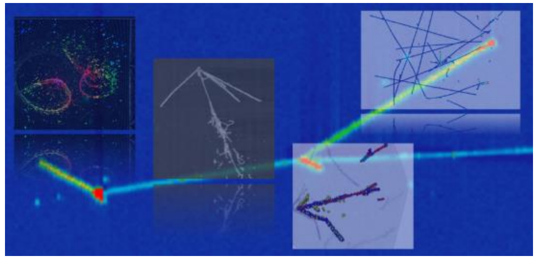
Enhancing Liquid Argon TPCs Performance in Low-Energy Physics Classification Problems with Quantum Machine Learning

Neutrino Physics and Machine Learning 2024

Roberto Moretti
roberto.moretti@mib.infn.it

Milano Bicocca University
INFN – Milano Bicocca

A. Giachero – UniMiB, INFN
F. Terranova – UniMiB, INFN
D. Guffanti – UniMiB, INFN
M. Grossi – CERN QTI



NPML24

Overview

- **LArTPC** sensitivity to $^{136}Xe\ 0\nu\beta\beta$.
- Background mitigation with **Convolutional** and **Transformer** Neural Networks.
- **Quantum Support Vector Machines**.
- Performance analysis.
- **Automatic optimization of Quantum Feature Maps.**
- Towards commercial Quantum Computers usage.
- Conclusions.

DUNE potential at few-MeV

DUNE: Deep Underground Neutrino Experiment
Several physics goals:

High-Energy sector
- Mass hierarchy
- CP violation
- Proton decay

Low-Energy sector
- Supernova neutrinos
- Solar neutrinos
- WIMPs
- $\mathbf{0\nu\beta\beta}$

proposals

$$\Gamma^{0\nu}_{\beta\beta} = \frac{1}{T^{0\nu}_{\beta\beta}} = G|M^{0\nu}|^2 m^2_{\beta\beta} \longrightarrow$$

phase space

nuclear matrix element

Majorana mass

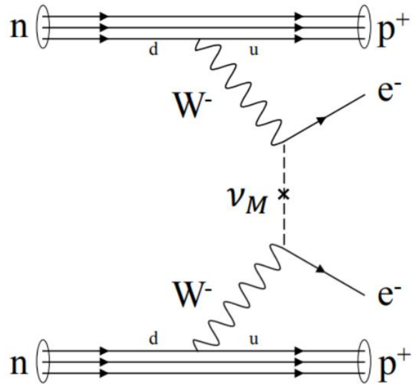
$$m_{\beta\beta} = \sum_i U^2_{ei} m_{\nu i}$$

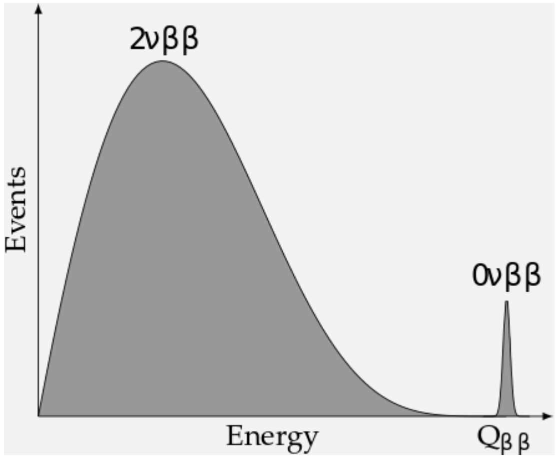
The Neutrinoless double beta decay ($0\nu\beta\beta$)
- Hypothetical BSM process
- Consequences:
  - Neutrinos are Majorana particles.
  - Lepton number is not conserved.

Candidate:

$$^{136}\mathrm{Xe}_{54} \rightarrow {}^{136}\mathrm{Ba}_{56} + 2e^- + \cancel{2\bar{\nu}_e}$$

$$Q^{136Xe}_{\beta\beta} = 2.458\ MeV$$



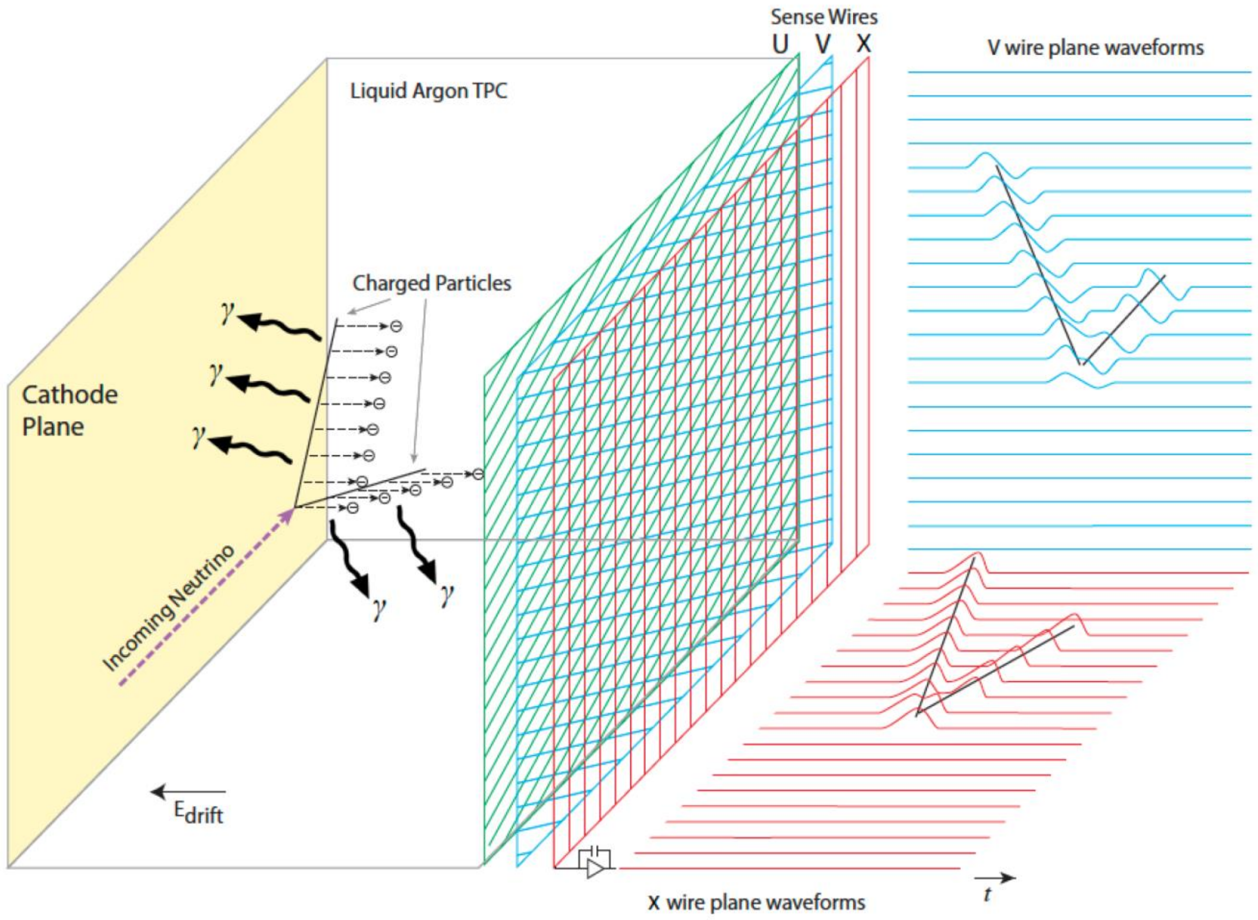


$$T^{0\nu}_{\beta\beta} > 1.07 \cdot 10^{26} y \text{ at } 90\%\ C.L.$$

DUNE LArTPC and track reconstruction

DUNE is composed of a Near Detector (ND) and Far Detector (FD) facilities.

- FD: four modules of 17kton **Liquid Argon Time Projection Chambers (LArTPCs)**.
- **Proposal:** an «*opportunity*» module with argon doped with xenon at 2% concentration for the search of the $^{136}\mathrm{Xe}\ 0\nu\beta\beta$ decay.
- Careful background studies (β, n, solar ν, etc ...)
  β from $^{42}\mathrm{Ar}$ dominates.



Goal: leverage TPC tracking for background mitigation. Challenging tasks at the MeV-scale in FD LArTPCs:

Opportunity to explore **Quantum Machine Learning** models (QML).

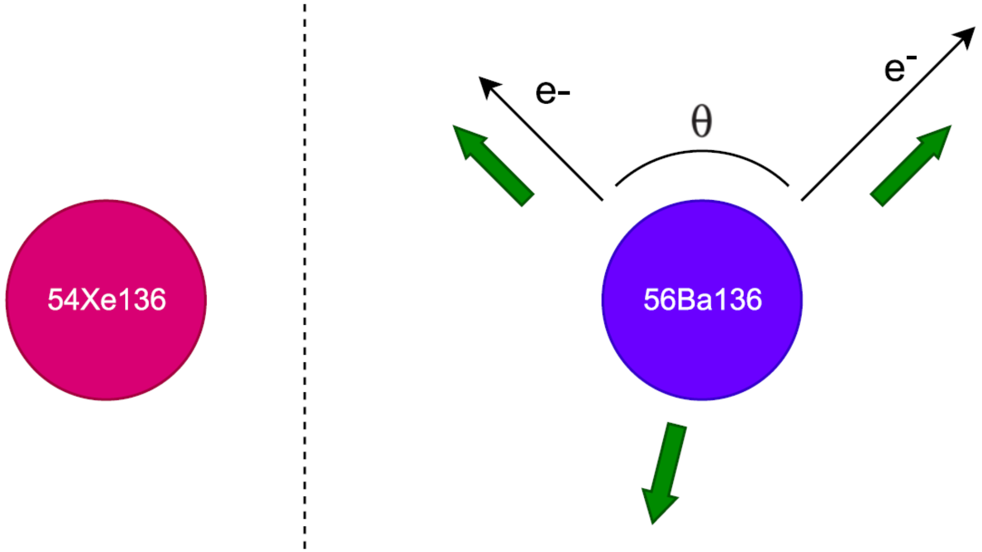
Topology-based classification

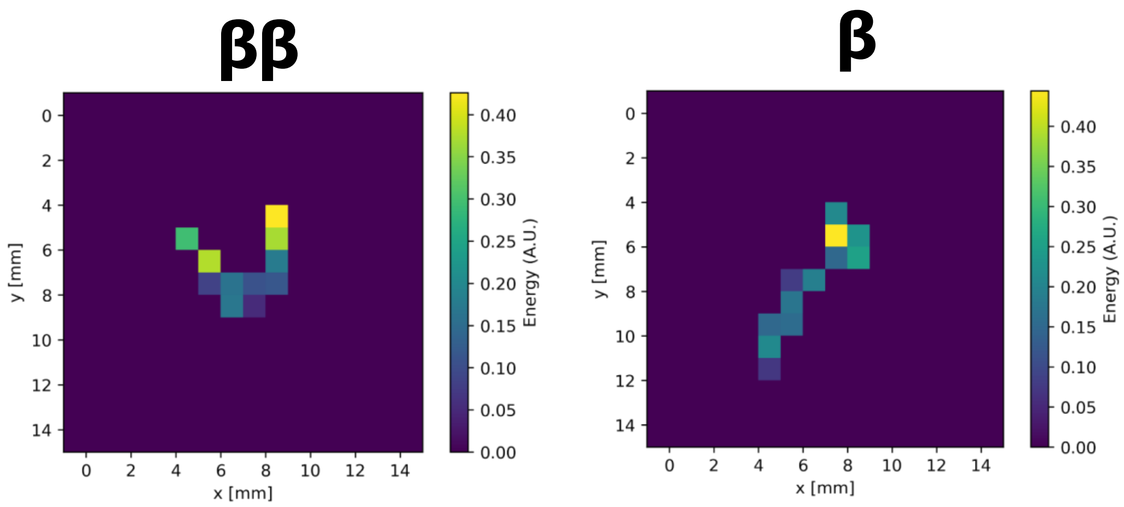
$\beta\beta$ topologies (signal): two electrons originating from the same position in space.

