Machine Learning for Event-Level Background Rejection in LAFOV PET

from Classification to Reconstruction and Uncertainty Quantification

Michał Obara

National Centre for Nuclear Research · Warsaw University of Technology

WMLQ 2026, 9th September, Warsaw

Supported by the IMPET project FENG.02.02-IP.05-0152/23 (FIRST TEAM FENG programme of the Foundation for Polish Science, co-financed by the European Union under FENG 2021-2027).

European Funds for Smart Economy

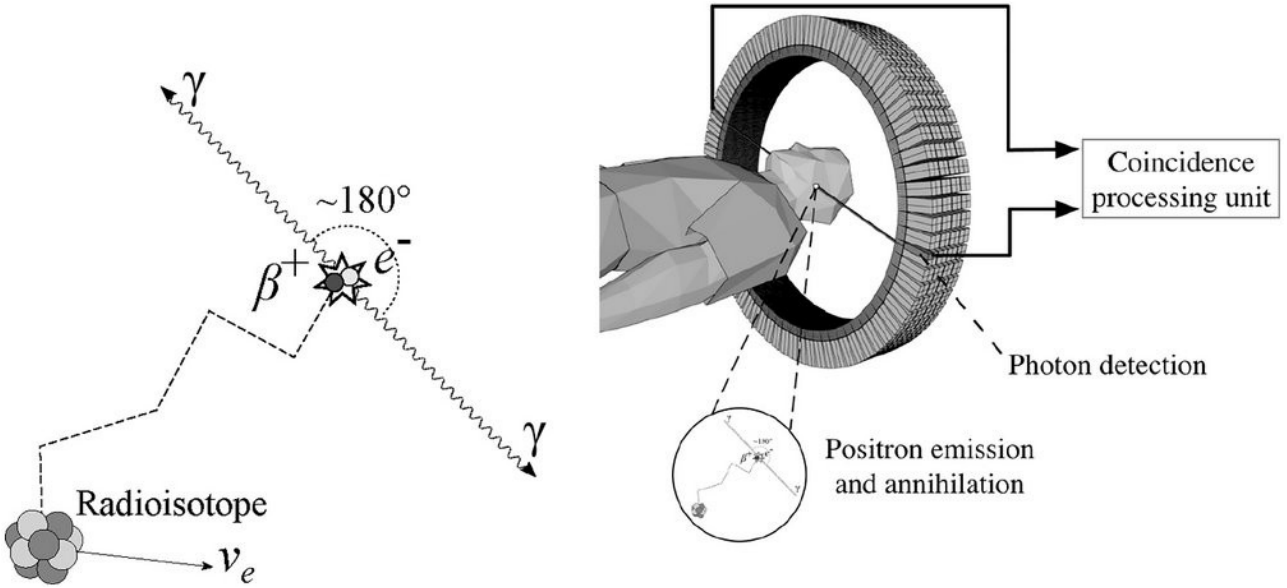
Republic of Poland

Co-funded by the European Union

Positron Emission Tomography

From annihilation photons to a reconstructed image

$$ {}^{A}_{Z}X \rightarrow {}^{A}_{Z-1}Y + {}^{0}_{+1}e + {}^{0}_{0}\nu $$



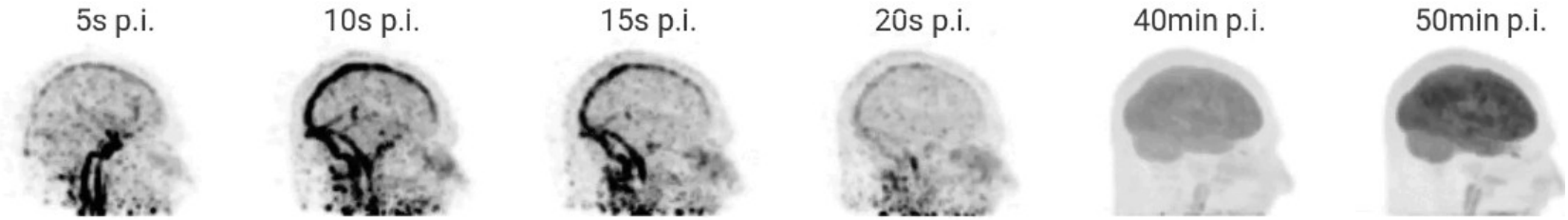
- Energies
- Times
- Positions

Source: N. Belcari et al., Riv. Nuovo Cim. 2024

What PET Is For

CT and MRI show structure - PET shows metabolism

- Visualises metabolic processes, not just anatomical structures
- The standard tracer is FDG, a labelled glucose
- Main clinical use is oncology; also cardiology and neurology
- Different tracers image different processes, and a process can be followed in time



