



Co-funded by  
the European Union



# Application of Conformal Inference in High Energy Physics

---

Jiri Franc

2024-06-28

Department of Mathematics, FNPSE, CTU in Prague,  
Neutrino Physics and Machine Learning 2024, ETH Zurich

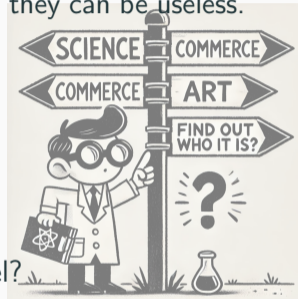
# Motivation: Decision making theory

## What to take from decision making theory.

- Point predictions are easy to compare, but without uncertainty, they can be useless.
- Decision making requires quantification of our confidence.

## What questions should be answered.

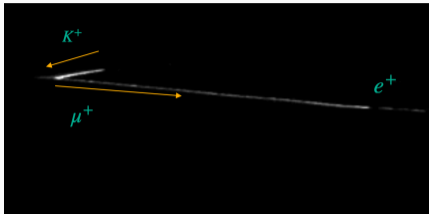
- **Regression:** How wide are my prediction intervals?
- **Binary classification:** Is my output score a probability?
- **Multi-label classification:** Do I have to choose only one label?
- **Time series prediction:** Do I want to lose money in the market, or do I prefer risk control?



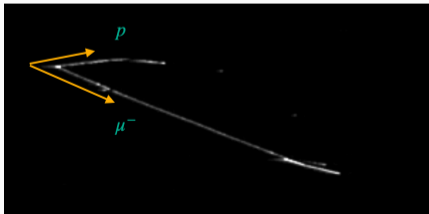
NPML says:

We pledge to welcome those interested in joining the community, and realize that including people with a variety of opinions and backgrounds will only serve to enrich our community.

# HEP Motivation: Proton decay analysis at DUNE



A MC  $p \rightarrow K^+ \bar{\nu}$  proton decay event.



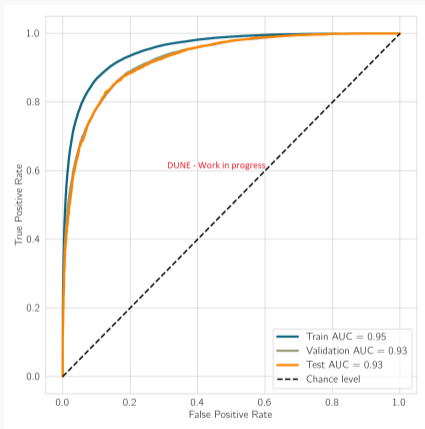
A MC background event mimicking the

- DUNE FD will exploit a number of complementary signatures for a broad range of nucleon decay channels. [2]
- The task of how a DNN model can handle this task is not the goal of this presentation.

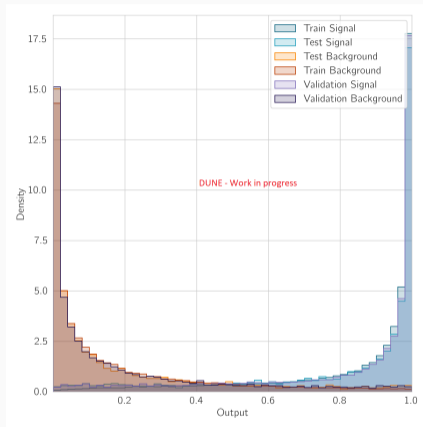
More in **Anna G. Gartman ICHEP Poster:**

Proton Decay Identification in DUNE with Multimodal Machine Learning Fusion Techniques [3]

# HEP Motivation: Proton decay analysis at DUNE



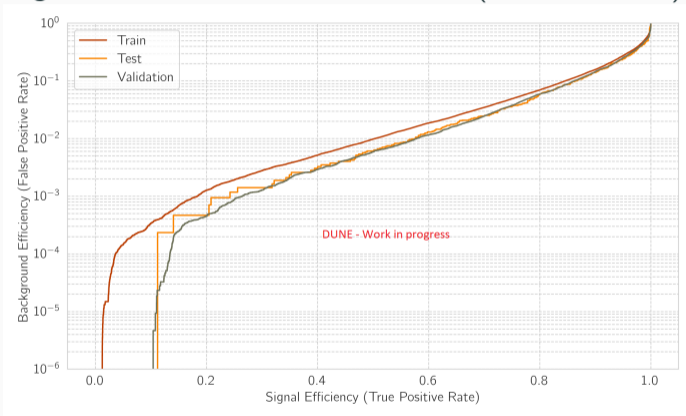
How do we choose the threshold?