β topology (background): one electron with an energy close to $Q_{\beta\beta}^{^{136}Xe} = 2.458\ \mathrm{MeV}$

Energy-angle distribution for $\beta\beta$:

$$\boxed{\frac{d\Gamma^{0\nu}_{\beta\beta}}{dE_1 d\cos\theta} \propto \frac{1}{16\pi^5} F(E_1, Z) F(E_2, Z) dE_1 d\cos\theta E_1 E_2 p_1 p_2 \left(1 - \frac{\vec{p}_1 \cdot \vec{p}_2}{E_1 E_2}\right)}$$





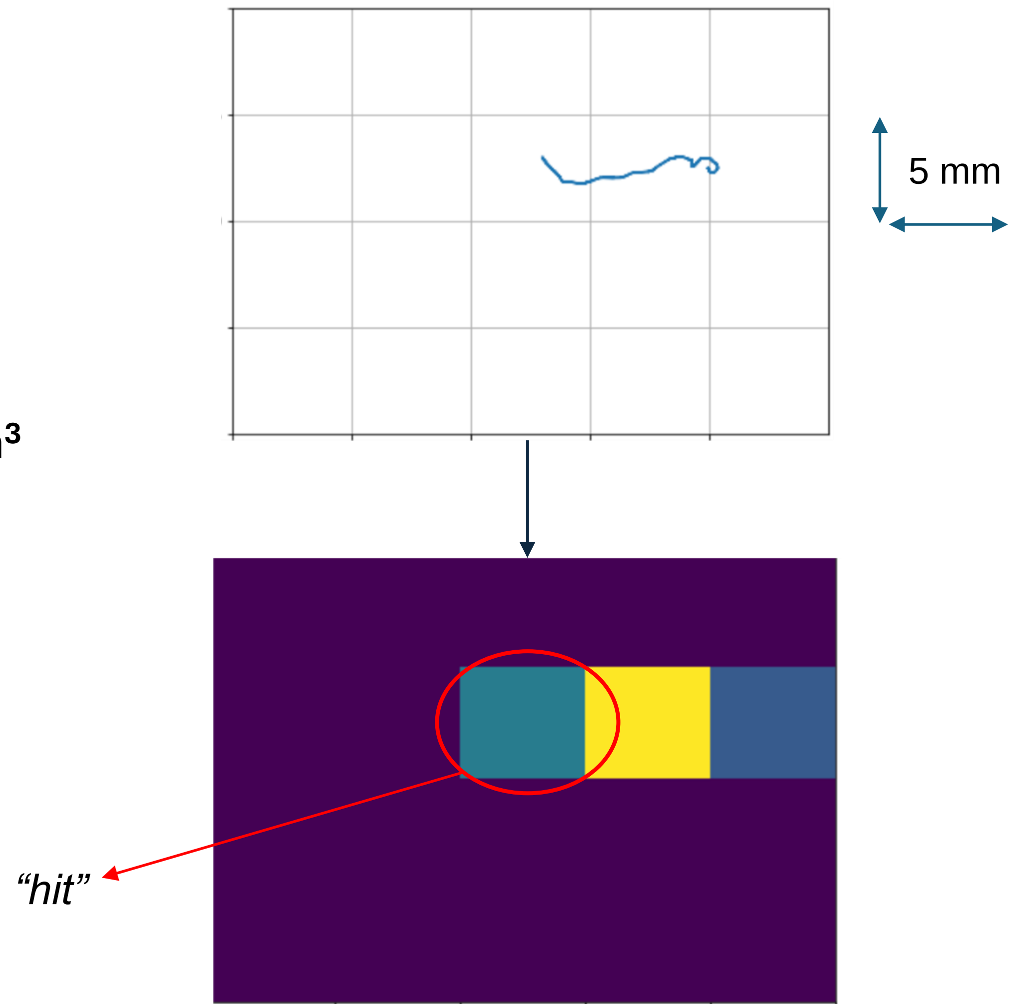
“Toy” dataset

Dataset:

- Geant4 propagated high-resolution **β** and **ββ** tracks in LAr at $E = Q_{\beta\beta}^{136Xe} = 2.458\ \mathrm{MeV}$.

- Tracks have been downsampled to **3D voxelized data** with variable detector granularity (bin-widths) of $[\boldsymbol{w} \times \boldsymbol{w} \times \mathbf{1}]$ **mm³** to simulate a DUNE-like granularity (or better).
- **Variable Energy threshold** from $10\ \mathrm{keV}$ to $200\ \mathrm{keV}$.
- Diffusion and recombination of ions and electrons are taken into account.

- Other detector effects were not considered.



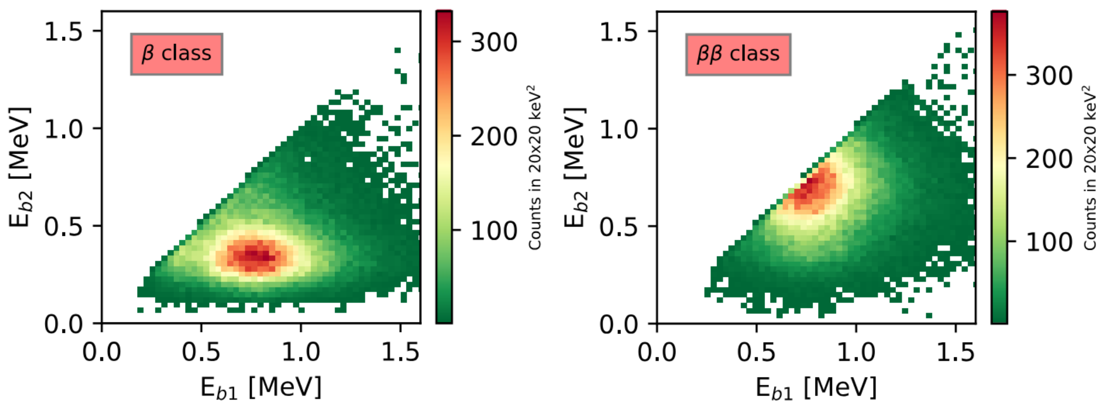
(XY) view (5 x 5) mm²

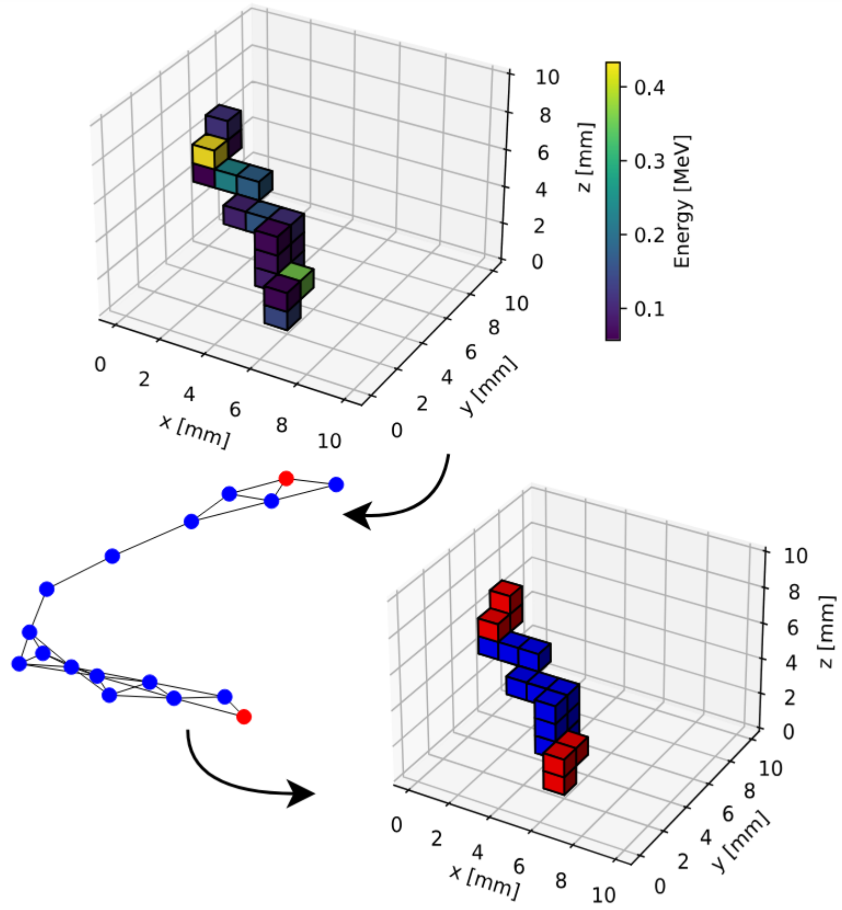
Classical approach: Blob

- **Graph representation**
  - Every hit is a node.
  - Nodes are connected if corresponding hits are neighbours.
- **Breadth-first search (BFS) algorithm**
  - Finds the «farthest» node pair
  - We expect to have a blob centroid there
- Compute the blob energies by **integrating within a radius**