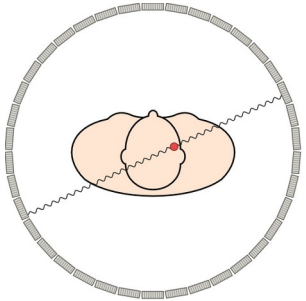
Distribution of [18F]FDG in a human brain over time - the tracer moves from the vessels into the tissue

Adapted from: Georg Schramm, KU Leuven

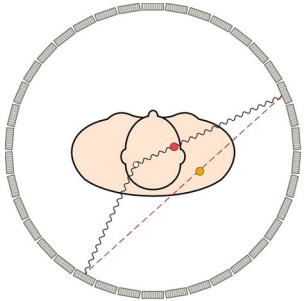
Coincidence Events: True and Background

Scatter and accidental events degrade the image

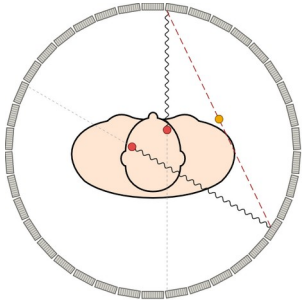
- True



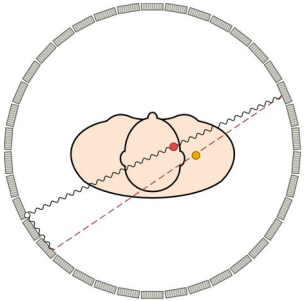
- Scatter

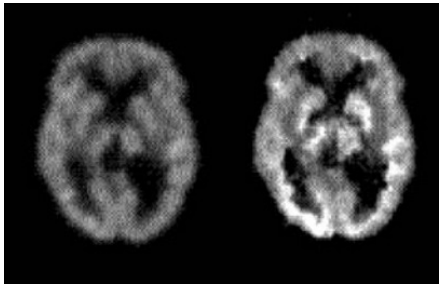


- Accidental



- Detector Scatter





- Left: uncorrected image
- Right: after correction - the contrast the background takes away

Why Machine Learning at the Event Level?

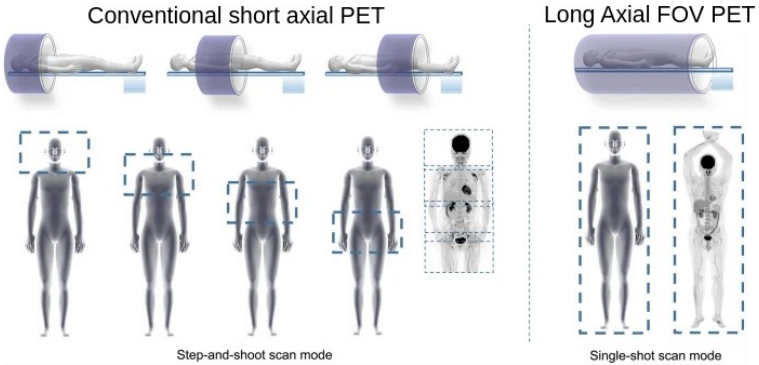
Corrections and ML alike work on images, not on single events

- Standard PET correction (single scatter simulation / delayed-window) estimate how much background sits in each bin – single events are never rejected
- Machine learning in PET usually does the same: an uncorrected image or sinogram goes in, a corrected one comes out
- A threshold would not separete it: a scattered photon can be detected with the same energy and timing as a true one