Do we have good train/test & S/B discriminant distribution agreement?

# HEP Motivation: Proton decay analysis at DUNE

LogBE vs SE for the late fusion model (illustrative case)



- Will we have enough observations in the cut-out area to apply the classical approach?
- DUNE TDR: Under some assumptions, observation of even one or two candidate

# Are probabilistic predictions the right way?

What should we keep in mind.

- Most ML methods fail to properly estimate the uncertainty of their predictions.
- Most ML and DL methods are not well calibrated as well.
- Scikit-learn `predict_proba` is not probability and needs to be calibrated.
- Histogram binning is simple but not very accurate calibration technique.

For useful probabilistic predictions, the estimated probabilities must be well-calibrated, i.e. they should accurately reflect the true likelihood of an event occurring.

For example, probability scores are calibrated if among all events with a score of 0.9, we find the signal 9 times out of 10 events.

# Conformal Prediction (CP)

CP is an emerging distribution-free and model-agnostic probabilistic forecasting approach that

- invented by V. Vovk and A. Gammerman in 2005, Royal Holloway in London [9].
- does not rely on the distribution of the data or the model used, no need any prior probabilities.
- transforms point predictions into prediction intervals.
- ensures that the prediction intervals have a probabilistic guarantee of covering the true outcome.
- uses residuals from a calibration dataset to create prediction intervals for a test dataset.
- easy to implement and can be applied to any ML and DL model.

I tried to look for CP in the HEP papers but I failed.

# Conformal Prediction (CP)

Are there any assumptions?

The only one is **exchangeability** (a weaker assumption than *i.i.d.*)

For random variables  $(X_1, X_2, \dots, X_n)$ , they are exchangeable if for any permutation  $(\pi)$  of the indices  $(\{1, 2, \dots, n\})$ , the joint distribution remains the same:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \mathbb{P}(X_{\pi(1)} = x_1, X_{\pi(2)} = x_2, \dots, X_{\pi(n)} = x_n).$$

Interested in Conformal Prediction? See [1] [5] [6] for example.



CP affects all possible areas; let's focus now only on classification.

- **Probabilistic prediction:** the goal is to obtain a **valid** predictor, i.e. the probability distributions from the predictor must perform well against statistical tests based on subsequent observation of the labels.
- **Binary classification (Venn-Aber Predictors [10], [11])**

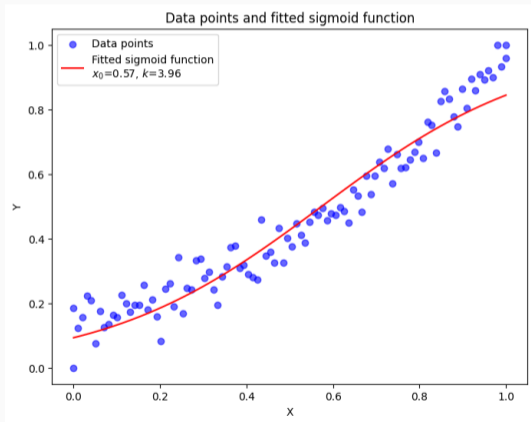
Let us say that a random variable  $P$  taking values in  $[0, 1]$  is perfectly calibrated for a random variable  $Y$  taking values in  $\{0, 1\}$  if

$$\mathbb{E}[Y|P] = P \text{ a.s.}$$

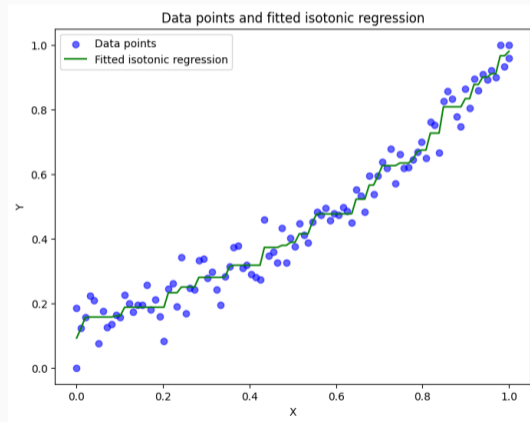
A probabilistic predictor for a random variable  $Y$  whose prediction  $P$  satisfies previous definition is said to be well calibrated.

# Existing approaches for calibration

**Platt scaling** using a sigmoid function to a calibration set. [8]



**Isotonic regression** uses non-decreasing step-wise calibration function. [12]



## Venn-ABERS Predictors

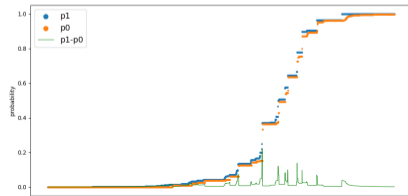
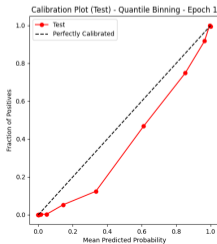
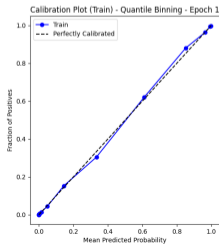
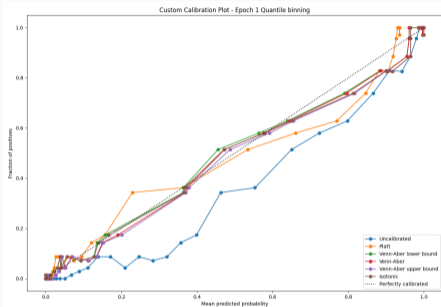
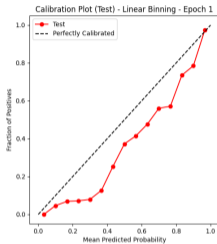
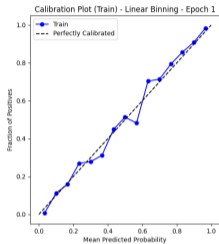
- Venn-ABERS predictors fit isotonic regression twice for each test object by adding it to the calibration set with both labels.
- This process results in two probabilities,  $p_0$  (lower bound) and  $p_1$  (upper bound), forming a prediction interval for the probability of class 1.
- The combined probability  $p = \frac{p_1}{1-p_0+p_1}$  (optimal  $p$  for log loss) is used for decision making and calibration.
- Venn-ABERS provides mathematically guaranteed prediction intervals without making assumptions about score distributions, avoiding overfitting issues.
- The width of the interval  $(p_0, p_1)$  indicates classification confidence, with narrower intervals for larger datasets and wider intervals for smaller datasets.

## Smaller 20k toy example

Training and Testing Metrics by Epoch with Modified Resnet18 classifier on 20k images, balanced dataset.

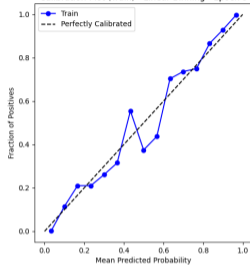
Epoch	Tr. Acc	Tr. Precision	Tr. Recall	Te. Acc	Te. Precision	Te. Recall
1	0.929	0.902	0.883	0.921	0.848	0.928
2	0.962	0.947	0.937	0.933	0.966	0.825
3	0.970	0.959	0.949	0.936	0.852	0.977
4	0.975	0.967	0.956	0.961	0.949	0.931
5	0.980	0.971	0.968	0.958	0.943	0.930
6	0.987	0.983	0.978	0.948	0.879	0.977
7	0.988	0.981	0.983	0.957	0.974	0.896
8	0.987	0.980	0.981	0.946	0.886	0.958
9	0.990	0.984	0.986	0.957	0.913	0.962
10	0.987	0.981	0.982	0.966	0.943	0.954

# Example of calibration binary classifier: after 2nd Epoch.

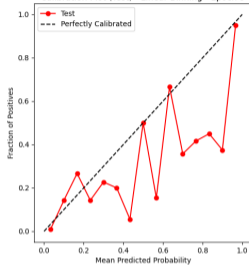


# Example of calibration binary classifier: after 10th Epoch.

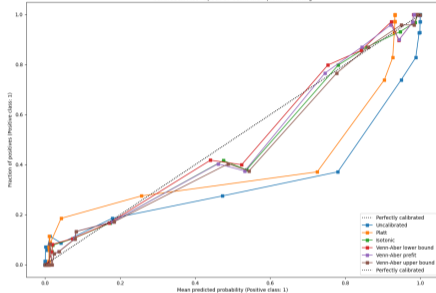
Calibration Plot (Train) - Linear Binning - Epoch 9



