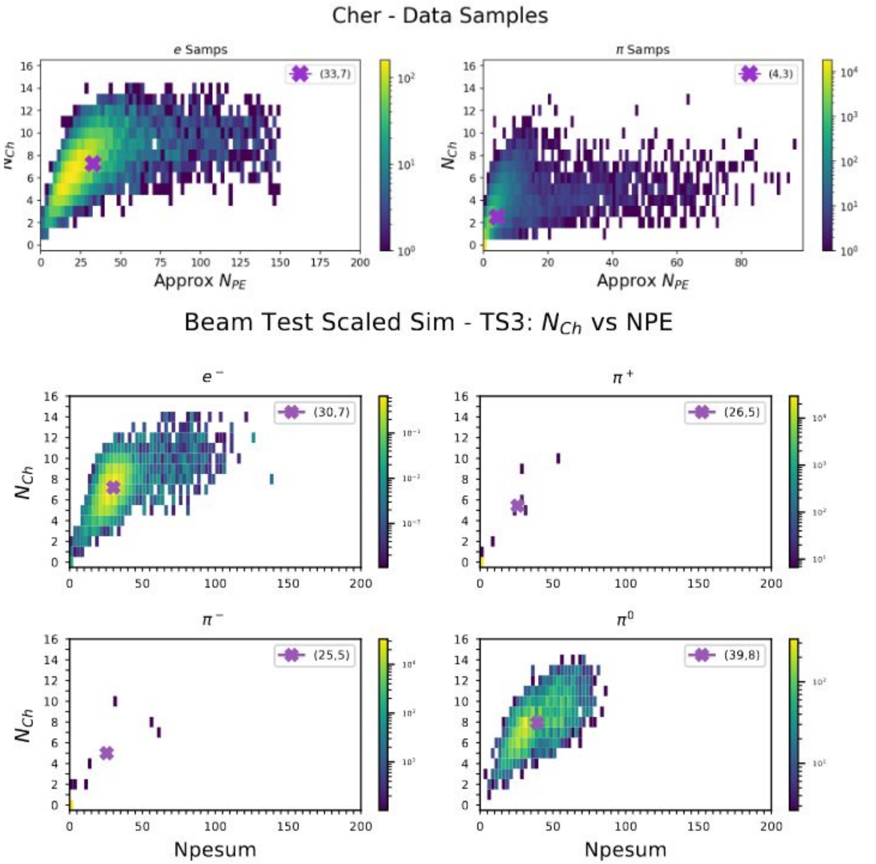
PID Using Machine-Learning Methods for SoLID Beam Test Analysis

Richard L. Trotta

Summary of Darren’s Report

- Integrate machine learning approaches within the SoLID collaboration, employing the ECal beam test to showcase their practical benefits.
- By utilizing simulated events for the beam test, we can create machine learning-assisted particle identification (ML-Assisted PID) methods to apply to the beam test data.
- Preliminary samples of electrons (e) and charged pions ($\pi^{+/-}$) have already been determined, yet further actions are necessary to enhance the agreement between simulation and experimental data.



Outline

1. Physics Introduction
2. ML Basics
3. Bringing it Together

1. Physics Introduction

What is SoLID?

- SoLID - Solenoidal Large Intensity Device
- Scientific Goals:
  - 3D Imaging of Nucleons: Provides insights into the three-dimensional structure of nucleons, including transverse momentum distributions (TMDs) and generalized parton distributions (GPDs).
  - Parity-Violation Experiments: Facilitates studies of parity-violating asymmetries in electron-nucleon scattering, contributing to our understanding of the weak force and electroweak interactions.
  - Hadron Spectroscopy: Helps explore the spectrum of hadrons, including exotic states and their properties.
  - SIDIS (Semi-Inclusive Deep Inelastic Scattering): Enables detailed studies of SIDIS processes, crucial for understanding the spin and flavor structure of the nucleon.
- Key Features:
  - Solenoidal Magnet: Utilizes a large solenoidal magnetic field to achieve high acceptance for charged particles.
  - High Luminosity: Capable of handling high luminosity beams, allowing for precise measurements and rare event detection.
  - Versatility: Modular design allows for a variety of detector configurations to suit different experimental needs.
  - High Acceptance: Designed to have a large acceptance for both forward and large-angle scattering, enabling comprehensive data collection.