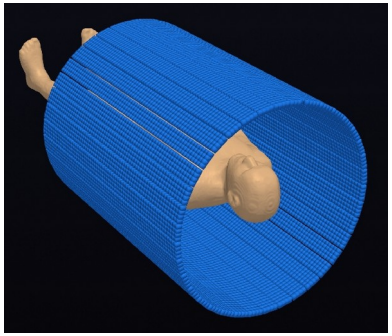
This work: per-event classification of RAW list-mode coincidences – a pre-reconstruction filter

Long Axial Field-of-View PET

Siemens Biograph Vision Quadra and simulated data



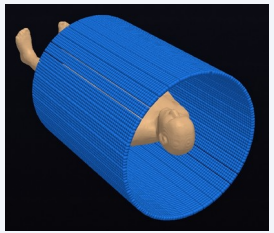
- **Siemens Biograph Vision Quadra PET/CT Scanner**



- 1060 mm axial FOV, 700 mm transaxial FOV
- 3.2 × 3.2 × 20 mm crystal dimensions
- 243,200 total number of crystals
- Energy window of 435 – 585 keV
- Coincidence time window of 4.7 ns

Pipeline

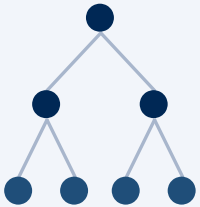
Monte Carlo gives us both the data and the truth to label it with



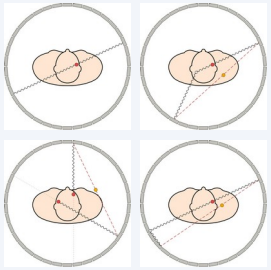
GATE Monte Carlo
Biograph Vision Quadra
50 MBq of [18F]FDG, 10 s acquisition

Post-processing
4-class labelling
from the Monte Carlo truth

Per-event features
Numbers describing
each coincidence



Classifier
trained on labelled
coincidences

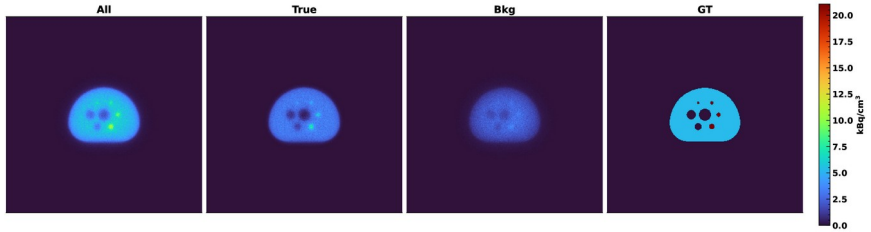


One class per event
true / phantom scatter /
detector scatter / accidental

Simulation is what makes this possible: it gives us the coincidences and, for each one, the interaction history that tells us what it really was.

Simulated Data and Ground-Truth Labels

GATE Monte Carlo and class labeling



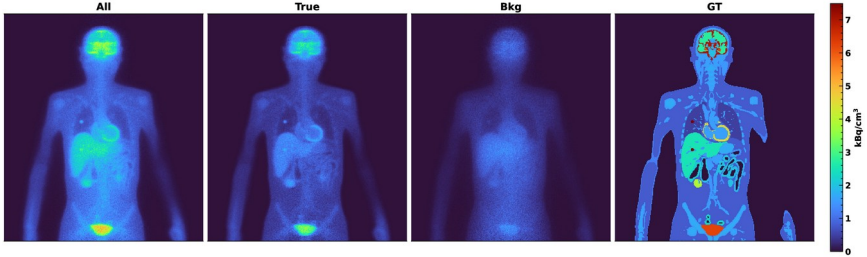
All **100%**

True **46.8%**

Background **53.2%**
phantom sc. **25.2%**
detector sc. **9.8%**
accidental **18.2%**

Ground truth reference

NEMA IEC phantom (QA standard)



All **100%**

True **49.0%**

Background **51.0%**
phantom sc. **24.5%**
detector sc. **10.2%**
accidental **16.2%**