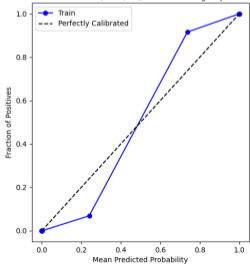
Calibration Plot (Test) - Linear Binning - Epoch 9



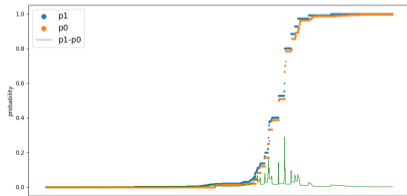
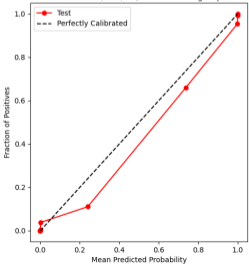
Calibration Plot - Epoch 9: scikit with quantile binning



Calibration Plot (Train) - Quantile Binning - Epoch 9



Calibration Plot (Test) - Quantile Binning - Epoch 9



## Closing Remarks

- Calibration does not have to improve metric like accuracy, recall or precision.
- The quality of the probability estimates can be evaluated using the

$$\text{log loss: } \lambda_{\log} = \begin{cases} -\log p & \text{if } y = 1 \\ -\log(1 - p) & \text{if } y = 0, \end{cases} \text{ or Brier loss: } \lambda_{Br} = (y - p)^2.$$

The empirical investigation showed that Venn-Abers prediction intervals are very tight, the probability estimates extremely well-calibrated and is always the most exact. [4]

- You can conformalize the ROC curve to obtain an uncertainty interval around ROC too. [7] [13].
- The concept is easy to use in multiple label classification too. See [1].




- **MAPIE**: A scikit-learn-compatible module for estimating prediction intervals.
- **CREPES**: Conformal classifiers, regressors, and predictive systems.
- **nonconformist**: Python implementation of the conformal prediction framework.
- **Venn-ABERS calibration package**.




---




Will Conformal Prediction find use in Neutrino and High Energy Physics?







Thank you for your attention

-  A. N. Angelopoulos and S. Bates.  
**A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022.**
-  DUNE Collaboration and B. A. et. al.  
**Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume I Introduction to DUNE.**  
*JINST*, 15(08):T08008, 2020.
-  A. G. Gartmann, J. Franc, and V. Pec.  
**Proton decay identification in dune with multimodal machine learning fusion techniques.**  
*ICHEP 2024, Contribution ID 1206*, 2024.

-  U. Johansson, T. Löfström, and H. Boström.  
*Calibrating Probability Estimation Trees using Venn-Abers Predictors,*  
pages 28–36.
-  V. Manokhin.  
*Practical Guide to Applied Conformal Prediction in Python: Learn and  
apply the best uncertainty frameworks to your industry applications.*  
Packt Publishing, 2023.
-  C. Molnar.  
*Introduction To Conformal Prediction With Python: A Short Guide For  
Quantifying Uncertainty Of Machine Learning Models.*  
Independent, 2023.

-  P. Novello, J. Dalmau, and L. Andeol.  
**Out-of-distribution detection should use conformal prediction (and vice-versa?), 2024.**
-  J. C. Platt.  
**Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.**  
*In Advances in large margin classifiers. MIT Press, pp. 61–74., 1999.*
-  G. S. V. Vovk, A. Gammerman.  
***Algorithmic Learning in a Random World.***  
Springer, 2005.

-  V. Vovk and I. Petej.  
**Venn-abers predictors, 2014.**
-  V. Vovk, I. Petej, and V. Fedorova.  
**Large-scale probabilistic predictors with and without guarantees of validity.**  
In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

-  B. Zadrozny and C. Elkan.  
**Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers.**  
*Proceedings of the 18th international conference on machine learning*, page 609–616, 2021.
-  Z. Zheng, B. Yang, and P. Song.  
**Quantifying uncertainty in classification performance: Roc confidence bands using conformal prediction, 2024.**

1

---

<sup>1</sup>This Presentation was co-funded by the European Union and supported by the Czech Ministry of Education, Youth and Sports (Project No. FORTE – CZ.02.01.01/00/22/\_008/0004632).