Scientific Goals (1)

- 3D Imaging of Nucleons
  - Transverse Momentum Distributions (TMDs):
    - TMDs provide information on the momentum of quarks and gluons inside the nucleon perpendicular to the direction of motion.
    - Essential for understanding the internal dynamics and spatial distribution of partons.
  - Generalized Parton Distributions (GPDs):
    - GPDs describe the simultaneous distribution of partons in both position and momentum space.
    - Crucial for creating a three-dimensional image of the nucleon, combining spatial and momentum information.

- Parity-Violation Experiments
  - Parity Violation:
    - Parity violation occurs when physical processes distinguish between left-handed and right-handed coordinate systems.
    - Studying these asymmetries provides insights into weak interactions and the structure of the weak force.

Scientific Goals (2)

- Hadron Spectroscopy
  - Objective:
    - To explore the spectrum of hadrons (particles made of quarks, such as protons, neutrons, and mesons).
  - Exotic States:
    - Investigating hadrons that do not fit into the traditional quark model, including tetraquarks, pentaquarks, and hybrids.
    - Helps understand the strong interaction and the binding mechanism of quarks and gluons.

- Semi-Inclusive Deep Inelastic Scattering (SIDIS)
  - Process:
    - SIDIS involves scattering a high-energy electron off a nucleon and detecting one or more of the resulting hadrons in the final state.
  - Significance:
    - Provides detailed information on the spin and flavor structure of the nucleon.
    - Allows the extraction of parton distribution functions (PDFs) and fragmentation functions, which describe how quarks and gluons are distributed and how they form hadrons.

Key Features (1)

- Solenoidal Magnet
  - Design and Purpose:
    - The solenoidal magnet creates a strong, uniform magnetic field along its axis.
    - This magnetic field bends the trajectories of charged particles, allowing their momentum to be measured precisely.
  - Benefits:
    - Enables effective momentum resolution, crucial for precise measurements.
- High Luminosity
  - Definition:
    - Luminosity measures the number of particles colliding per unit area per unit time, directly impacting the rate of events that can be studied.
  - Advantages:
    - High luminosity allows for the detection of rare processes by providing a large dataset.
    - Enhances statistical precision, reducing uncertainties in measurements.

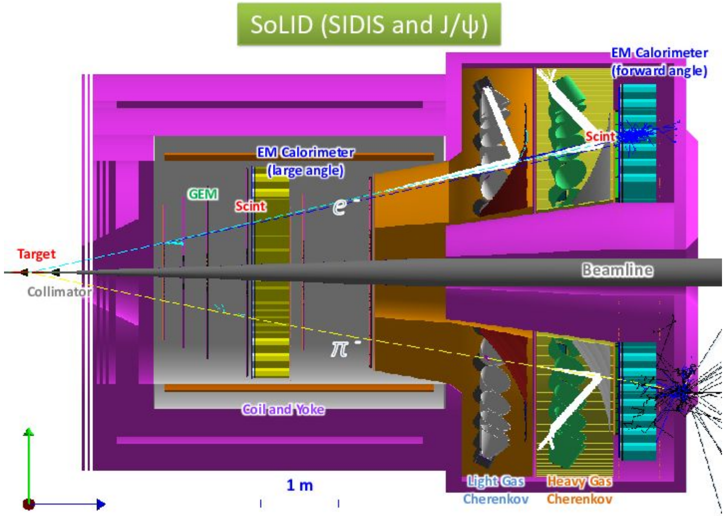
Key Features (2)

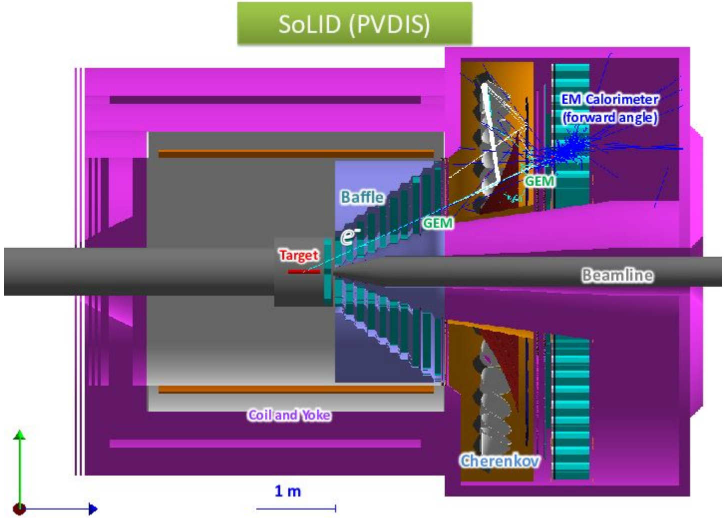
- Versatility
  - Modular Design:
    - SoLID's configuration can be adjusted based on the specific requirements of different experiments.
    - Allows the integration of various detectors (tracking, calorimetry, Cherenkov) to optimize performance for specific research goals.
  - Flexibility:
    - Can accommodate multiple experimental setups, including different target materials and beam energies.