Pros:
- physics-informed
- easy implementation
- deterministic

Cons:
- does not use all track infomation,
- can't handle track discontinuities





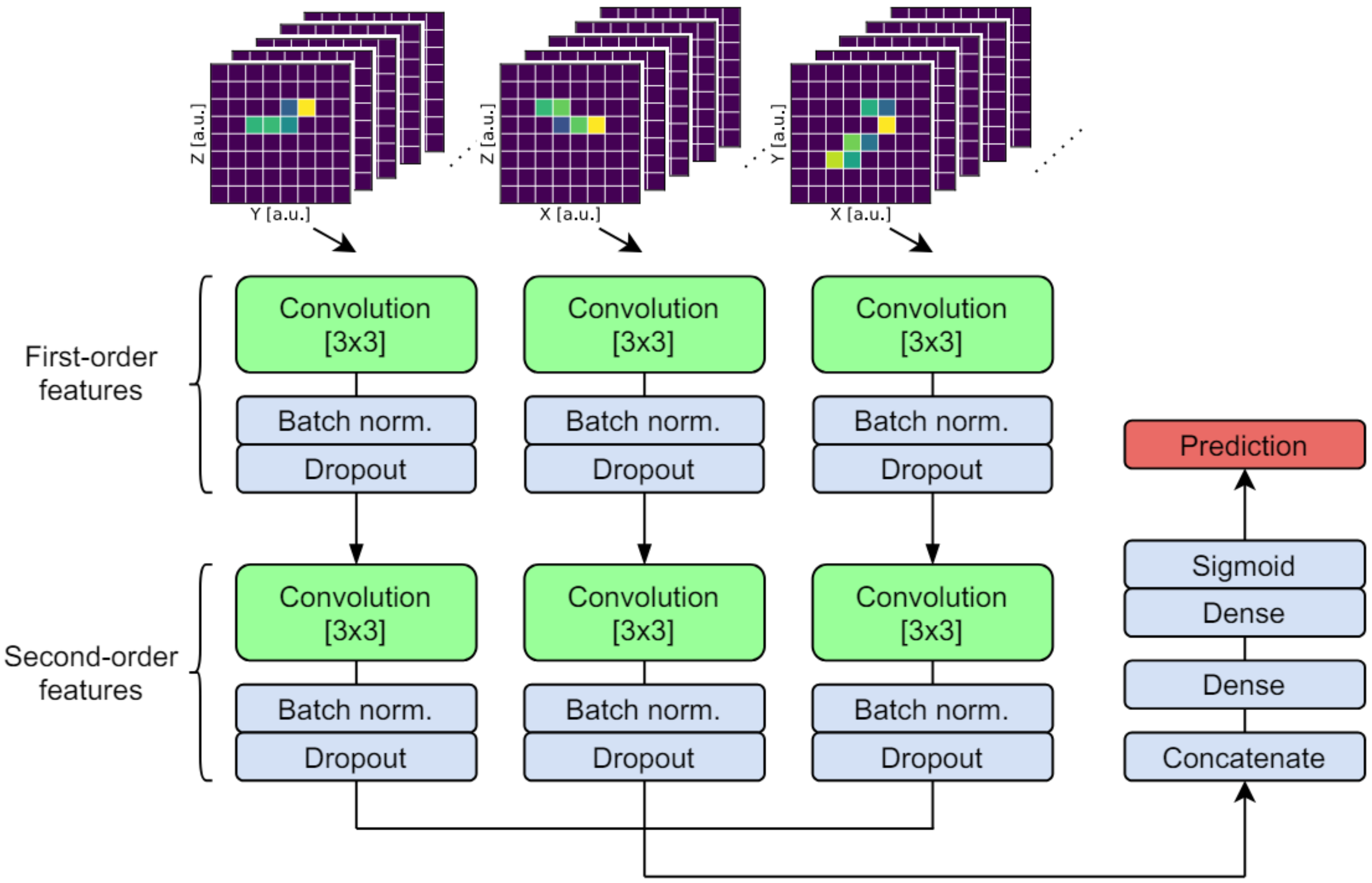
R. Moretti et al. (2024)
EPJP *10.1140/epjp/s13360-024-05287-9*
E-print: arXiv:2305.09744v2

DL approach: CNN

- Feed parallel convolutional branches with three planar track projections.

Pros:
- captures complicated track features.

Cons:
- can become memory-inefficient, especially at high resolution.
- By projecting in 2D, some information is lost.



DL approach: Transformer

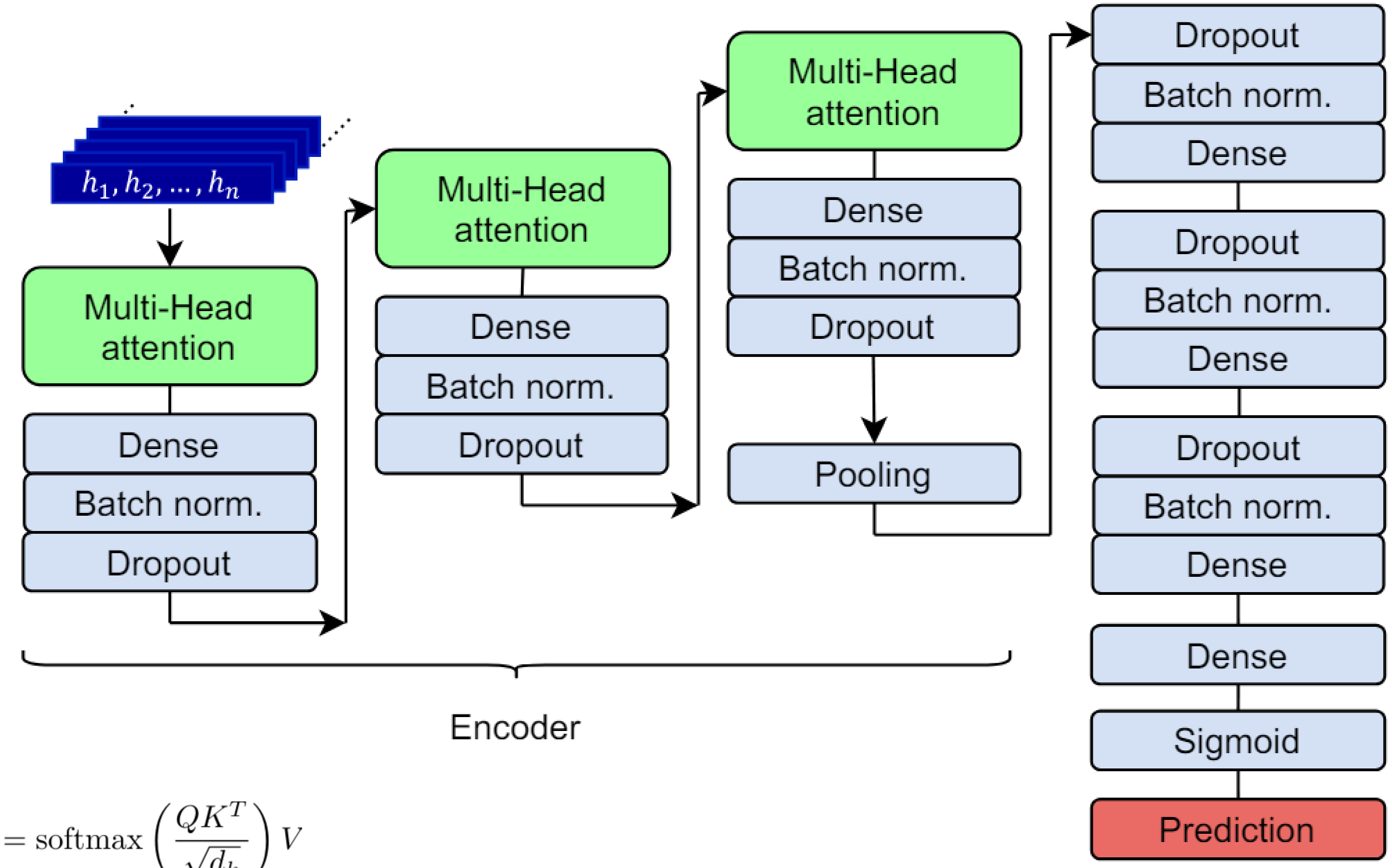
- Feed tracks as lists of hit energies and spatial coordinates.
- Only the «Encoder» part of a typical Transformer is used.

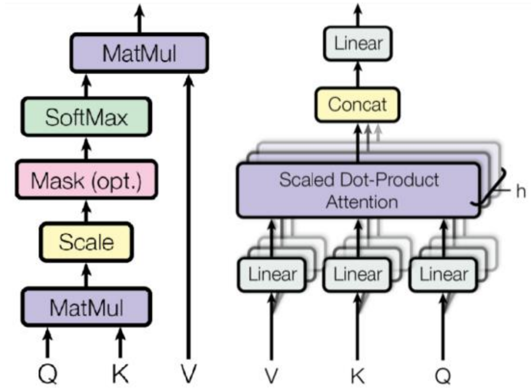
Pros:

- memory-efficient, uses the full track information.

Cons:

- harder to interpret, more complex structure.



