Distributed Training and HPO at the Centre of Excellence on AI- and Simulation-based Engineering at Exascale

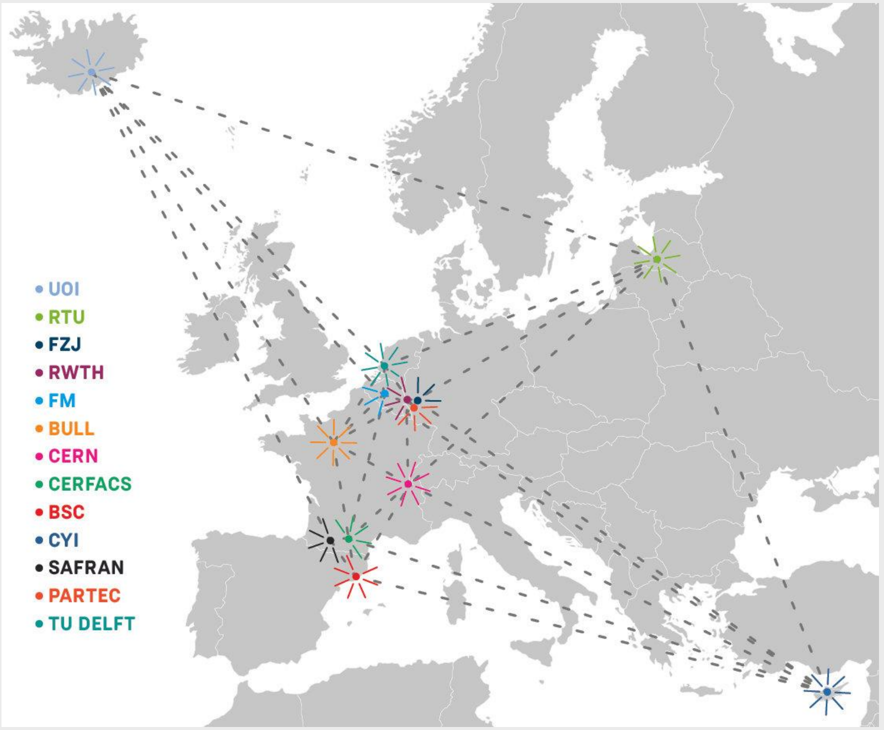
Eric Wulff [1,2], Maria Girone[1,2],
Juan Pablo García Amboage[1,2], Joosep Pata[3]
[1]CERN, [2]CoE RAISE WP4
[3]NICPB

With material from the CMS Collaboration

CoE RAISE

- CoE RAISE [1]: Center of Excellence for Research on AI- and Simulation-based Engineering at Exascale
  - Develop novel, scalable Artificial Intelligence technologies
- CERN (Dr. M. Girone) leads WP4: ***Data-Driven Use-Cases towards Exascale*** [2]
  - Including Task 4.1 (E. Wulff): *Event reconstruction and classification at the CERN HL-LHC*, which we'll see more details on later
- UOI (Prof. M. Riedel) leads WP2: ***AI- and HPC-Cross Methods at Exascale*** [3]
  - Provides expert support on HPC and AI methods to use cases in WP4

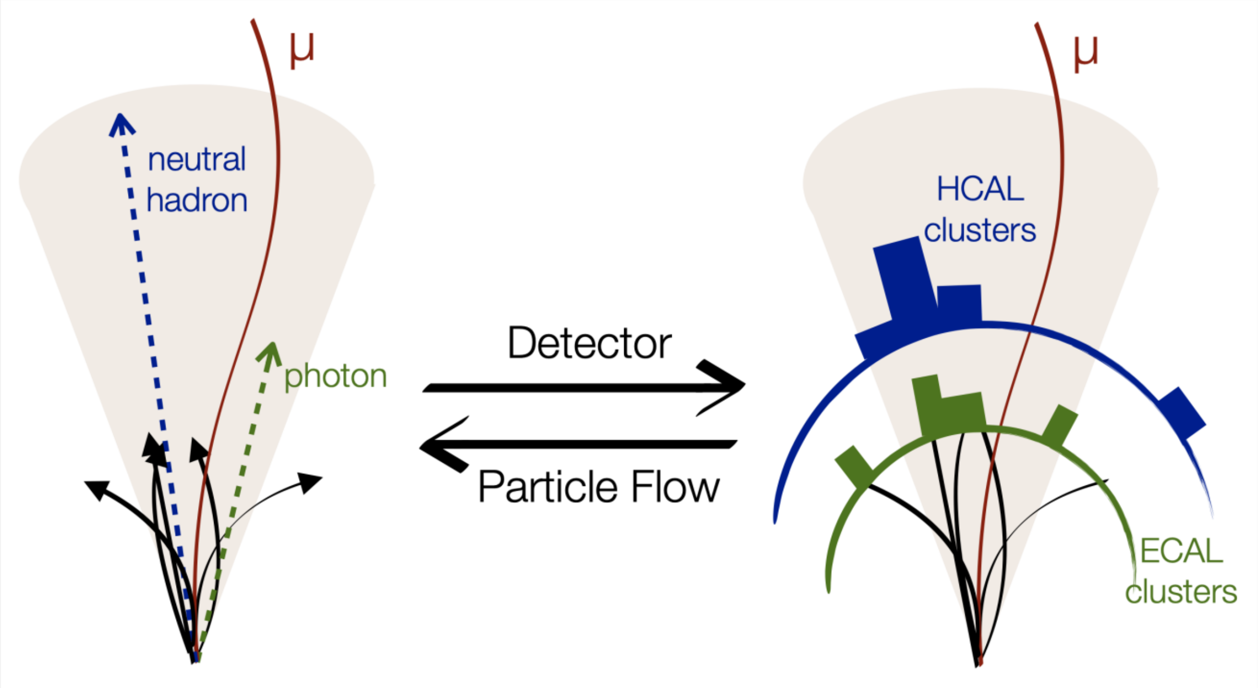
CoE RAISE Partners



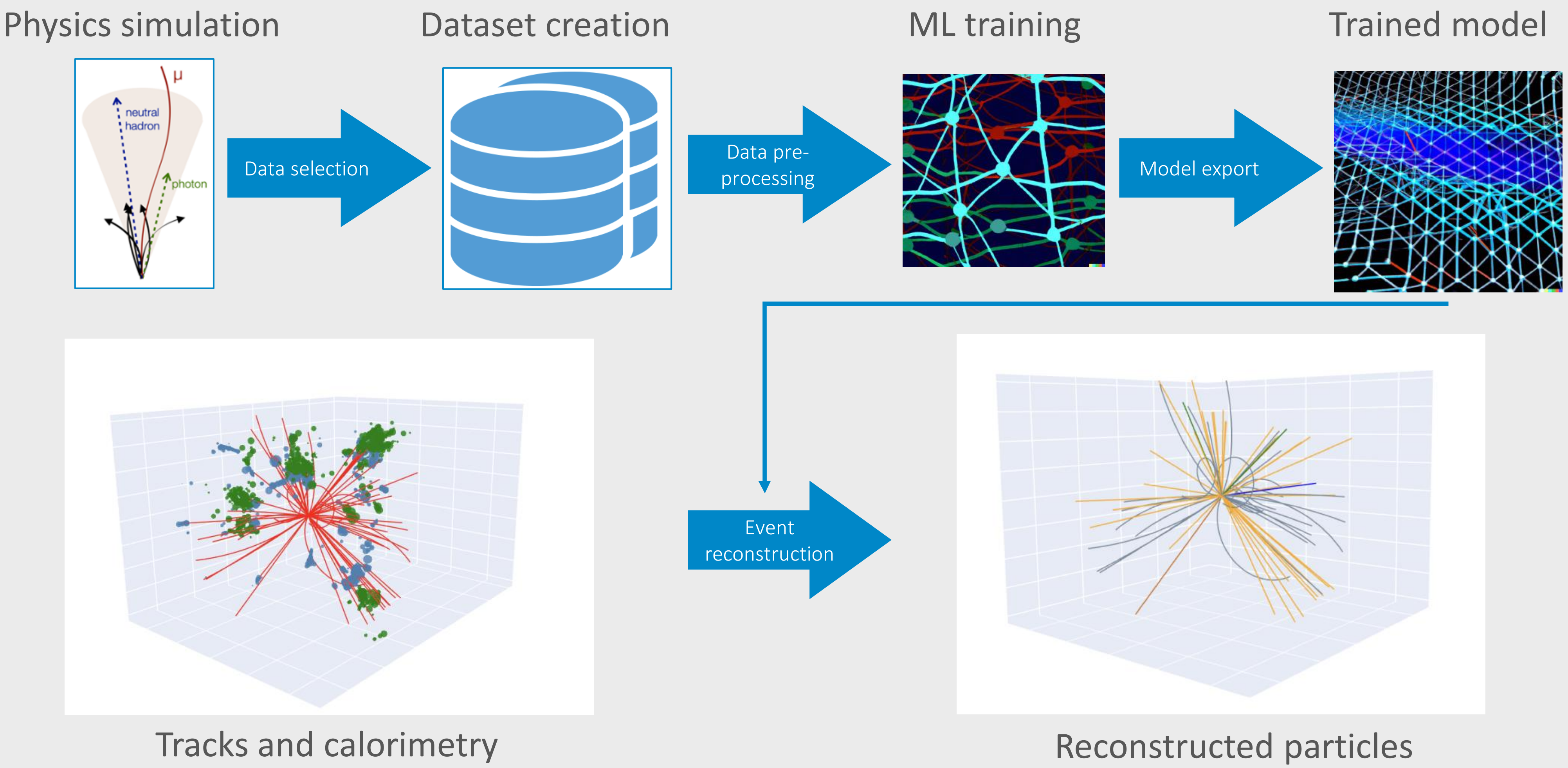
RAISE example use-case:
Event reconstruction and classification at the CERN HL-LHC

Event reconstruction

- Event reconstruction attempts to solve the inverse problem of particle-detector interactions, i.e., going from detector signals back to the particles that gave rise to them
- Particle-flow (PF) reconstruction takes tracks and clusters of energy deposits as input and gives particle types and momenta as output



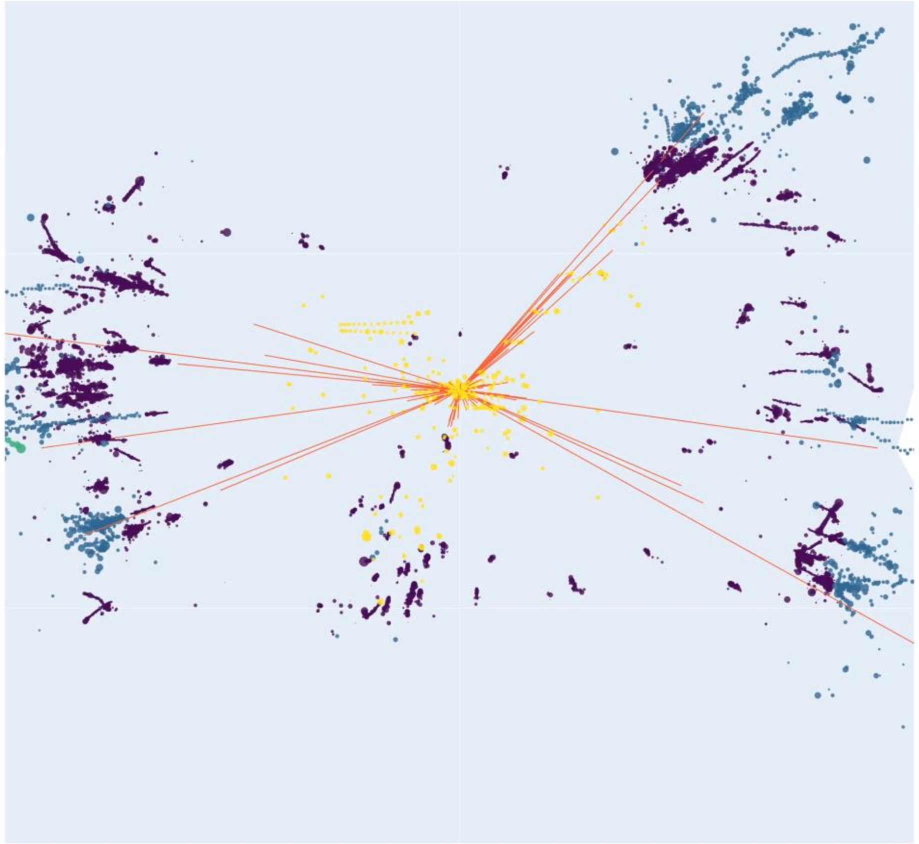
AI-based particle flow reconstruction workflow



New CoE RAISE open data

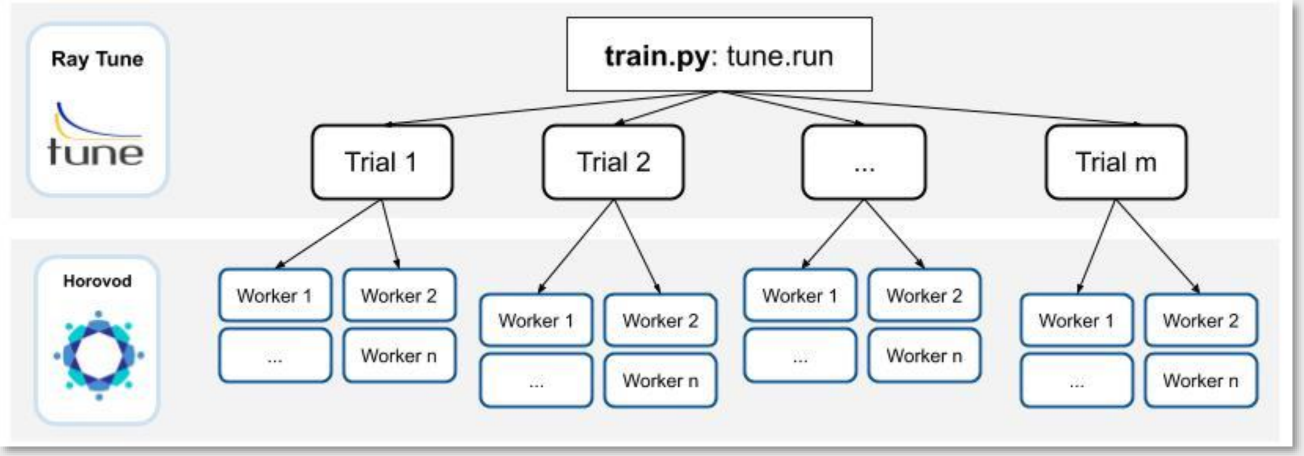
- https://www.coe-raise.eu/od-pfr
- An extensive open dataset of physics events with full GEANT4 [1] simulation, suitable for PF reconstruction, available in the EDM4HEP [2] format
- ~2.5 TB before pre-processing
- The dataset contains
  - Reconstructed tracks, calorimeter hits and clusters
    - We use these as inputs
  - All generator particles
    - We use these as targets
  - Reconstructed particles by the Pandora algorithm [3,4,5]
    - We use these as a baseline for comparison
- A mixture of $t\bar{t}$, $q\bar{q}$, ZH and WW events

3D visualization of a single event



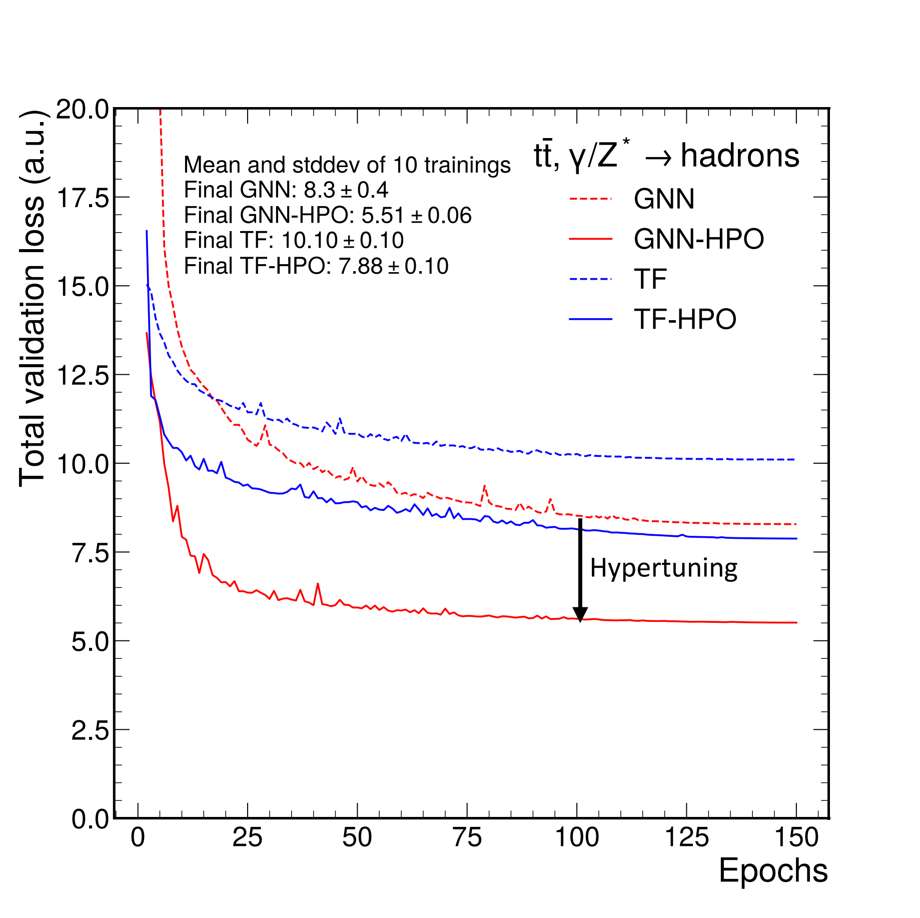
Improvements from large-scale distributed hyperparameter optimization (HPO)

Distributed HPO



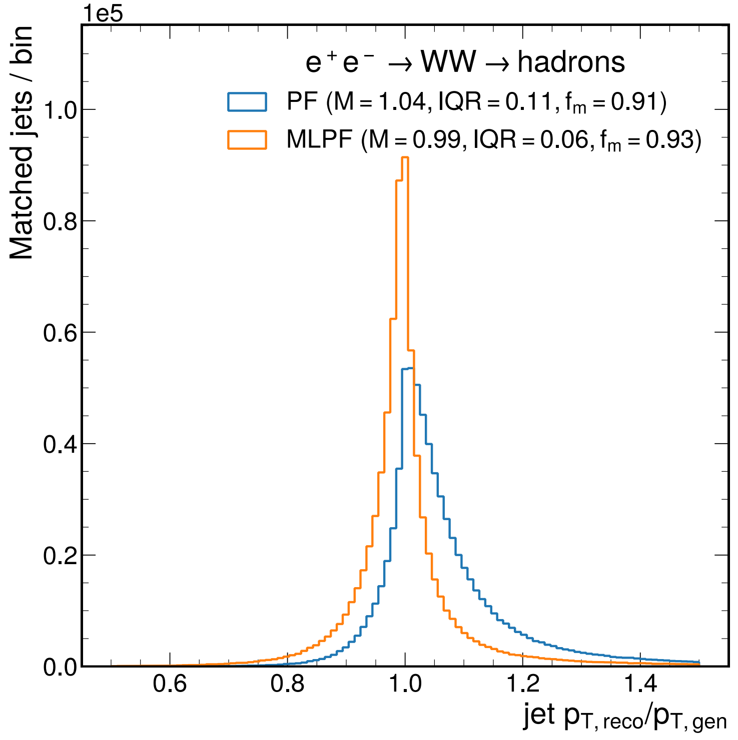
- Two levels of parallelization
- Using ASHA + Bayesian Optimization for HPO
- Final validation loss decreased by ~34% giving a significant performance improvement from HPO

Hyperparameter tuning results

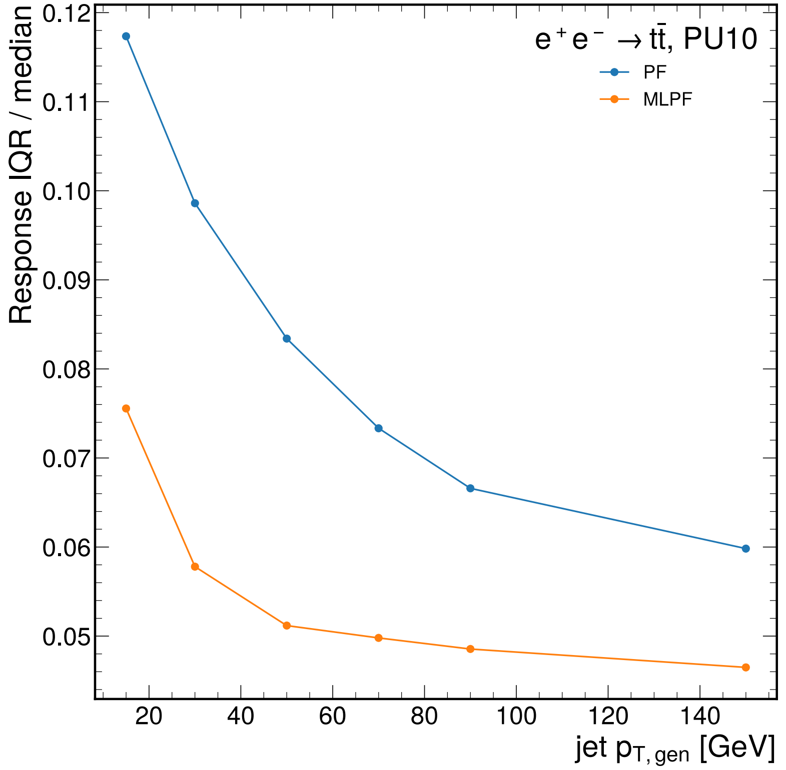


Improvements over the baseline: Jet resolution

- Using data never seen in training
- Almost **50% improvement** in jet response width over the baseline
- Consistent improvement over the entire p_T spectrum



[1]

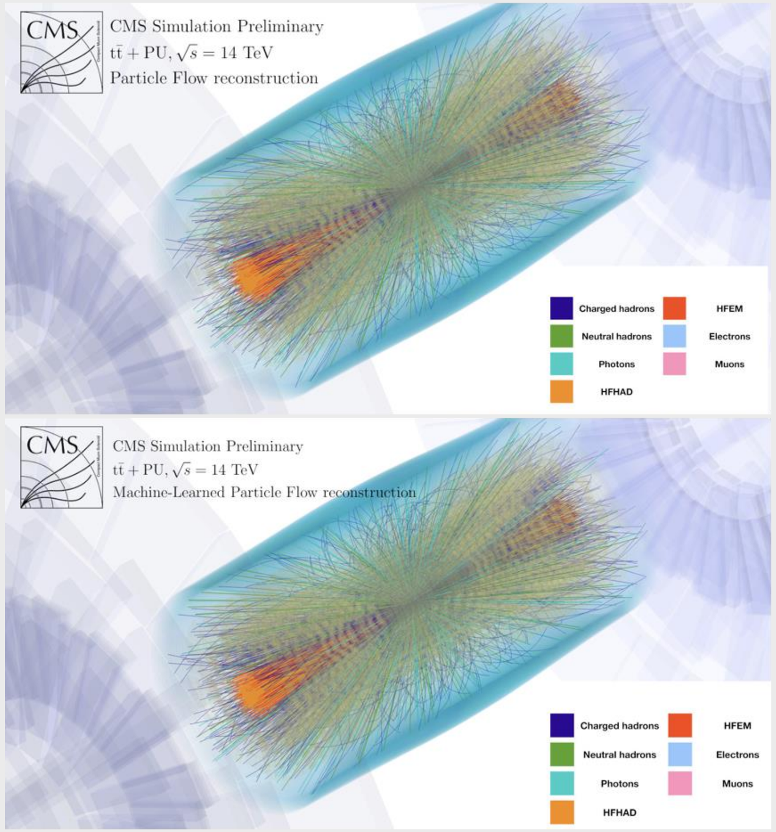


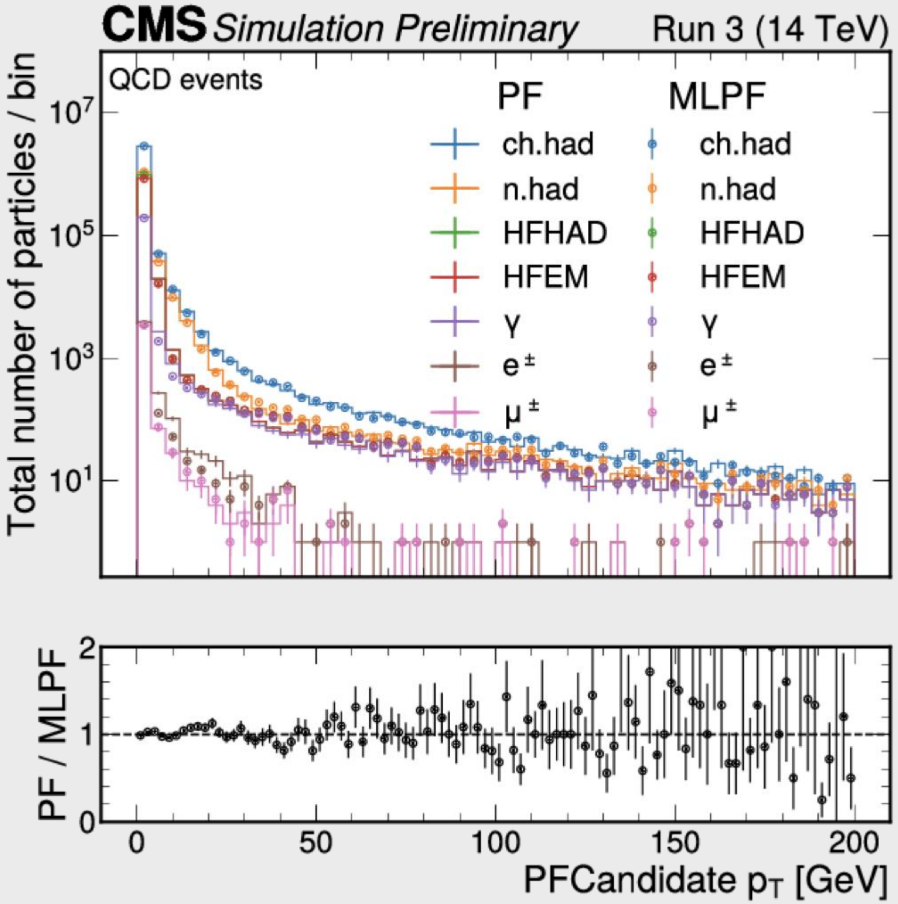
[1]

[1] Joosep Pata, Eric Wulff, Farouk Mokhtar, David Southwick, Mengke Zhang, Maria Girone, Javier Duarte. *Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors, (in press) Commun Phys, (2024)* https://arxiv.org/abs/2309.06782

Tested in a real detector

- This approach was also tested in a real detector (CMS) in 2022
- We plan to update the model for CMS in 2024

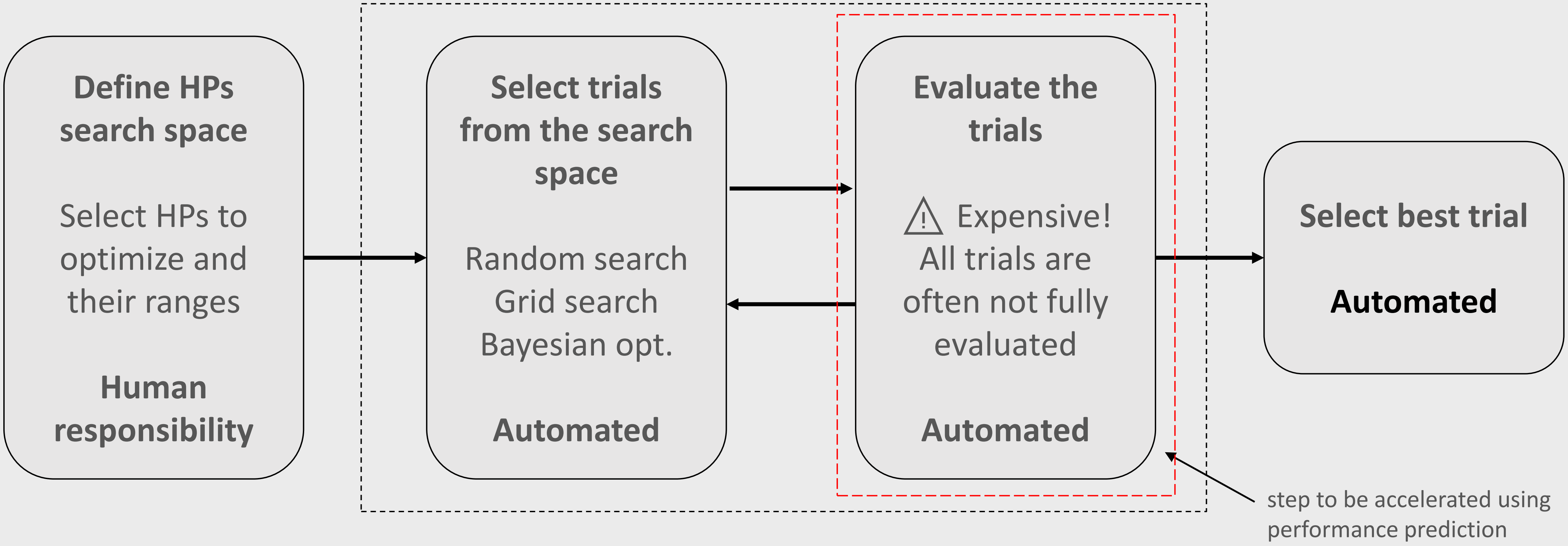




JP, Javier Duarte, Farouk Mokhtar, Eric Wulff, Jieun Yoo, Jean-Roch Vlimant, Maurizio Pierini, Maria Girone. Machine Learning for Particle Flow Reconstruction at CMS. ACAT 2021. https://doi.org/10.48550/arXiv.2203.00330, http://cds.cern.ch/record/2792320

Quantum-SVR for model performance prediction in HPO

The hyperparameter optimization process

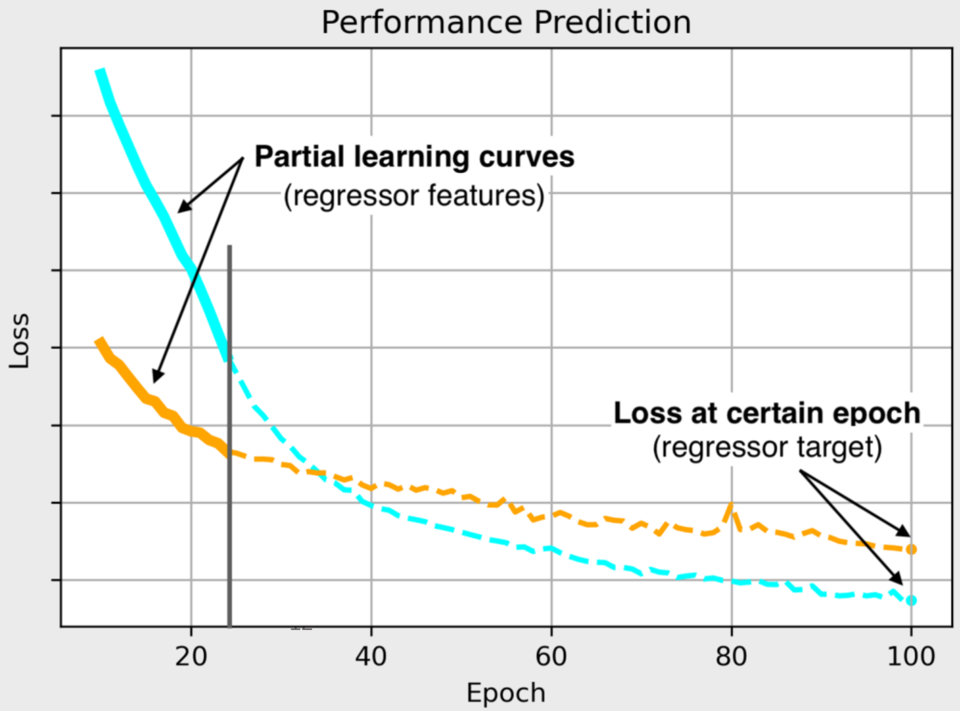


Model performance prediction

- Using performance prediction can accelerate the evaluation step in HPO.
  - Use a meta-model which provides a cheap approximated evaluation of the target model

- The performance predictor
  - Must be fast to train
  - The training samples come from previously fully trained trials

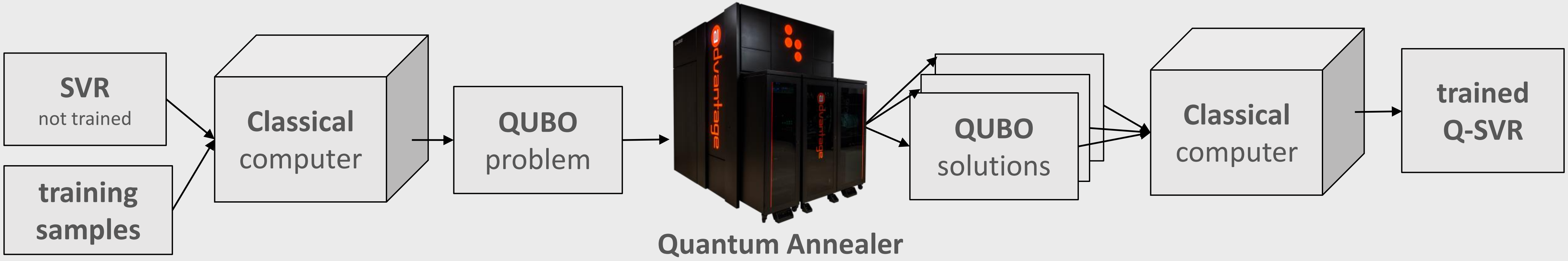
- We use a Quantum Annealer to train a Q-SVR as out model performance predictor



Saved 75 **epochs** of the target model!

Quantum SVR

- Q-SVR: re-formulation of SVR model that can be trained in a Quantum Annealer. (Pasetto et al.)
- In theory: Q-SVR training is O(N) and SVR is O(N^3), N=#training samples. (Date et al.)
- In practice:
  - Currently no time advantage from Q-SVR.
  - Limited training size: ~20 samples.



We used the D-Wave Advantage™ system JUPSI at the Jülich Supercomputer Centre

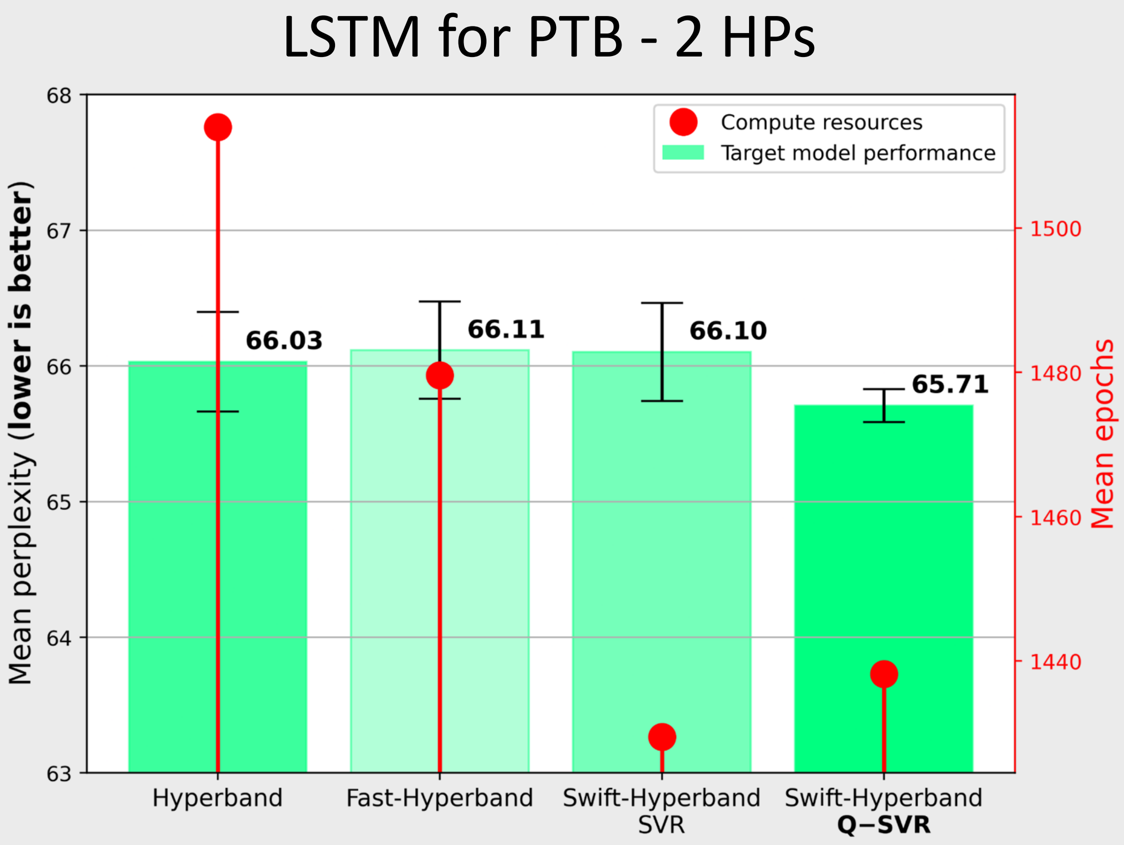
Swift-Hyperband

- Fast-Hyperband: not suitable for integration with Q-SVRs.
- Swift-Hyperband: new approach to combine performance prediction with Hyperband.

Fast-Hyperband		Swift-Hyperband	
Multiple decision points inside each round	Trains **many** SVRs	Only 1 decision point inside each round	Trains **few** SVRs
Estimates σ for every SVR		No need to estimate σ	
Sequential	**Not suitable for Q-SVRs**	**Easily parallelizable**	**Suitable for Q-SVRs**

HPO algorithm comparison

- Green bars show performance of the best found trial on the validation set
- Red markings show consumed compute resources
- Lower is better in both cases



Summary

Summary

- CoE RAISE develops novel, scalable AI methods towards Exascale
  - Use-cases from a wide range of sciences and industry
- New open dataset available on the CoE RAISE website
- Large-scale distributed HPO significantly increased model performance in the example use-case of Machine-Learned Particle Flow (MLPF)
- Swift-Hyperband integrates performance prediction with Hyperband and runs in a hybrid Quantum-Classical manner

drive. enable. innovate.

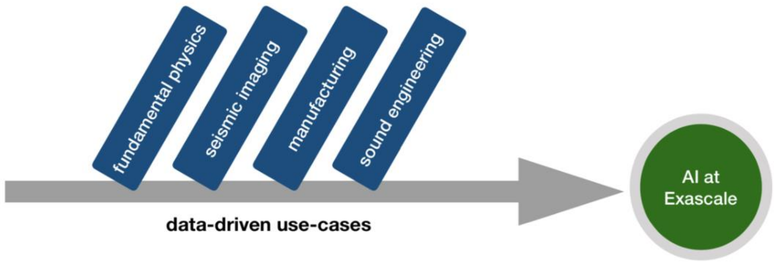
RAISE Center of Excellence

The CoE RAISE project have received funding from
the European Union's Horizon 2020 –
Research and Innovation Framework Programme
H2020-INFRAEDI-2019-1 under grant agreement no. 951733

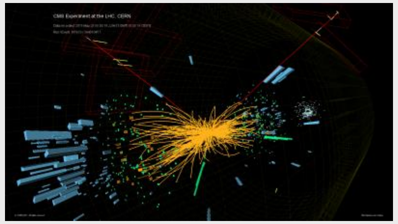
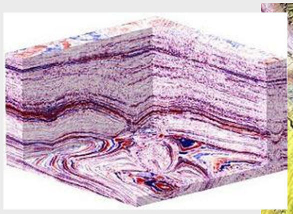
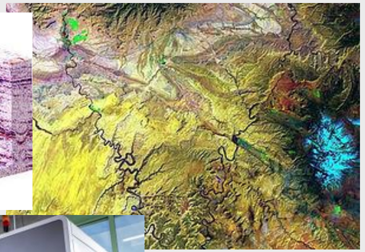
Follow us:

Backup

Data-driven use-cases



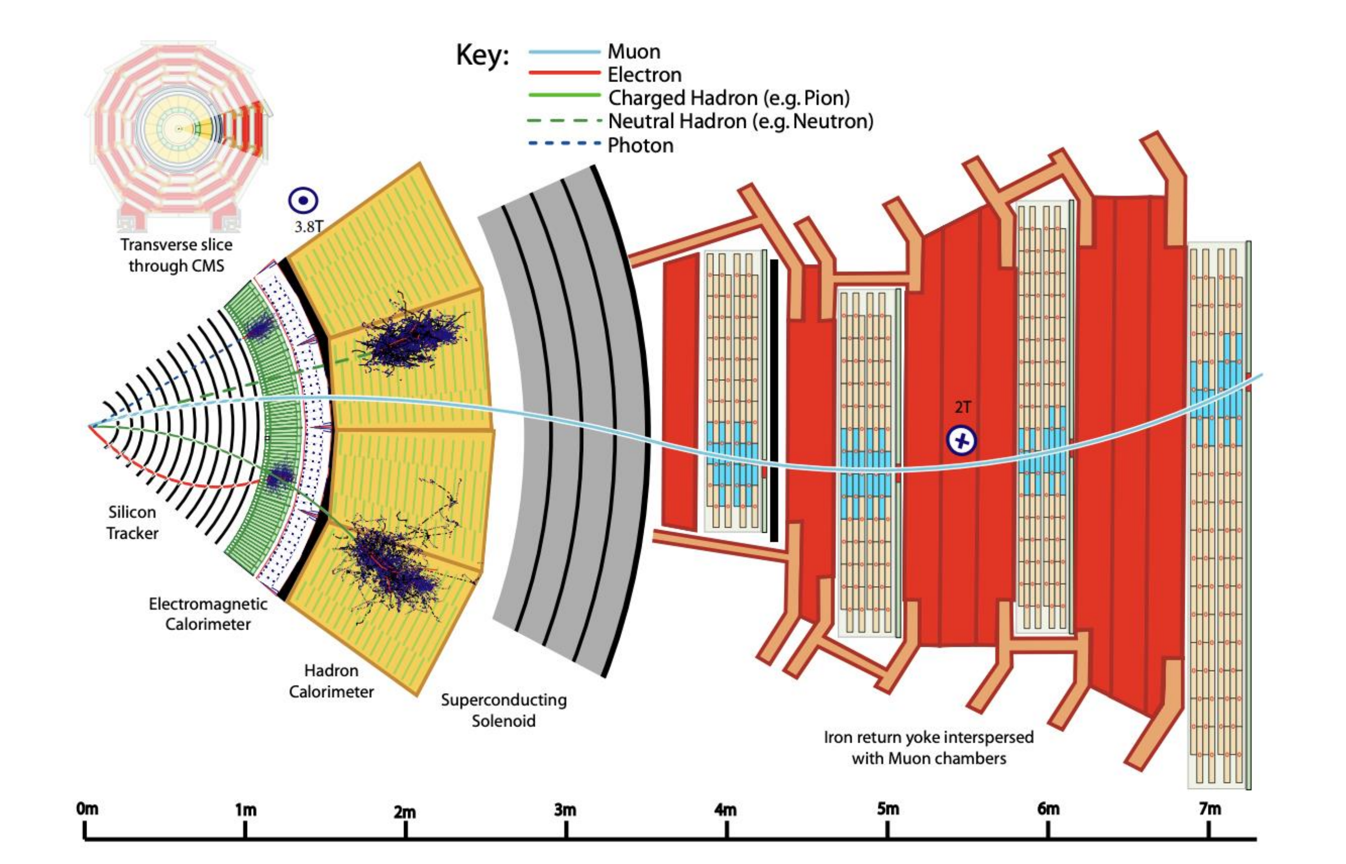
- Representative use-cases from research and industry/SMEs, which have a strong focus on ***data-driven*** technologies, i.e., analyzing data-rich descriptions of physical phenomena
  - *Event reconstruction and classification at the CERN HL-LHC (CERN, RTU)*
    - develop novel approaches for HL-LHC collision event reconstruction replacing traditional algorithms with AI-driven techniques towards HPC-to-Exascale
  - *Seismic imaging with remote sensing for energy applications (FZJ, UOI, CYI)*
    - optimize seismic imaging and remote sensing, enabling AI approaches, combining satellite and airborne data with seismic imaging
  - *Defect-free metal additive manufacturing (UOI, FM)*
    - develop prediction models that detect porosity inside metal parts such that the information is exploited to improve the product quality in additive manufacturing
  - *Sound engineering (FZJ, UOI)*
    - develop a deep-learning-based algorithm that associates individual anatomy to a head-related transfer function (HRTF), for use in spatial audio systems

Event reconstruction at the LHC

- Particle detectors at the LHC are extremely complex, with many subdetectors
- Particles interact with the detectors and leave **tracks** and **energy deposits**
- **Information** from subdetectors are **combined** to produce a **particle-level interpretation** of the event
- **Event reconstruction** is the process of inferring higher-level physics objects from detector signals

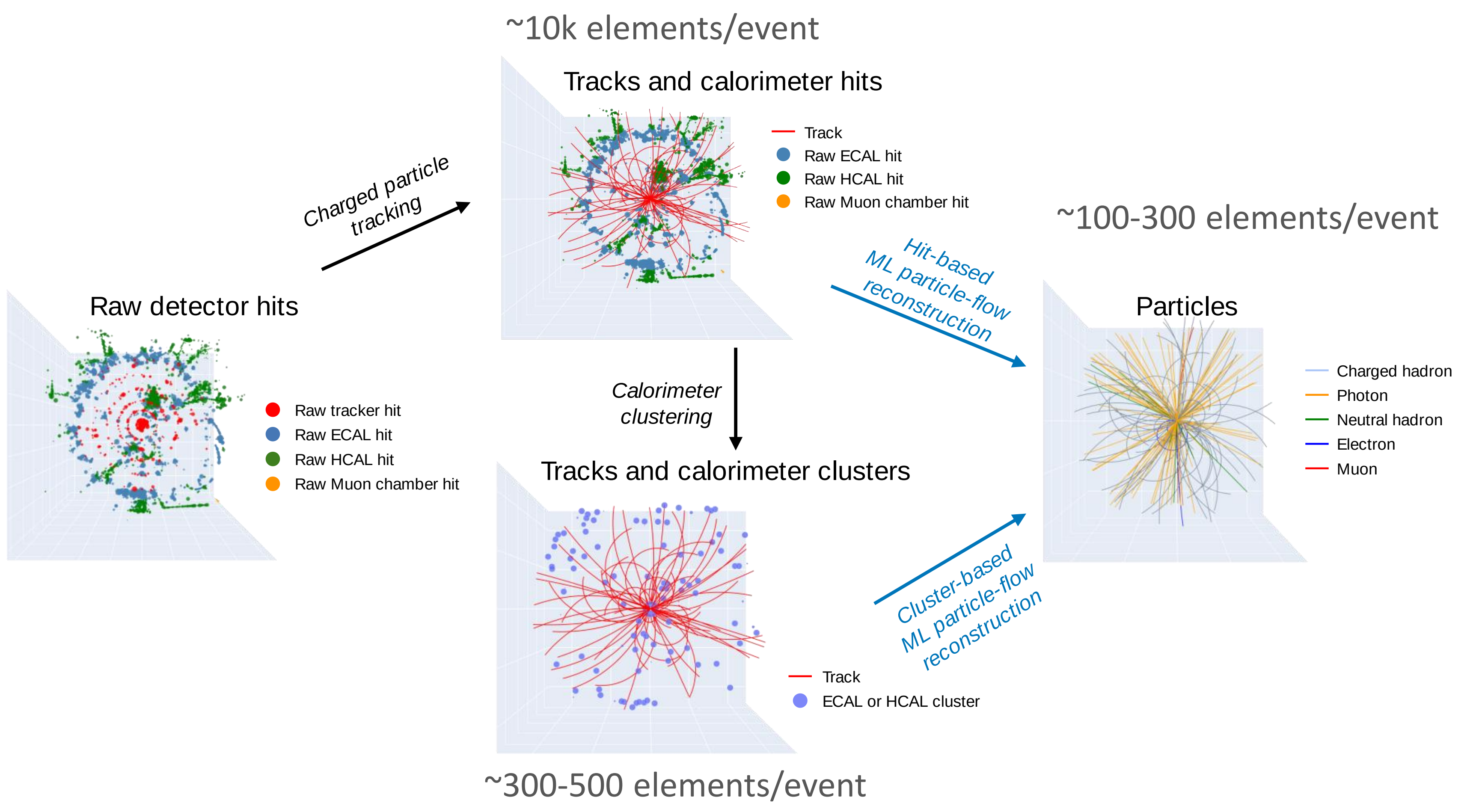
Transverse slice through the CMS detector



JINST 12 (2017) P10003

New open dataset for supervised learning

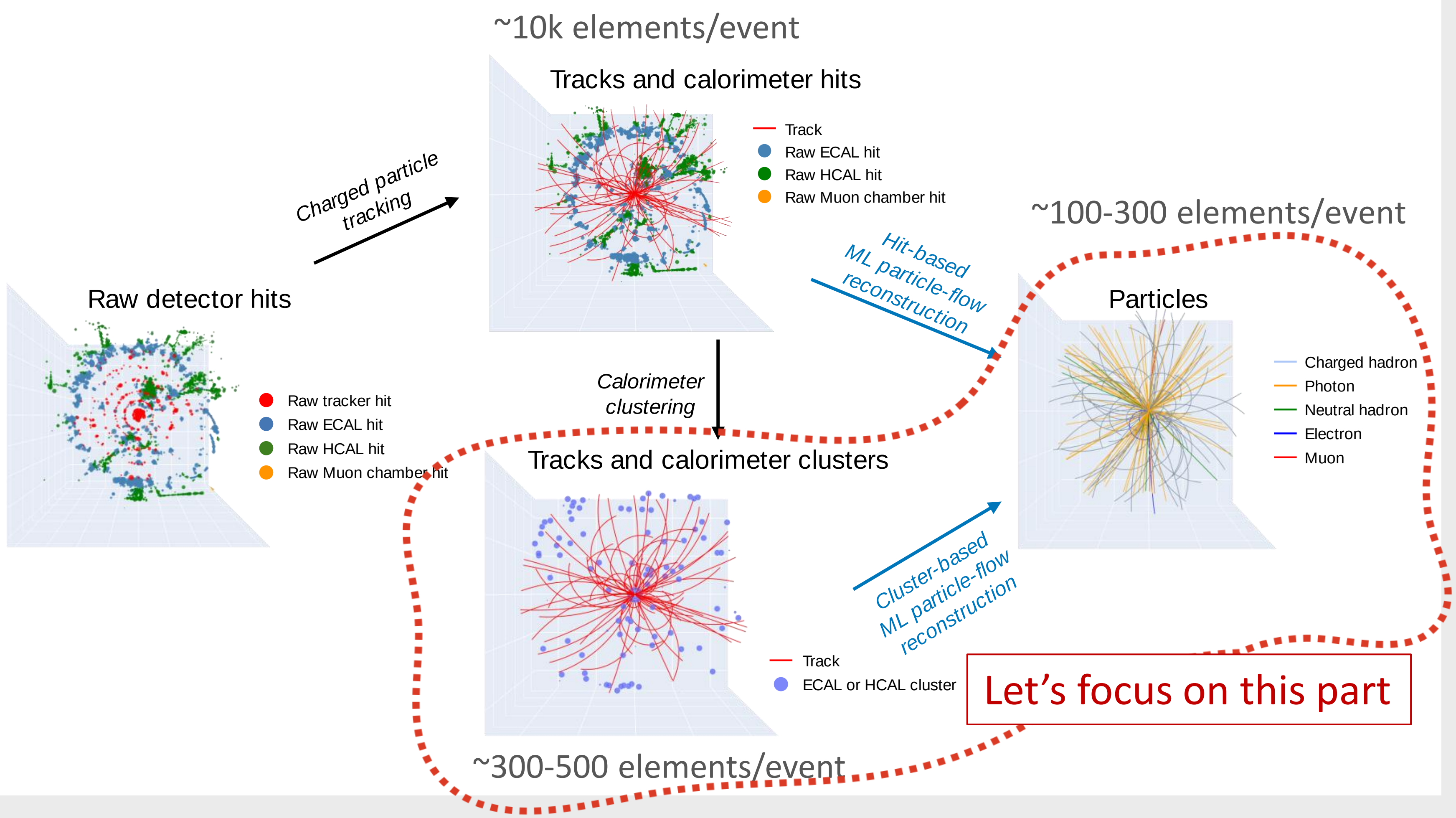
- Full detector simulation using GEANT4
- Electron-positron collision in CLIC detector geometry
- Dataset contains
  - Calorimeter and tracker hits
  - Tracks and calorimeter clusters
  - Generator-level particles (ground truth for supervised learning)
  - Baseline reconstructed particles (from a non-ML PF algo)



Joosep Pata, Eric Wulff, Farouk Mokhtar, David Southwick, Mengke Zhang, Maria Girone, Javier Duarte. *Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors, (in press) Commun Phys, (2024)* https://arxiv.org/abs/2309.06782

New open dataset for supervised learning

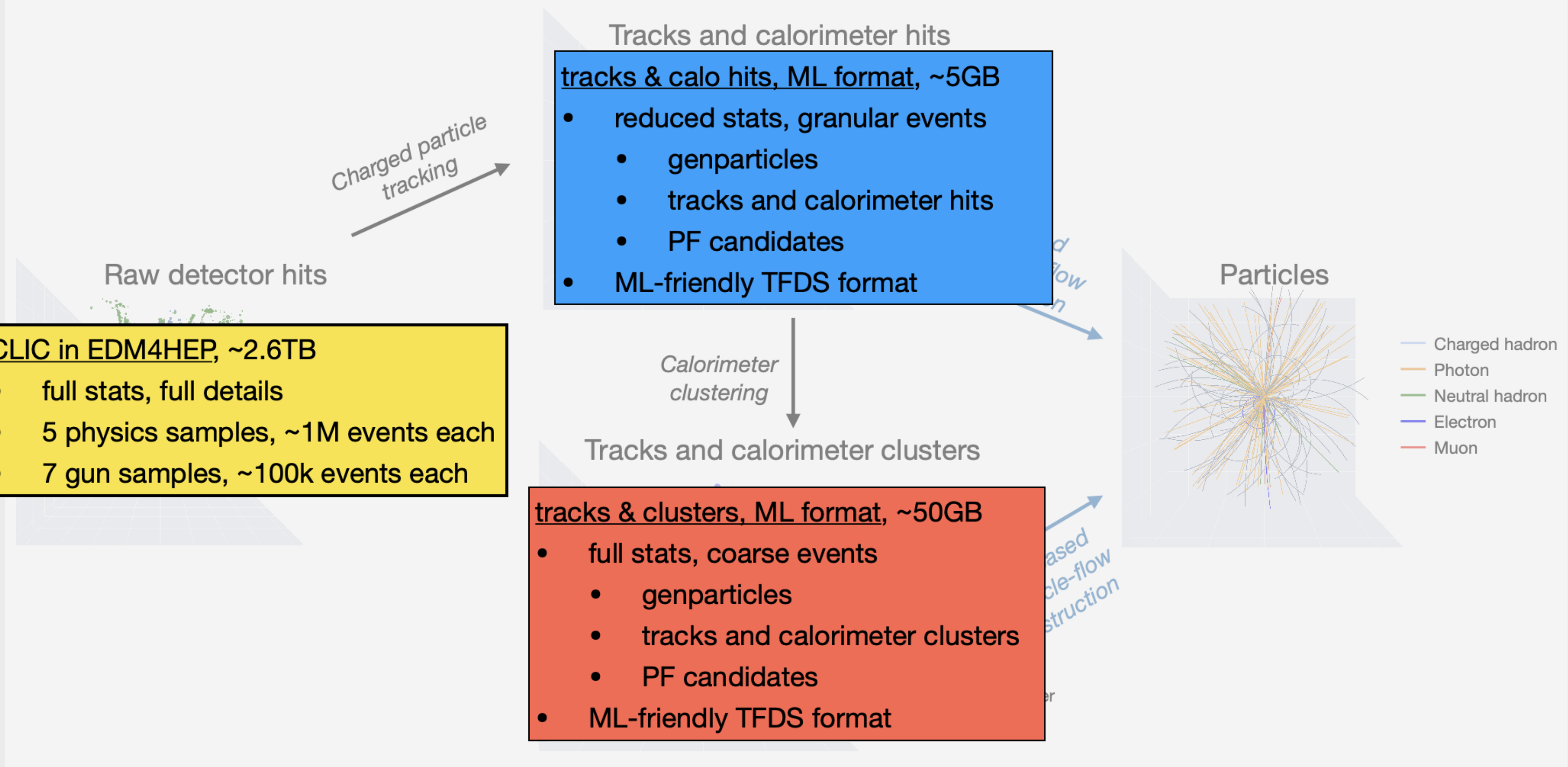
- Full detector simulation using GEANT4
- Electron-positron collision in CLIC detector geometry
- Dataset contains
    - Calorimeter and tracker hits
    - Tracks and calorimeter clusters
    - Generator-level particles (ground truth for supervised learning)
    - Baseline reconstructed particles (from a non-ML PF algo)



Joosep Pata, Eric Wulff, Farouk Mokhtar, David Southwick, Mengke Zhang, Maria Girone, Javier Duarte. *Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors, (in press) Commun Phys, (2024)* https://arxiv.org/abs/2309.06782

Open datasets

RAISE Center of Excellence



CLIC in EDM4HEP, ~2.6TB
- full stats, full details
- 5 physics samples, ~1M events each
- 7 gun samples, ~100k events each

tracks & calo hits, ML format, ~5GB
- reduced stats, granular events
    - genparticles
    - tracks and calorimeter hits
    - PF candidates
- ML-friendly TFDS format

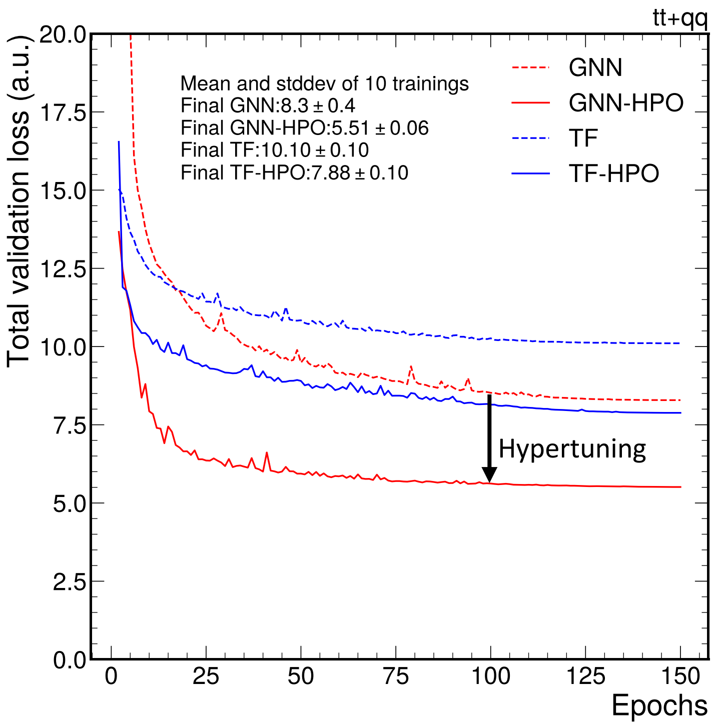
tracks & clusters, ML format, ~50GB
- full stats, coarse events
    - genparticles
    - tracks and calorimeter clusters
    - PF candidates
- ML-friendly TFDS format

- https://doi.org/10.5281/zenodo.8260741
- https://doi.org/10.5281/zenodo.8414225
- https://doi.org/10.5281/zenodo.8409592

Improvement in training from HPO

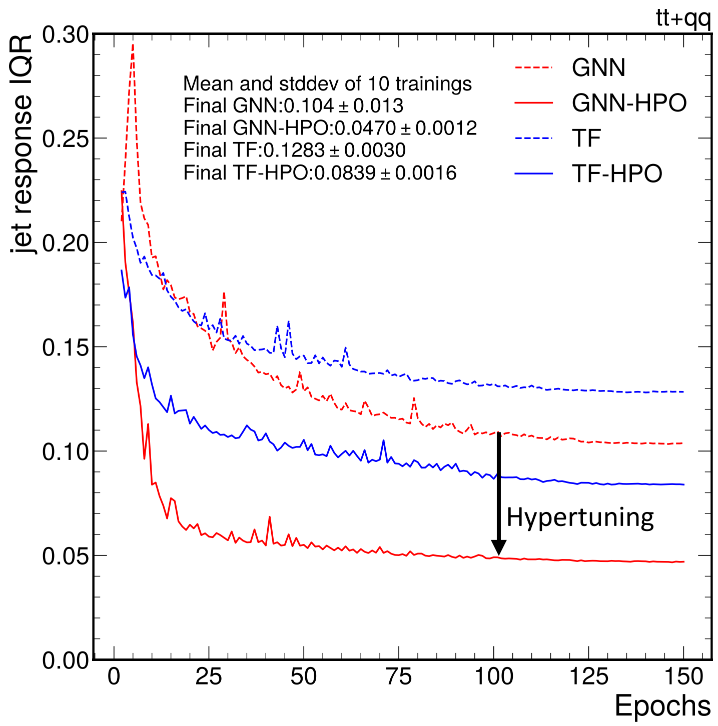
- HPO significantly improved model performance for both the GNN-based and the transformer-based MLPF models
- GNN outperforms transformer

Validation loss



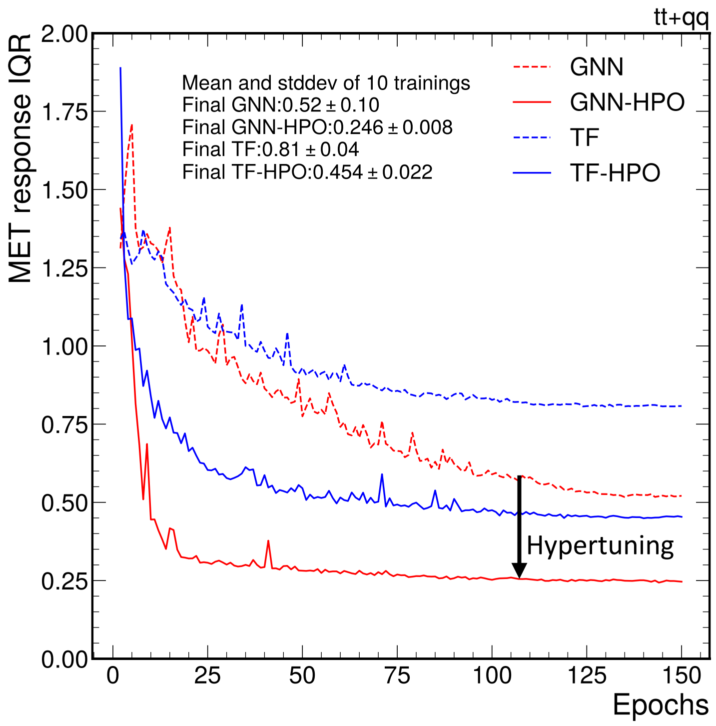
[1]

Jet resolution



[1]

MET resolution



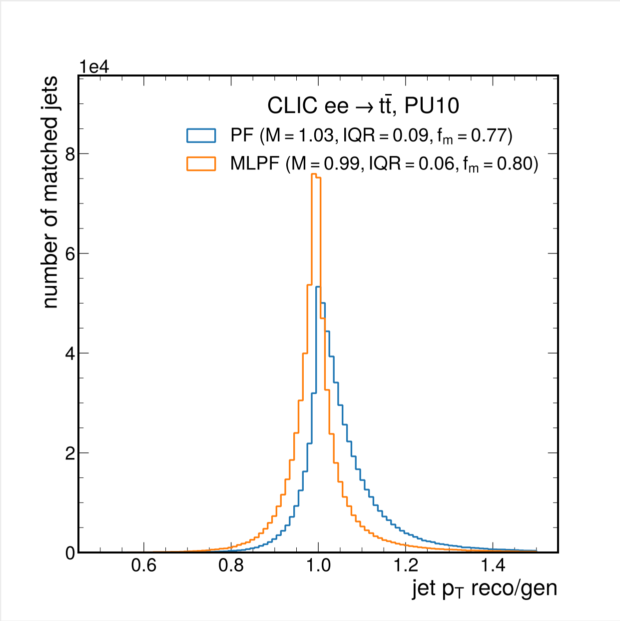
[1]

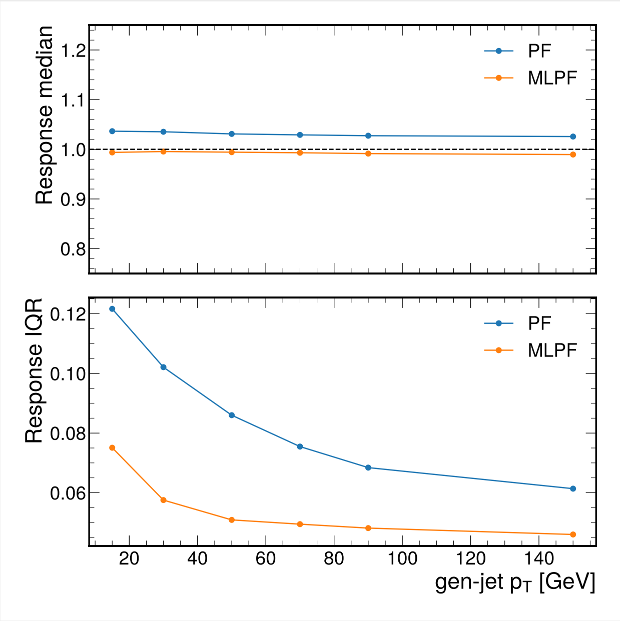
[1] Joosep Pata, Eric Wulff, Farouk Mokhtar, David Southwick, Mengke Zhang, Maria Girone, Javier Duarte. *Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors*, (in press) Commun Phys, (2024) https://arxiv.org/abs/2309.06782

Jet and MET in ttbar + PU10 test data

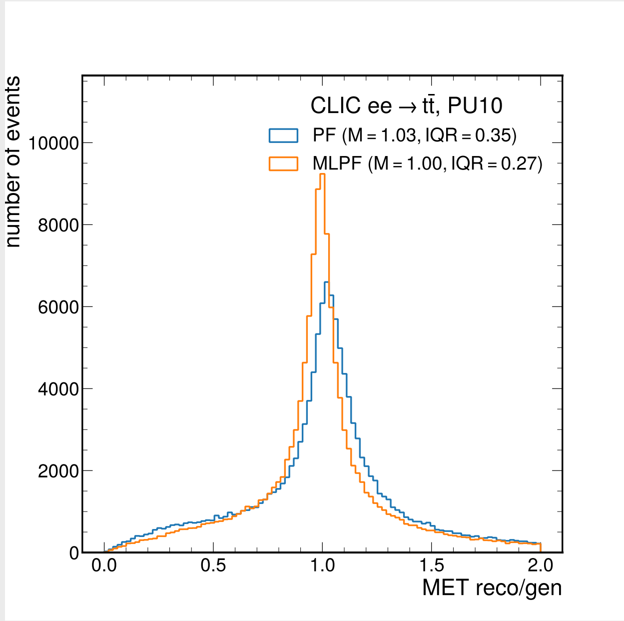
- For all test samples MLPF outperforms PF in Jet and MET reconstruction in terms of response width (quantified by median and interquartile range (IQR))
- MLPF also outperforms PF in terms of fraction of reconstructed jets ($^{n_{reco\ jets}}/_{n_{ground-truth\ jets}}$)
- Very similar results are seen in ZH and WW events

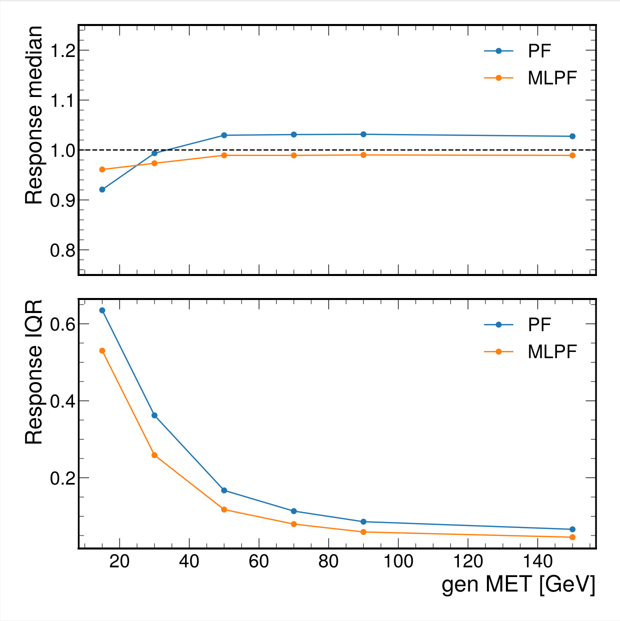
Jets resolution





MET resolution

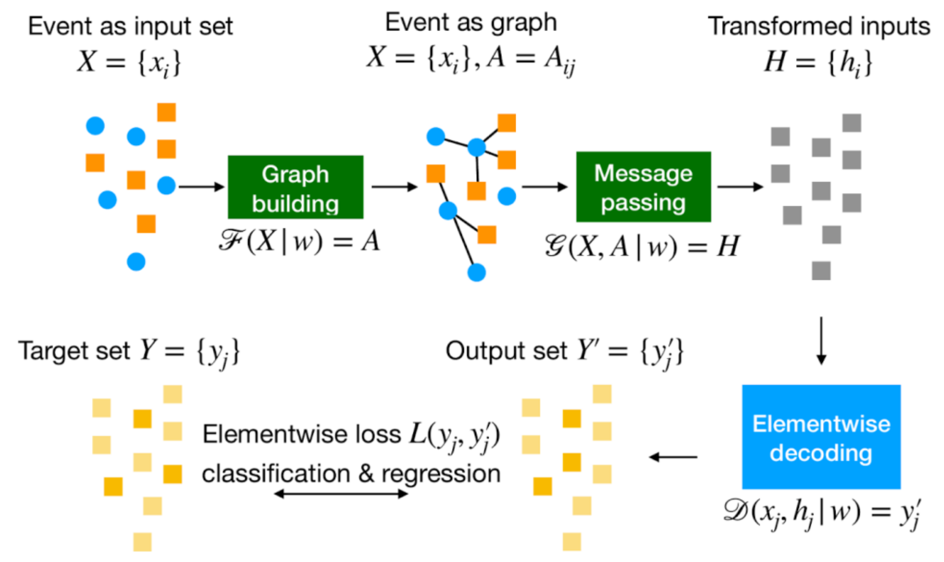




Machine-Learned Particle-Flow (MLPF)

- The Particle Flow (PF) Algorithm [1]
  - Tries to identify and reconstruct all stable individual particles from collision events by combining information from different subdetectors (tracks, calorimeter clusters)
- Machine-Learned Particle-Flow (MLPF) [2]
  - GPU accelerated, GNN-based algorithm for PF
  - Code available on GitHub
  - See ACAT2021 talk by J. Pata (and proceedings) for more MLPF model details and ACAT 2021 talk by E. Wulff (and proceedings) for more details on the hypertuning of MLPF
  - ACAT2022 poster

The MLPF model



$$x_i = [\text{elem. type}, p_T, E_{ECAL}, E_{HCAL}, \eta, \phi, \eta_{outer}, \phi_{outer}, q, \dots]$$

$$y_j = [\text{PID}, p_T, E, \eta, \phi, q], \quad \text{PID} \in \{\text{none, charged hadron, neutral hadron}, \gamma, e^\pm, \mu^\pm, \dots\}$$

$$h_i \in \mathbb{R}^{N_{hidden}}$$

Trainable neural networks: $\mathcal{F}, \mathcal{G}, \mathcal{D}$

● - track, ■ - calorimeter cluster, ■ - encoded element

■ - target (predicted) particle, ■ - no target (predicted) particle

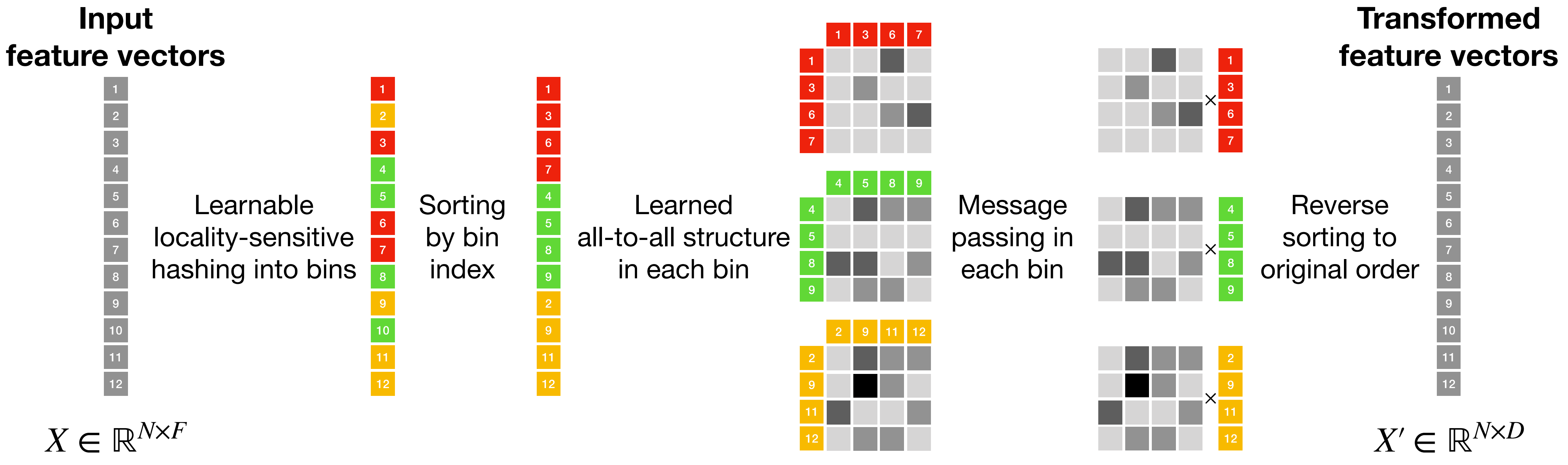
Based on Eur. Phys. J. C 81, 381 (2021)
https://arxiv.org/abs/2101.08578

[1] CMS Collaboration https://cds.cern.ch/record/1194487?ln=en

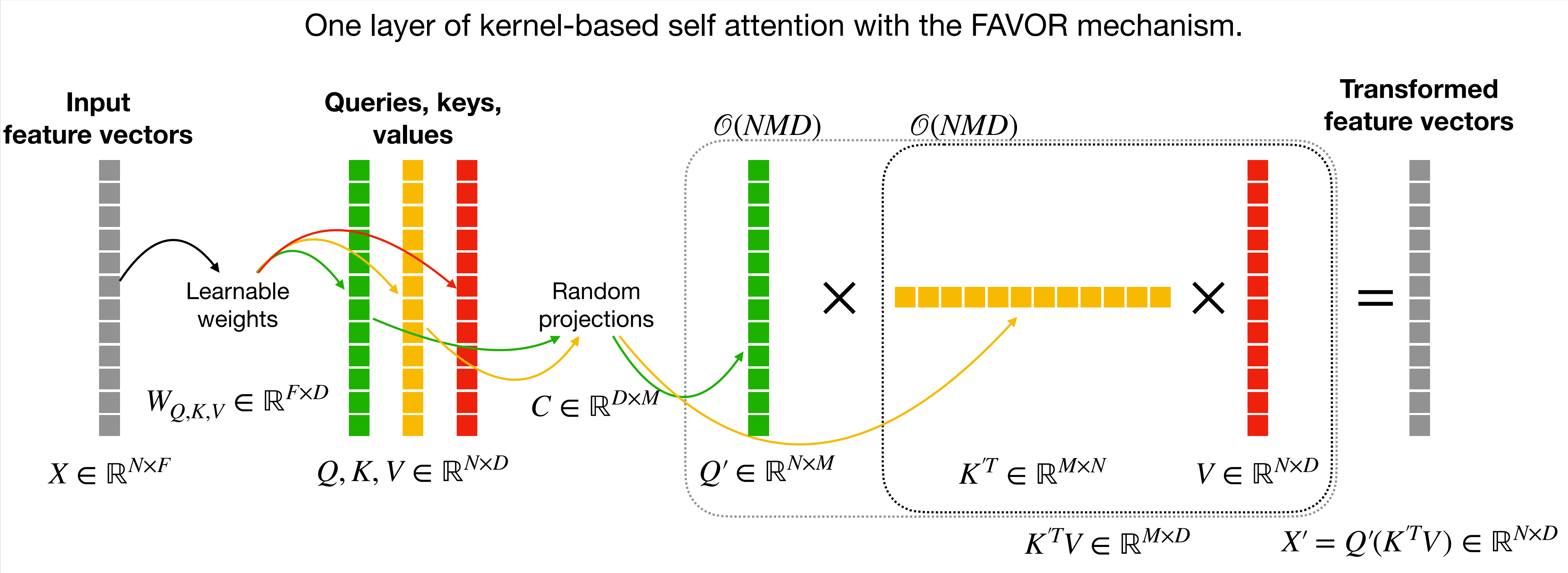
[2] Pata, J., Duarte, J., Vlimant, JR. *et al.* MLPF: efficient machine-learned particle-flow reconstruction using graph neural networks. *Eur. Phys. J. C* **81,** 381 (2021). https://doi.org/10.1140/epjc/s10052-021-09158-w

Graph Neural Network (GNN) with Locality Sensitive Hashing (LSH)

One layer of learnable graph building with locality sensitive hashing and message passing



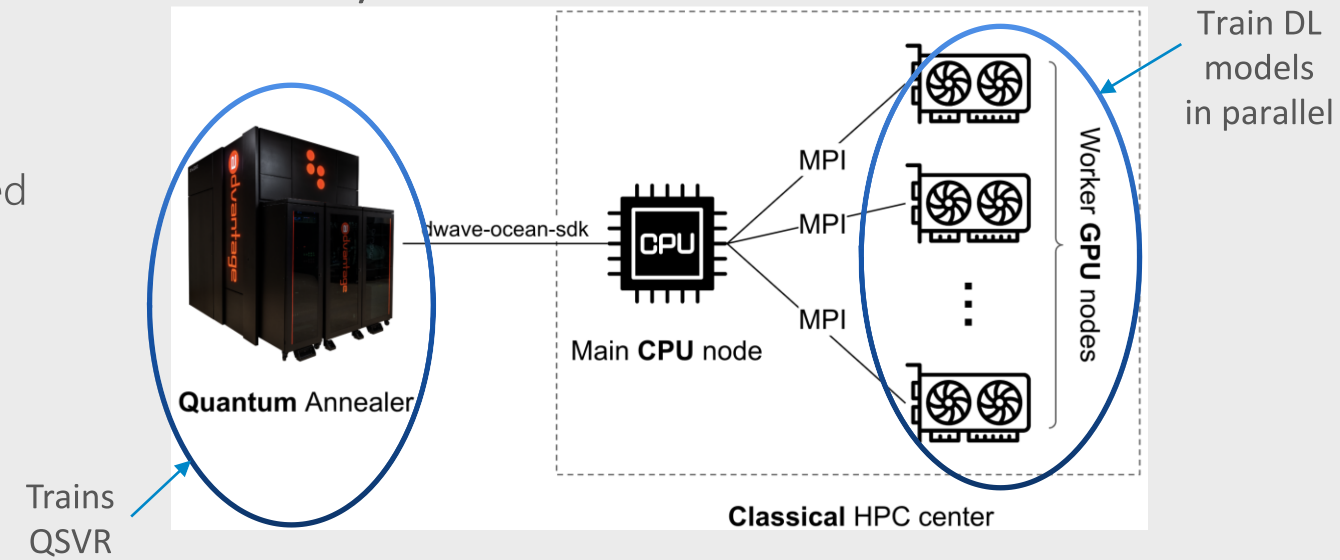
Kernel-based self-attention Transformer



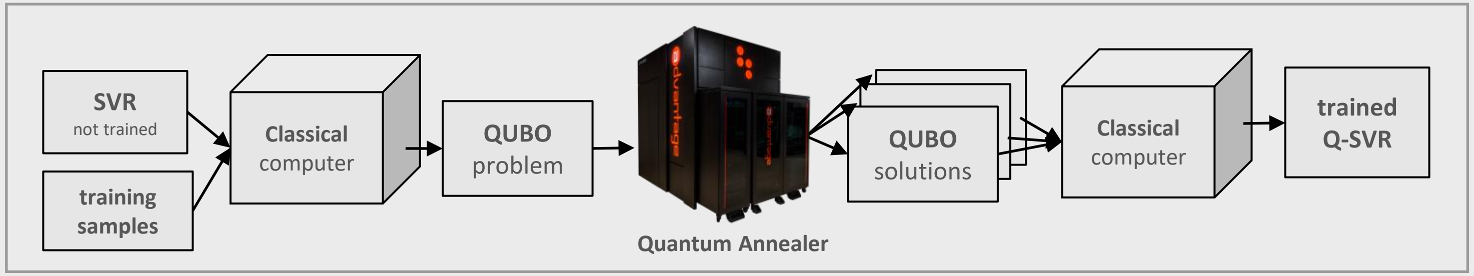
A Hybrid Quantum-Classical workflow for HPO

- Distributed Hybrid Quantum-Classical Model Performance Prediction for Hyperparameter Optimization (HPO) of Deep Learning (DL) Models
- Quantum Annealer (QA) aids classical GPU-accelerated HPC cluster in performing HPO
- GPU cluster trains DL models
- QA trains Quantum-SVR (QSVR) used to aid the HPO process
- Promising results
- This work was shown at QTML at CERN 19th-24th November 2023 and continues the effort based on the following previous works:
    - ACAT 2022, E. Wulff, J.P García Amboage, David Southwick, Maria Girone, Eduard Cuba
    - CHEP 2023, E. Wulff, J.P García Amboage, David Southwick, Maria Girone, Eduard Cuba
    - ISC 2023, M. Aach, E. Wulff, E. Pasetto, A. Delilbasic, R. Sarma, E. Inanc, M. Girone, M. Riedel, A. Lintermann

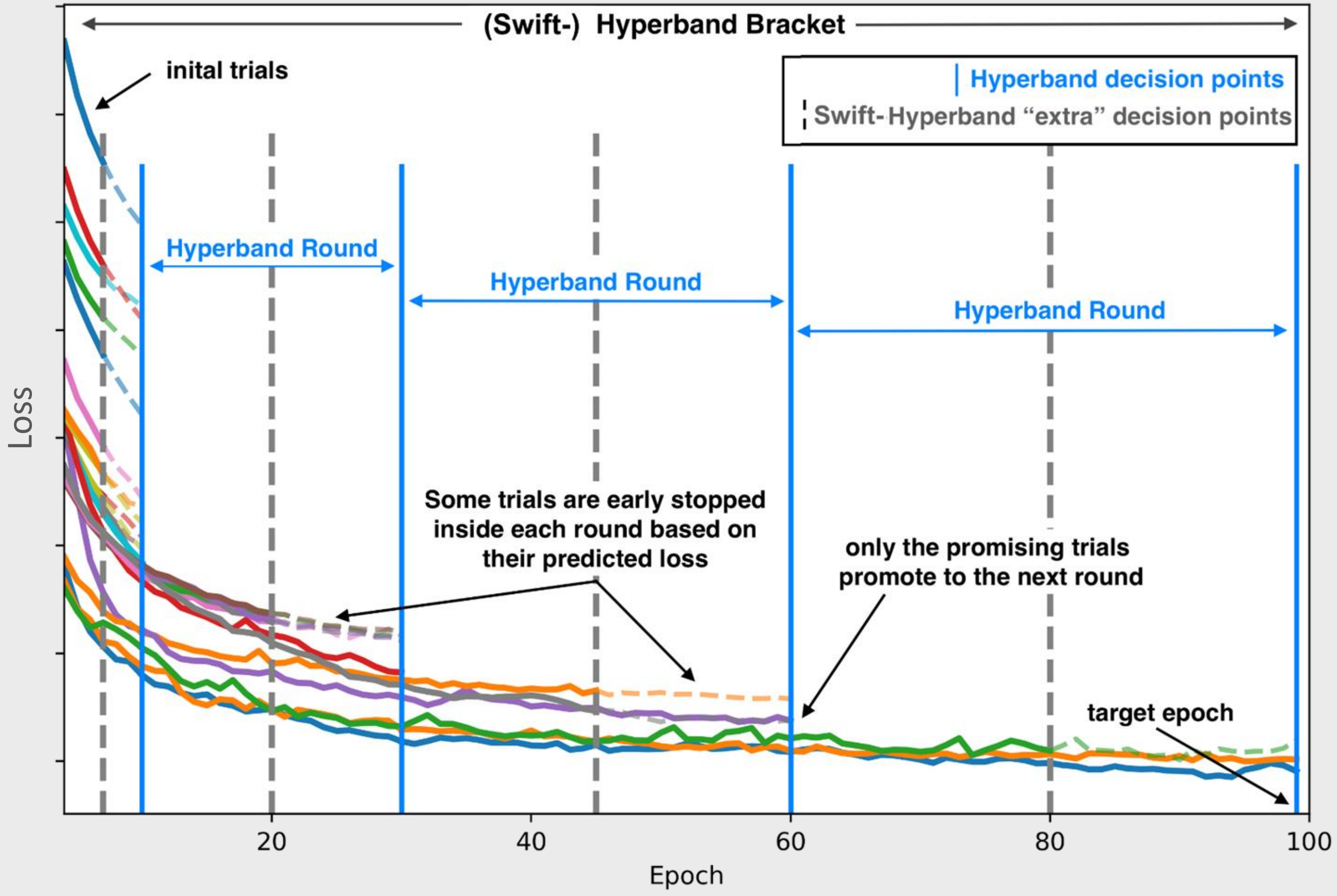
Hybrid Quantum-Classical Workflow



Quantum-SVR training workflow



Swift-Hyperband



- One extra decision point inside each round
- At the beginning of the round some trials are fully trained to define a threshold.
- The other trials are partially trained
- If their predicted loss is lower than the threshold the trials are stopped before completing the round.

Trainings are done in parallel

From SVR to Q-SVR formulation (Pasetto *et al.*)

Classical SVR primal formulation

1

$$\underset{w,b,\xi_i,\xi_i^*}{\text{minimize:}} \quad \frac{1}{2}||w||^2 + C\sum_{i=0}^{N-1}(\xi_i + \xi_i^*)$$

$$y_i - w^\top x_i - b \leq \epsilon + \xi_i^* \quad \forall i \in \{0, ..., N-1\}$$
$$w^\top x_i + b - y_i \leq \epsilon + \xi_i \quad \forall i \in \{0, ..., N-1\}$$
$$\xi_i, \xi_i^* \geq 0 \quad \forall i \in \{0, ..., N-1\}$$

predictions: $y = w^\top x + b$

Classical SVR dual formulation

2

$$\underset{\alpha,\hat{\alpha}}{\text{minimize:}} \quad \frac{1}{2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}(\alpha_n - \hat{\alpha}_n)(\alpha_m - \hat{\alpha}_m)k(x_n, x_m) - \epsilon\sum_{n=0}^{N-1}(\alpha_n + \hat{\alpha}_n) + \sum_{n=0}^{N-1}(\alpha_n - \hat{\alpha})y_n \tag{2.7}$$

$$\sum_{n=0}^{N-1}(\alpha_n - \hat{\alpha}_n) = 0 \tag{2.8}$$

$$0 \leq \alpha_n, \hat{\alpha}_n \leq C \quad \forall n \in \{0, ..., N-1\} \tag{2.9}$$

predictions:

$$y = \sum_{0}^{N-1}(\alpha_n - \hat{\alpha}_n)k(x_n, x_m) + b$$

calculate b:

$$b = y_n - \epsilon - \sum_{m=1}^{N}(\alpha_m - \hat{\alpha}_m)k(x_n, x_m)$$

Resulting problem with binary variables and without restrictions ✓

QUBO formulation

3

$$\underset{a}{\text{minimize:}} \quad f_Q(a) = a^\top Q a = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} Q_{ij}a_i a_j$$

Add restriction as penalty term

$$\xi\left(\sum_{n=0}^{N-1}(\alpha_n - \hat{\alpha}_n)\right)^2$$

Encode SVR variables using binary variables

$$\alpha_n = \sum_{k=0}^{K-1} B^{k-k_0} a_{Kn+k} \quad , \qquad \hat{\alpha}_n = \sum_{k=0}^{K-1} B^{k-k_0} a_{K(N+n)+k}$$

"Ignore" the 2nd restriction

$$\sum_{k=0}^{K-1} B^{k-k_0} \leq C.$$

4

$$\underset{a}{\text{minimize:}} \quad \sum_{n,m}^{N-1}\sum_{i,j=0}^{K-1}\sum_{s,t=0}^{1} a_{K(sN+n)+i}\tilde{Q}_{K(sN+n)+i,K(tN+m+j)}a_{K(tN+m)+j}$$

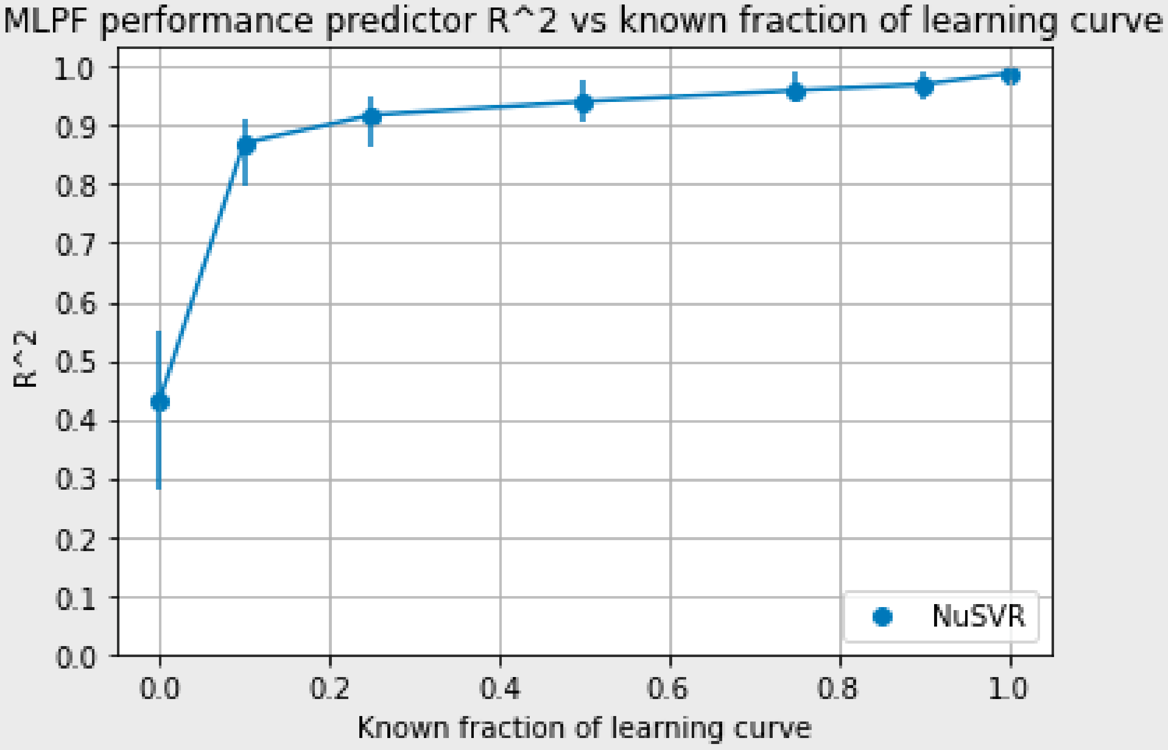
$$\tilde{Q}_{K(sN+n)+i,K(tN+m+j)} = (-1)^{(1-\delta_{st})}B^{i+j-2k_0}\left(\frac{1}{2}k(x_n, x_m) + \xi\right) + \delta_{nm}\delta_{ij}B^{i-k_0}\delta_s t(\epsilon + (-1)^{(1-s)(1-t)}y_n)$$

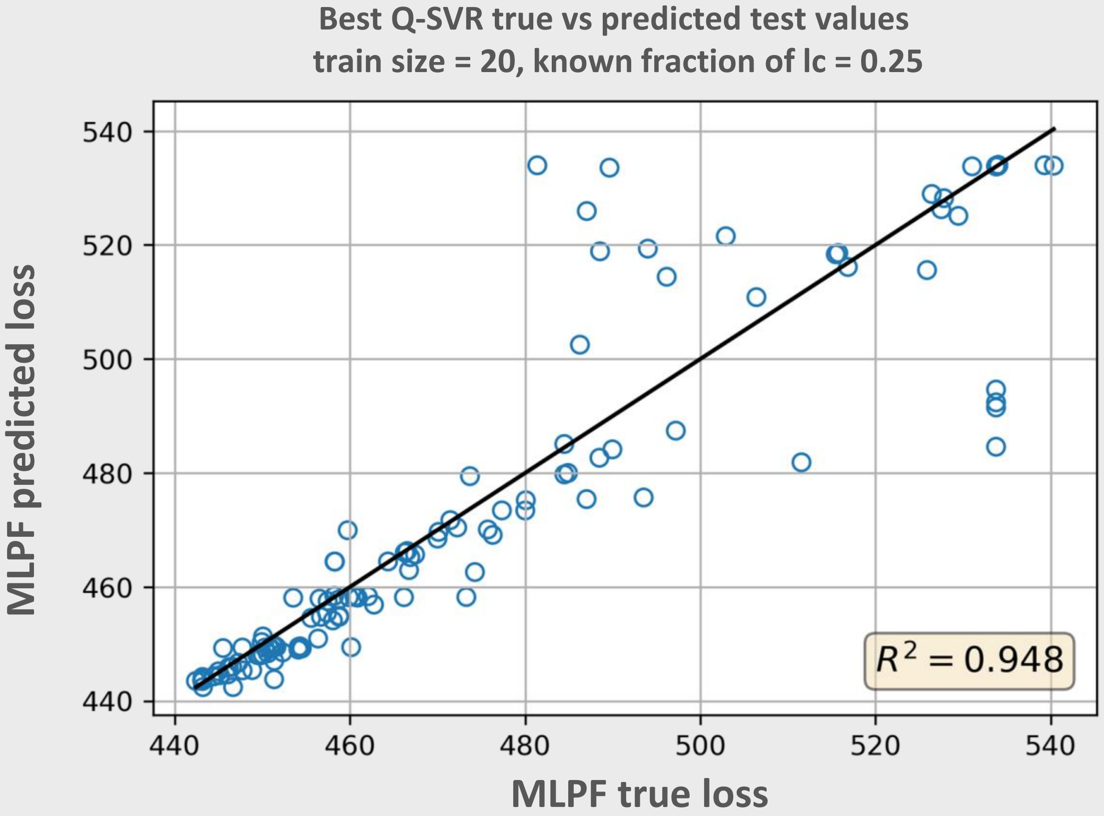
QUBO matrix for the canonical formulation:

$$Q_{ij} = \begin{cases} \tilde{Q}_{ij} + \tilde{Q}_{ji} & \text{si } i < j \\ \tilde{Q}_{ij} & \text{si } i = j \\ 0 & \text{si } i > j \end{cases}$$

Performance prediction of MLPF

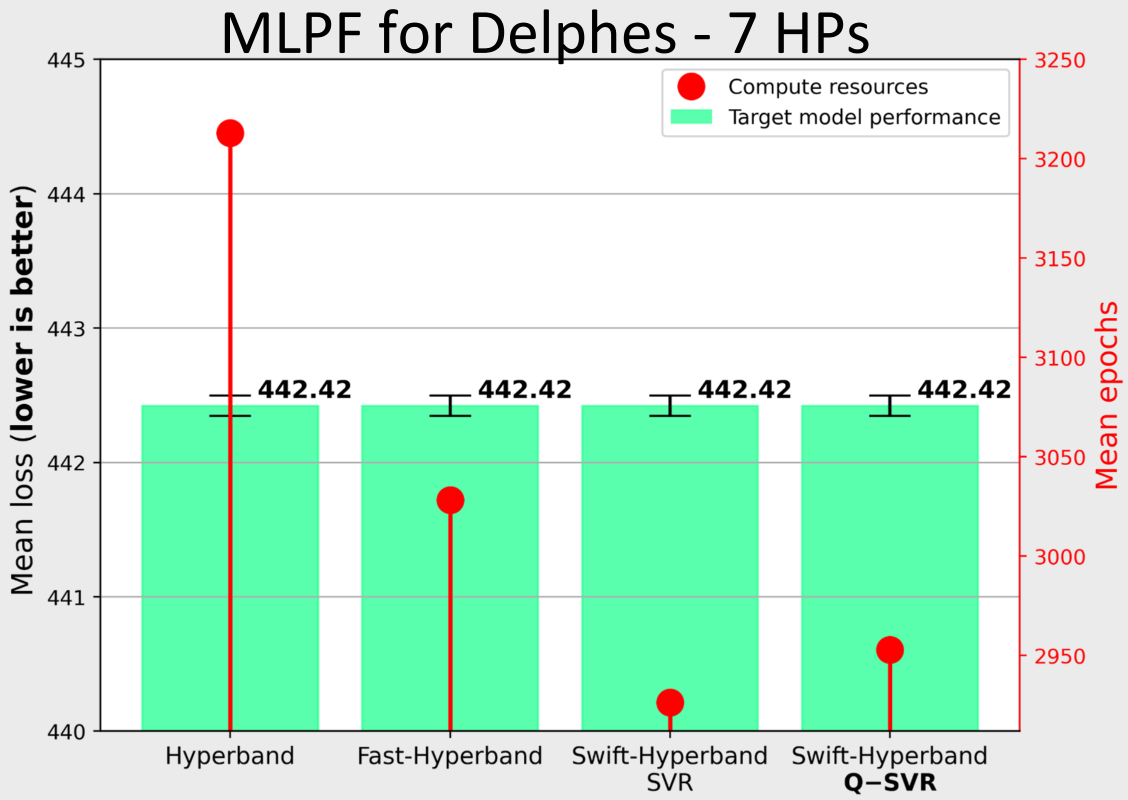
- Very promising results for Q-SVR and SVR.

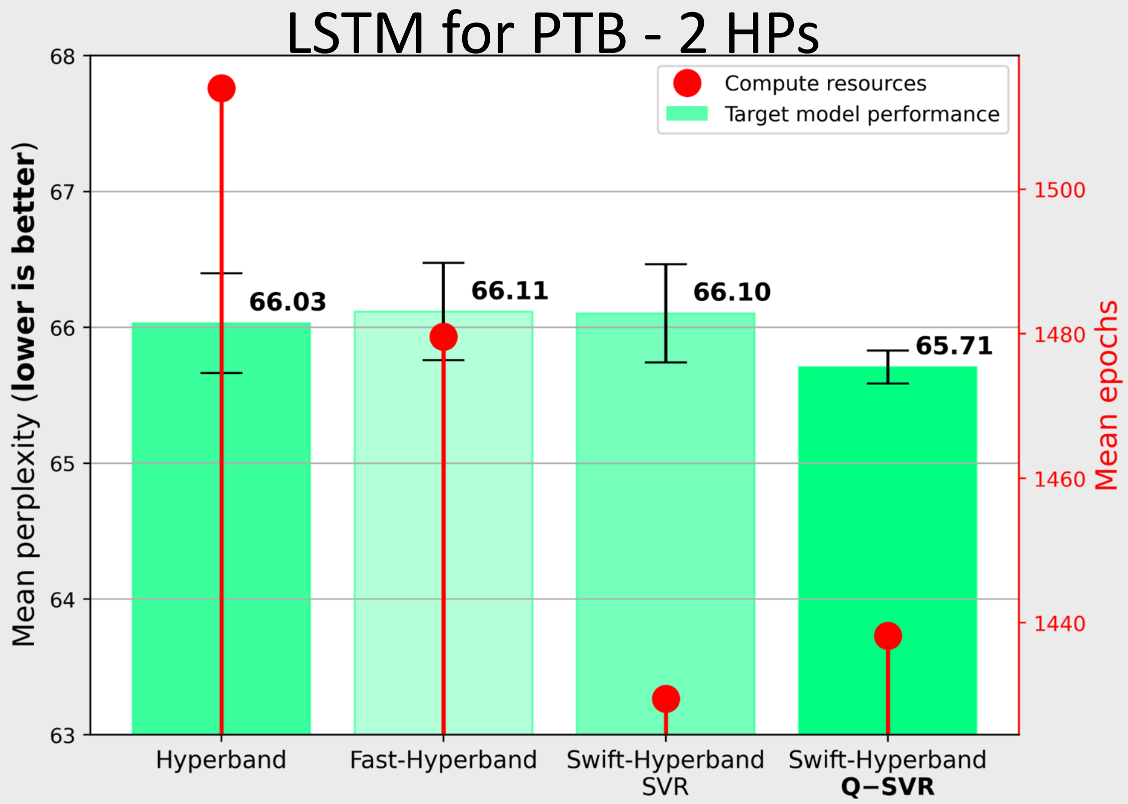




Algorithm Comparison

- Simulated results using learning curve datasets

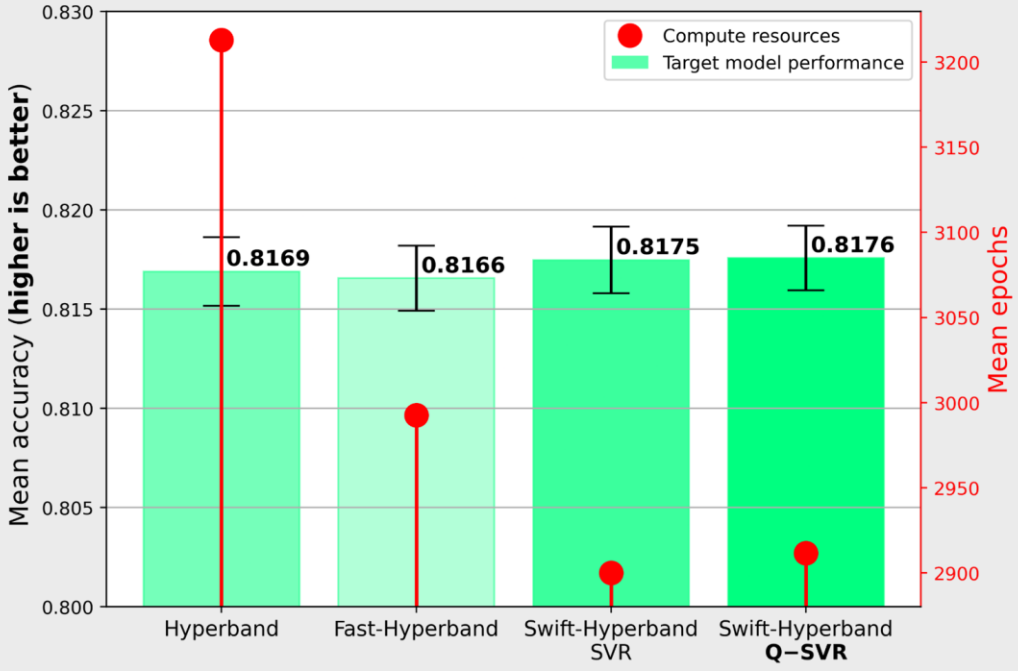




Algorithm Comparison

- Simulated results using learning curve datasets

CNN for CIFAR-10 - 5 HPs



CNN for SVHN - 9 HPs