- High Acceptance
  - Definition:
    - Acceptance refers to the fraction of the total solid angle around the interaction point that the detector can observe.
  - Importance:
    - High acceptance ensures comprehensive data collection from a wide range of scattering angles and particle types.

SoLID Breakdown

- Think of SoLID as a high-tech surveillance system with multiple components working together to provide a comprehensive understanding of events:
  - GEM Detectors: Like motion-tracking cameras that follow every movement with high precision.
  - Calorimeters: Similar to radar guns and energy meters that measure the speed and energy of moving objects.
  - Scintillators: Comparable to security cameras with timestamps that record when events happen and measure their intensity.
  - Cherenkov Detectors: Like specialized cameras that can identify and differentiate between different types of objects based on specific characteristics.
  - Baffle: Similar to noise-canceling barriers that block out irrelevant sounds and movements, focusing only on what's important.





SoLID's Technical Components: Tracking Detectors

- Tracking Detectors
  - Purpose:
    - Track the paths of charged particles produced in collisions.
    - Determine the momentum and trajectories of these particles.
  - How They Work:
    - Basic Principle: As charged particles move through a medium, they ionize the atoms along their path, creating a trail that can be detected.
  - Types:
    - Gas Electron Multiplier (GEM) Detectors:
      - Structure: Consists of thin foils with microscopic holes, layered between the electrodes.
      - Operation:
        - When a charged particle passes through the GEM detector, it ionizes the gas in the detector.
        - Electrons produced by ionization are amplified through the holes in the foils due to the high electric field.
        - This amplification process results in a measurable charge that can be detected and recorded.
      - Benefit: High spatial resolution and fast response time. Capable of operating at high particle rates without significant degradation of performance.

SoLID's Technical Components: Calorimeters

- Calorimeters
    - Purpose:
        - Measure the energy of particles produced in collisions.
    - How They Work:
        - Electromagnetic Calorimeters (ECal):
            - Target Particles: Electrons and photons.
            - Structure: Layers of high-density material interspersed with sensors.
            - Operation: When an electron or photon enters, it produces a shower of secondary particles. These secondary particles generate light or electrical signals in the sensors, proportional to the energy of the initial particle.
            - Benefit: Provides precise energy measurements for electromagnetic particles.
        - Hadronic Calorimeters:
            - Target Particles: Hadrons (such as protons, neutrons, and pions).
            - Structure: Similar to electromagnetic calorimeters but made of materials better suited to absorb hadrons.
            - Operation: Hadrons produce more complex showers of particles when they hit the calorimeter. The energy of these showers is measured to determine the energy of the initial hadron.
            - Benefit: Measures the energy of particles that interact via the strong nuclear force.

SoLID's Technical Components: Scintillators

- Scintillators
  - Purpose:
    - Detect particles and measure their energy.
    - Provide timing information to determine the speed of particles.
  - How They Work:
    - Basic Principle: Scintillators emit light (scintillation) when a charged particle passes through them.
  - Components:
    - Scintillating Material: Typically a plastic or crystal that emits light when excited by a charged particle.
    - Photomultiplier Tubes (PMTs) or Photodiodes: Convert the light emitted by the scintillator into an electrical signal.
  - Operation:
    - As a charged particle passes through the scintillating material, it excites the atoms in the material, causing them to emit photons (light).
    - The emitted light is collected and converted into an electrical signal by PMTs or photodiodes.
    - The intensity of the signal is proportional to the energy of the particle.
  - Benefit:
    - Provides precise timing information, essential for determining the speed of particles.
    - Can cover large areas and detect a wide range of particle types..