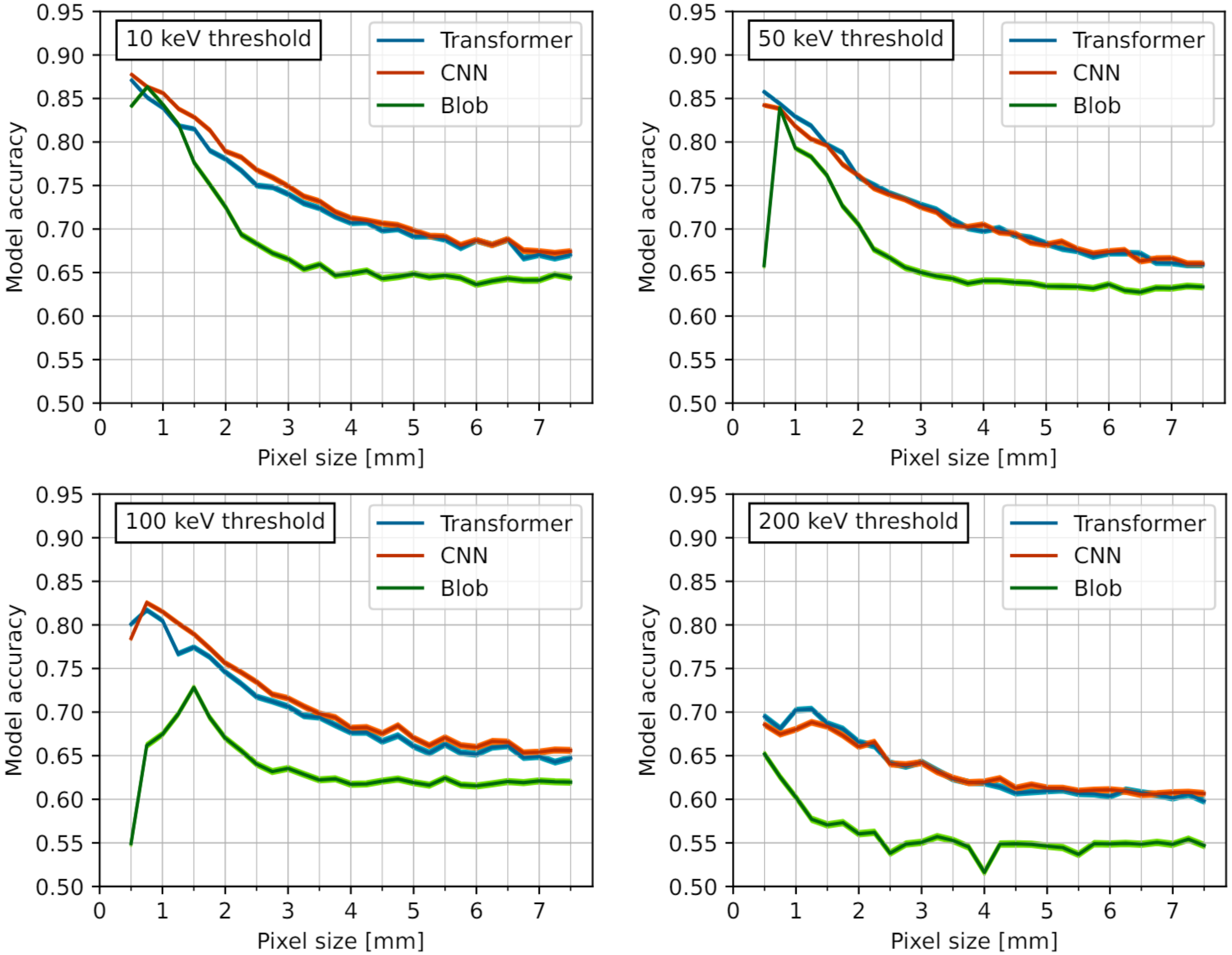


$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Classification comparison

- We trained each model for several pixel size (detector granularity) and hit-energy threshold.
- **Neural Networks outperforms Blob** in almost any configuration.
- No decisive «winner» between CNN and Transformer.

Training size: 140×10^3
Validation size: 30×10^3
Test set size: 30×10^3



Quantum Computing – in theory

A qubit is a **2-level quantum system** described by the wavefunction:

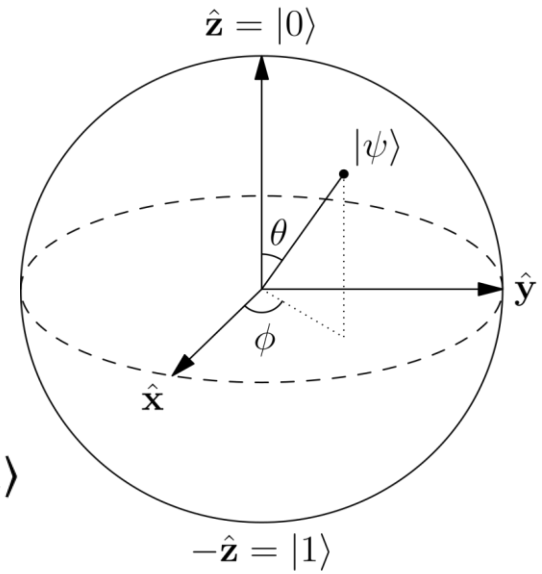
$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$

$$|\alpha|^2 + |\beta|^2 = 1 \qquad \alpha; \beta \in \mathbb{C}$$

- Fundamental unit of quantum computation.
- $|0\rangle$ and $|1\rangle$ are the two computational basis, in analogy with 0 and 1 of classical computing.

Qubit states can be visualized as points on a sphere's surface.

Bloch Sphere representation



$$|\psi\rangle = \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\phi}\sin\left(\frac{\theta}{2}\right)|1\rangle$$

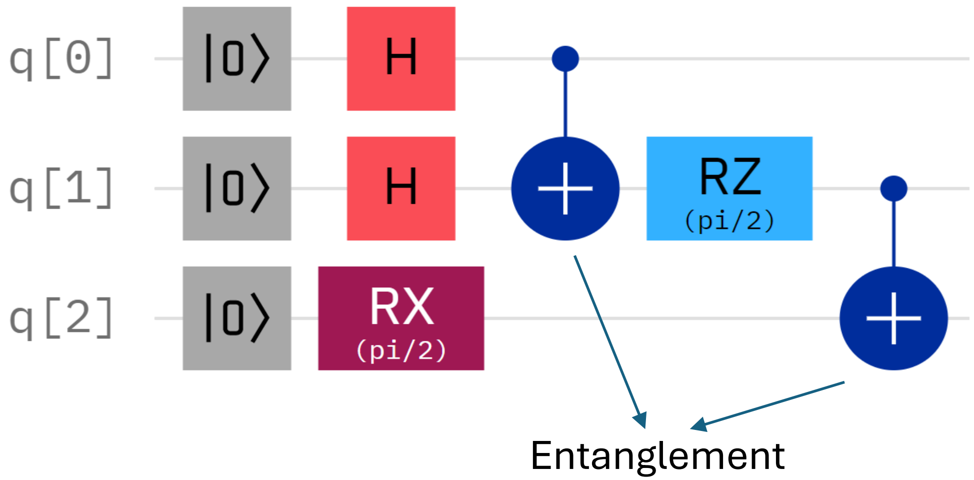
Qubits are controlled by unitary operators called **quantum gates**, organized in **quantum circuits**.

H $\quad H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \quad H|1\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$ Hadamard

$CX|\psi_0\rangle|\psi_1\rangle = |\psi_0\rangle|\,\psi_0 \oplus \psi_1\rangle$ CNOT

RZ RX RY $\quad R_i(\theta) = e^{-\frac{i\theta}{2}\vec{\sigma}_i}$ Pauli rotations

⋮



Support Vector Machine

- Well-known Machine Learning model suited for binary and multilabel classification.
- Useful for signal/background discrimination.

Task: binary classifications of feature vectors $\vec{\boldsymbol{x}} \in \mathbb{R}^{\boldsymbol{n}}$
i.e. predicting the class outcome $y \in \{-1;\ +1\}$.

Idea: given a **feature map** $\phi(\vec{x}),\ \phi(\vec{x}_i) \in M{:}\, \dim(M) = m > n$,
finding the best linear decision boundary $\vec{w}^{\boldsymbol{T}} \boldsymbol{\phi}(\vec{\boldsymbol{x}}) - \boldsymbol{b} = \boldsymbol{0}$
by maximizing:

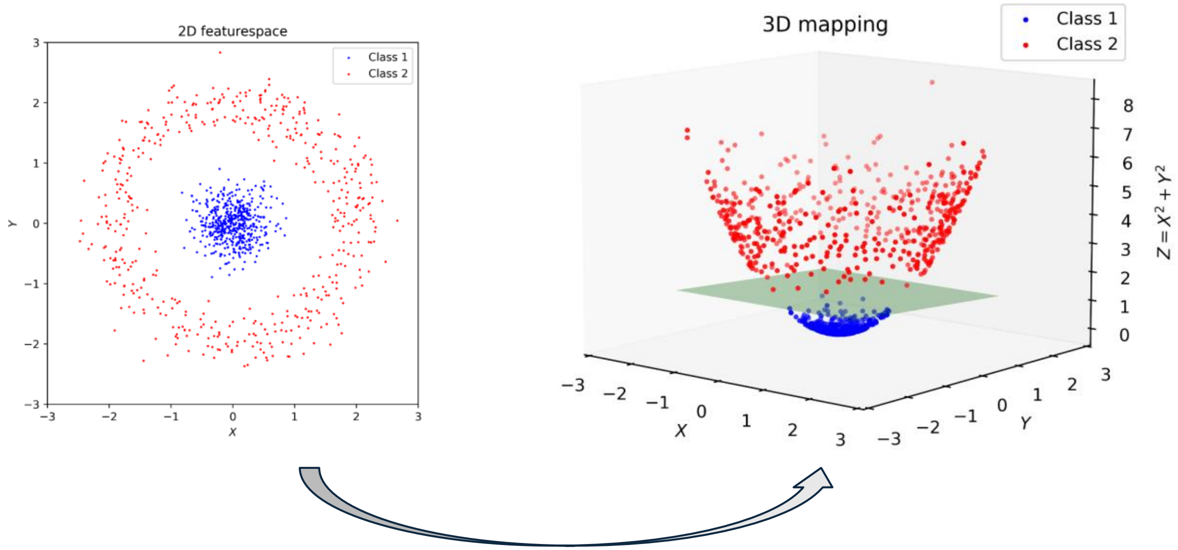
$$f(c_1, c_2, \dots, c_n) = \sum_i c_i - \frac{1}{2} \sum_{ij} y_i c_i y_j c_j \underbrace{\langle \phi(\vec{x}_i), \phi(\vec{x}_j) \rangle}_{\boldsymbol{k}(\vec{\boldsymbol{x}}_{\boldsymbol{i}}, \vec{\boldsymbol{x}}_{\boldsymbol{j}})}$$

with $\vec{w} = \sum_i \boldsymbol{c_i y_i \phi}(\vec{\boldsymbol{x}}_{\boldsymbol{i}})$.

$\downarrow$

Kernel function

When projecting on the original feature space, the decision boundary will be generally nonlinear.



