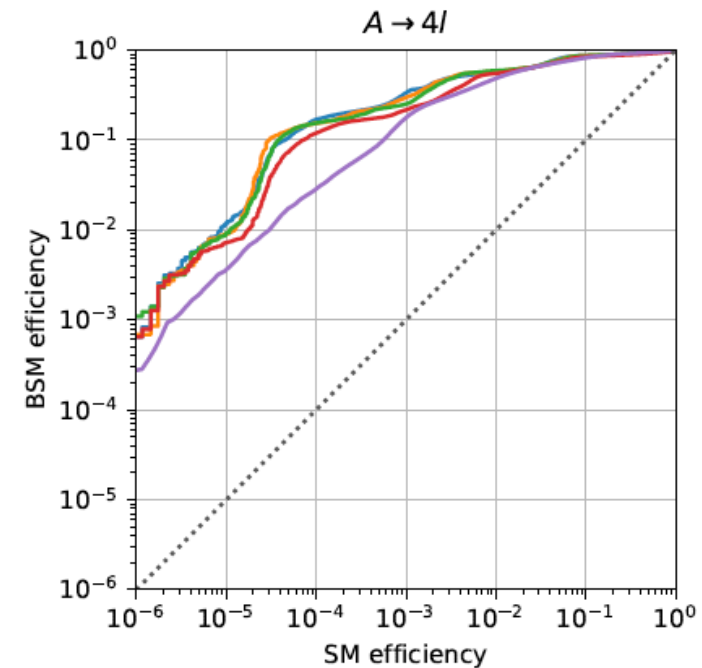
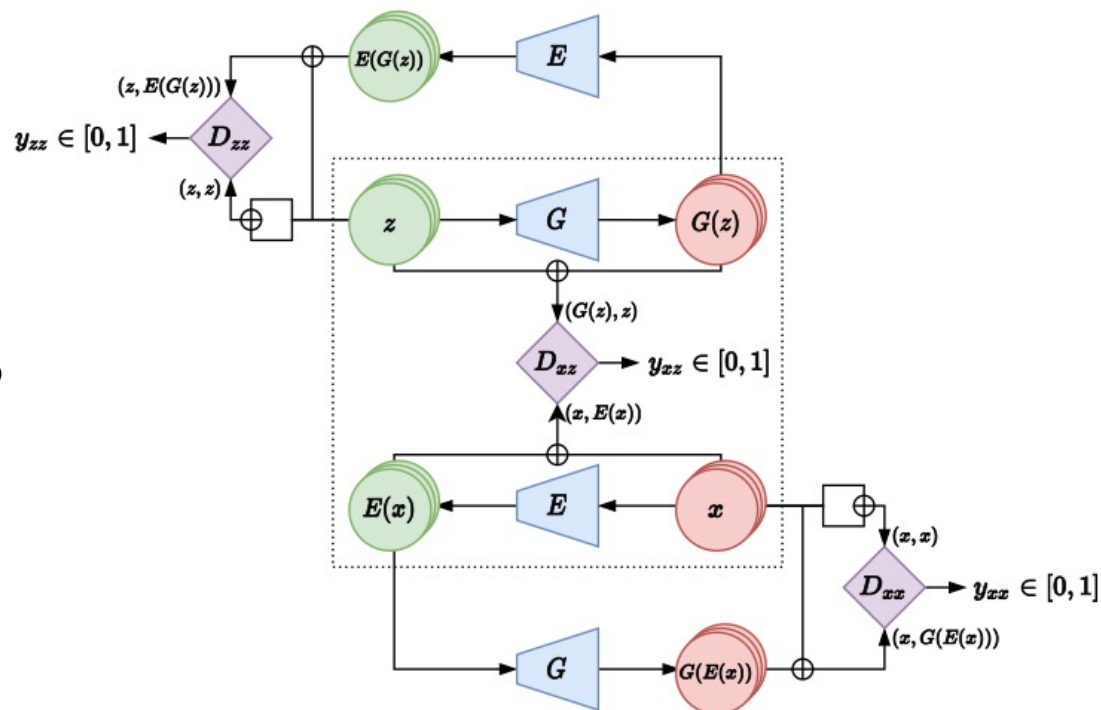
Adaptability to Multi-Tasking:

From image generation to regression

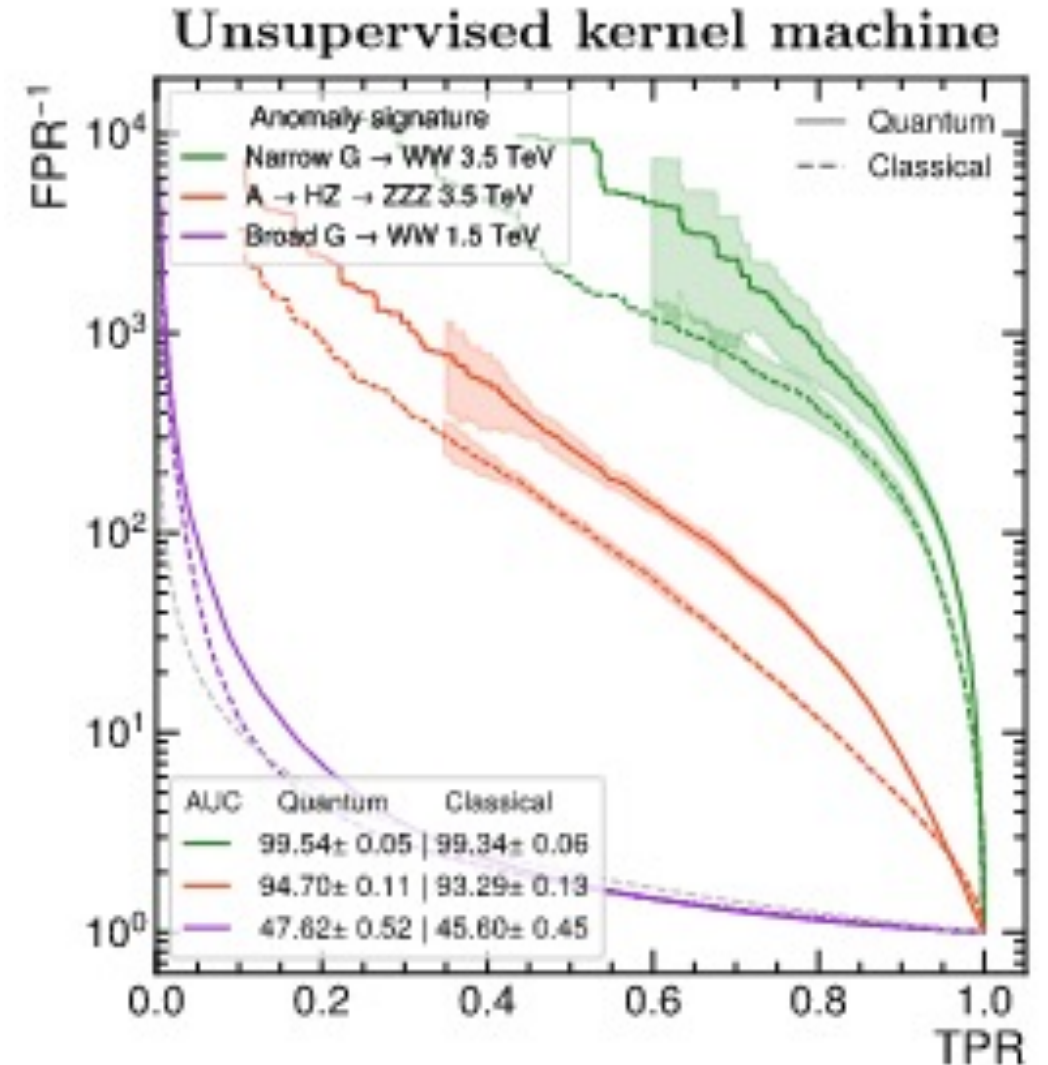
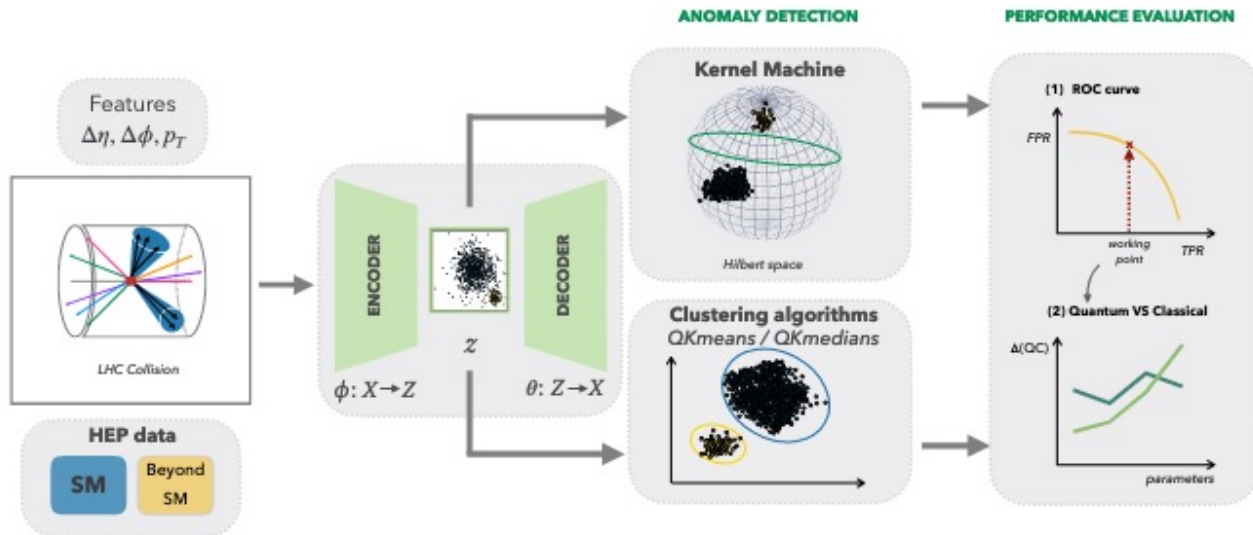


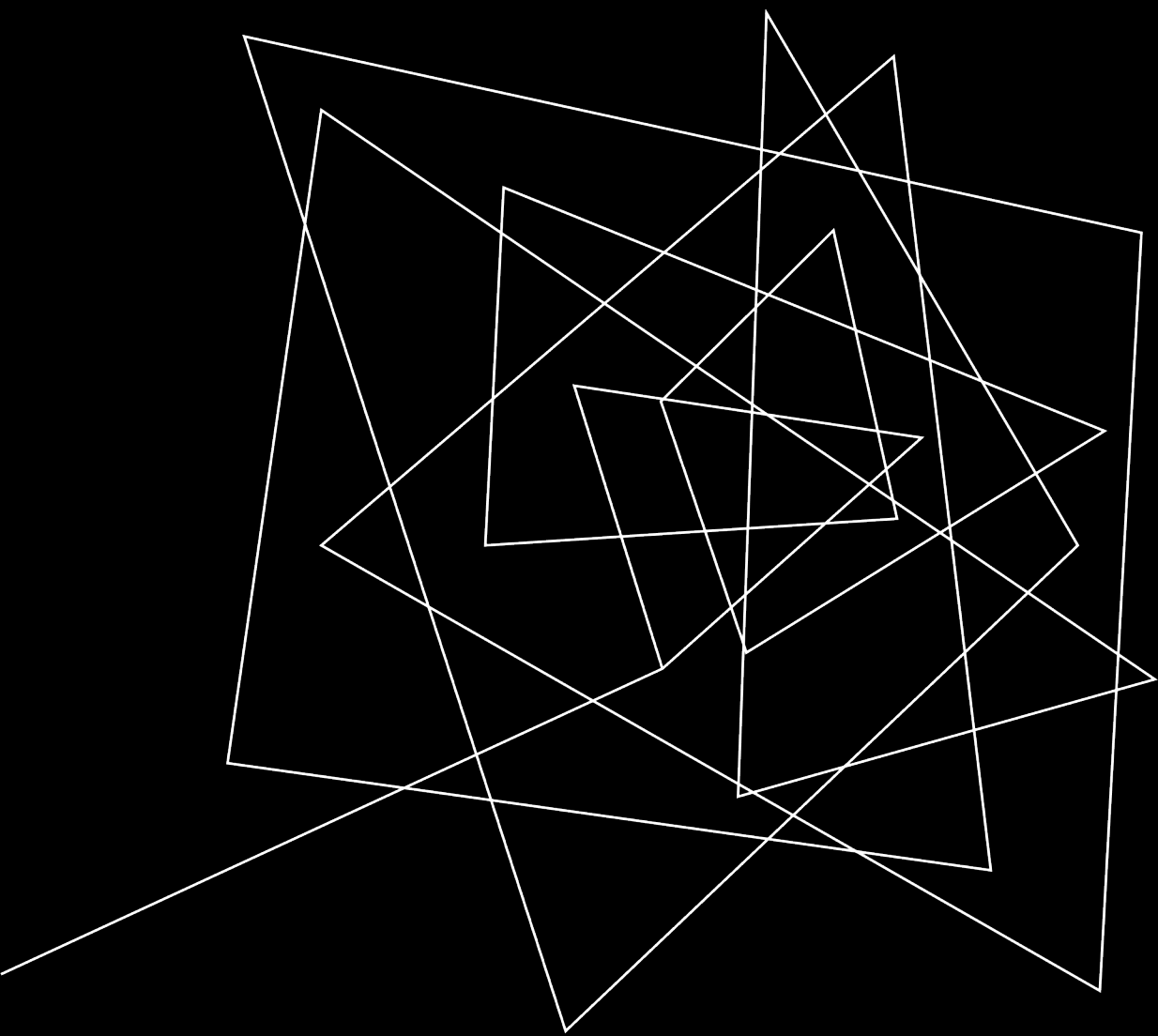
PHYSICS MINING AS ANOMALY DETECTION

- Classical strategy uses very **loose selection**
 - 1M Standard Model (“known physics”) events per day
- Train **Variational AD models** on known physics



DATA COMPRESSION + UNSUPERVISED SVM



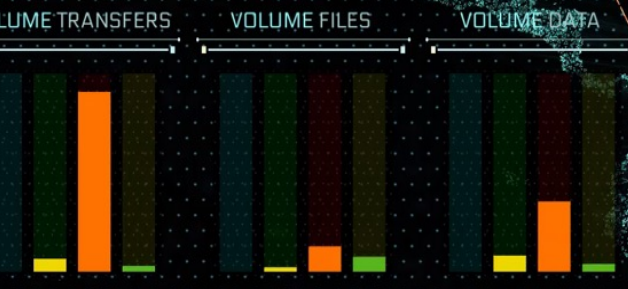
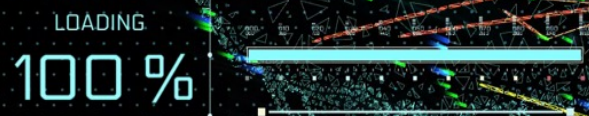


INFRASTRUCTURE OPTIMIZATION



LAST DATA UPDATE

9.7 MB Downloaded Wednesday, 11 September 2019 14:05:12
Last transfer was on : Monday, 29 July 2019 08:00:00



DATA TRANSFER CONSOLE

405847605 From UFlorida-HPC To UMissHEP Monday, 29 July 2019 04:04:50
 0 From UCS072 To INFN-T1 Monday, 29 July 2019 04:05:40
 0 From Vanderbilt To Nebraska Monday, 29 July 2019 04:06:08
 165672773 From INFN-CG To INFN-BARI Monday, 29 July 2019 04:07:31
 4938009 From FLHIP_T2 To CERN-PRDD Monday, 29 July 2019 04:08:20
 76591123.5 From INFN-T1 To GLOW Monday, 29 July 2019 04:08:36
 132252923.125 From INDIACMS-T1FR To pic Monday, 29 July 2019 04:08:43
 1827625179.6667 From CERN-PRDD To KR-KNU-T3 Monday, 29 July 2019 04:09:29
 1874048 From MIT_CMS To FLHIP_T2 Monday, 29 July 2019 04:09:54
 502051950 From INFN-T1 To CIT_CMS_T2 Monday, 29 July 2019 04:10:11
 264700 From CERN-PRDD To SWF Monday, 29 July 2019 04:11:04
 0 From UNI-SOUTHGRID-RALPP To GLOW Monday, 29 July 2019 04:12:05
 166839772 From INFN-T1 To JINR-T1 Monday, 29 July 2019 04:12:10
 1276779676.33333 From CSCS-LCG2 To INFN-LNL-2 Monday, 29 July 2019 04:12:10
 2905786385 From SPRACE To JINR-T1 Monday, 29 July 2019 04:12:20
 0 From INFN-LNL-2 To CSCS-LCG2 Monday, 29 July 2019 04:12:25
 224432295.855556 From IN2P3-CC To praguec2 Monday, 29 July 2019 04:13:03
 4918992.261851667 From UNI-SOUTHGRID-IOX-HEP To CERN-PRDD Monday, 29 July 2019 04:13:11
 0 From BelgGrid-UCL To CIT_CMS_T2 Monday, 29 July 2019 04:14:30
 0 From Vanderbilt To UCS072 Monday, 29 July 2019 04:14:57
 33666768.3792114 From RU-Protvino-IHER To CERN-PRDD Monday, 29 July 2019 04:15:10
 169449714 From CSCS-LCG2 To RU-Protvino-IHER Monday, 29 July 2019 04:15:45

The Worldwide LHC Computing Grid (WLCG)

About 1 million processing cores

170 data centres in 42 countries

>1000 Petabytes of CERN data stored worldwide

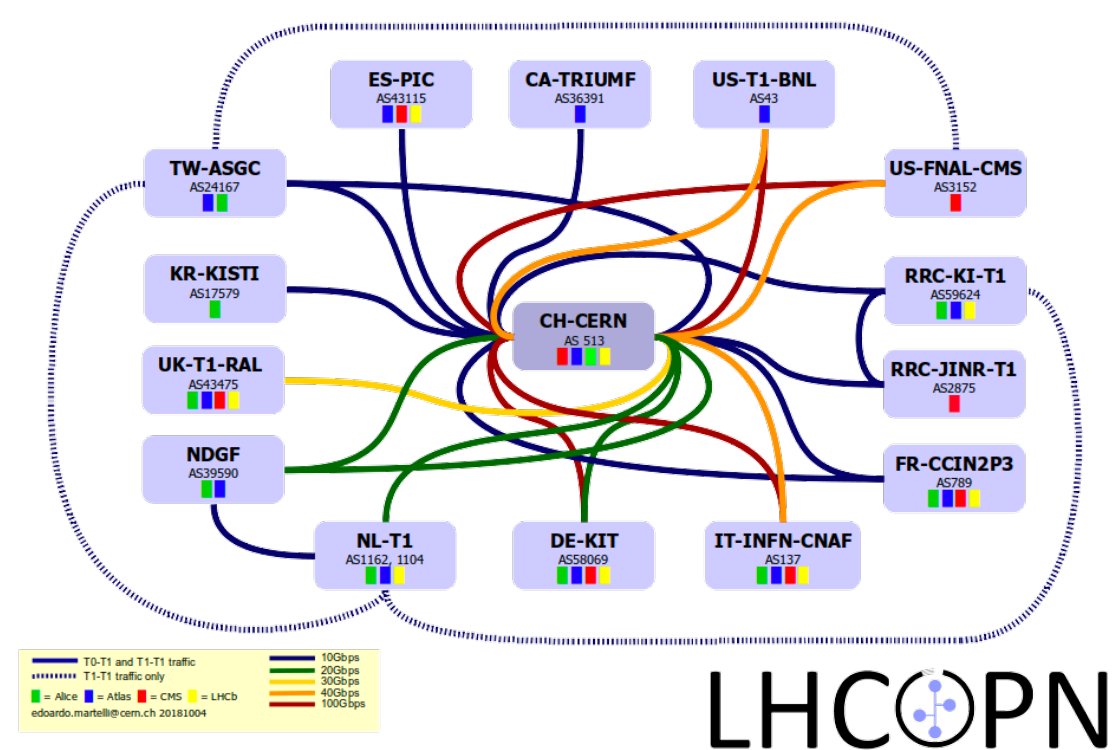
NETWORK TRAFFIC PREDICTION

LHCOPN (Large Hadron Collider Optical Private Network) topology

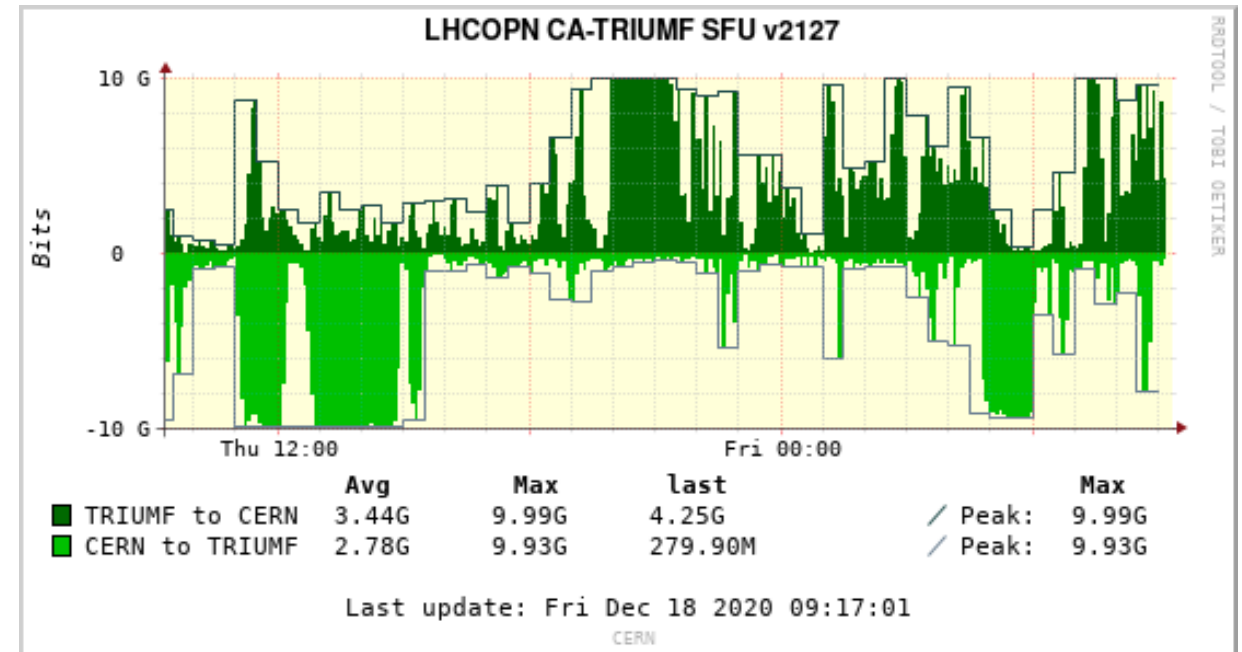
Network traffic on CERN – TRIUMF link

Predict saturation (can occur in both directions)

Optimise transfer: automatically modify network devices configuration (SDNC)

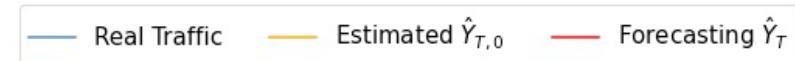


LHCOPN



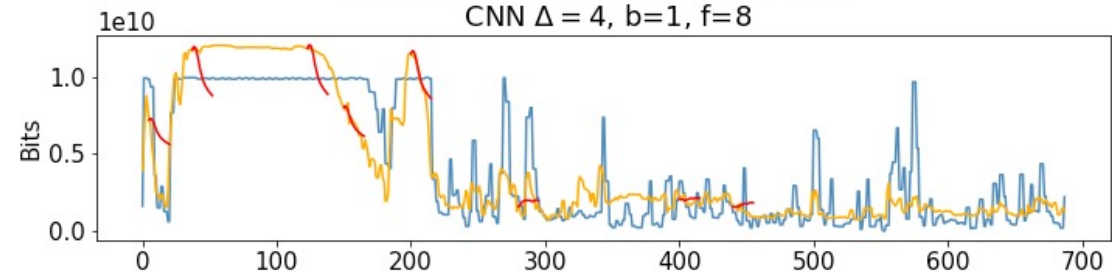
PERFORMANCE

Compare CNN, LSTM and hybrid architectures

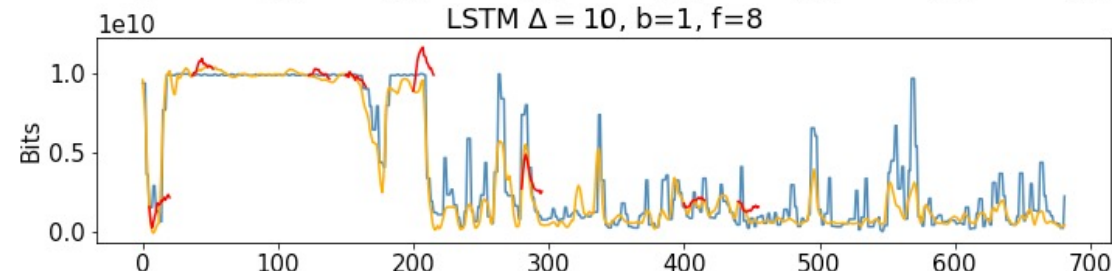


Transfers from TRIUMF to Tier0/Tier1

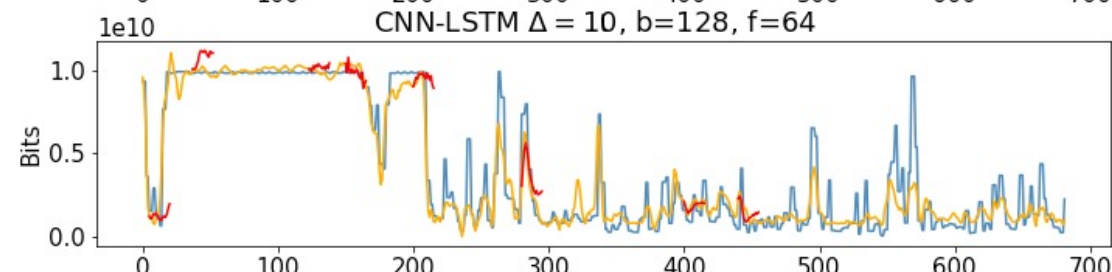
CNN $\Delta = 4, b=1, f=8$



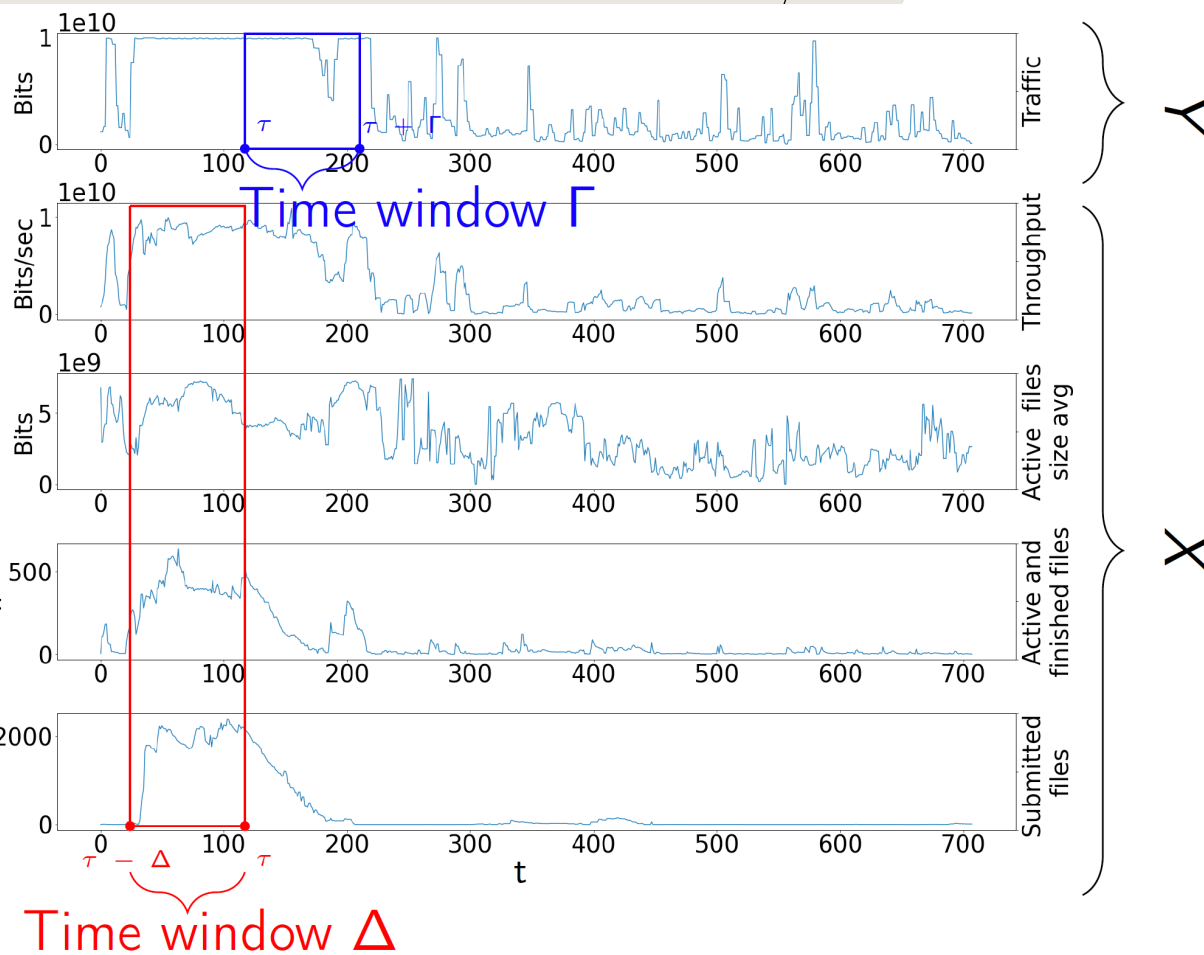
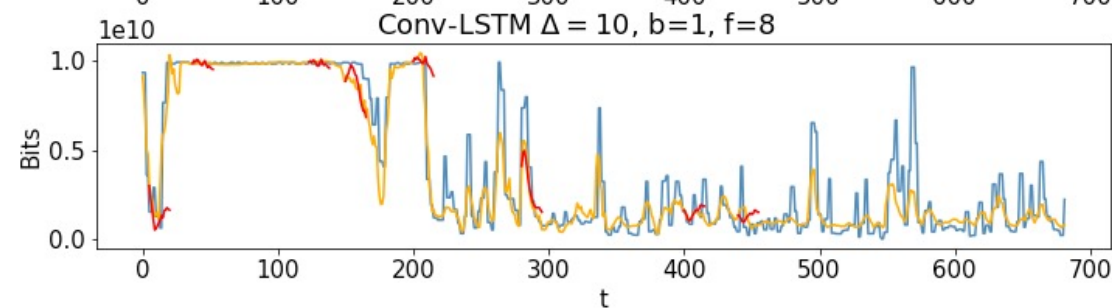
LSTM $\Delta = 10, b=1, f=8$



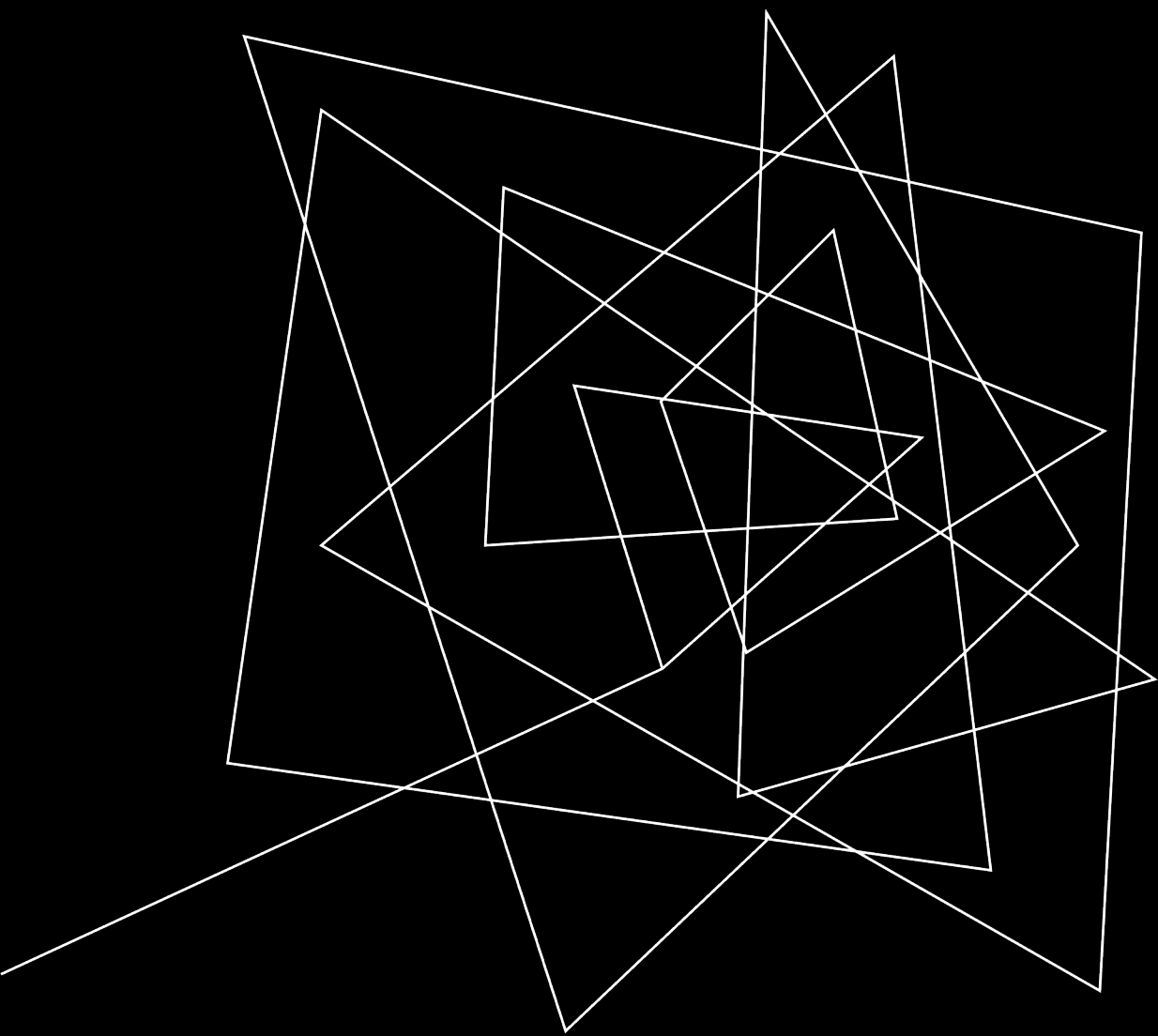
CNN-LSTM $\Delta = 10, b=128, f=64$



Conv-LSTM $\Delta = 10, b=1, f=8$



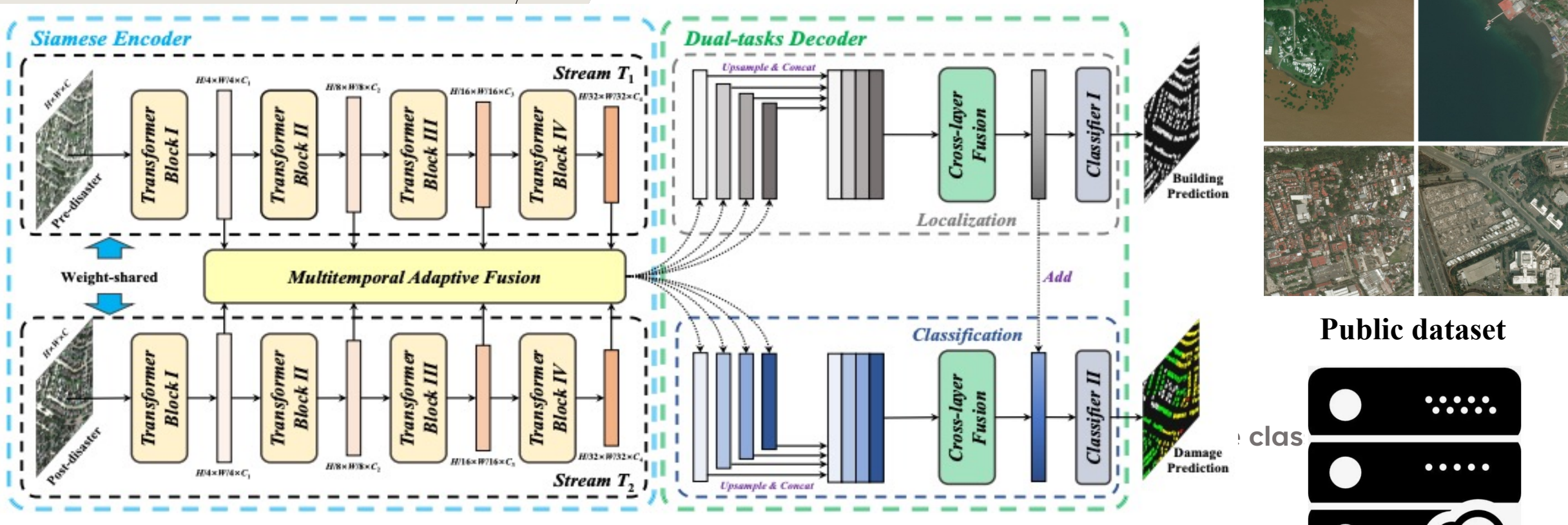
Time window Δ



AI FOR
SUSTAINABILITY

BUILDING DAMAGE DETECTION

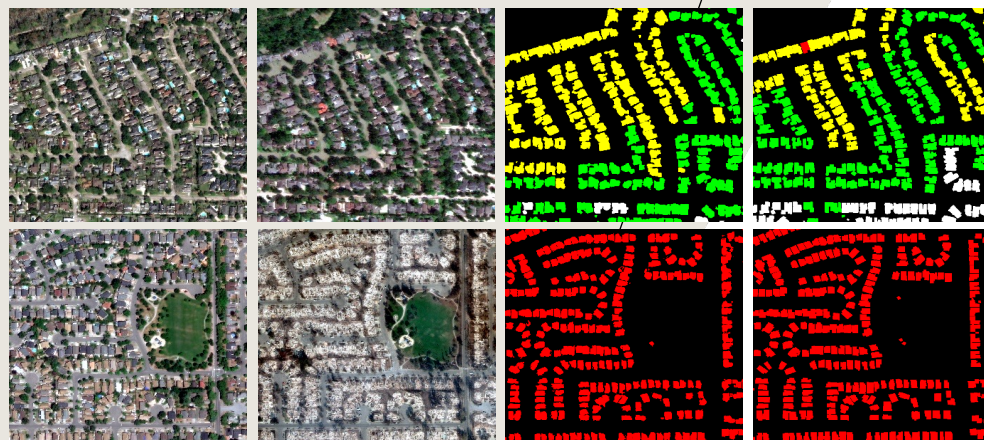
Training



Deploying

PERFORMANCE

xBD dataset (<https://xview2.org/dataset>) from [Maxar Open Data Program](https://www.maxar.com/)



T₁ (Pre-damage)

T₂ (Post-damage)

Prediction

Reference

Transfer learning



Pre-damage



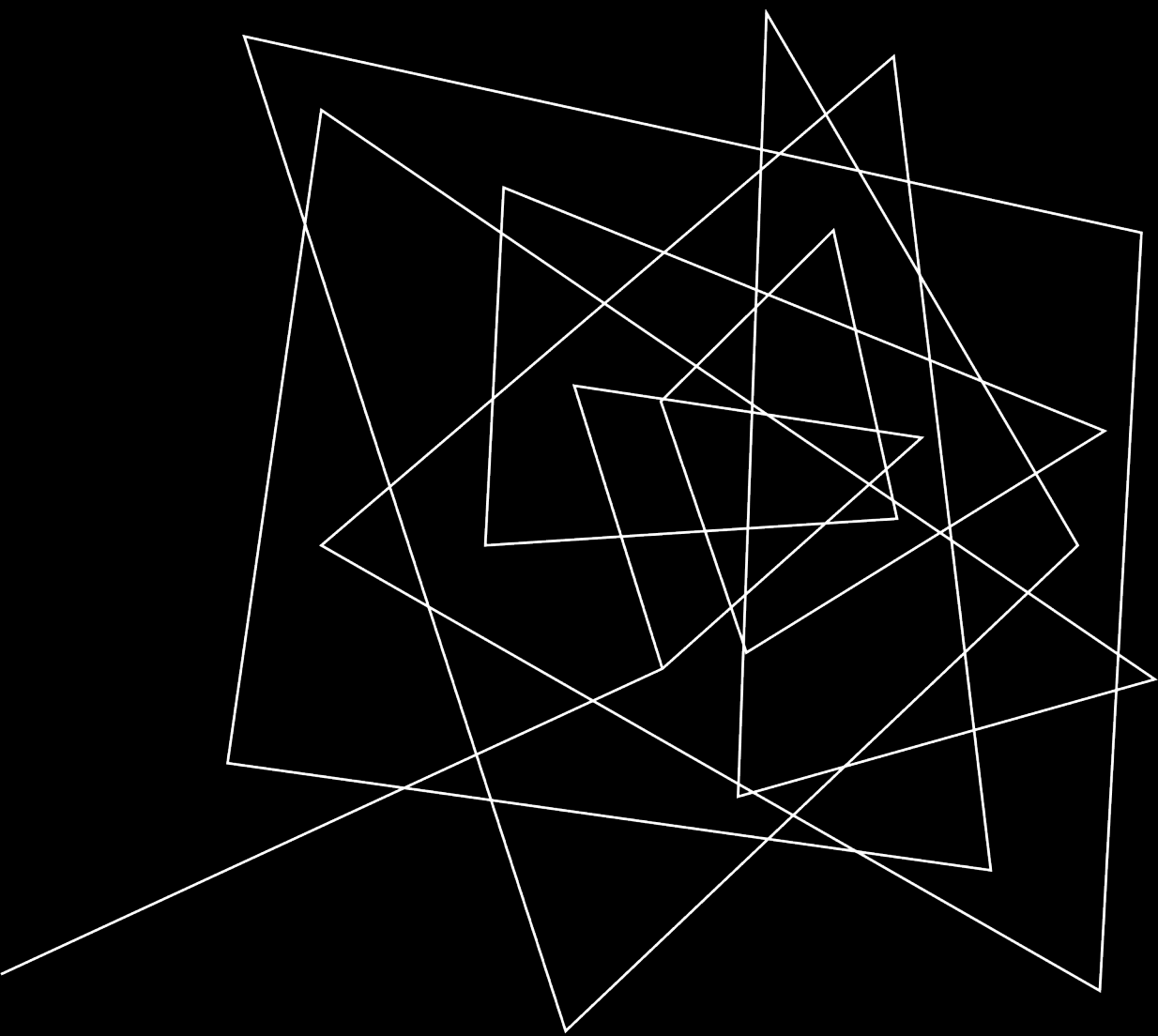
Post-damage

| Method | F_1^{oa} | F_1^{loc} | F_1^{dam} | Damage F_1 per class | | | |
|------------------|--------------|--------------|--------------|------------------------|--------------|--------------|--------------|
| | | | | No | Minor | Major | Destroyed |
| xView2 Baseline | 26.54 | 80.47 | 3.42 | 66.31 | 14.35 | 0.94 | 46.57 |
| Siamese-UNet | 71.68 | 85.92 | 65.58 | 86.74 | 50.02 | 64.43 | 71.68 |
| MaskRCNN | 74.10 | 83.60 | 70.02 | 90.60 | 49.30 | 72.20 | 83.70 |
| ChangeOS | 75.50 | 85.69 | 71.14 | 89.11 | 53.11 | 72.44 | 80.79 |
| DamFormer | 77.02 | 86.86 | 72.81 | 89.86 | 56.78 | 72.56 | 80.51 |

Natural & Man-made damages

Use modified F_1 score combining localization, damage assessment scores*

* suggested in the "CV for Building damage assessment challenge" (<https://www.xview2.org/>)



SUSTAINABLE AI ?

SUSTAINABLE AI ?

ML/DL inference can be more energy efficient than classical algorithms

Training Energy cost can be very high

Contribute to AI community efforts to design best practices¹

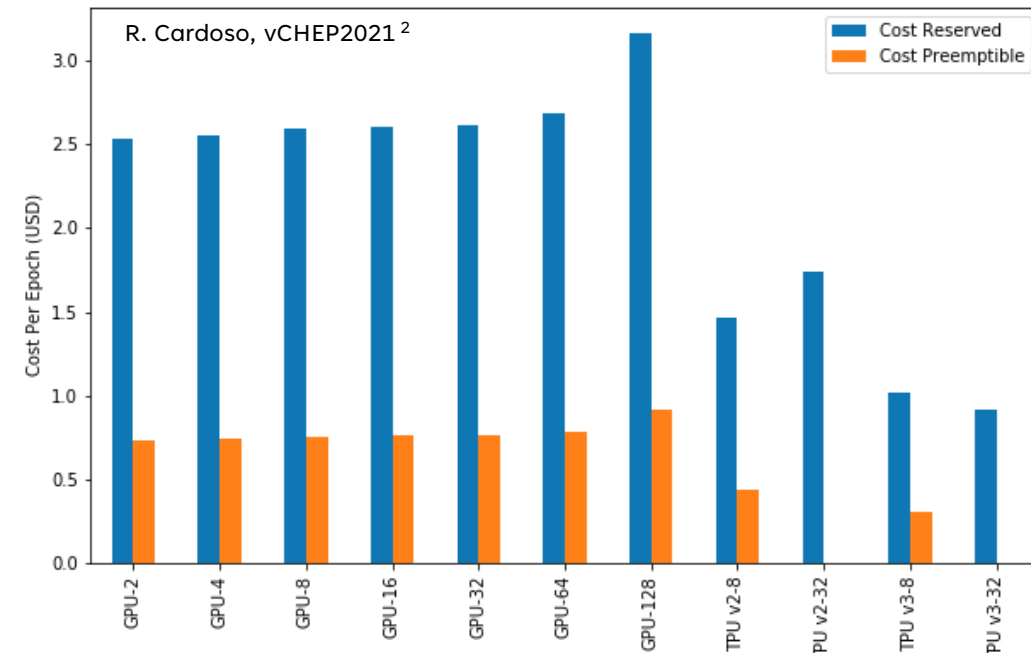
Efficient ML architectures

Processors and systems optimized for ML training, versus general-purpose processors

Centralised computing ? (Cloud vs on prem)

Efficient training strategies (Self-supervision, few-short learning, pre-training)

New hardware ? (neuromorphic, quantum)

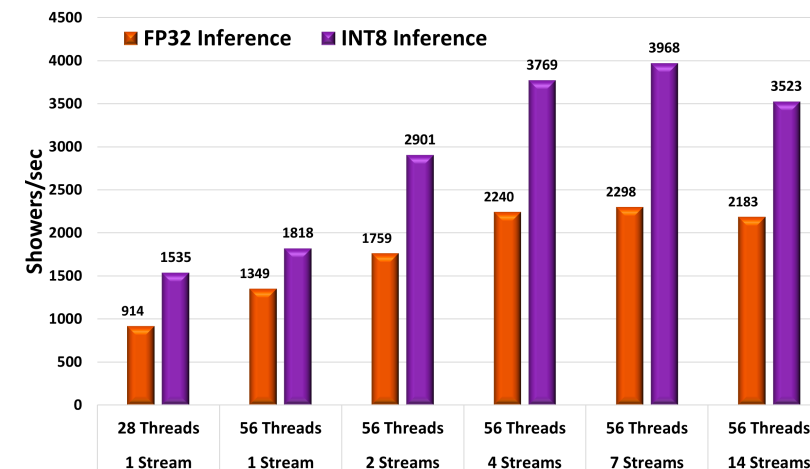


¹ Patterson, David, et al. "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink." (2022).

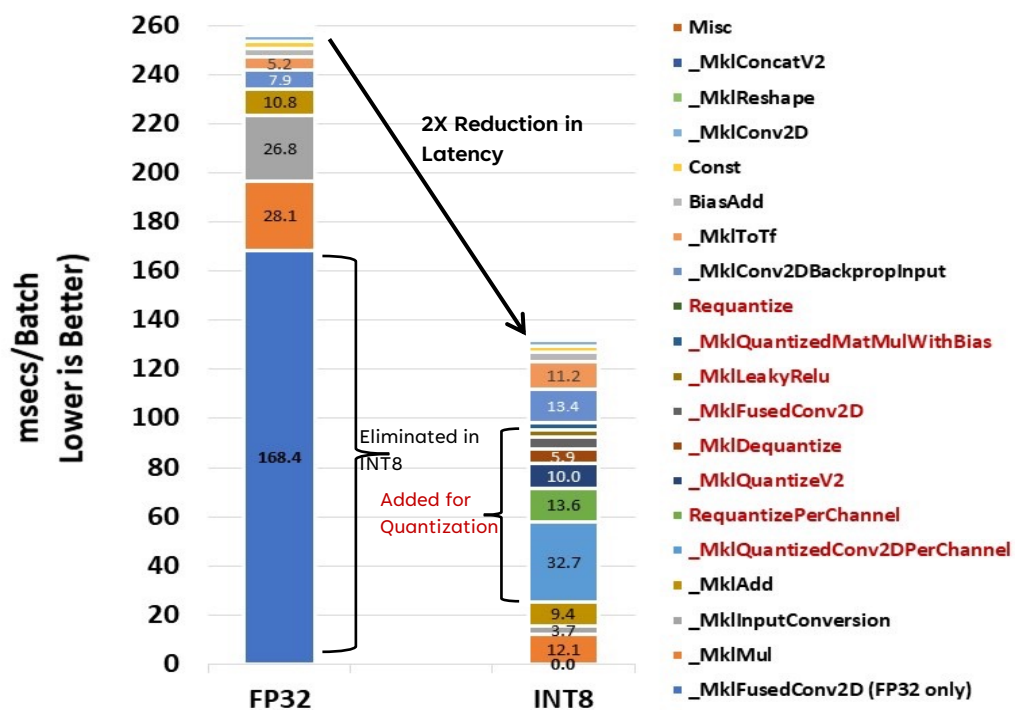
² Cardoso, Renato, et al. "Accelerating GAN training using highly parallel hardware on public cloud." EPJ Web of Conferences. Vol. 251. EDP Sciences, 2021.

FASTER THEN MONTE CARLO?

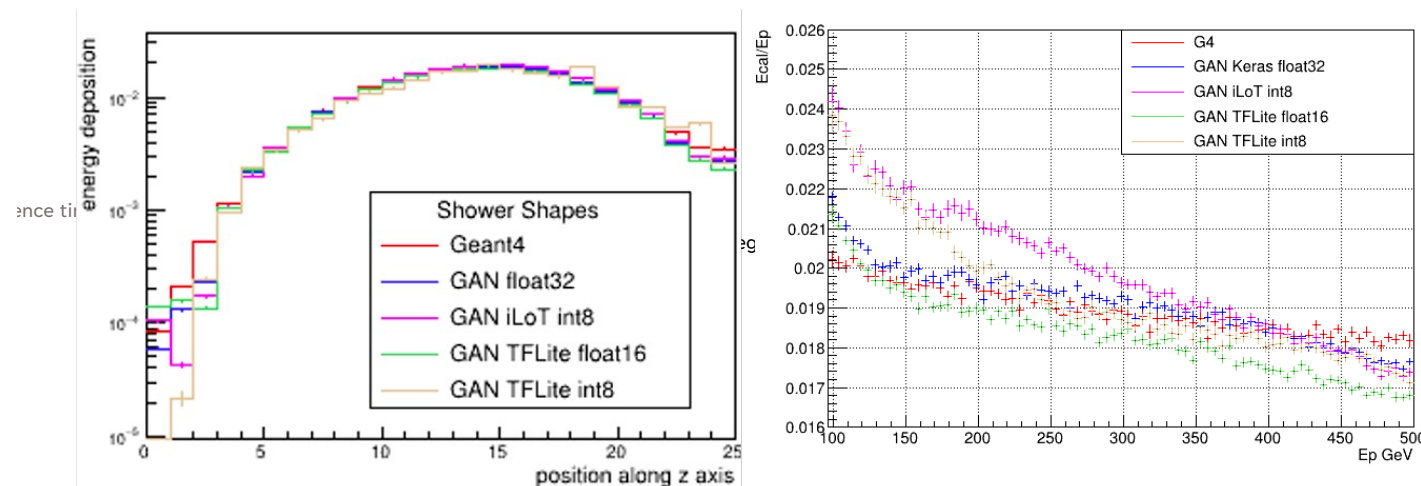
Post training quantization (INT8) using Intel DLBoost and iLoT tool



CERN 3D-GANS Inference FP32 & INT8 (DL Boost) Operation Times per Batch on 15 Intel(R) Xeon(R) Scalable Processor 8280



FP32: 3DGAN is **38000x** faster than Monte Carlo
INT8: quantized 3DGAN is **68000x** faster than Monte Carlo



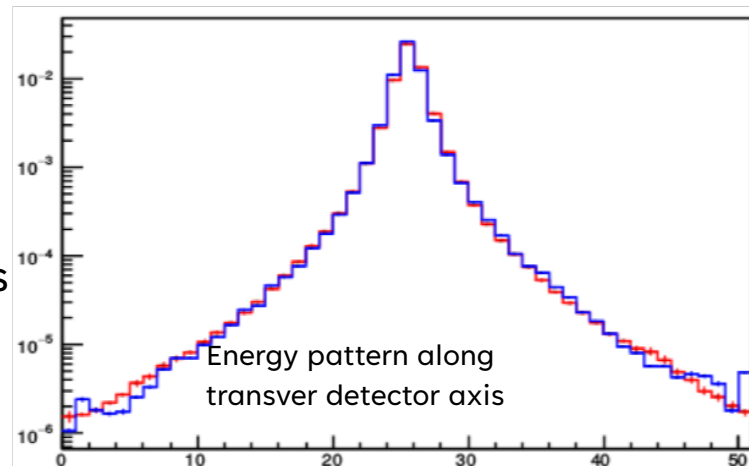
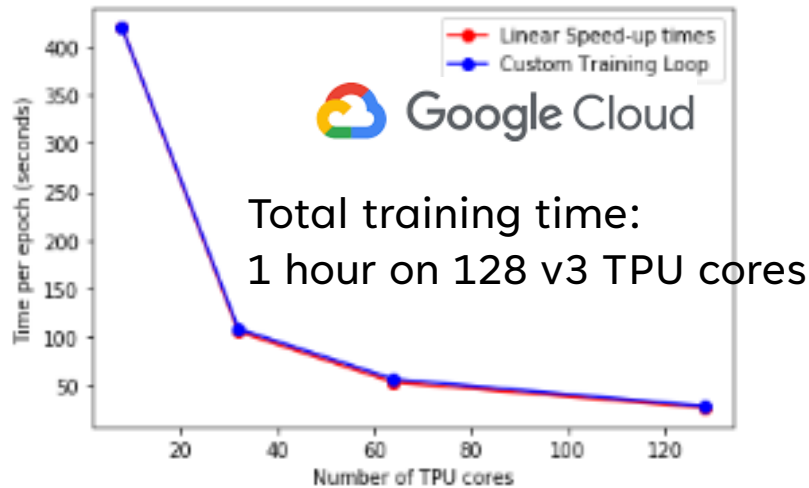
OPTIMIZED TRAINING

Training 3DGAN (3M parameters) takes ~7 days on a GPU

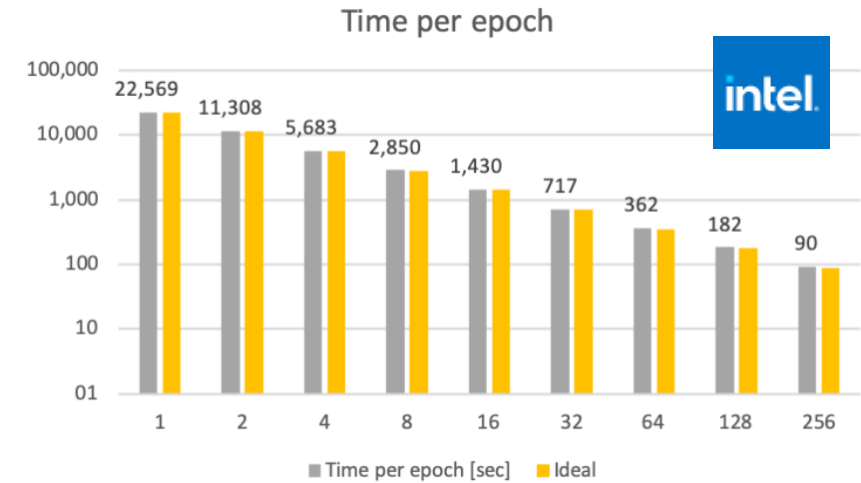
Distributed training is essential

Keep physics under control

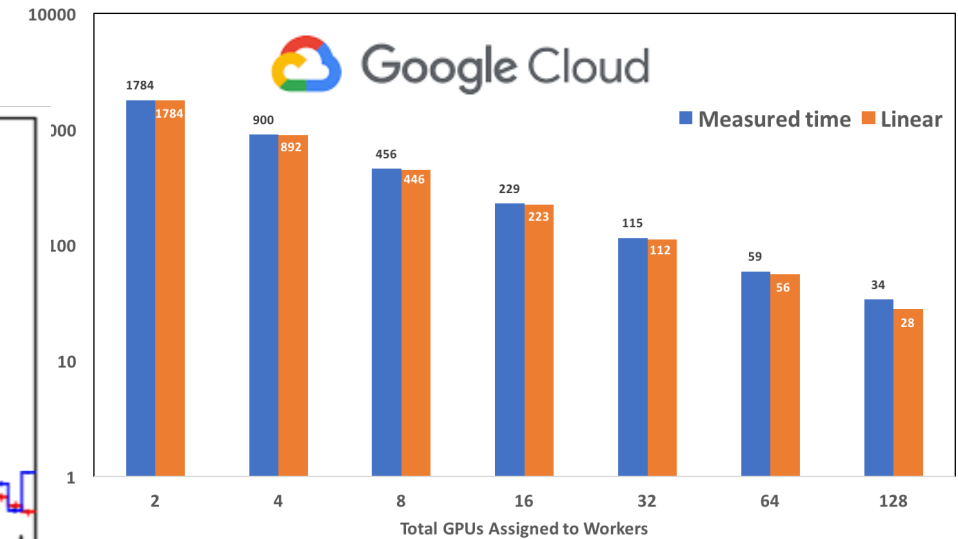
Optimise costs



Total training time: 3 hours on 256 Intel Xeons



Total training time: 1 hour on 128 V100 GPUs





THANK YOU

Sofia Vallecorsa

Sofia.Vallecorsa@cern.ch