SoLID's Technical Components: Cherenkov Detectors

- Cherenkov Detectors
  - Purpose:
    - Identify different types of charged particles by measuring the Cherenkov radiation they emit.
  - How They Work:
    - Basic Principle: When a charged particle moves through a medium faster than the speed of light in that medium, it emits a cone of light called Cherenkov radiation.
  - Components:
    - Radiator Medium: The material through which particles travel, typically gas or liquid.
    - Photon Detectors: Sensors that detect the Cherenkov light emitted by the particles.
  - Operation:
    - As particles pass through the radiator, they emit Cherenkov radiation if they are moving fast enough.
    - The emitted light forms a cone, and the angle of this cone depends on the particle's speed and type.
    - Photon detectors capture the Cherenkov light and measure the angle and intensity of the emitted light.
  - Benefit: Different particles (e.g., electrons, pions, kaons) emit Cherenkov light at different angles and intensities, allowing them to be distinguished from each other.

SoLID's Technical Components: Baffle

- Baffle
  - Purpose:
    - Reduce background noise and unwanted particles from reaching the detectors.
    - Improve the signal-to-noise ratio in the measurements.
  - How They Work:
    - Structure: Consists of absorbing materials or geometric structures placed strategically to block unwanted particles.
    - Operation:
      - The baffle is designed to allow particles from the interaction region to pass through while blocking particles that are not of interest.
      - It reduces the amount of stray particles and radiation that can create noise in the detectors.
  - Benefit:
    - Enhances the quality of the data collected by reducing background noise.
    - Improves the accuracy and precision of the measurements.

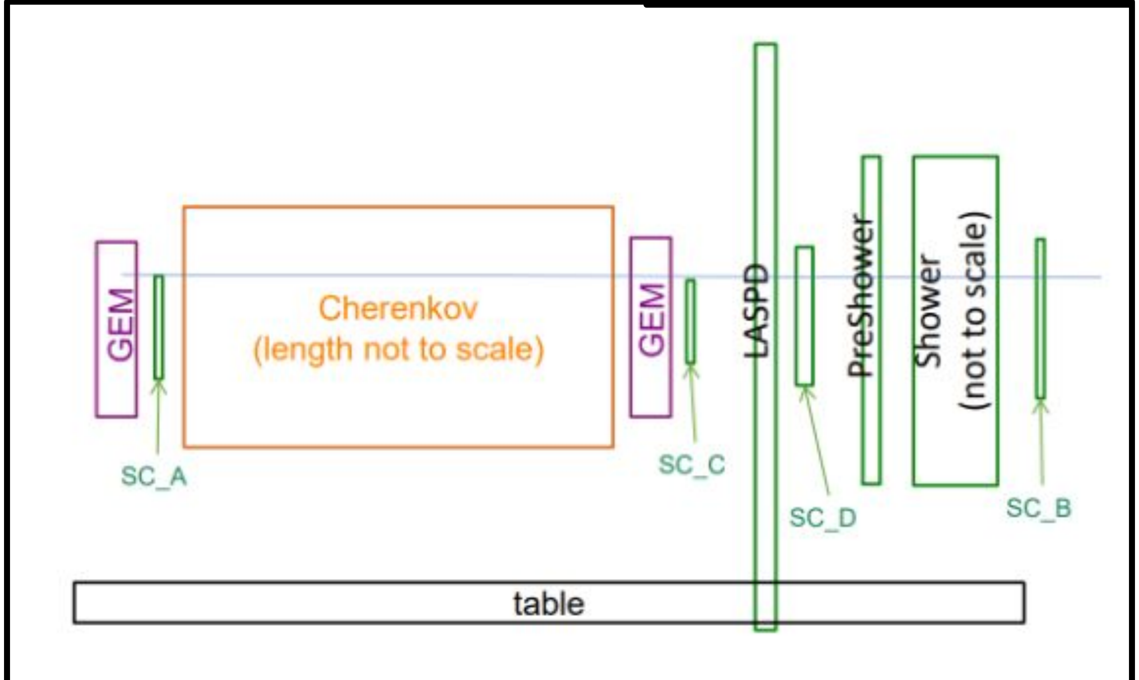
SoLID's Technical Components: Baffle

- Baffle
  - Purpose:
    - Reduce background noise and unwanted particles from reaching the detectors.
    - Improve the signal-to-noise ratio in the measurements.
  - How They Work:
    - Structure: Consists of absorbing materials or geometric structures placed strategically to block unwanted particles.
    - Operation:
      - The baffle is designed to allow particles from the interaction region to pass through while blocking particles that are not of interest.
      - It reduces the amount of stray particles and radiation that can create noise in the detectors.
  - Benefit:
    - Enhances the quality of the data collected by reducing background noise.
    - Improves the accuracy and precision of the measurements.