Ground truth reference

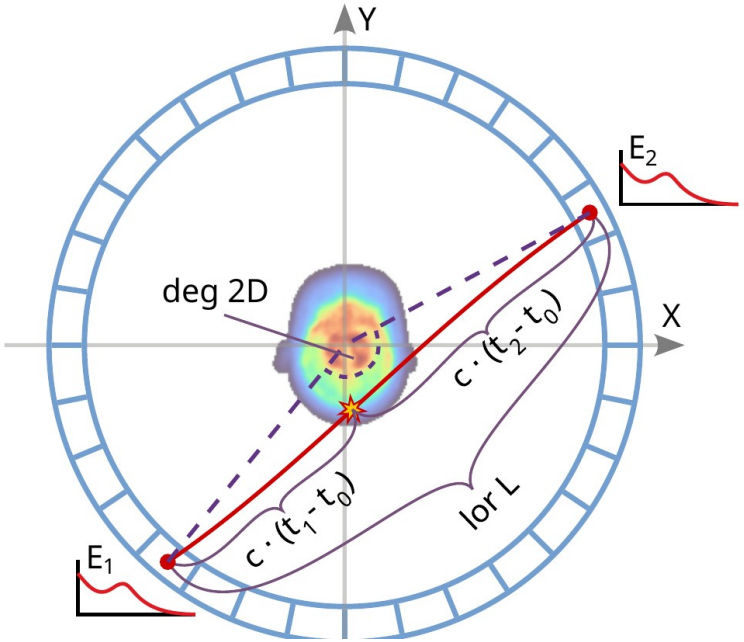
XCAT phantom (anthropomorphic)

- GATE 9.4.1 Monte Carlo of the Biograph Vision Quadra – Monte Carlo truth known for every coincidence
- Post-processing: time smearing to CRT 219 ps, hit positions discretised to crystal centres, geometric cut ($r < 35$ cm, $|z| < 53$ cm)
- Hierarchical 4-class labeling from the MC truth: True / Phantom Scatter / Detector Scatter / Accidental [Obara et al., IEEE EUROCON 2025]
- About half of the accepted events are still background

Input Data: Per-Event Features

Two feature sets: 4F (four features) and 6F (six features)

- 4F – Four-Feature set (geometry-independent):
  - AF – attenuation integral along the LOR
  - dt – photon time difference
  - eSum, eDiff – energy sum and difference
- 6F – Six-Feature set = 4F + topology:
  - lorL – length of the line of response
  - deg2D – 2D opening angle



Models and the Cross-Phantom Test

XGBoost, AdaBoost, feed-forward NN – Bayesian hyperparameter search

- Gradient-boosted trees: XGBoost (best: depth 9, 175 trees), AdaBoost
- Feed-forward NN: 2 hidden layers × up to 300 neurons, dropout rate optimised to 0, batch ≈500 (Bayesian-optimised)
- scikit-optimize Bayesian search: 240 iterations, 3-fold stratified CV, per phantom
- Trained on 3M sampled events; inference on the full ≈100M sample

Cross-phantom generalisation test: train on one phantom, evaluate on the other (e.g. trained on NEMA IEC → tested on XCAT). The model must survive a geometry it has never seen.

How Good Is the Model? Level 1: Global Metrics

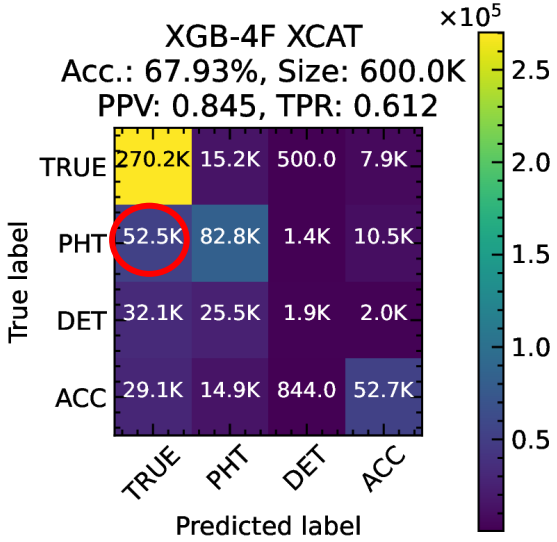
One number per model – convenient, but is it informative?