Feature embedding

Common kernel choices:

Linear	$K(\vec{x}_i\ ,\ \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j$
Polynomial	$K(\vec{x}_i, \vec{x}_j) = (\gamma\, \vec{x}_i \cdot \vec{x}_j + r)^d$
RBF	$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\gamma \|\vec{x}_i - \vec{x}_j\|^2 + C\right)$

Quantum Support Vector Machine

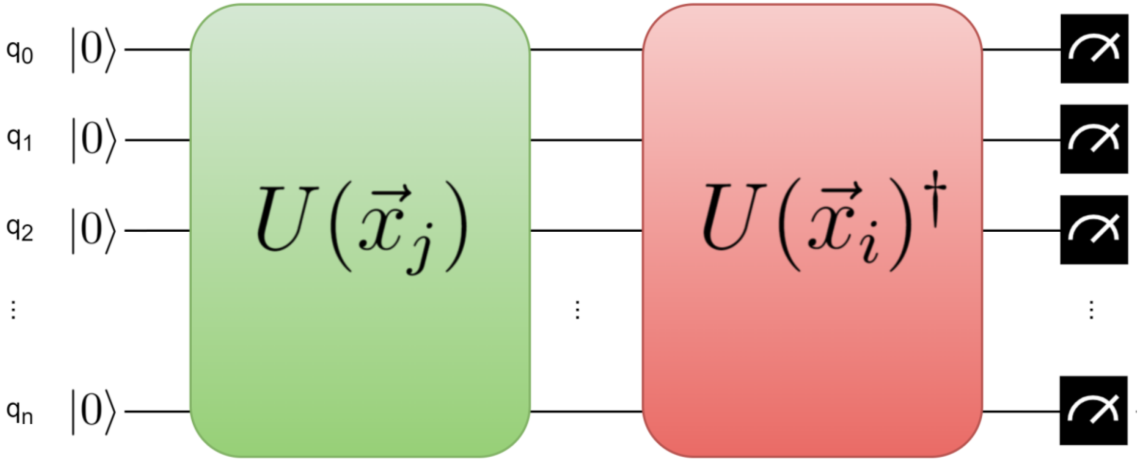
Promoting the classical feature mapping to a quantum state:

$$\phi(\vec{x}) \rightarrow |\phi(\vec{x})\rangle\langle\phi(\vec{x})| =$$

$$= U(\vec{x})|0\rangle\langle 0|U(\vec{x})^\dagger$$

$$K(\vec{x}_i, \vec{x}_j) = \left|\langle 0|U(\vec{x}_i)^\dagger U(\vec{x}_j)|0\rangle\right|^2$$

- Feature maps are still implicitly defined.
- Kernel function is still a measure of similarity between different samples.



Quantum circuits of with this structure are suitable kernels.

Pros:

- Hilbert space grows rapidly with qubit's number
    - **Expressive classifiers.**
- Quantum kernels are generally hard to compute classically
    - **No classical counterpart.**
- Good results even with small sized circuits
    - **Is a NISQ-era algorithm.**

Cons:

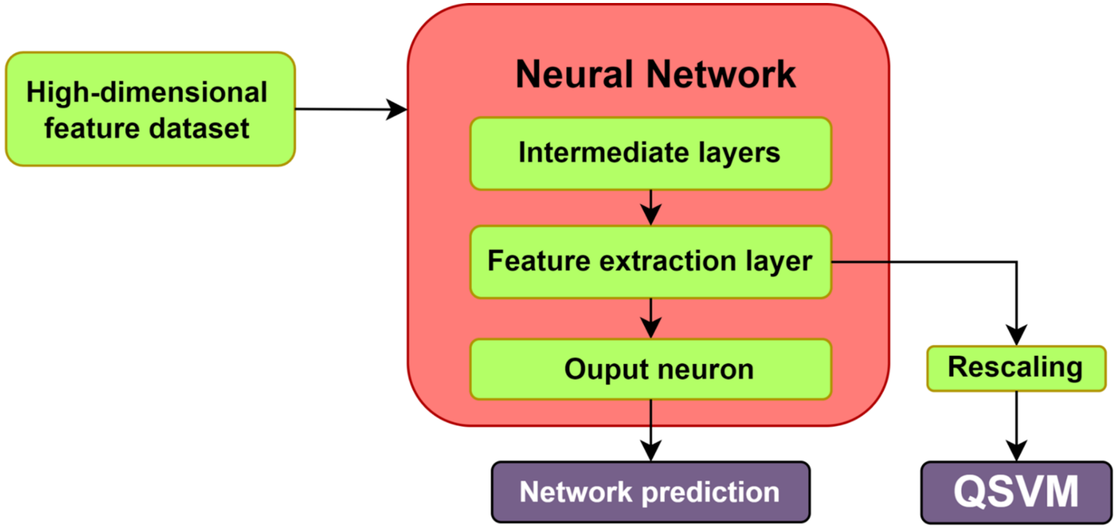
- Lack of featuremap explainability
    - Unintuitive relation between circuit and outcome.
- Usually set arbitrarily
    - Problem of chosing a good Quantum Kernel.

Room for quantum advantage.

Hybrid model – 2 qubits

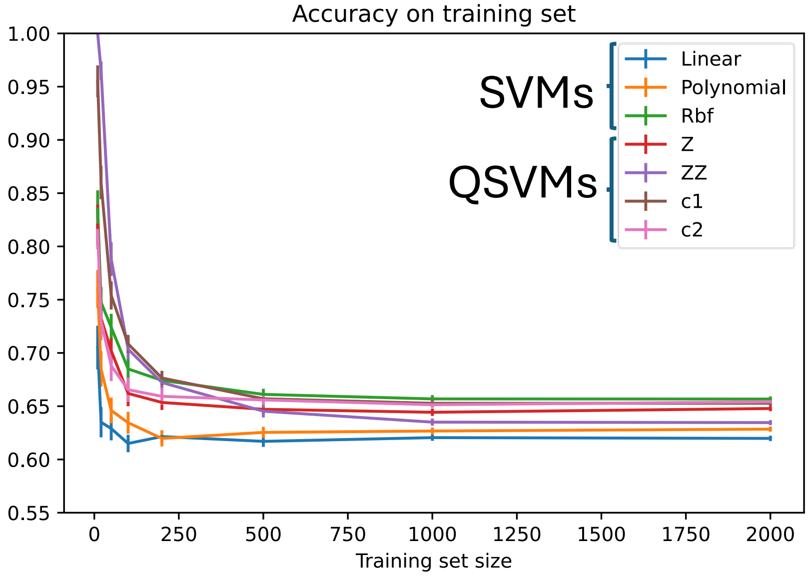
For implementing the NISQ Quantum Support Vector Machine (QSVM) with LArTPC measurements, the input features must be reduced, while maintaining useful informative content.

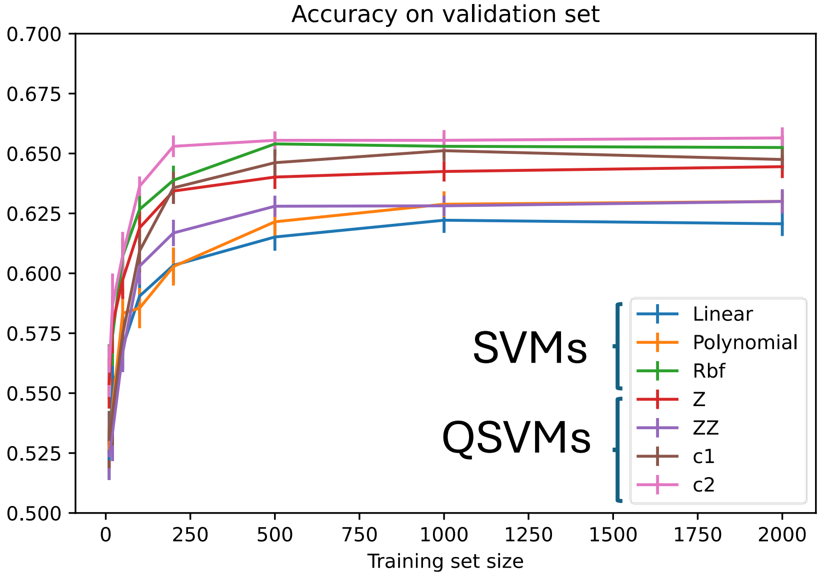
Proposed approach: training Neural Networks as *standalone* classifiers, while defining specific *feature extraction layers* for the QSVM input.



Simulating a QSVM is computationally expensive, and complexity still scales as $O(n^2)$

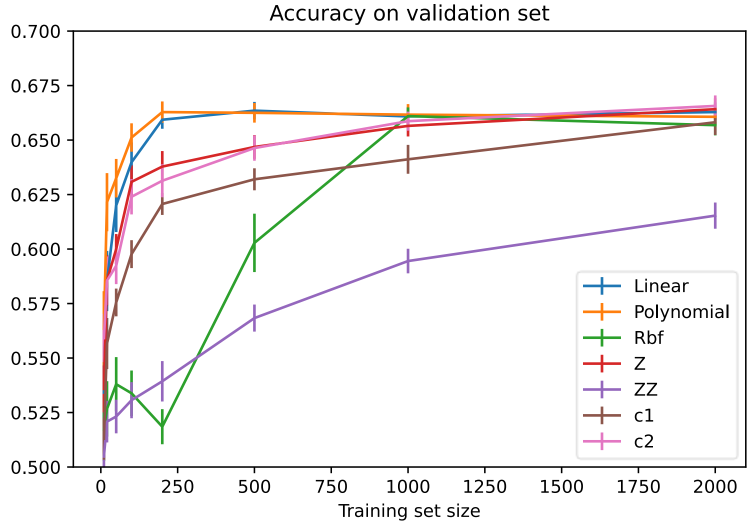
- Limited training set size.
- Still good results, comparable with SVM using Rbf.



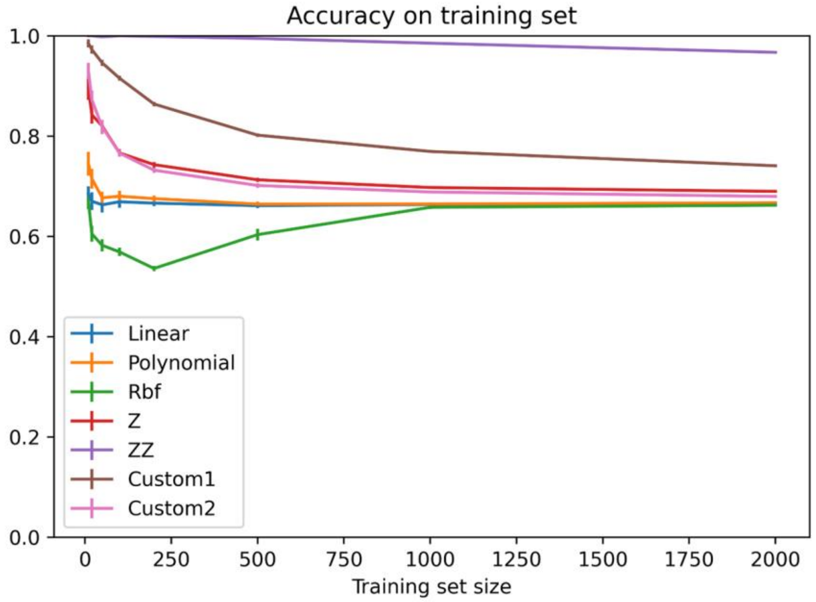


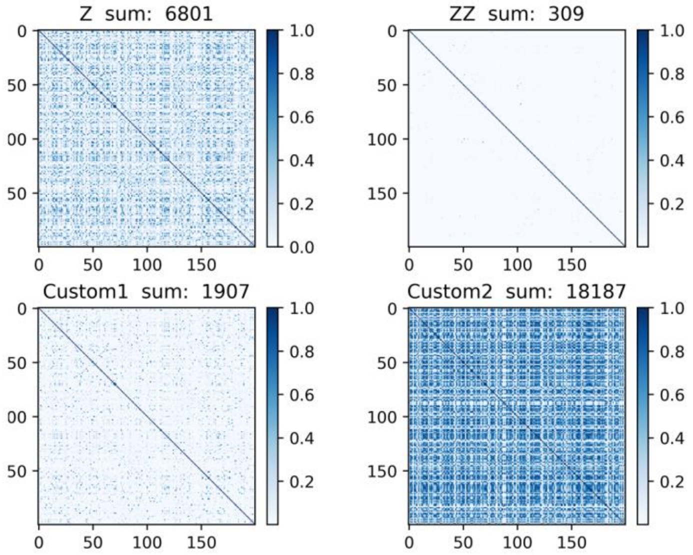
Hybrid model – more qubits

- Increasing the qubit number up to 10 leads to **overfitting** with quantum kernels that use entangling gates.
- Strong correlation between amount of overfitting and kernel density.



Low kernel density also implies that noise will affect more the QSVM outcome:
→ hard to run on NISQ hardware.

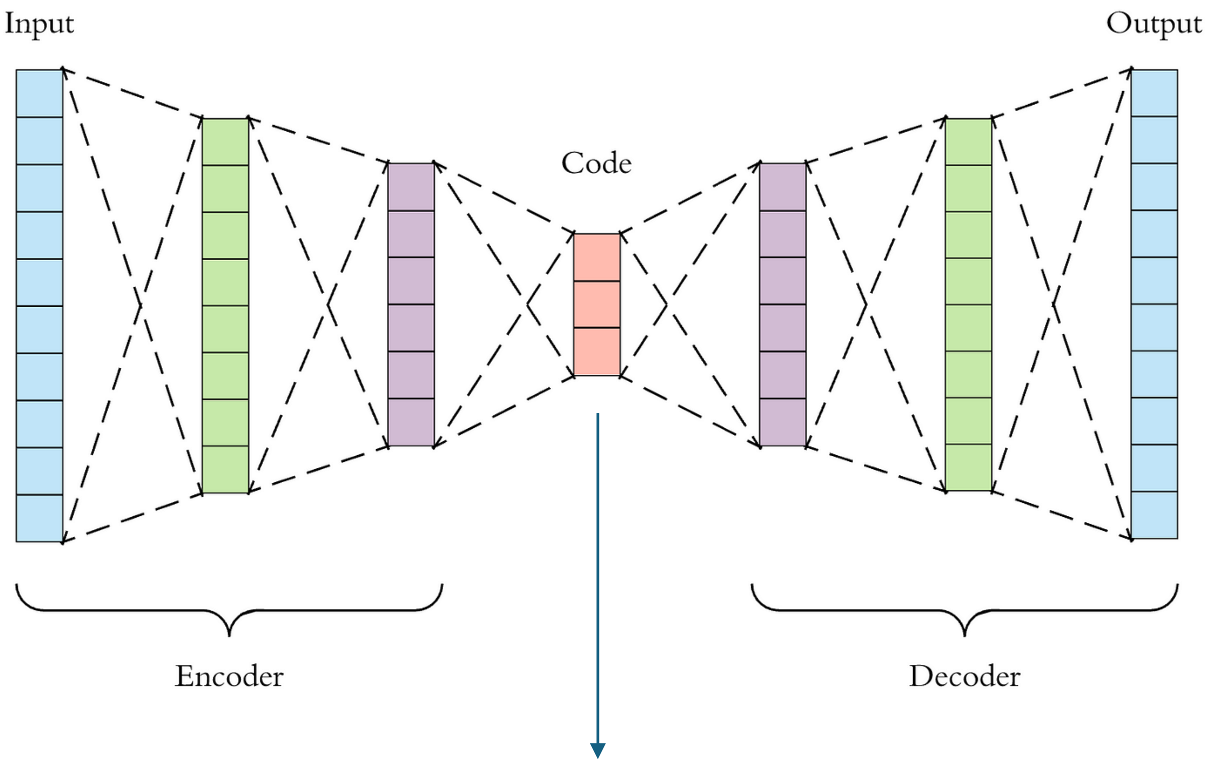




Autoencoder as feature extractor

We used a feature reduction algorithm that is completely agnostic to the classification task (labels), i.e. an **autoencoder:**

- Stack of **feed-forward layers** divided into an Encoding and a Decoding part.
- Input and output should match as closely as possible.
- The hidden layer that produces the reduced feature distribution is called «Bottleneck».
- Training cost function minimizes the information loss (Mean/Absolute Square Error).

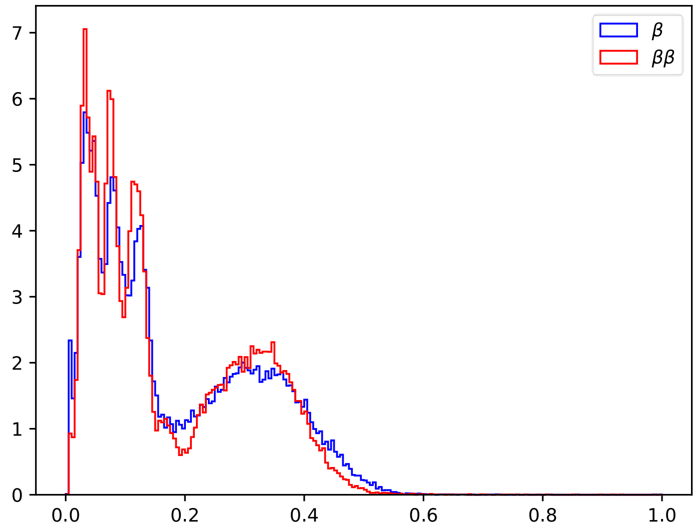


Reduced features
76 -> 18

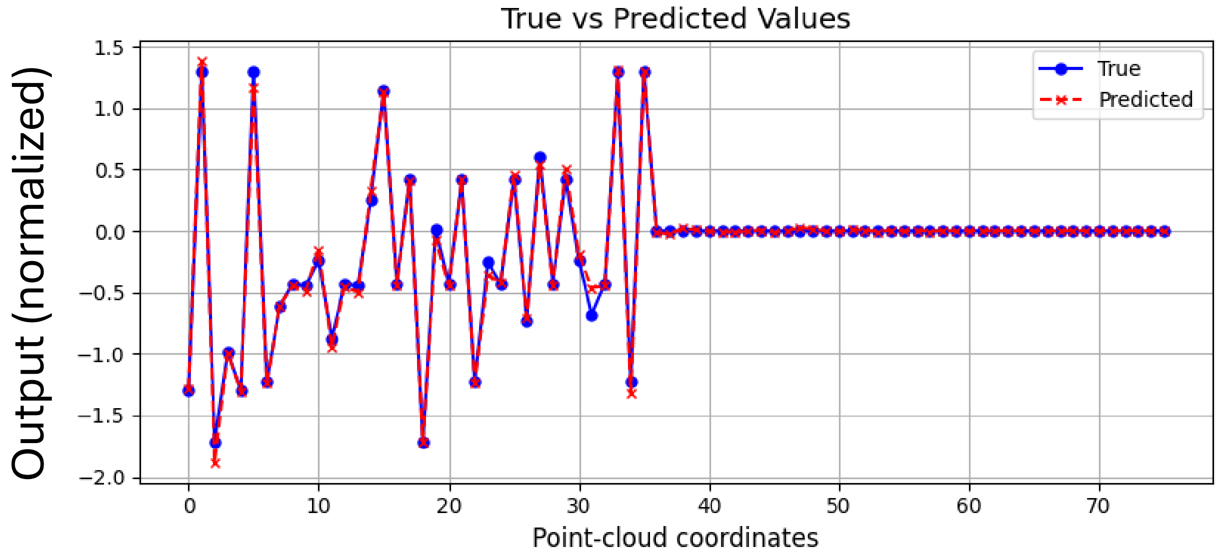
Feature distributions

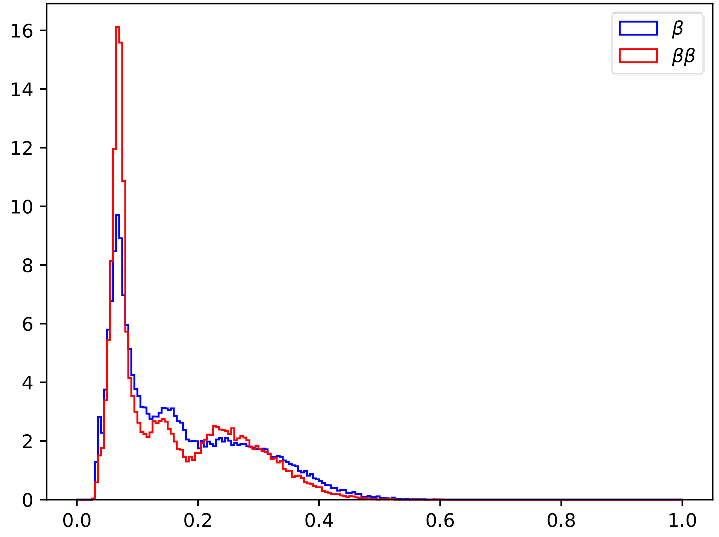
Features are not well-separated for $\beta/\beta\beta$ classes and not all of them have «gaussian-like» distributions.

We expect low accuracy overall, but it is a viable testbed for the genetic optimization.

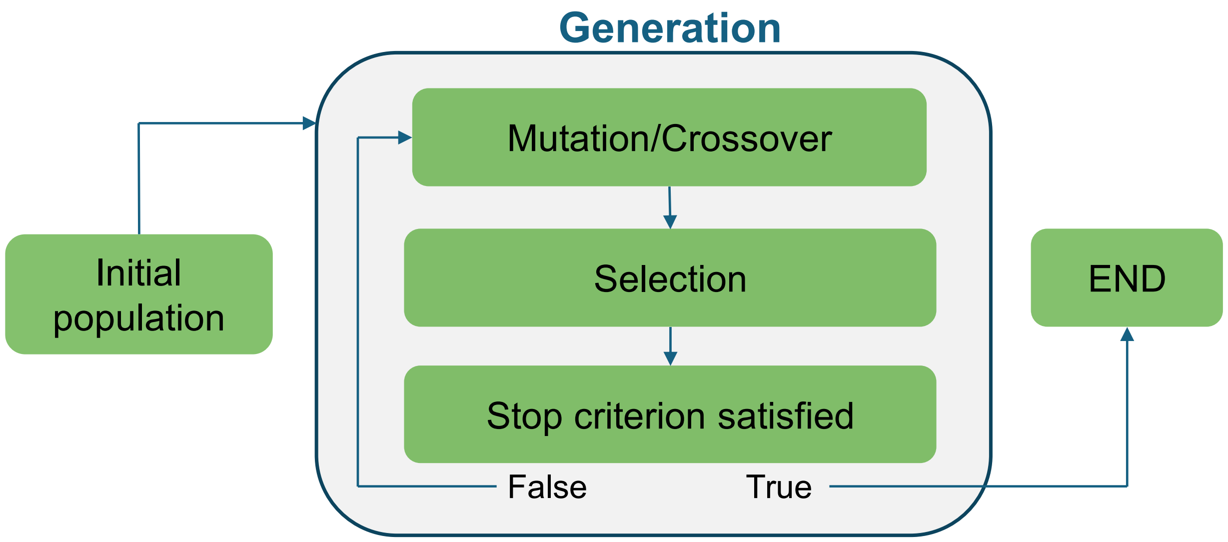


- Some features are multi-modal.
- Big overlap between distributions.





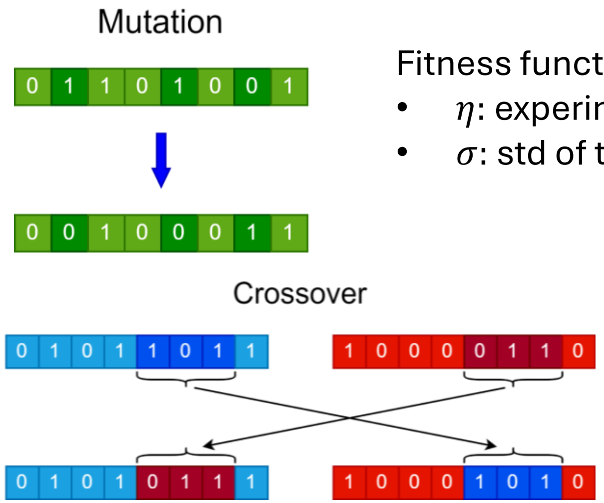
Featuremap automatization



Meta-heuristic approach – Genetic algorithm

- *Fitness function* – quantifies the goodness of a kernel.
- *Mutation* and *Crossover* operators – introduce variability through generations.
- A *parent/offspring* selection criteria.
- Initial population – Generation zero.

Goal: specialize the kernel population for the given classification task.



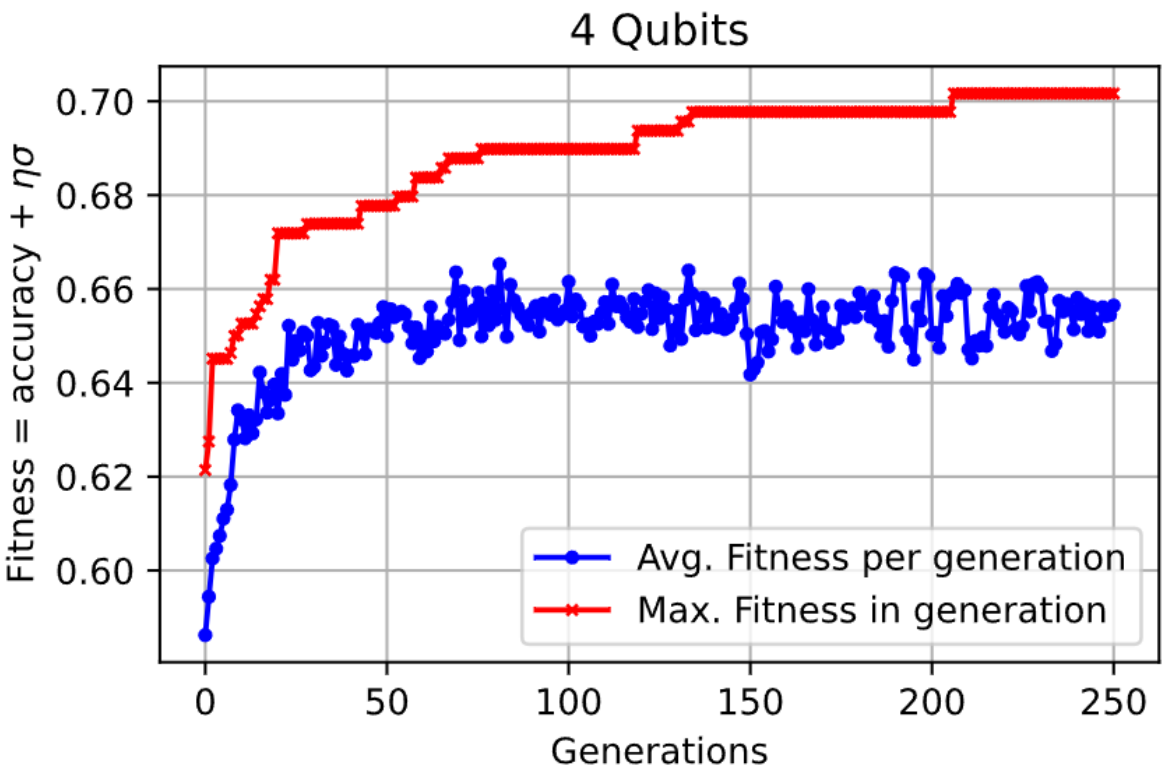
Fitness function $= accuracy + \eta \times \sigma$

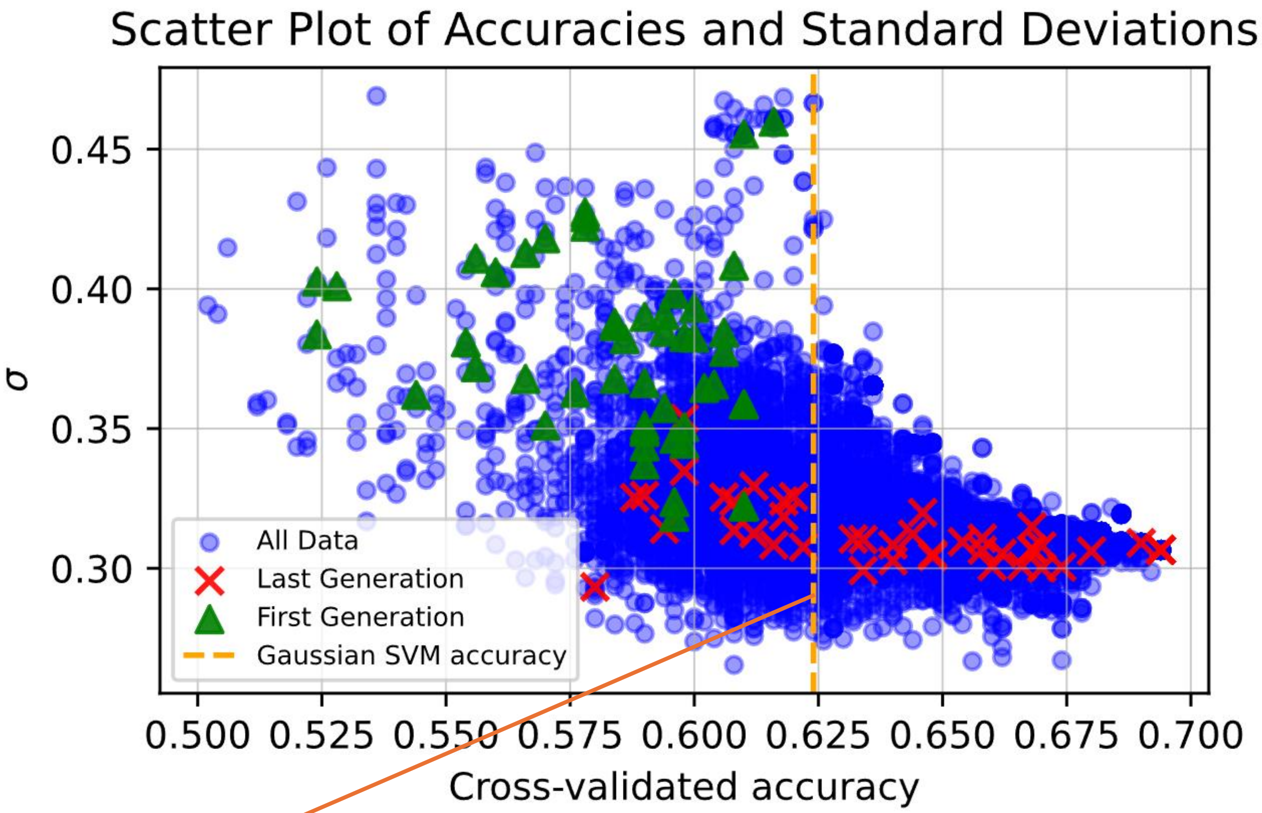
- η: experimentally-tuned value set to 0.025
- σ: std of the kernel matrix

6 integers per quantum gate

	Min	Max	Examples
Gate-type	0	custom	I, H, X, RX, RY, RZ, CX, S, CRX, …
Feature index	0	custom	Feature to use as a primary gate argument
Second feature index	0	custom	Feature to use as a secondary argument
Featuremap type	0	2	Linear, quadratic, trigonometric, …
Multi-feature per gate	0	1	Use one or two features in a gate rotation angle
Target qubit index	0	custom	Target qubit index

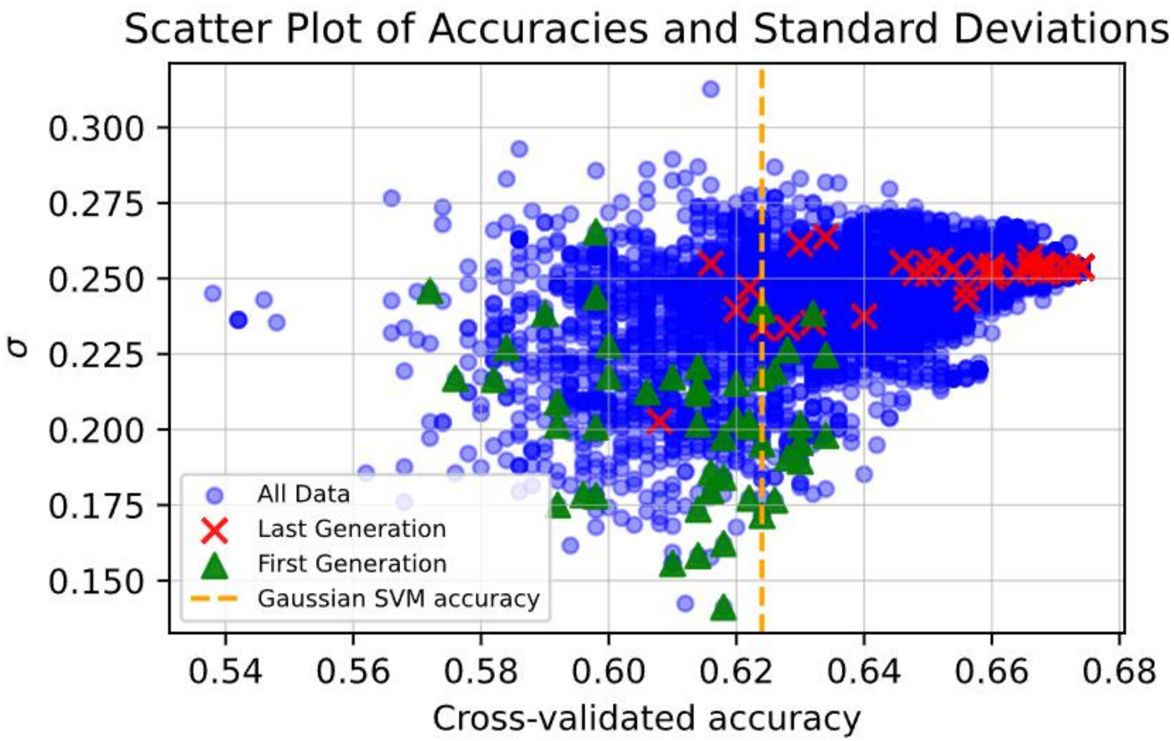
Featuremap automatization





Best classical SVM achieved

Featuremap automatization



Speedup through backend parallelization

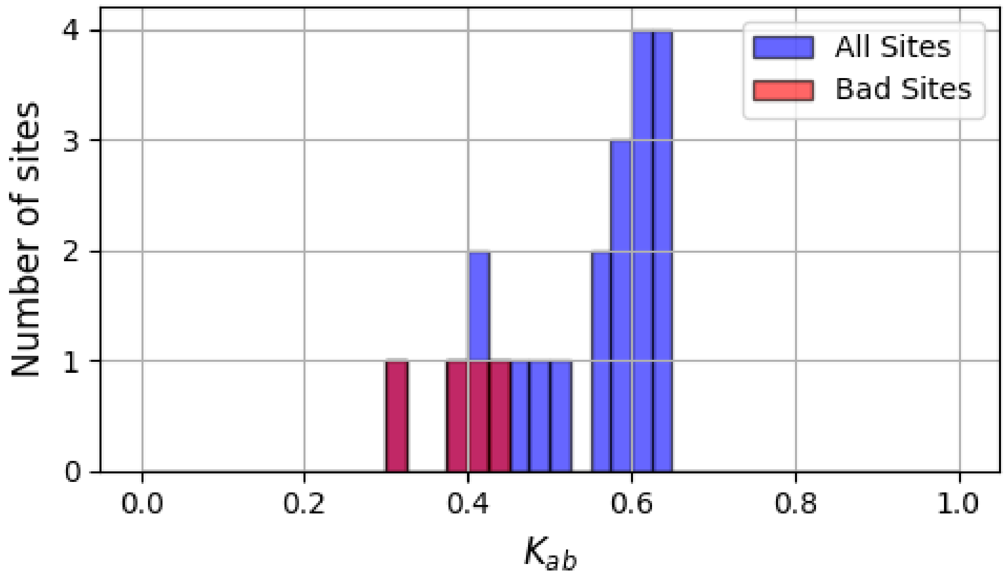
Backend: IBMQ Nazca (127)
Parallel training of 21 4-qubits QSVM

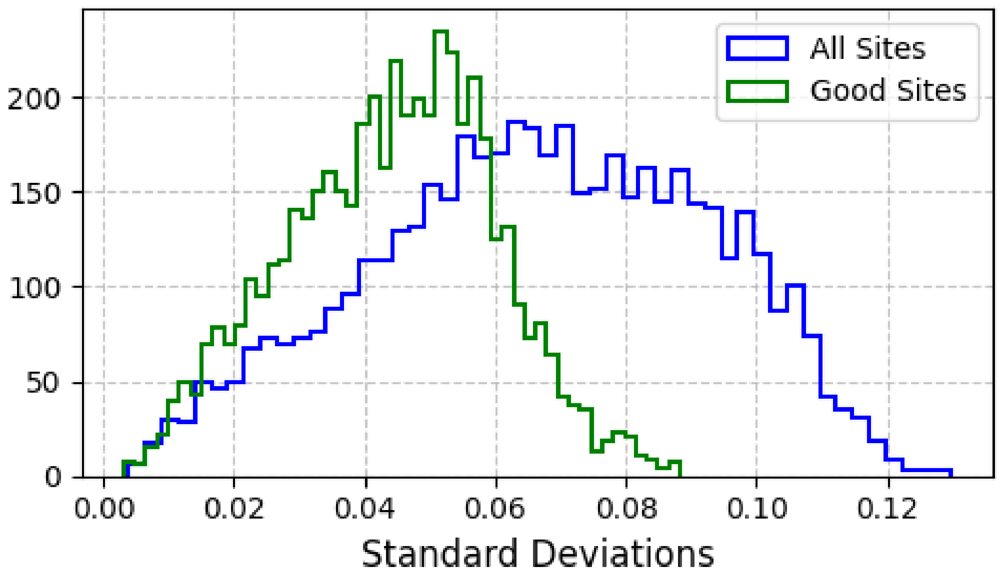
We can retrieve the kernel entries for each site and estimate the output spread due to the QPU noise.

It turns out that **some sites are less performant than other**. We can discard them based on how much they differ from the avgerage matrix.

By discarding only 4 sites out of 21, the spread halves.

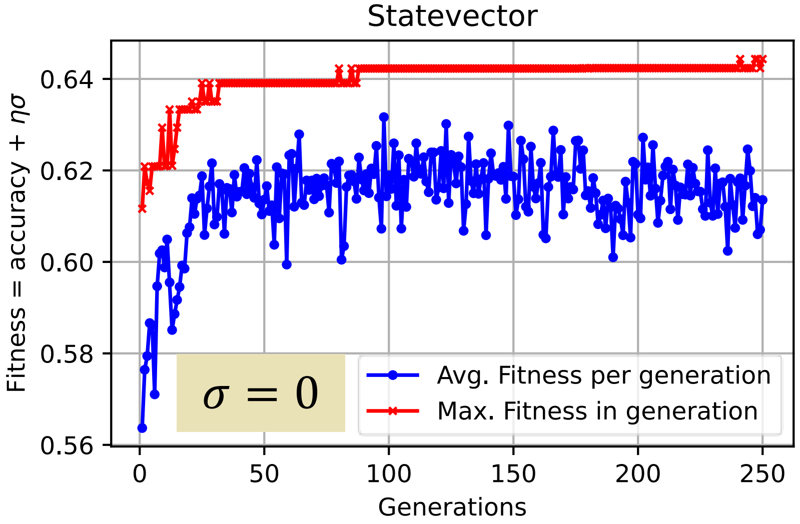
- On average the good sites std is **0.044.**
- The kernel matrix std between all sample entries is **0.241.**



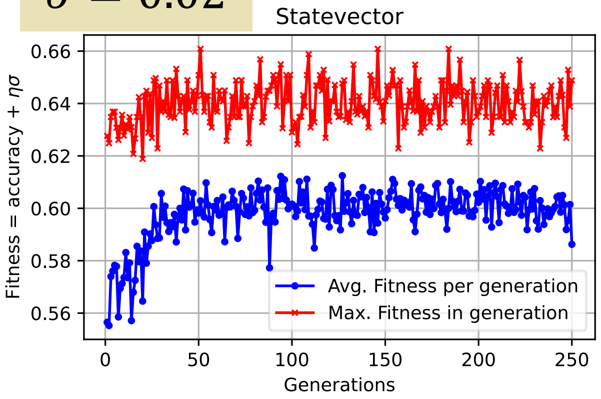


Data-driven spread effect on genetic runs

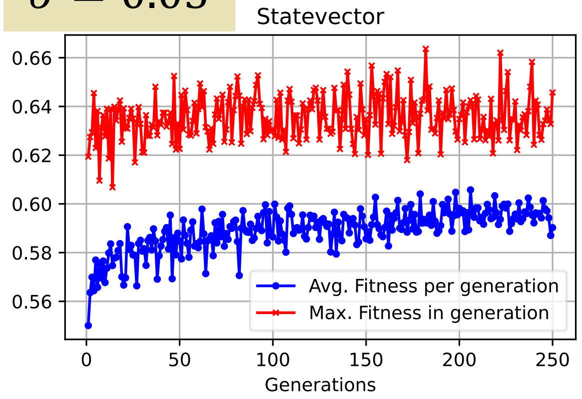
We simulated a gaussian noise on the kernel matrix entries up to the dispersion estimated from experimental data ($\sigma = 0.044$). Up to $\sigma = 0.03$, the genetic optimization succeeds exhibiting a positive trend throughout generations.



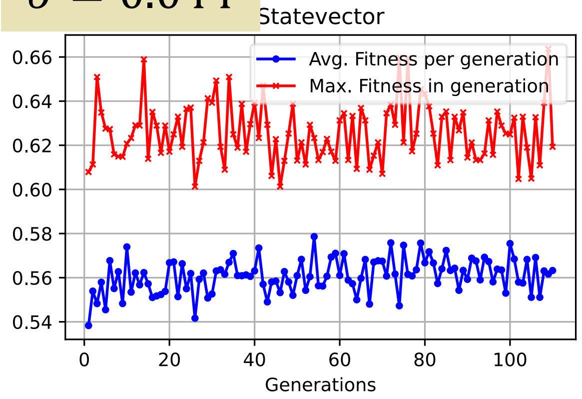
$\sigma = 0.02$



$\sigma = 0.03$



$\sigma = 0.044$



Conclusions

Physics

- **Modest $\beta\beta$ vs β classification accuracy overall** $(\sim 65\%)$ for an ideal $5 \times 5\mathrm{mm}^2$ pixel-size LArTPC.
- → **Depleted argon and better spatial resolution are mandatory.**
- Energy threshold heavily affects performances.
- Interesting technique for **other low energy physics channels in DUNE.**

Deep Learning

- CNN and Transformer performances are equally good for most granularity/threshold conditions, despite the different data-handling.

Quantum classifier – QSVM

- The use of QSVMs have been demonstrated for this dataset.
- Although Quantum Advantage can't be claimed, simulated, genetic-optimized QSVMs exhibit promising performances thanks to genetic optimization.
- Commercially available NISQ hardware is likely to be exploitable for running the QSVMs we developed.