Detectors for SoLID Beam Test

- **GEM**
- **Scintillator**
- **Cherenkov**
- **ECal**

Hardware configuration for beam test



2. ML Basics

Machine Learning Overview

- Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on the development of algorithms that enable computers to learn from and make predictions or decisions based on data.
- The primary goal of ML is to develop systems that can automatically improve with experience without being explicitly programmed for each task.

ML Key Concepts

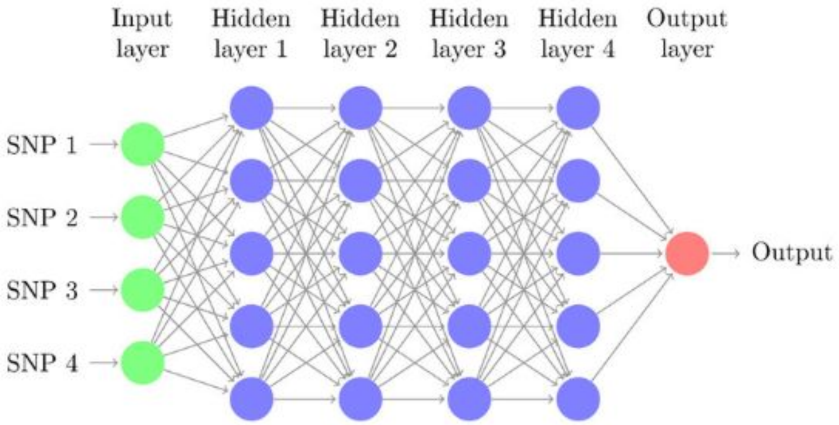
- Data: The foundation of ML. It can be structured (like databases) or unstructured (like text and images).
- Features: Individual measurable properties or characteristics of a phenomenon being observed.
- Labels: The target outcomes or categories that the algorithm is learning to predict (used in supervised learning).
- Model: A mathematical representation of a real-world process. It’s trained on data to recognize patterns.
- Training: The process of learning from data to adjust the model’s parameters.
- Testing: Evaluating the model’s performance on a separate dataset not seen during training.

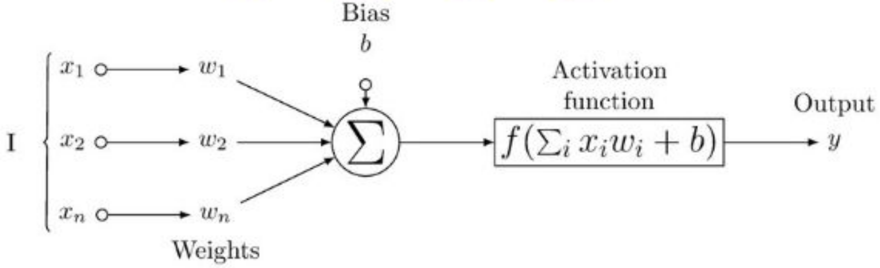
Type of ML

- Supervised Learning:
  - In supervised learning, the algorithm learns from labeled data, where each example is paired with a label or outcome.
  - It aims to learn a mapping function from input variables to output variables.
  - Common algorithms include linear regression, logistic regression, decision trees, random forests, support vector machines, and neural networks.
- Unsupervised Learning:
  - Unsupervised learning deals with unlabeled data, where the algorithm tries to find patterns or relationships in the data without explicit guidance.
  - Clustering and dimensionality reduction techniques such as k-means clustering, hierarchical clustering, principal component analysis (PCA), and t-distributed stochastic neighbor embedding (t-SNE) are examples.
- Reinforcement Learning:
  - Reinforcement learning involves an agent learning to make decisions by interacting with an environment to achieve a goal.
  - The agent receives feedback in the form of rewards or penalties based on its actions.
  - Examples include Q-learning, deep Q-networks (DQN), and policy gradient methods.

Neural Networks

- Neural networks are a class of models inspired by the structure and functioning of the human brain.
- They consist of interconnected nodes (neurons) organized in layers, including an input layer, one or more hidden layers, and an output layer.
- Each connection between nodes has an associated weight, and neurons apply an activation function to the weighted sum of their inputs.
- Common activation functions include sigmoid, tanh, ReLU, and softmax.