Model (XCAT eval)	Trained on	Accuracy
XGB-6F	XCAT	69.1%
XGB-4F	XCAT	67.9%
XGB-6F-X	**NEMA IEC**	**56.1%**
XGB-4F-X	**NEMA IEC**	**64.7%**

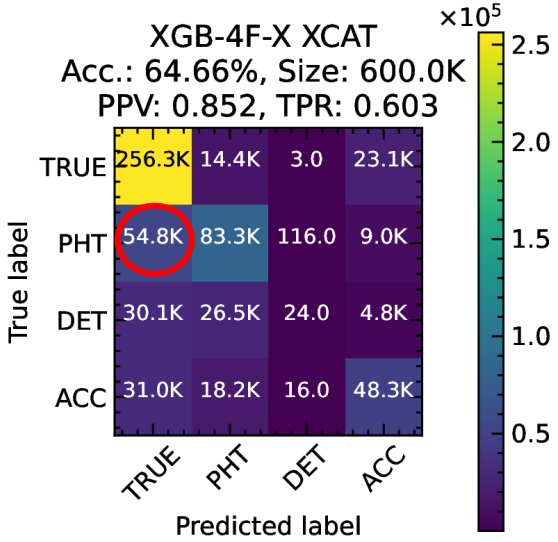
- Same-phantom: 6F slightly better (ADA/NN within ≈0.5 pp of XGB)
- Cross-phantom: 6F collapses (−13 pp) – topology overfits the geometry
- 4F drops only −3.3 pp
- A single number says nothing about **where** or **why** the model fails

Level 2: Confusion Matrices

Which classes suffer



XGB-4F: trained on XCAT (Acc 67.9%)

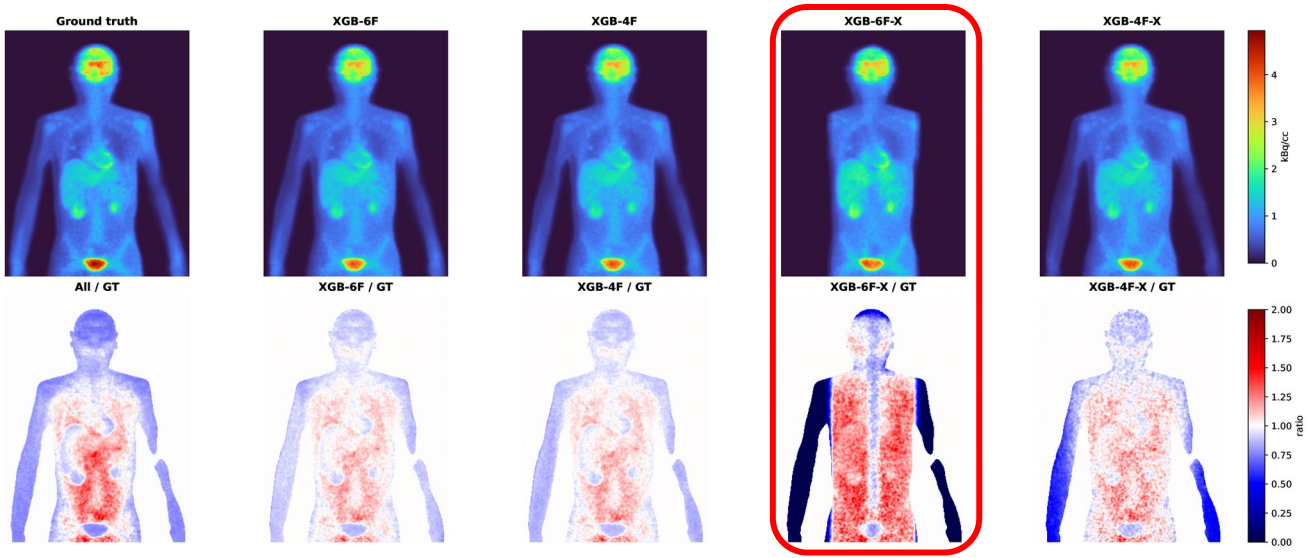


XGB-4F-X: trained on NEMA IEC (Acc 64.7%)

- Cross-phantom: ≈3× more true events misclassified as accidental; detector scatter is the class the model essentially cannot find - 3% recovered on its own phantom, almost none across phantoms
- We now know **which** classes suffer – but still not **where** in the body

Level 3: Impact on the Reconstructed Image

Does it show in the image? – naive TOF reconstruction, ratio to GT



Top row: naive TOF reconstruction (each event placed along its LOR → weighted 3D histogram); left panel = ground truth.
Bottom row: per-voxel ratio to ground truth – blue = underestimated, red = overestimated; 'All / GT' = unfiltered data, before any classifier.

- XGB-6F-X: the arms vanish - the NEMA training cylinder has nothing at that radius, so the topology features never saw such events
- XGB-4F-X: the same effect, weaker - the whole −3.3 pp comes from this one region, not from uniform noise a global number could hide

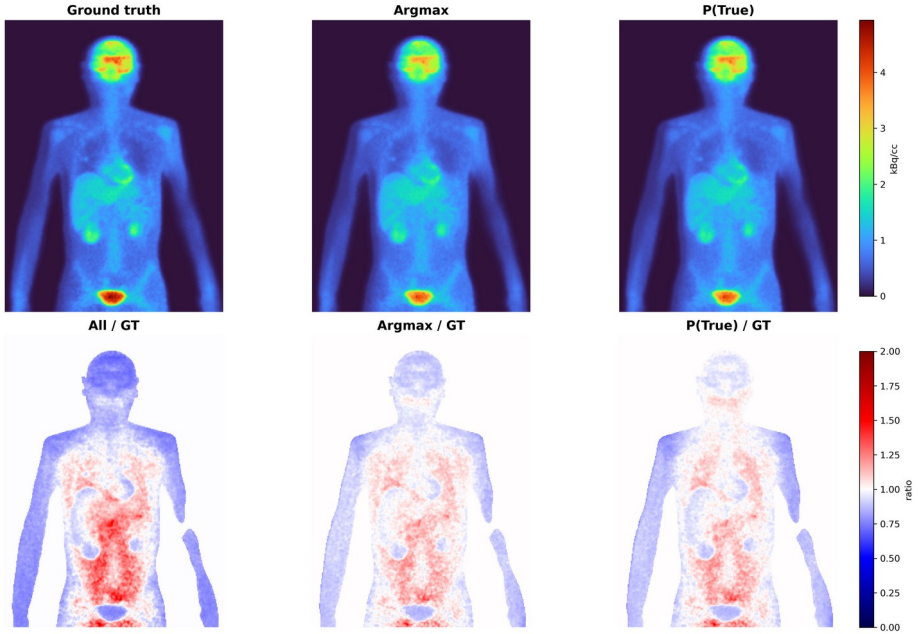
Soft Probability Weighting

Each event contributes in proportion to P(True)

- So far we used the classifier through argmax only – keep or reject
  - an event with P(True) = 0.51 counts exactly as much as one with 0.99
- Hard weighting keeps or drops each event: $$w_i^{\text{hard}} = \mathbf{1}\left[\arg\max_c p_c(x_i) = \text{True}\right]$$
- Soft weighting uses the full true-class probability: $$w_i^{\text{soft}} = \hat{p}_{\text{True}}(x_i) \in [0, 1]$$
- An event the classifier is unsure about is not discarded – it enters the image with weight P(True)

Result: Soft Weighting Lowers the Spatial Error

17% lower voxel-wise MSE than hard filtering (equal-activity comparison)



Top row: reconstructed activity (kBq/mL) – ground truth, hard top-1, soft P(True).
Bottom row: per-voxel ratio to ground truth – blue = underestimated, red = overestimated.