Multi-Layer Perceptron (MLP)

- MLP was used in Darren's analysis
  - TensorFlow with the vanilla MLP classifiers
- MLP is a type of feedforward neural network that consists of multiple layers of nodes, including an input layer, one or more hidden layers, and an output layer (see previous slide).
- It is called "multi-layer" because it has more than one layer of neurons between the input and output layers.
- MLPs are capable of learning complex non-linear relationships in data.
- Training an MLP involves forward propagation of inputs through the network to produce predictions, followed by backward propagation of errors to adjust the weights using techniques like gradient descent and backpropagation.
- Despite their successes, machine learning algorithms, including MLPs, face challenges such as overfitting, interpretability, and scalability.

ML Challenges (1)

- Overfitting:
  - Overfitting occurs when a machine learning model learns the training data too well, capturing noise or random fluctuations rather than the underlying pattern.
  - This leads to poor generalization performance, where the model performs well on the training data but fails to generalize to unseen data.
  - Common techniques to mitigate overfitting include:
    - Regularization: Adding a penalty term to the loss function to discourage overly complex models, such as L1 and L2 regularization.
    - Cross-validation: Splitting the data into multiple subsets for training and evaluation to assess the model's performance on different data partitions.
    - Early stopping: Monitoring the model's performance on a validation set during training and stopping when performance starts to degrade.
- Interpretability:
  - Interpretability refers to the ability to understand and explain how a machine learning model makes predictions.
  - Deep neural networks, including MLPs with multiple hidden layers, are often referred to as "black box" models because their decision-making process can be complex and opaque.
  - Lack of interpretability can be a significant barrier to the adoption of machine learning models, particularly in domains where transparency and accountability are crucial, such as healthcare and finance.
  - Techniques to enhance interpretability include:
    - Feature importance: Analyzing the contribution of input features to model predictions, such as through feature importance scores or permutation importance.
    - Visualization: Visualizing model internals, such as activation patterns in hidden layers or decision boundaries in feature space, to gain insights into model behavior.
    - Simplification: Simplifying complex models into more interpretable forms, such as using decision trees or linear models as approximations of neural networks.

ML Challenges (2)

- Scalability:
  - Scalability refers to the ability of a machine learning algorithm to handle increasingly large datasets or computational demands.
  - As datasets grow in size and complexity, traditional machine learning algorithms, including MLPs, may struggle to scale efficiently.
  - Challenges in scalability include memory constraints, computational resources, and algorithmic efficiency.
  - Techniques to address scalability issues include:
    - Distributed computing: Distributing computation across multiple machines or clusters to parallelize training and inference tasks.
    - Batch processing: Processing data in batches rather than individually to optimize memory usage and computational efficiency.
    - Model compression: Reducing the size of machine learning models through techniques such as pruning redundant connections, quantization of weights, and knowledge distillation from larger models.
    - Hardware acceleration: Leveraging specialized hardware, such as GPUs (Graphics Processing Units) or TPUs (Tensor Processing Units), to accelerate training and inference tasks for deep learning models like MLPs.

ML Terminology (1)

- Bias-Variance Tradeoff:
  - The bias-variance tradeoff is a fundamental concept in machine learning that describes the balance between bias (underfitting) and variance (overfitting) in a model. A model with high bias may oversimplify the data, while a model with high variance may capture noise.
- Cross-Validation:
  - Cross-validation is a technique used to assess the performance of a machine learning model by splitting the data into multiple subsets for training and evaluation. It helps to estimate how well the model will generalize to unseen data.
- Loss Function:
  - The loss function is a mathematical function that measures the difference between the predicted output of a machine learning model and the true output (label) for a given input. It quantifies the model's performance during training and is used to adjust the model's parameters.
- Regularization:
  - Regularization is a technique used to prevent overfitting in machine learning models by adding a penalty term to the loss function.
  - It encourages the model to learn simpler patterns and avoid fitting noise in the training data.
  - Common regularization techniques include L1 regularization (lasso), L2 regularization (ridge), and dropout (a form of regularization specific to neural networks).

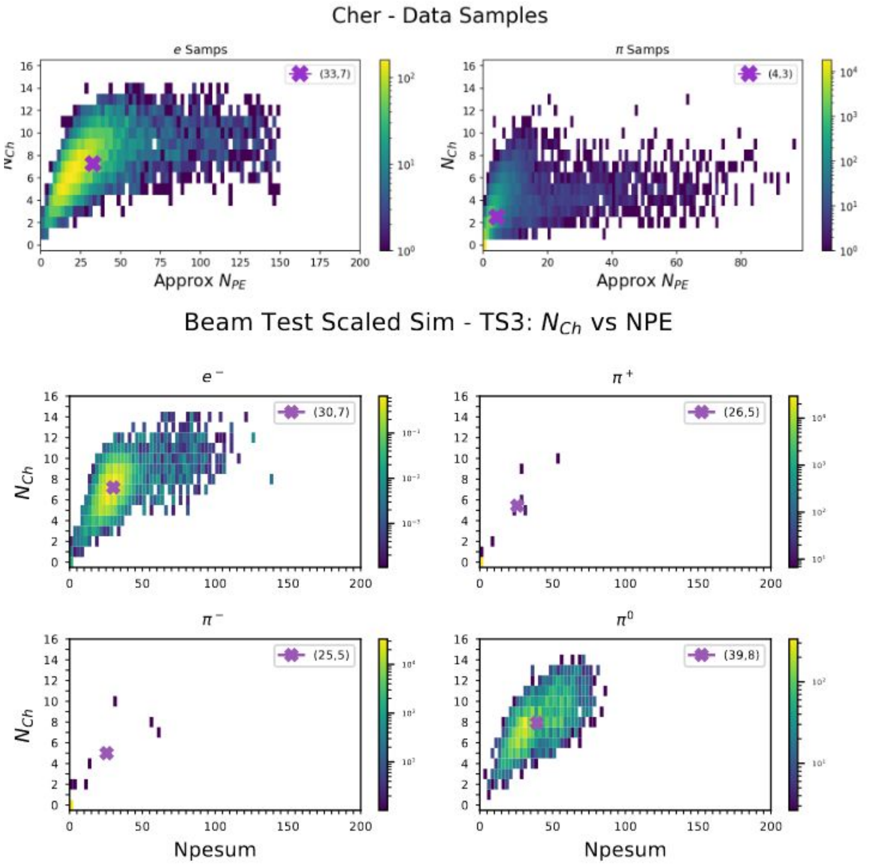
ML Terminology (2)

- Hyperparameters:
  - Hyperparameters are configuration settings or parameters that are set before the training process begins. They control aspects of the learning algorithm's behavior, such as learning rate, regularization strength, and network architecture (for neural networks).
- Gradient Descent:
  - Gradient descent is an optimization algorithm used to minimize the loss function and train machine learning models. It iteratively updates the model's parameters in the direction of the steepest descent of the loss function gradient.
- Learning Rate:
  - The learning rate is a hyperparameter that controls the size of the steps taken during the optimization process (e.g., gradient descent) to update the parameters of a machine learning model.
  - It determines how quickly or slowly the model learns from the training data.
  - A high learning rate may cause the model to converge quickly but risk overshooting the optimal solution, while a low learning rate may converge slowly but provide more stable updates.

3. Bringing it Together

Summary of Darren's Report

- Integrate machine learning approaches within the SoLID collaboration, employing the ECal beam test to showcase their practical benefits.
- By utilizing simulated events for the beam test, we can create machine learning-assisted particle identification (ML-Assisted PID) methods to apply to the beam test data.
- **Preliminary samples of electrons (e) and charged pions ($\pi^{+/-}$) have already been determined, yet further actions are necessary to enhance the agreement between simulation and experimental data.**



Initial Steps

- Reproduce Darren's results
- Together brainstorm ideas
  - Iterative Approach:
    - Why is iterating on our analysis process and continuously refining techniques beneficial?
  - Constraints for Event Selection:
    - What criteria should we consider when selecting events for PID studies?
    - How might imposing tighter constraints on event selection parameters impact PID results?
  - Trigger Optimization:
    - How can we adjust trigger parameters to capture relevant events while minimizing noise?
    - What considerations are important when designing triggers?
  - Tracking Efficiency:
    - What techniques can improve the efficiency of GEM tracking?
    - How does precise tracking enhance PID accuracy?
  - Integration of Simulation Feedback:
    - How can insights from simulation analysis be integrated into our machine learning approach?
    - What benefits arise from incorporating simulation data into our training dataset?
  - Machine Learning Model Optimization:
    - Which parameters should we focus on tuning to enhance the performance of our machine learning models?
    - How do different neural network architectures contribute to PID improvement?
  - Exploring Advanced ML Techniques:
    - What are the potential benefits of using CNNs or RNNs for PID?
    - How do we determine the best ML techniques for analyzing our data?