Reconstruction pipeline – identical for all three images

1. each event placed along its LOR using TOF → weighted 3D histogram
2. Gaussian smoothing + 2×2×2 median filter
3. ÷ attenuation+sensitivity map → activity (kBq/mL)
4. normalise to equal total activity
5. mask to the phantom volume

Voxel-wise MSE vs ground truth	
Hard top-1	5.44×10^{-3}
Soft P(True)	4.50×10^{-3}
Reduction	**−17.4 %**

- Equal total activity → the metric isolates the spatial error
- Before normalisation the gap looks larger (67–84 %), but hard filtering also keeps far fewer events – gap mostly reflects the difference in event counts, not spatial accuracy

XGBoost 4-class model, 6 features, ≈100M events

From Validation to Uncertainty

The maps so far all needed the truth - conformal prediction does not

- Everything so far - the spatial maps, the ratio to ground truth - needed the Monte Carlo truth
  - patient data does not have it
- So the question is:
  - can we say where the model is unsure, without any labels at all?

Conformal prediction

Labels are needed once, and only to fix one number - a score threshold.

After that the method needs none

Calibrating the Threshold

Split conformal prediction: sort, add down to the truth, take a percentile

Split conformal prediction: Angelopoulos & Bates 2021. APS score: Romano et al., NeurIPS 2020.

1. Take 1,000,000 events the classifier never saw with their ground-truth class.
2. For each: sort its class probabilities and add them up from the top, stopping at the class that was actually true. That sum is its score.
3. Sort all 1,000,000 scores and take the one 90% of the top the list - its score value is the threshold.

sorted probabilities	score
0.44 0.32 0.23 0.01	**0.44**
0.71 0.20 0.08 0.01	**0.71**
0.55 **0.36** 0.08 0.01	**0.55 + 0.36 = 0.91**

bold = the class that was actually true

where all 1,000,000 scores land

90% of the scores fall in the blue part.

Spending the Threshold: Prediction Sets

Without the true class, the threshold decides where to stop

one event, no label

P(True)	0.63
P(Phantom Sc.)	0.29
P(Accidental)	0.07
P(Detector Sc.)	0.01

add from the top until the sum crosses q

True	0.63
+ Phantom Sc.	**0.92**

0.92 crosses q ≈ 0.90: stop, and keep the class that crossed

prediction set C(x)

90%: q ≈ 0.90
C(x) = { True, Phantom Sc. }

95%: q ≈ 0.95
C(x) = { True, Phantom Sc., Accidental }

- Confident event → C(x) = { True } alone (a singleton)
- Ambiguous event → a bigger set
- Ask for more confidence → q rises

Guarantee (coverage): over all events, the true class is in C(x) at least 90% of the time.

Conformal Bounding: Events → Voxels

Turning per-event prediction sets into a per-voxel uncertainty map

- Count two groups of the events in each voxel:
  - exactly {True} → only True is left
  - True in the set → signal not excluded
- As fractions of the voxel's events:
  f_{min} (only True left) ≤ f_{max} (True not excluded)
  the difference between them is Δf
- mean set size |C| in the voxel – average number of classes kept

$$A_{\min}(v) = \sum_{i \in v} \mathbf{1}[C(x_i) = \{\text{True}\}]$$

$$A_{\max}(v) = \sum_{i \in v} \mathbf{1}[\text{True} \in C(x_i)]$$

$$f_{\min}(v) = \frac{A_{\min}(v)}{N(v)}, \qquad f_{\max}(v) = \frac{A_{\max}(v)}{N(v)}$$

$$\Delta f(v) = f_{\max}(v) - f_{\min}(v)$$

Example – one voxel with 1000 events
100 events with set {True} → f_{min} = 0.10
700 events with True in the set → f_{max} = 0.70
Δf = 0.60 → most of this voxel is unresolved

Δf is the voxel's ambiguity: 0 → fully resolved, large → unreliable region.

Result 2: Choosing the Confidence Level

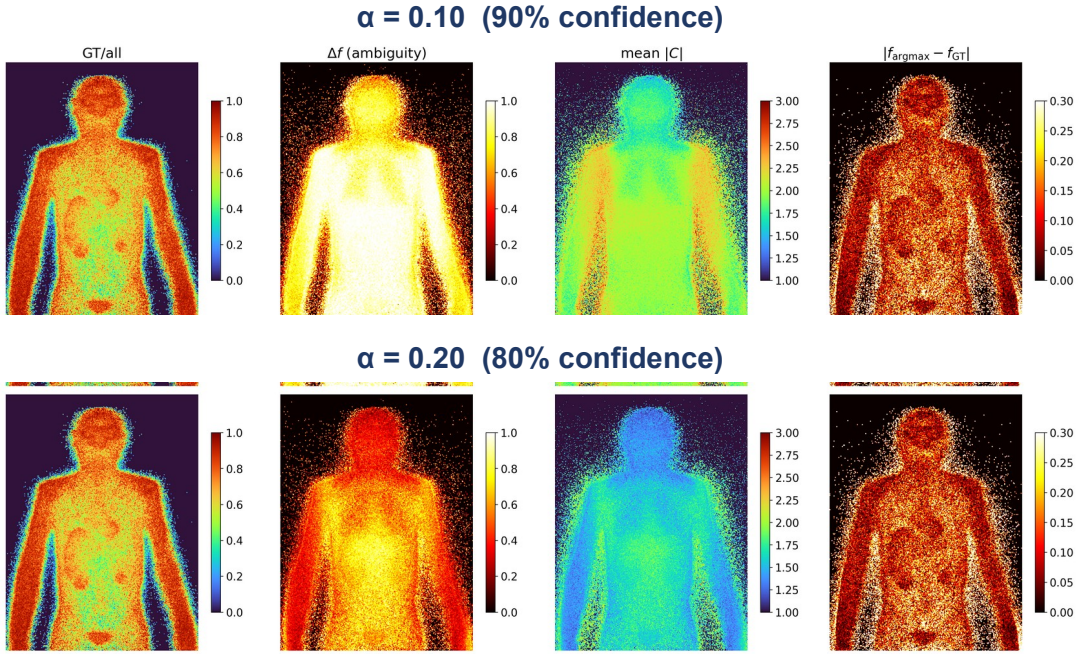
Ask for more confidence and the sets grow - until the map goes flat

α (confidence)	C(x) = {True} % of events	mean set size |C|	ambiguity Δf phantom mean
0.05 (95%)	0.3	2.40	1.00
0.10 (90%)	7.8	1.99	0.91
0.15 (85%)	23.4	1.74	0.73
0.20 (80%)	35.9	1.57	0.57
0.30 (70%)	51.7	1.35	0.34
0.40 (60%)	60.7	1.18	0.16

- At 95% confidence almost no event is left with only True – the sets stay large and Δf saturates
- Lower α, more confidence → larger sets, fewer events with C(x) = {True}
- Pick α so that Δf still varies across the image

Voxel-wise Uncertainty Maps

Ambiguity Δf and mean set size |C| – no ground truth needed



- Columns: GT signal fraction | Δf | mean |C| | argmax (top-1) error. Torso ambiguous, head and arms confident
- mean |C| already shows where the model is unsure at α = 0.10, where Δf is saturated; where CP says ‘certain’, the error is low

Conclusions

Classify, validate spatially, weight - and say how sure we are

Event classification

Klimaszewski et al., arXiv:2607.25732

- Classical classifiers on a handful of per-event features remove most of the background
  - what survives is the background that looks true
- Validation must go beyond a global number
  - a small global drop can hide a whole region failing
  - topology features overfit the training geometry
- Endpoint of this route: argmax filtering, keep or reject

Beyond hard classification

Obara et al., SPA 2026 (accepted)

- Use the probabilities, not just argmax
  - soft P(True) weighting lowers the spatial error against hard filtering
- Conformal prediction says where not to trust the corrections
  - per-voxel ambiguity Δf and mean set size $|C|$ - without any ground truth

Future work

- CNN classifiers on a spatial encoding of each coincidence, and a wider search over model architectures
- Weighting inside iterative reconstruction algorithms

Thank you!

European Funds
for Smart Economy

Republic
of Poland

Co-funded by the
European Union

This work was completed with resources provided by the Świerk Computing Centre at the National Centre for Nuclear Research. The study was partially supported by the project „IMPET – Industrial Multiphoton PET Tomography FENG.02.02-IP.05-0152/23” carried out within the FIRST TEAM FENG programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG).