Machine learning-based discrimination of bulk and surface events of germanium detectors for light dark matter detection

Peng Zhang

Department of Engineering Physics, Tsinghua University

中国锦屏地下实验室
China Jinping Underground Laboratory
清华大学・雅砻江流域水电开发有限公司

中国暗物质实验
China Dark matter EXperiment

COUSP 2024

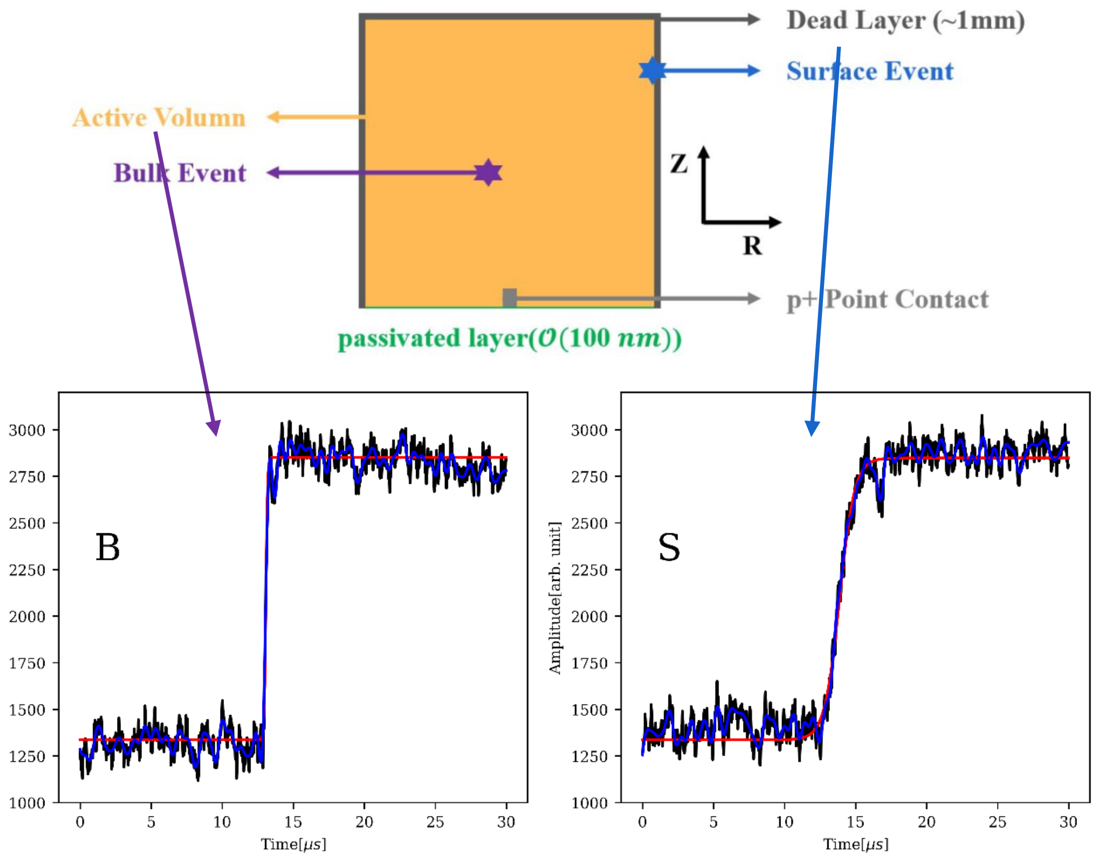
Outline

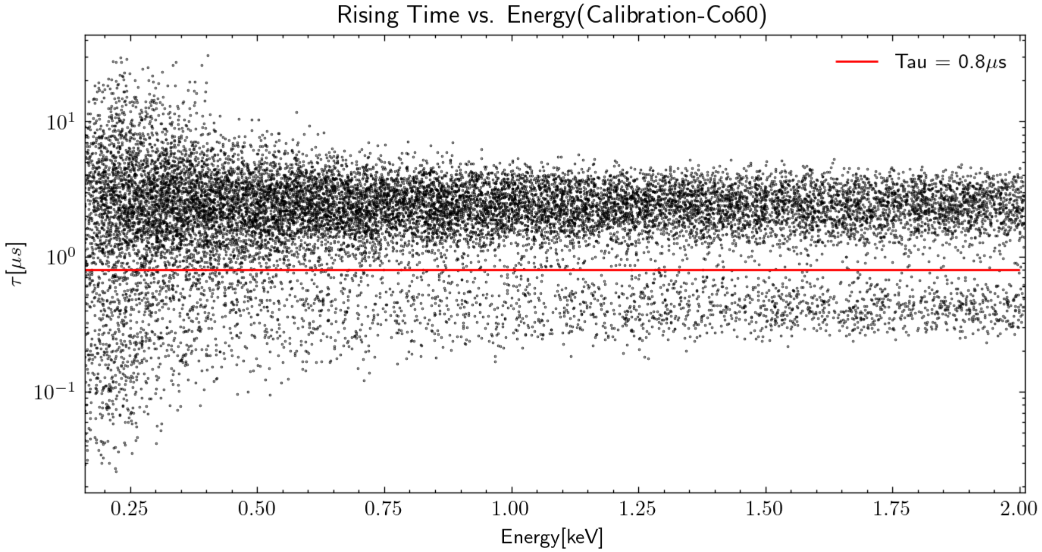
1. Introduction

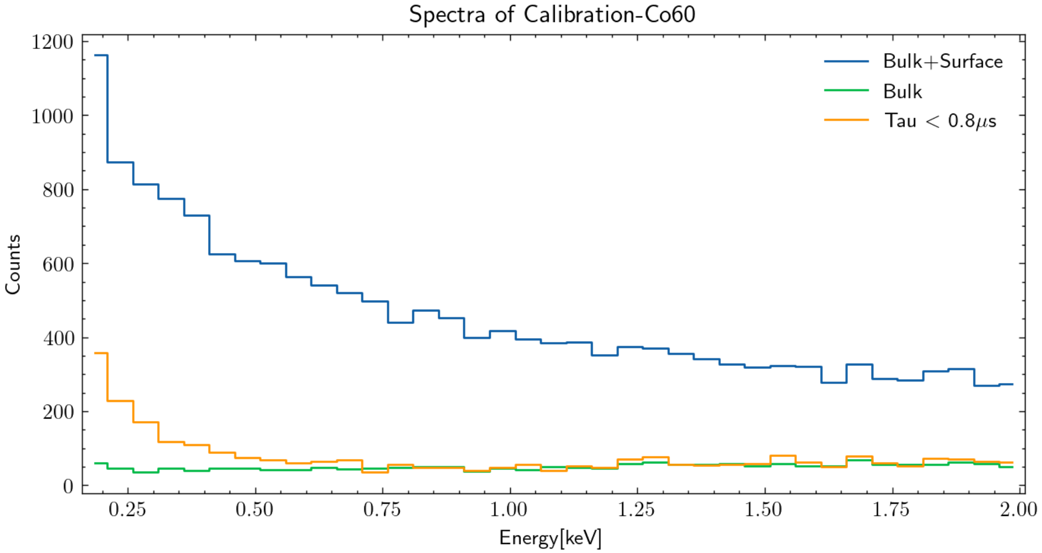
Introduction: Bulk & Surface Event

What are bulk and surface events?

The schematic diagram of a typical pPCGe detector







Introduction: ‘Band’ Method

Tsinghua University

Analysis Procedure

- B_m = counts ($\tau < 0.8$ μs), S_m = counts ($\tau > 0.8$ μs)
- Two efficiencies of this cut:
  B-signal retaining(ϵ_{BS}) , S-background suppression (λ_{BS})
- Solving these equations, we can get the true Bulk and Surface counts(denoted by B_r, S_r).

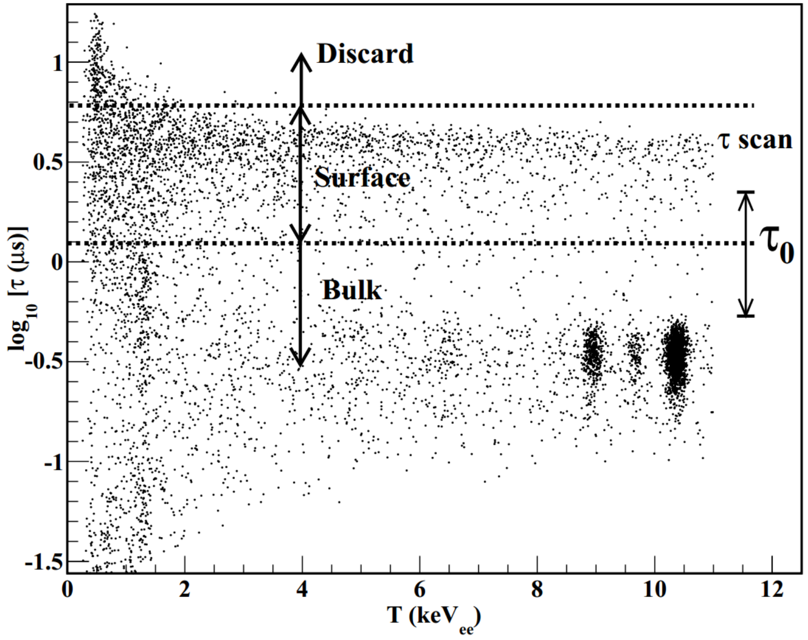
$$B_m = \epsilon_{BS} B_r + (1 - \lambda_{BS}) S_r$$
$$S_m = \lambda_{BS} S_r + (1 - \epsilon_{BS}) B_r$$
$$B_m + S_m = B_r + S_r$$

$$B_r = \frac{\lambda_{BS} B_m - (1 - \lambda_{BS}) S_m}{\epsilon_{BS} + \lambda_{BS} - 1}$$
$$S_r = \frac{\epsilon_{BS} S_m - (1 - \epsilon_{BS}) B_m}{\epsilon_{BS} + \lambda_{BS} - 1}$$

- Two efficiencies are calculated by comparing the Bm and the Br of different sources. The Br is obtained through Geant4 simulation

Summary

- Based on the cut of rising time, limited by fitting, and the choice of cut brings systematic error
- Too dependent on the simulation, which brings uncertainty, such as geometry, source term, etc



Introduction: ‘Ratio’ Method

Tsinghua University

Analysis Procedure

- After a thorough check, we find that the pdf of rising time is consistent between different sources term.
- Briefly, we consider an energy region $\mathrm{E} = [\mathrm{e}_0, \mathrm{e}_1]$, a rising time region $\tau_\mathrm{i} = [\mathrm{t}_\mathrm{i}, \mathrm{t}_{\mathrm{i}+1}]$. We consider two sample X. And we denote total counts in E is $\mathrm{N}_{\mathrm{X,i}}$, bulk counts in E is $\mathrm{B}_{\mathrm{X,i}}$($\mathrm{B}_X = \sum_i B_{X,i}$), surface counts in E is $\mathrm{S}_{\mathrm{X,i}}$ ($\mathrm{S}_X = \sum_i S_{X,i}$). We denote the probability of bulk events in τ_i is $\mathrm{P(B,i)}$, the probability of surface events in τ_i is $\mathrm{P(S,i)}$. Hence,

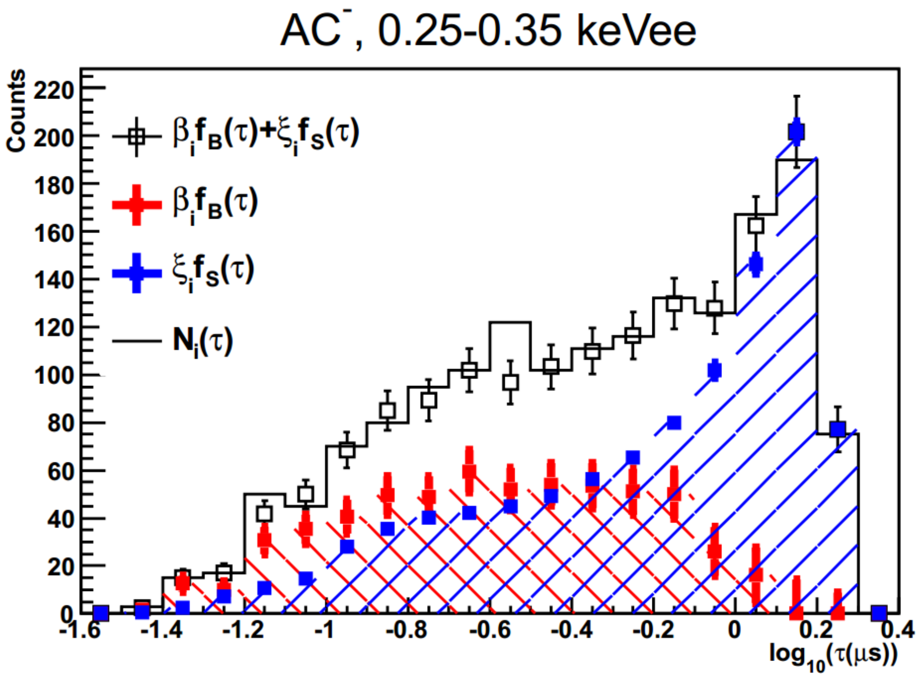
$$\begin{aligned} N_{X,i} &= B_{X,i} + S_{X,i} = P(B,i) \times B_X + P(S,i) \times S_X \\ N_{Y,i} &= B_{Y,i} + S_{Y,i} = P(B,i) \times B_Y + P(S,i) \times S_Y \end{aligned}$$

$$\Rightarrow \quad B_{X,i} = P(B,i) \times B_X = \frac{N_{Y,i} - \frac{S_Y}{S_X} \times N_{X,i}}{\frac{B_Y}{B_X} - \frac{S_Y}{S_X}}$$

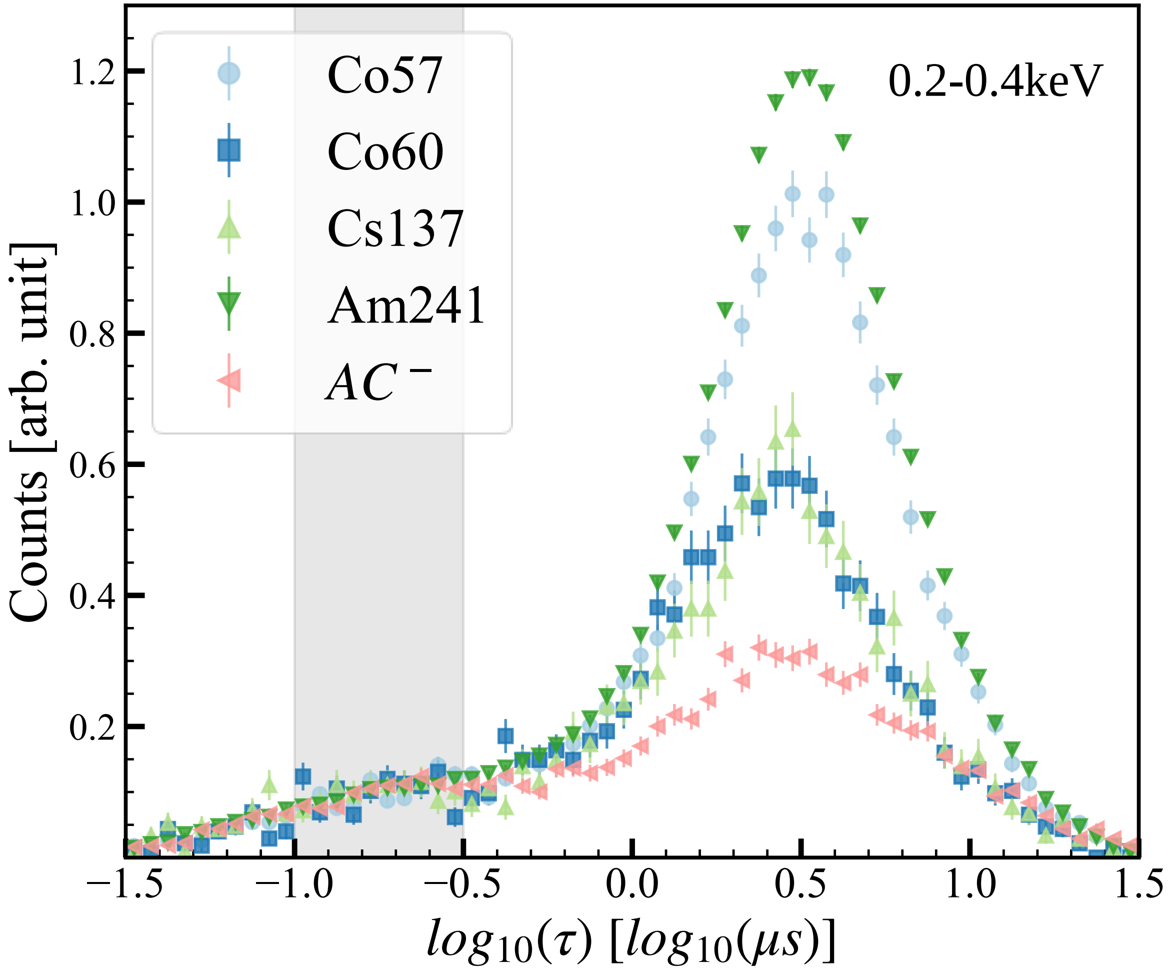
- S_Y, S_X is unknown, but in some τ region(say region C_S), we can obtain pure surface. And S_Y/S_X==$S_{Y,C_S}/S_{X,C_S}$. The same for bulk.
- With multiple iterations, we can get the accurate B_X, S_X, B_Y, S_Y.

Summary

- A statistical method (A binned analysis). Difficult to handle small sample.



Why using Machine Learning?

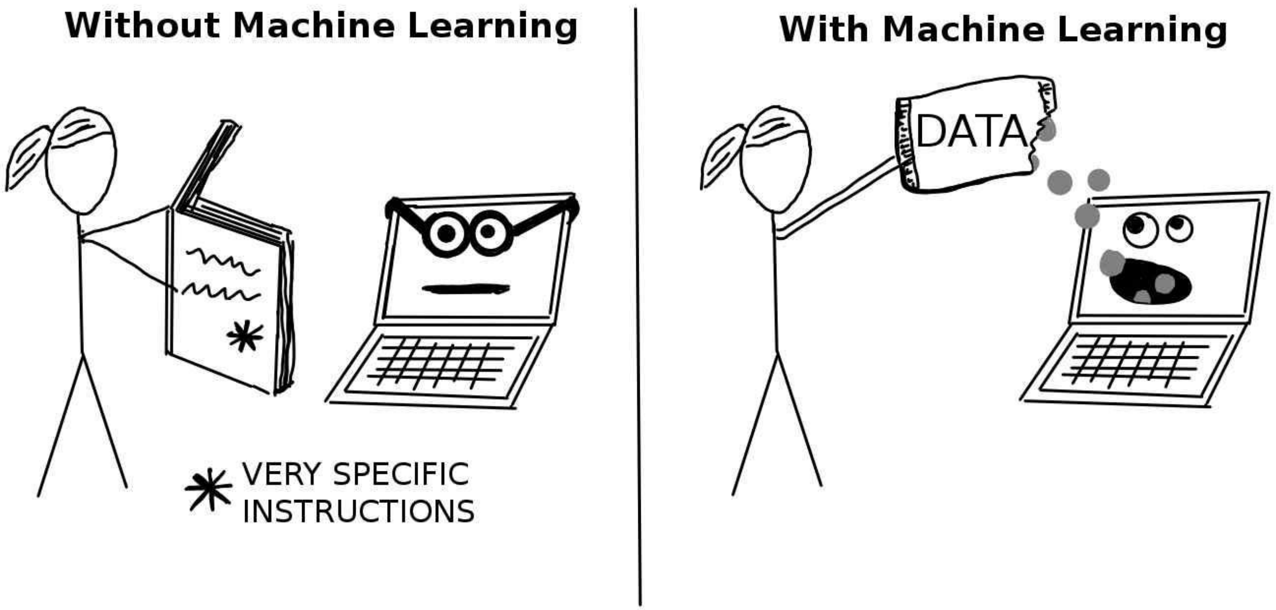


- Current features used for B/S discrimination is not good enough
- We want to find a map function to separate BE/SE further

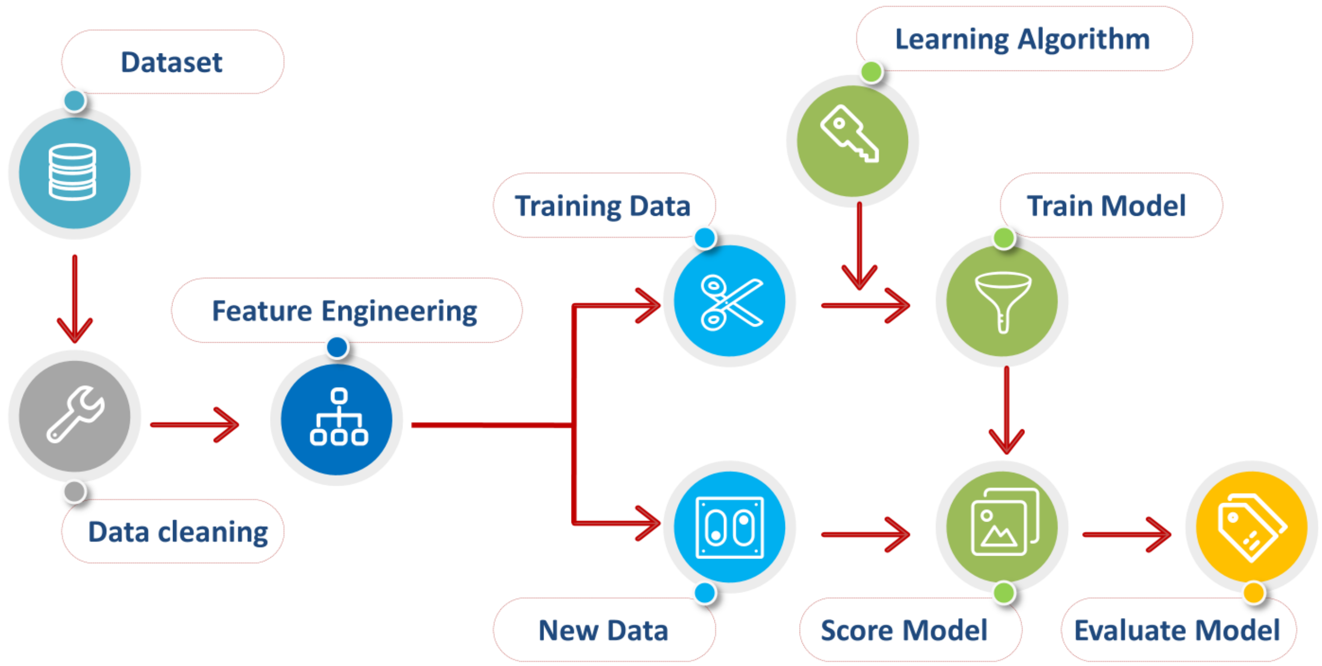
Introduction: Machine Learning

Why using Machine Learning?

- Able to automatically dig the information which is hard to find for human.
- We use supervised ML (with labels)



<u>Work Flow of Supervised Machine Learning-Based Model</u>



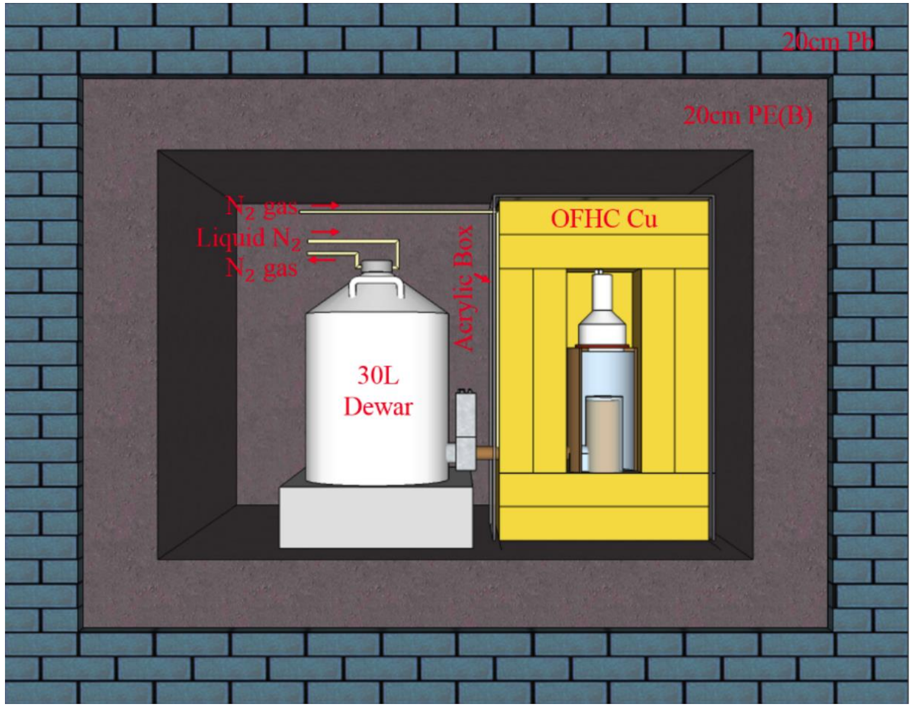
T: True
F: False
P: Positive
N: Negative

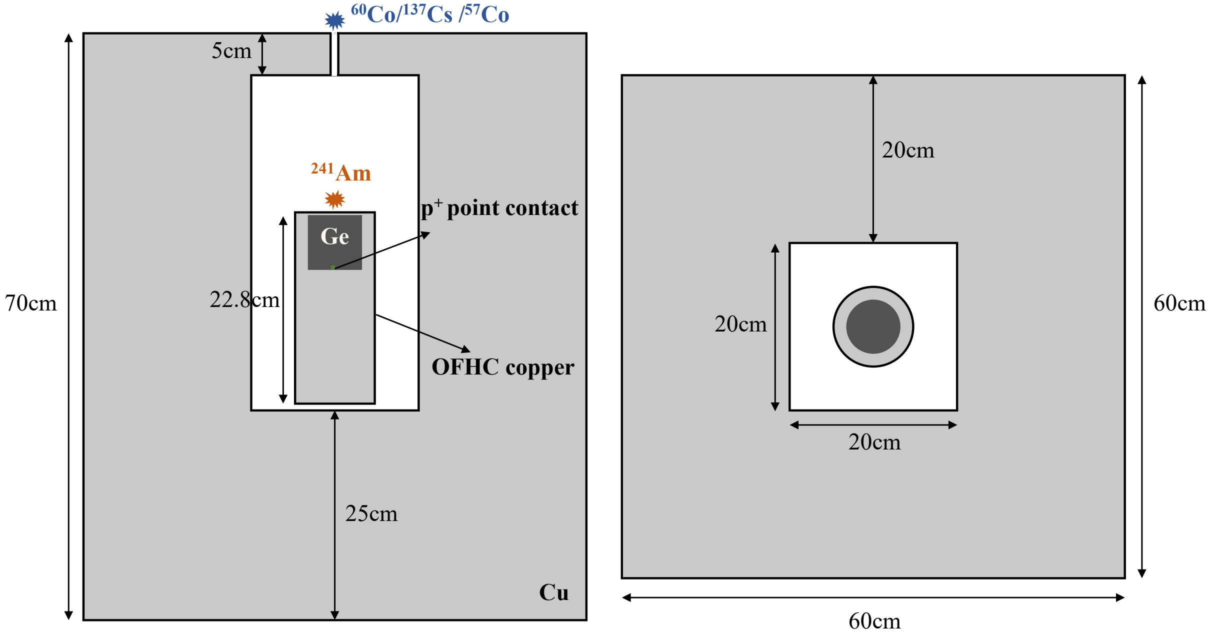
Confusion Matrix		Assigned class	
		Positive	Negative
Real class	Positive	TP	FN
	Negative	FP	TN

2. Experimental setup

Experimental Setup

- A pPCGe detector installed at China Jinping underground Laboratory (CJPL)
  - Crystal size: 62.3 mm × 62.3 mm (Φ × H)
  - Dead layer thickness: 0.88±0.12 mm





Experimental Setup : Dataset Construction

Tsinghua University

***Bulk events:* Positive—label 1**

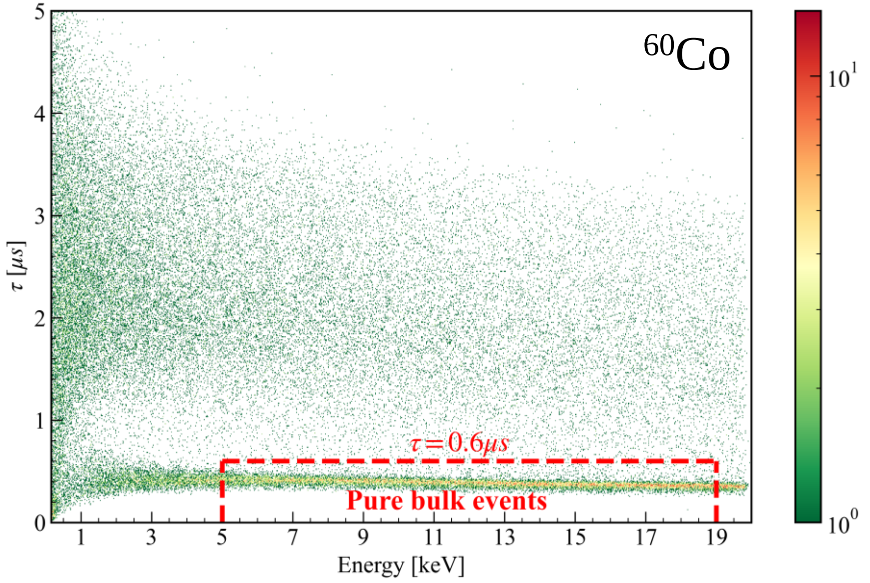
- Calibration pulses at high energy region (^{60}Co/^{137}Cs/^{57}Co), scaled to lower energy, then added with noise. Noise of original pulse is negligible after scale

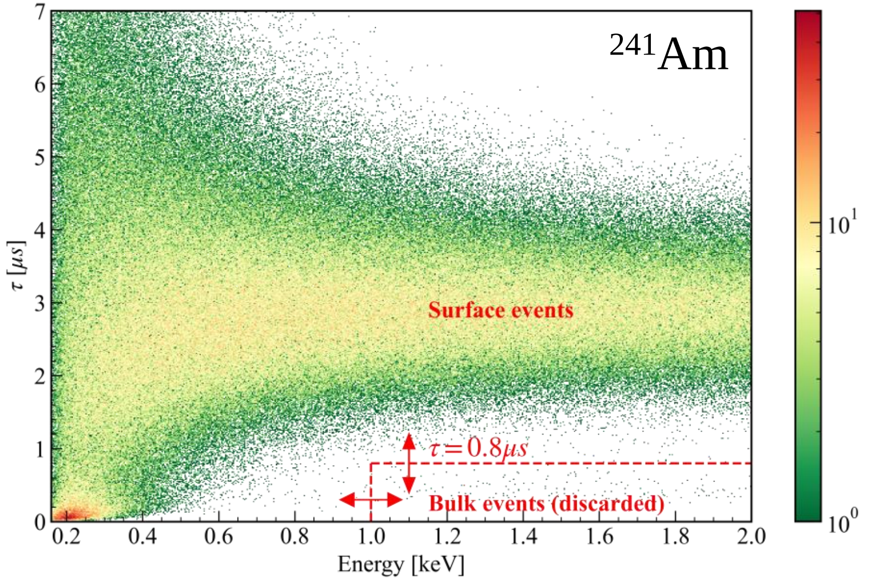
***Surface events:* Negative—label 0**

- ^{241}Am Events, as pure surface events

Dataset Splitting:

- Train-set: Train and tune the classification algorithms
- Correction-set: Calculation of efficiencies
- Test-set: Independent verification



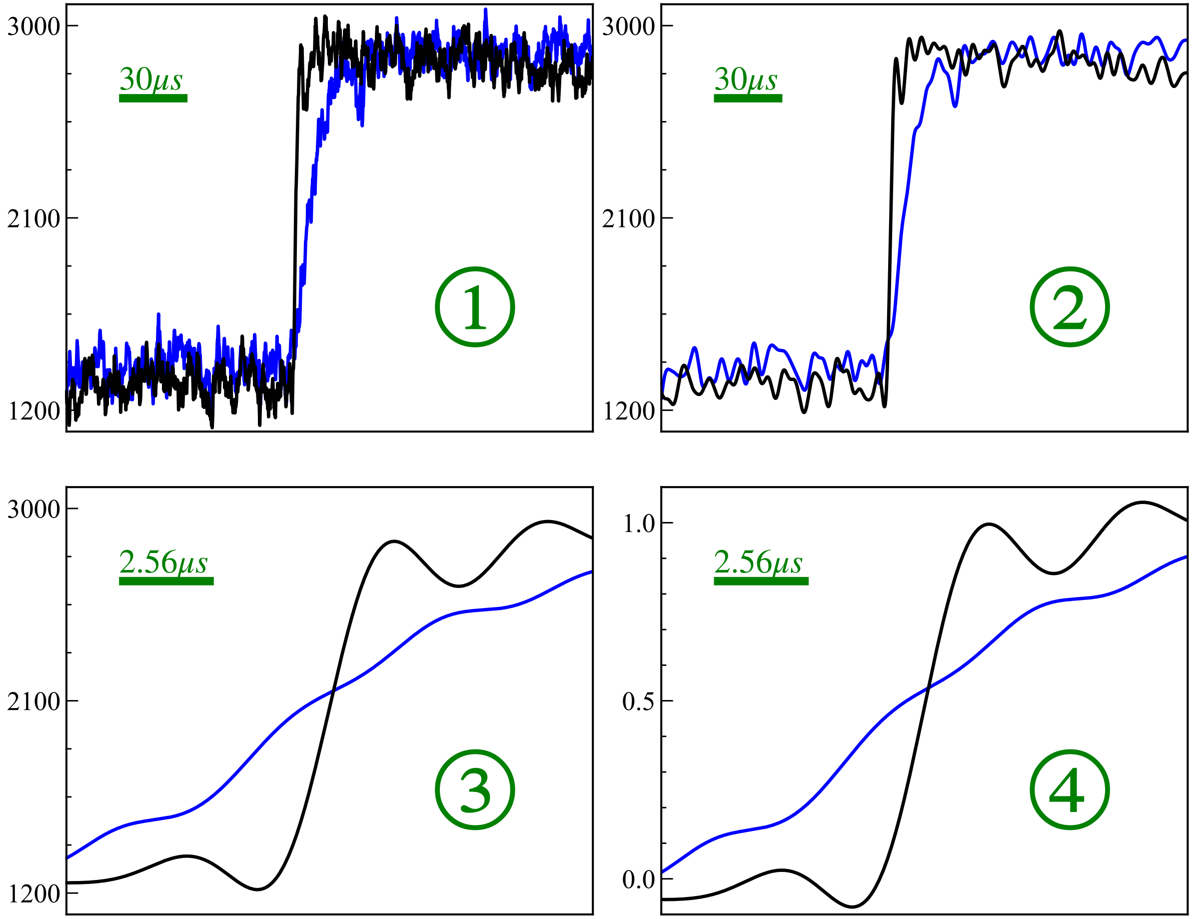


3. Method

Method: Pulse preprocessing

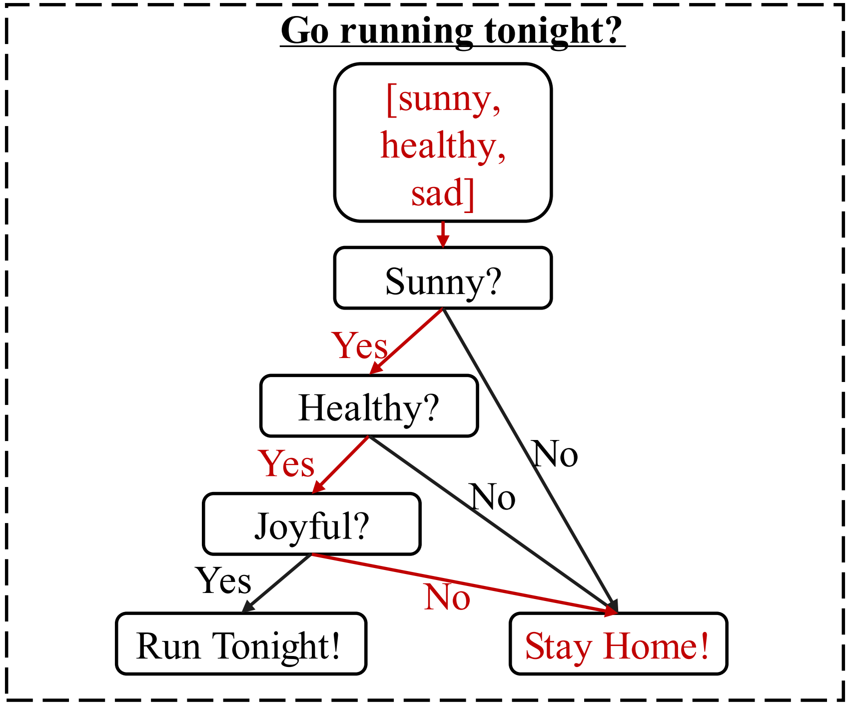
Pulse preprocess procedure:

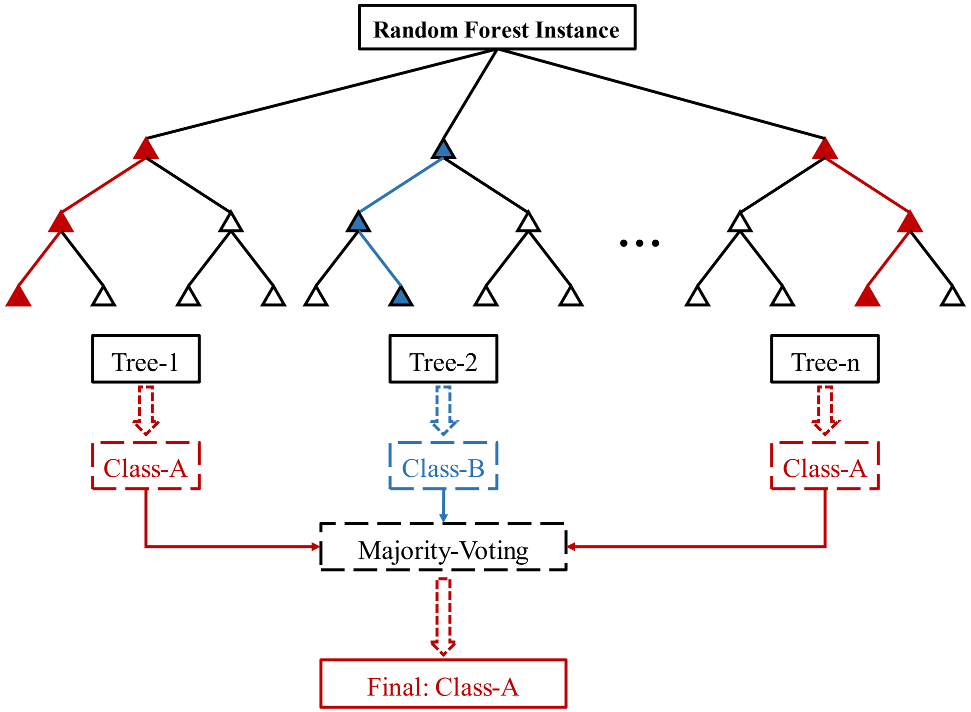
- Filtration and fitting
- Alignment and trim
- Normalization



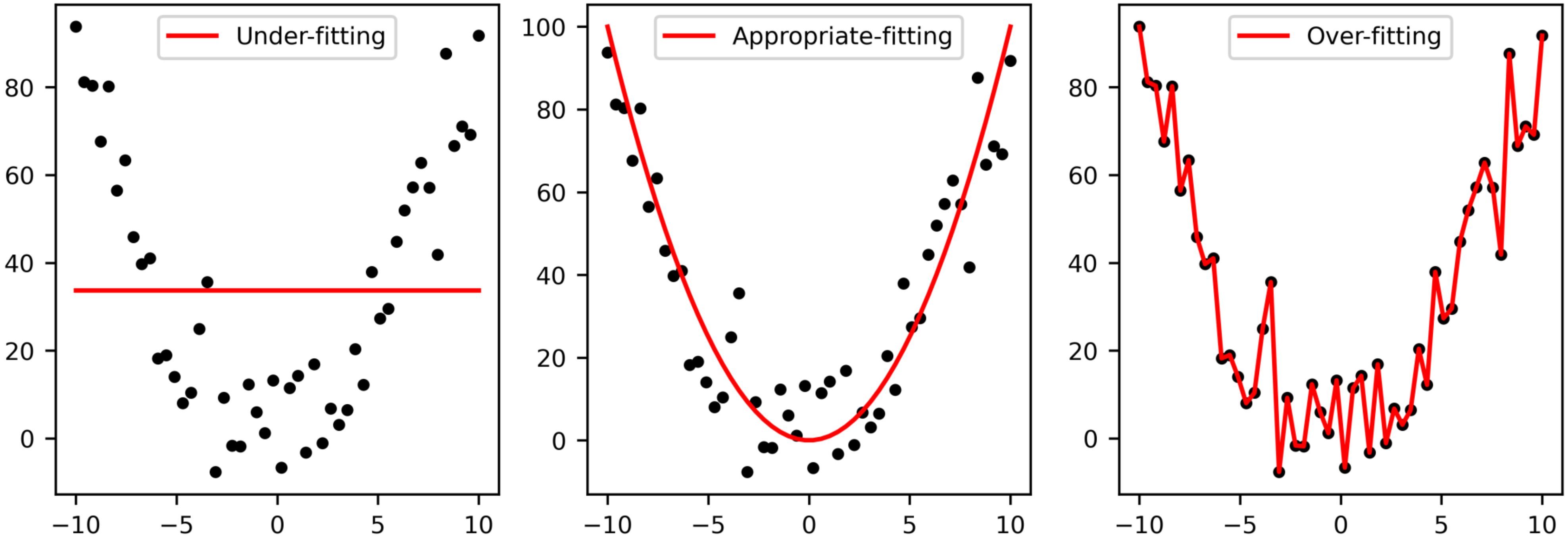
Method: Algorithm and training

Random Forest = Cart (Classification And Regression Tree) +Bagging(Bootstrap AGgregation)





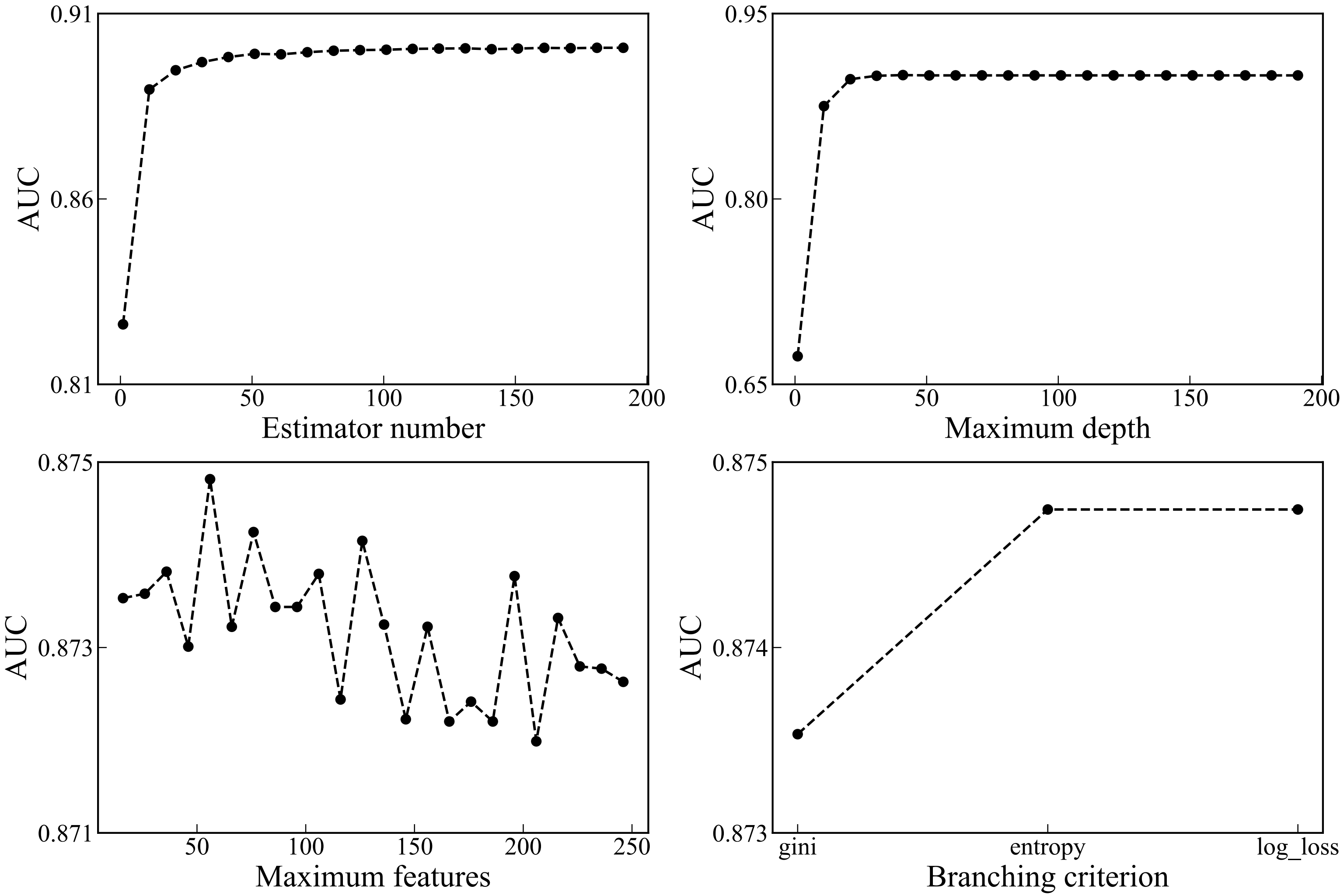
Principle of training and tuning



Method: Algorithm and training

Training and Tuning process

Tuning criterion: AUC score



Method: Correction-‘Band’ method

Tsinghua University

‘Band’ method (For a specific cut x):

Bulk Retaining: ϵ_{BS}

Suface Suppressing: λ_{BS}

$Br \rightarrow True\ Bulk\ Counts,\quad Sr \rightarrow True\ Surface\ Counts$

$B_m, S_m \rightarrow Classifier\ output$

$$B_m = \epsilon_{BS} \times B_r + (1 - \lambda_{BS}) \times S_r$$

$$N = B_r + S_r = B_m + S_m$$

Correction Formula: $B_r(x) = \frac{B_m(x) - (1-\lambda_{BS}(x)) \times N}{\epsilon_{BS}(x) + \lambda_{BS}(x) - 1}$

Using the BE/SE output PDF

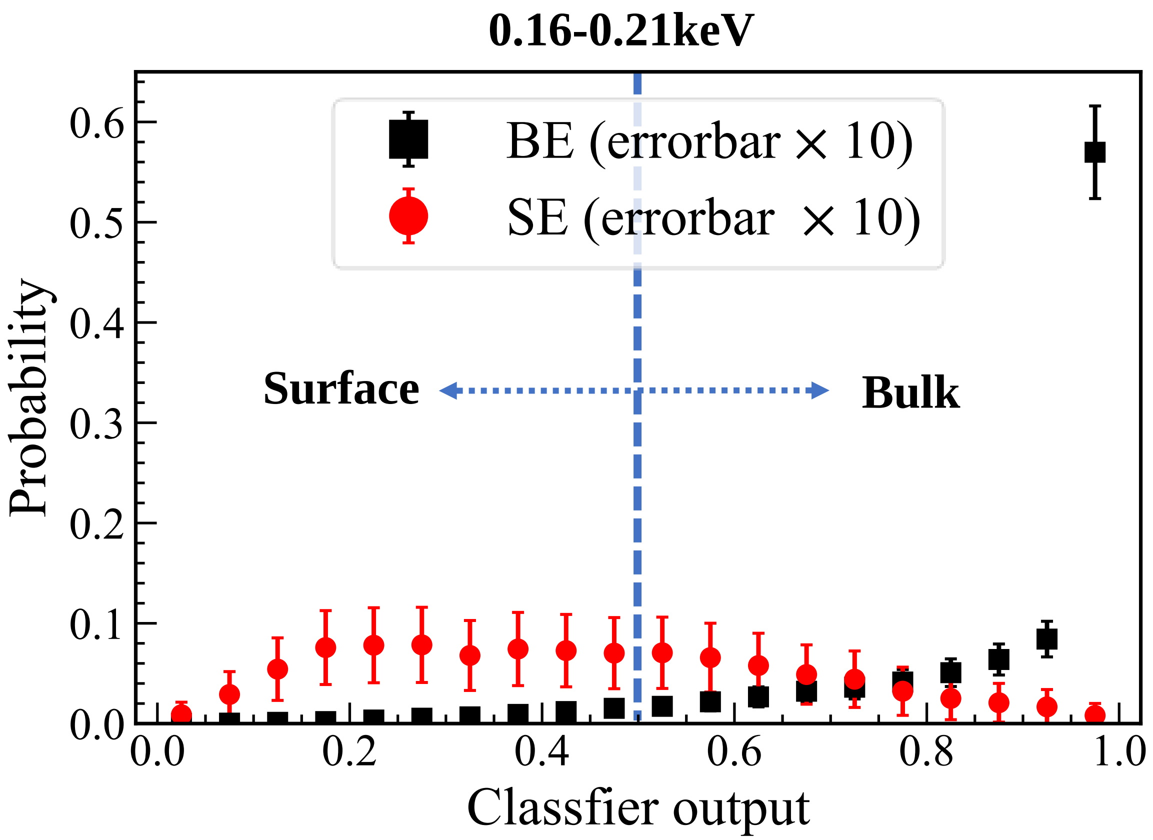
Bulk, Surface distribution: $f_B(x), f_S(x)$

Sample distribution(total counts: N): $f(x)$

$$f(x) = f_B(x) \times B_r + f_S(x) \times S_r$$

$$B_m = N \int_{cut}^{1} f(x)dx,\ N = N \int_{0}^{1} f(x)dx$$

$$\epsilon_{BS} = \int_{cut}^{1} f_B(x)dx,\ \lambda_{BS} = \int_{0}^{cut} f_S(x)dx$$



Not making full use of all information!
(Only use two bins)

Method: Correction-Integral correction

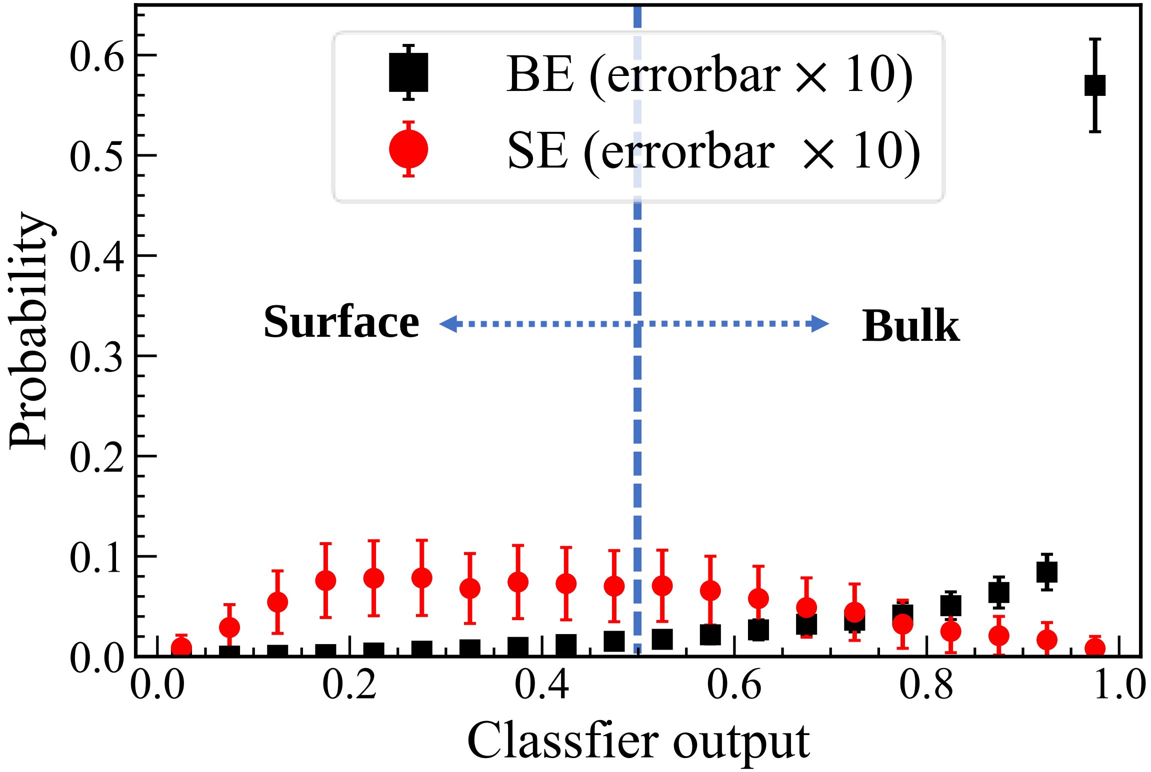
Integral Correction

$$B_r(x) = \frac{B_m(x) - (1 - \lambda_{BS}(x)) \times N}{\epsilon_{BS}(x) + \lambda_{BS}(x) - 1}$$

$$B_r = \int_0^1 B_r(x)dx \approx \frac{1}{N_x}\sum_x B_r(x)$$

$$\sigma_{B_r} = \sqrt{\frac{1}{N_x^2}\sum_i\sum_j cov(i,j)}$$

0.16-0.21keV



Method: Correction-Integral correction

Tsinghua University

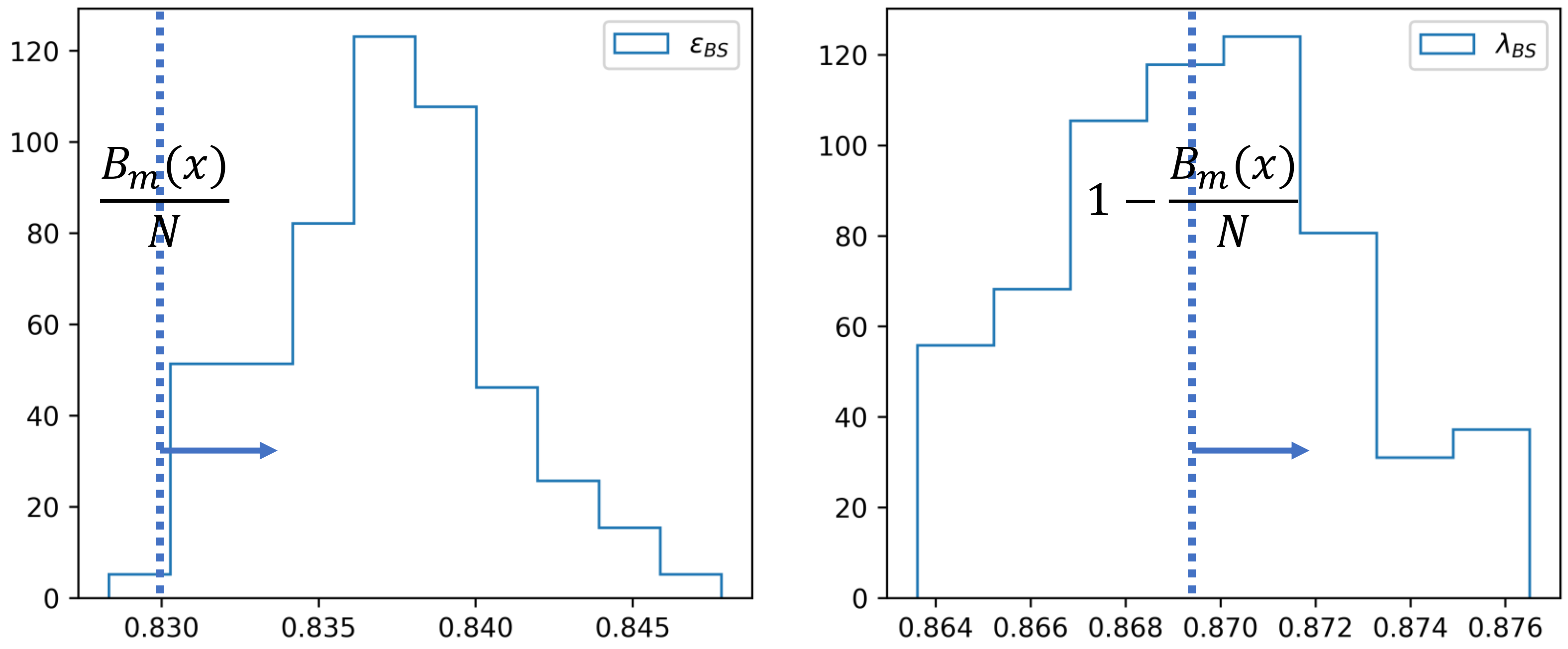
Bayesian inference:

B: 0~N

$$B_r = \frac{B_m - (1-\lambda_{BS}) \times N}{\epsilon_{BS} + \lambda_{BS} - 1}$$

$$f(B|B_m, B, \epsilon_{BS}, \lambda_{BS}) = \frac{\sum_{\epsilon_{BS},\lambda_{BS}} f(B_m|\epsilon_{BS}, \lambda_{BS}, B) f(\epsilon_{BS}, \lambda_{BS}, B)}{\sum_{B,\epsilon_{BS},\lambda_{BS}} f(B_m|\epsilon_{BS}, \lambda_{BS}, B) f(\epsilon_{BS}, \lambda_{BS}, B)}$$

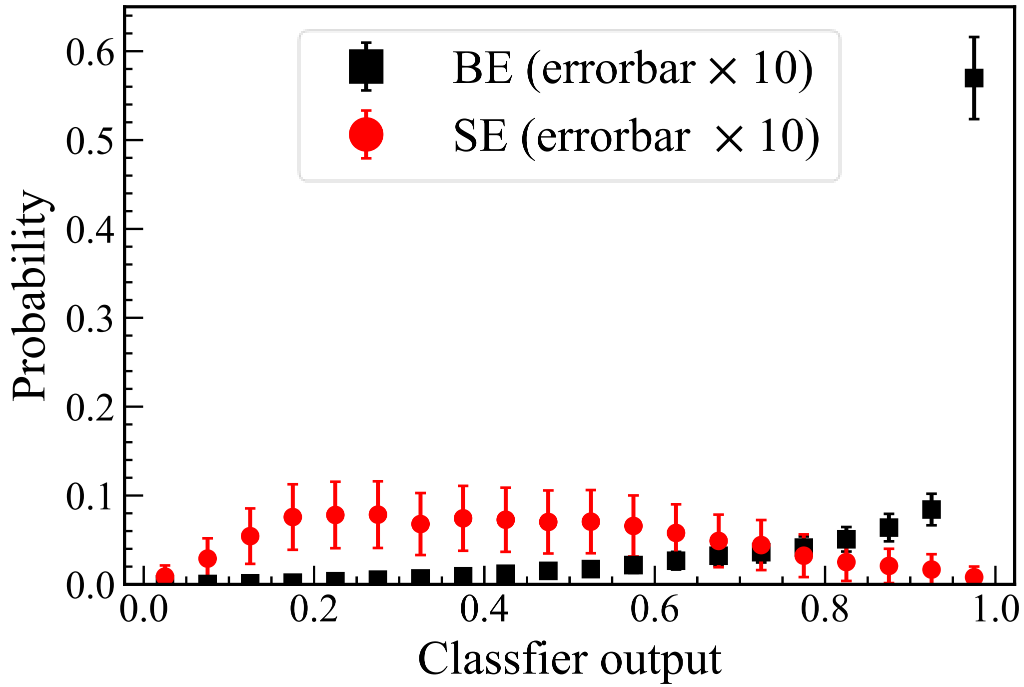
Distribution of efficiencies

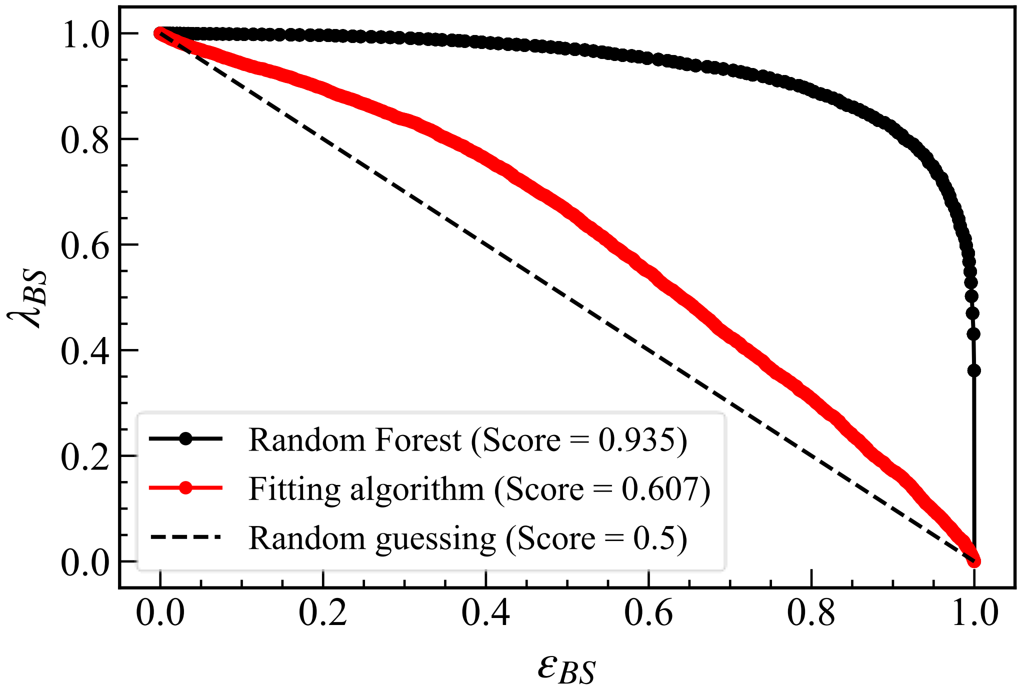


4. Results

Results: Classifier output and efficiencies

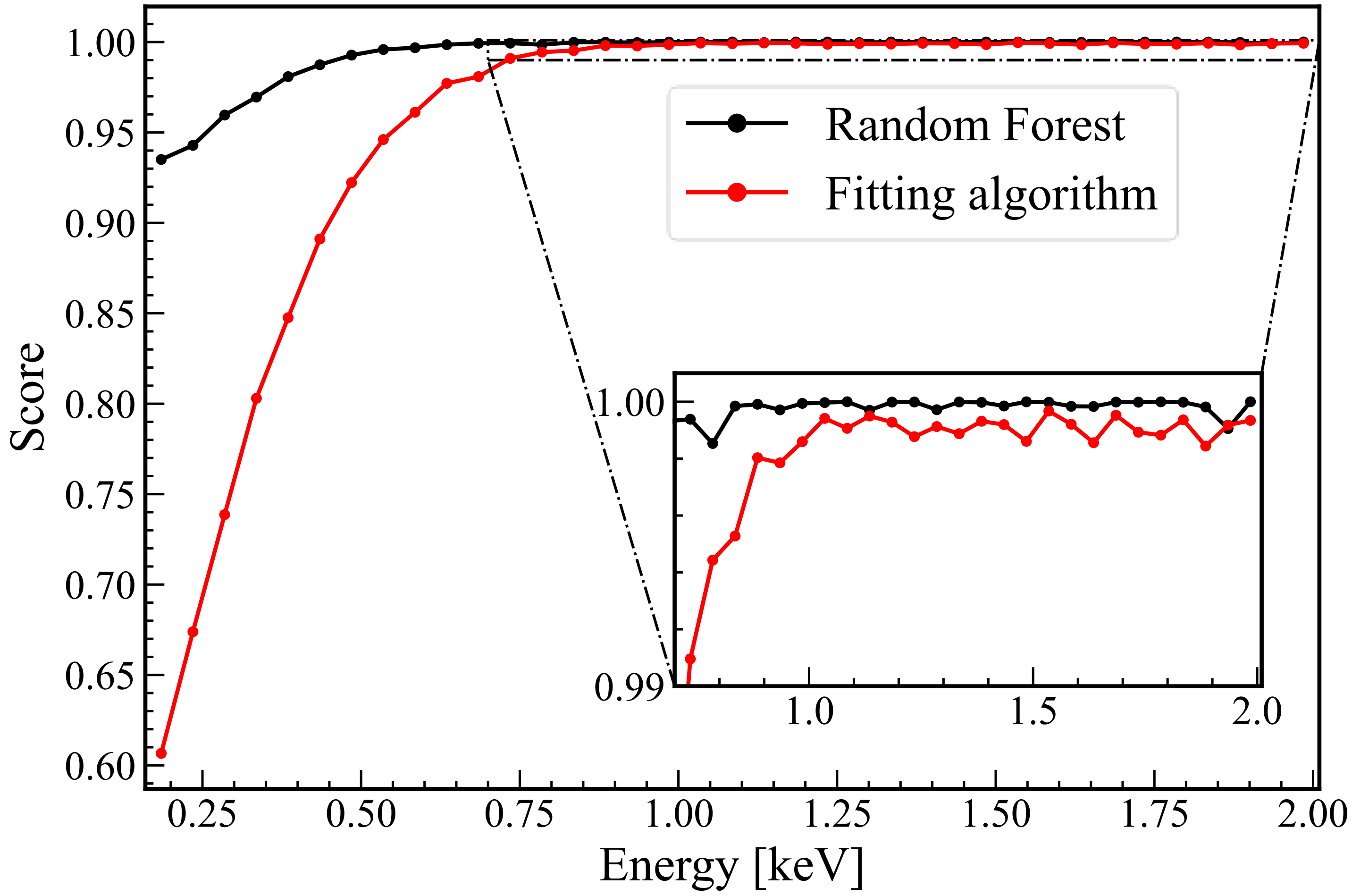
- ✓ **Classifier output distribution and efficiencies of different cuts (0.16-0.21keV)**
  - **Random forest gives a much better results than the fitting algorithm**



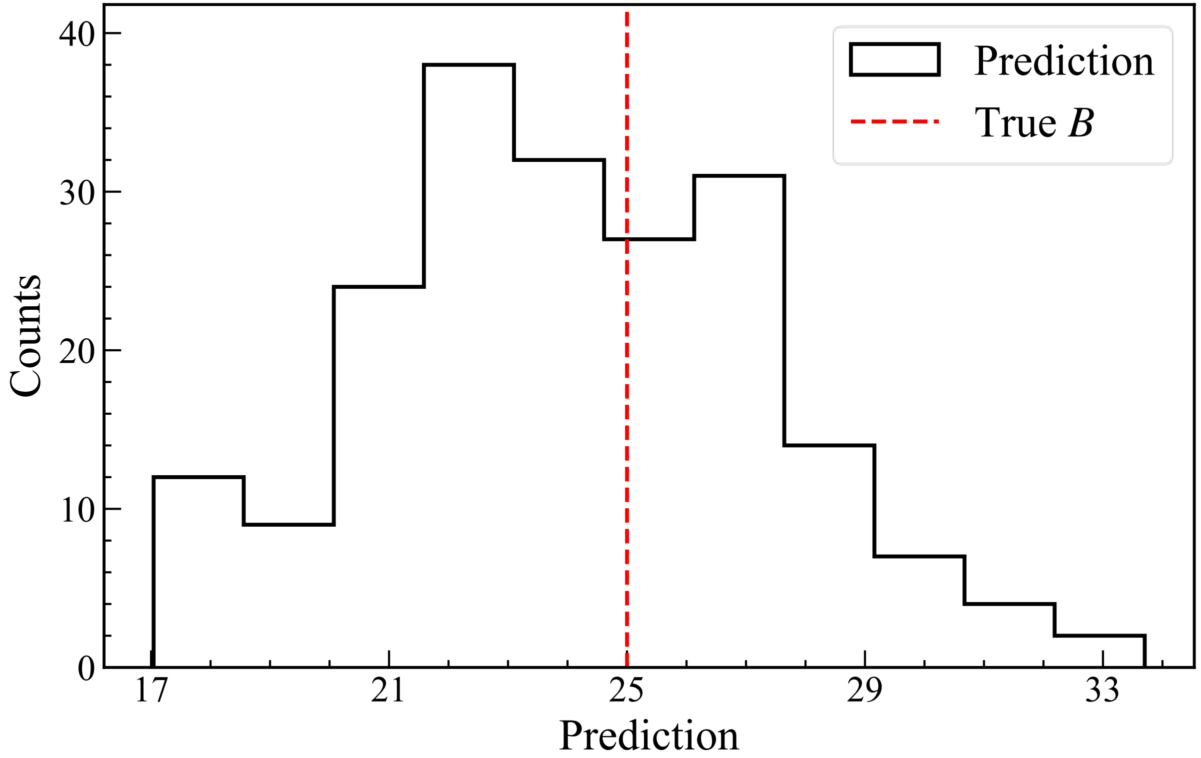


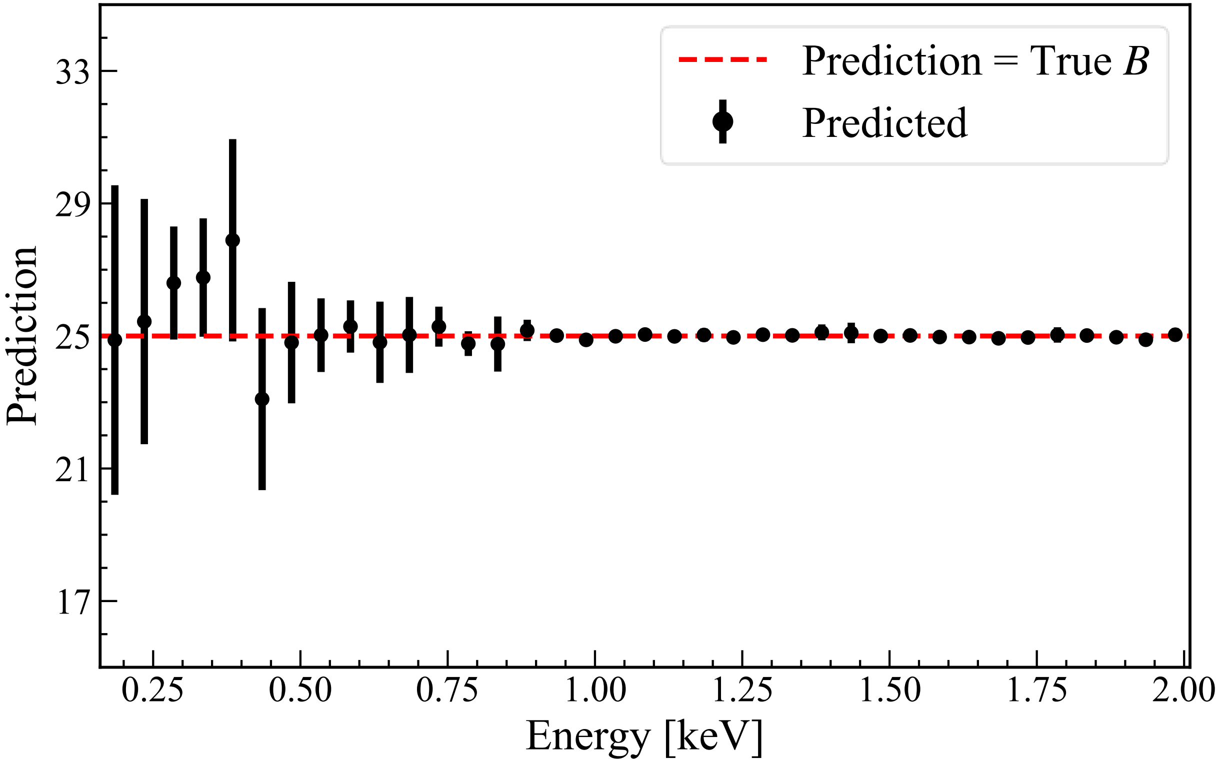
Tsinghua University

✓ **AUC scores of two algorithms vs. Energy**



Results: Verification with test-set

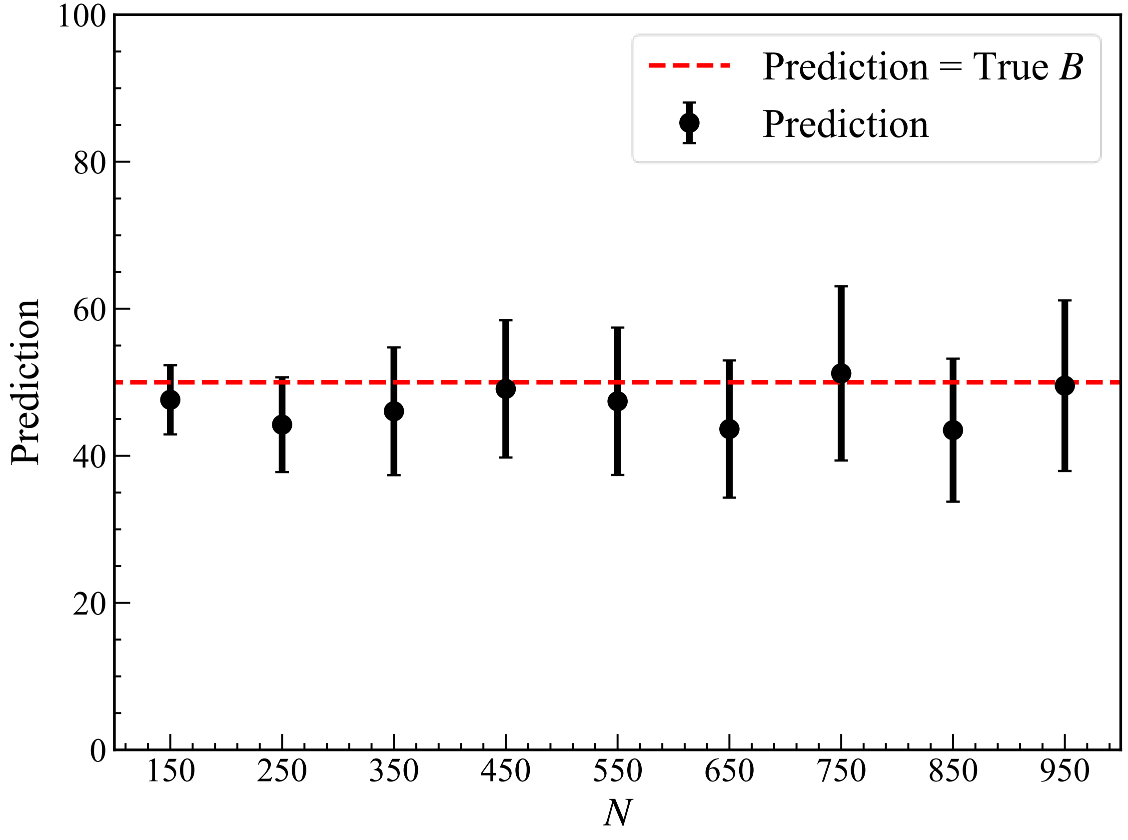




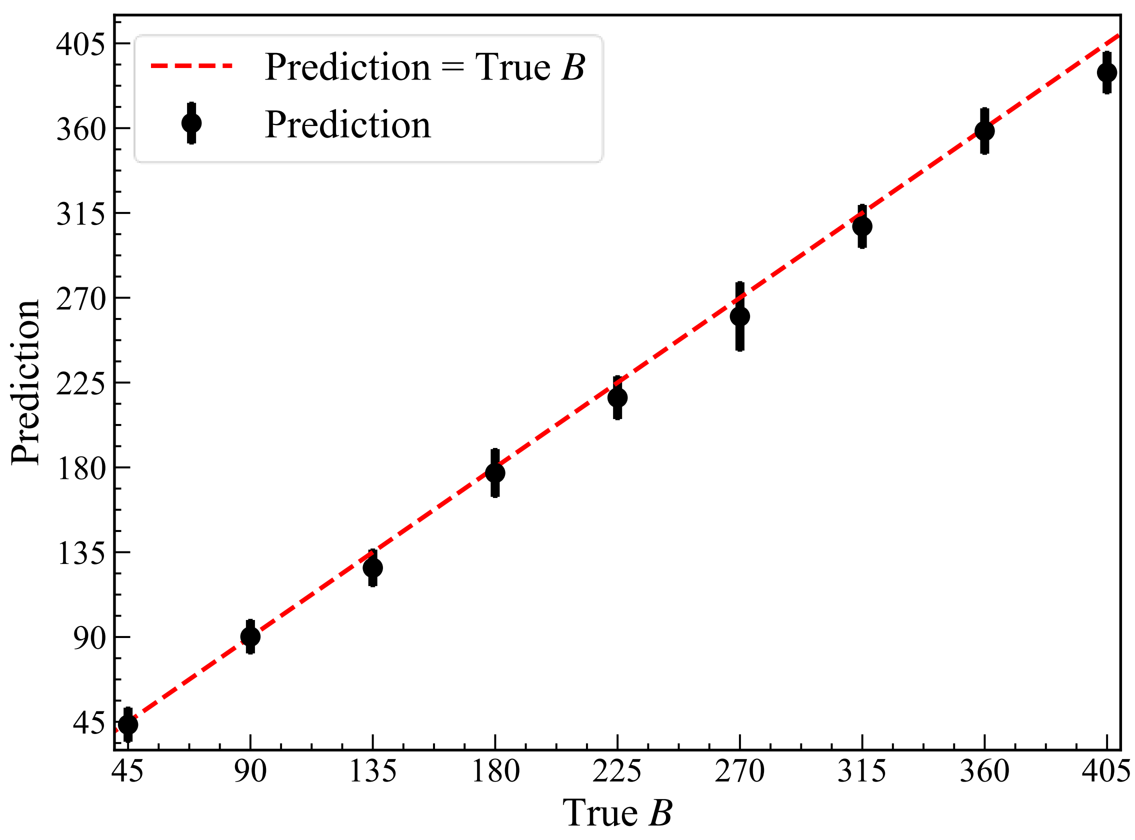
B=25, N=50

Results: Verification with test-set

Fix Bulk counts, change total counts



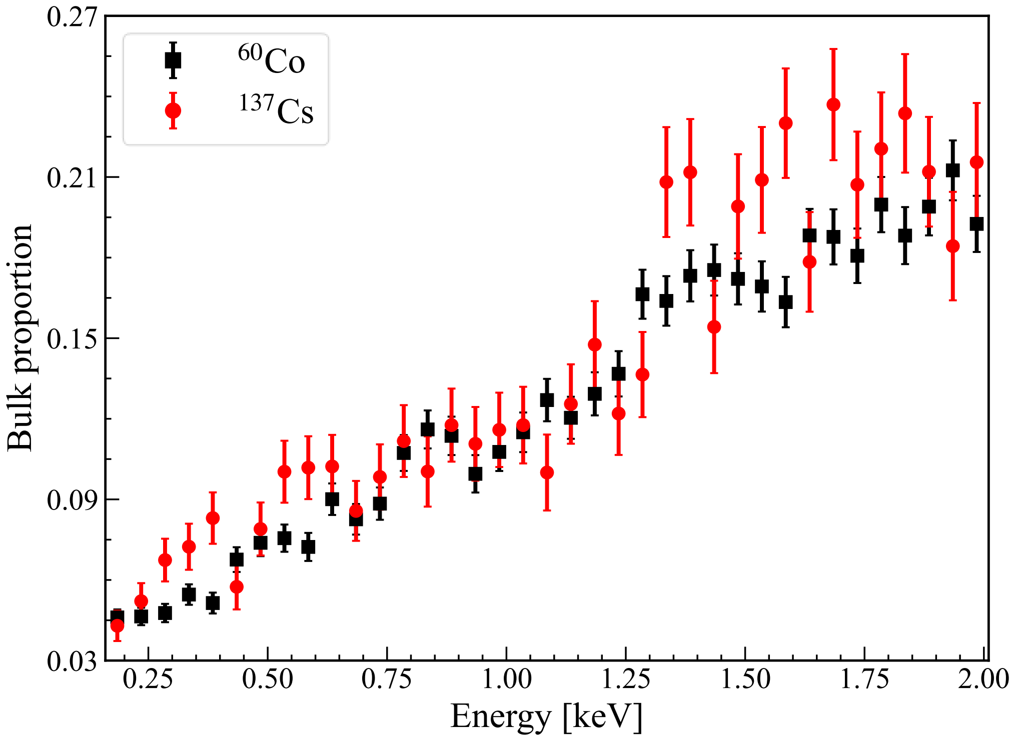
Fix Total counts, change Bulk counts

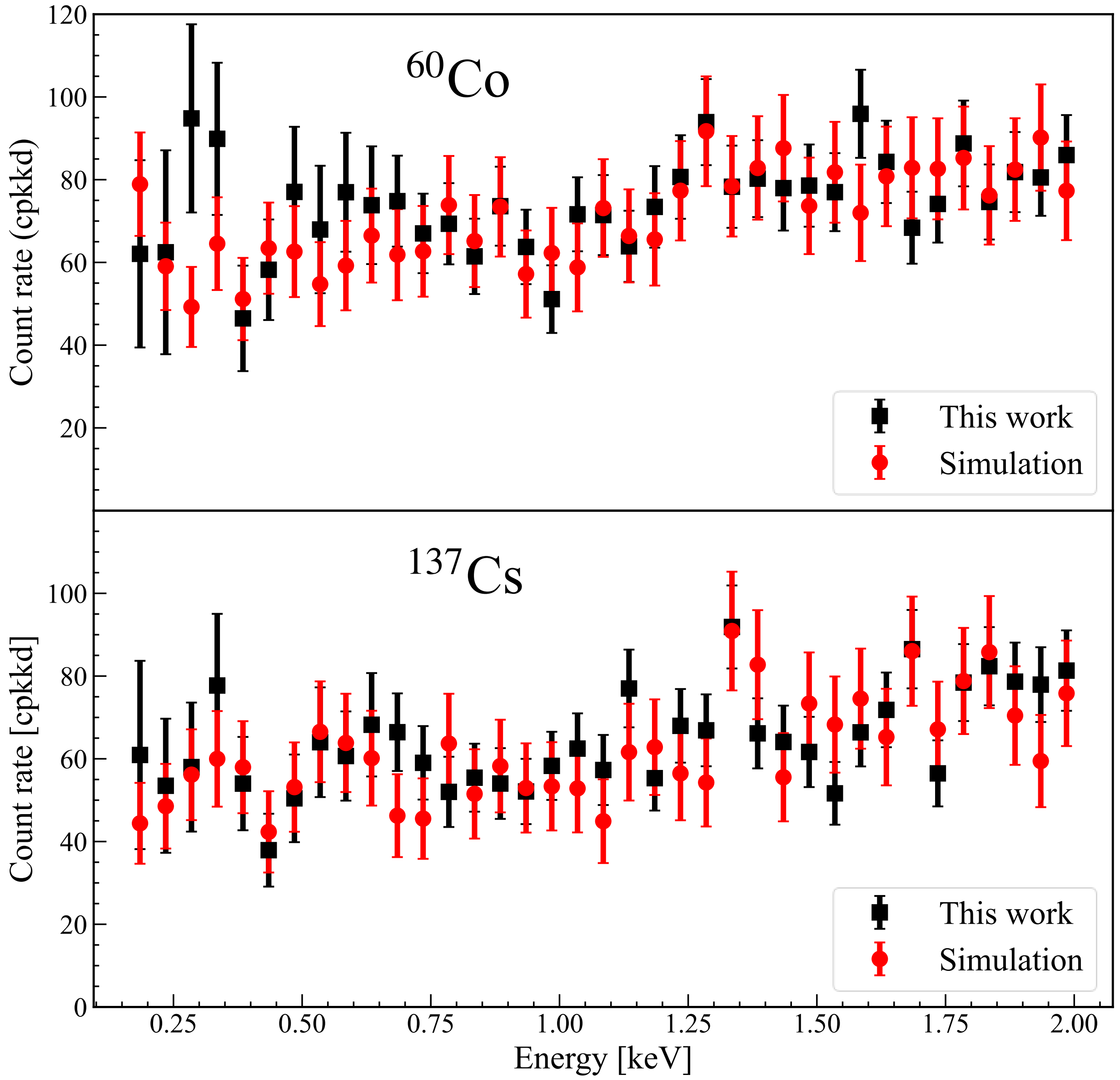


Results: Verification with exp data

Tsinghua University

- **Exp Data used in this part: high-energy calibration data**
- **Reference: simulated $P_B \times$ true N**

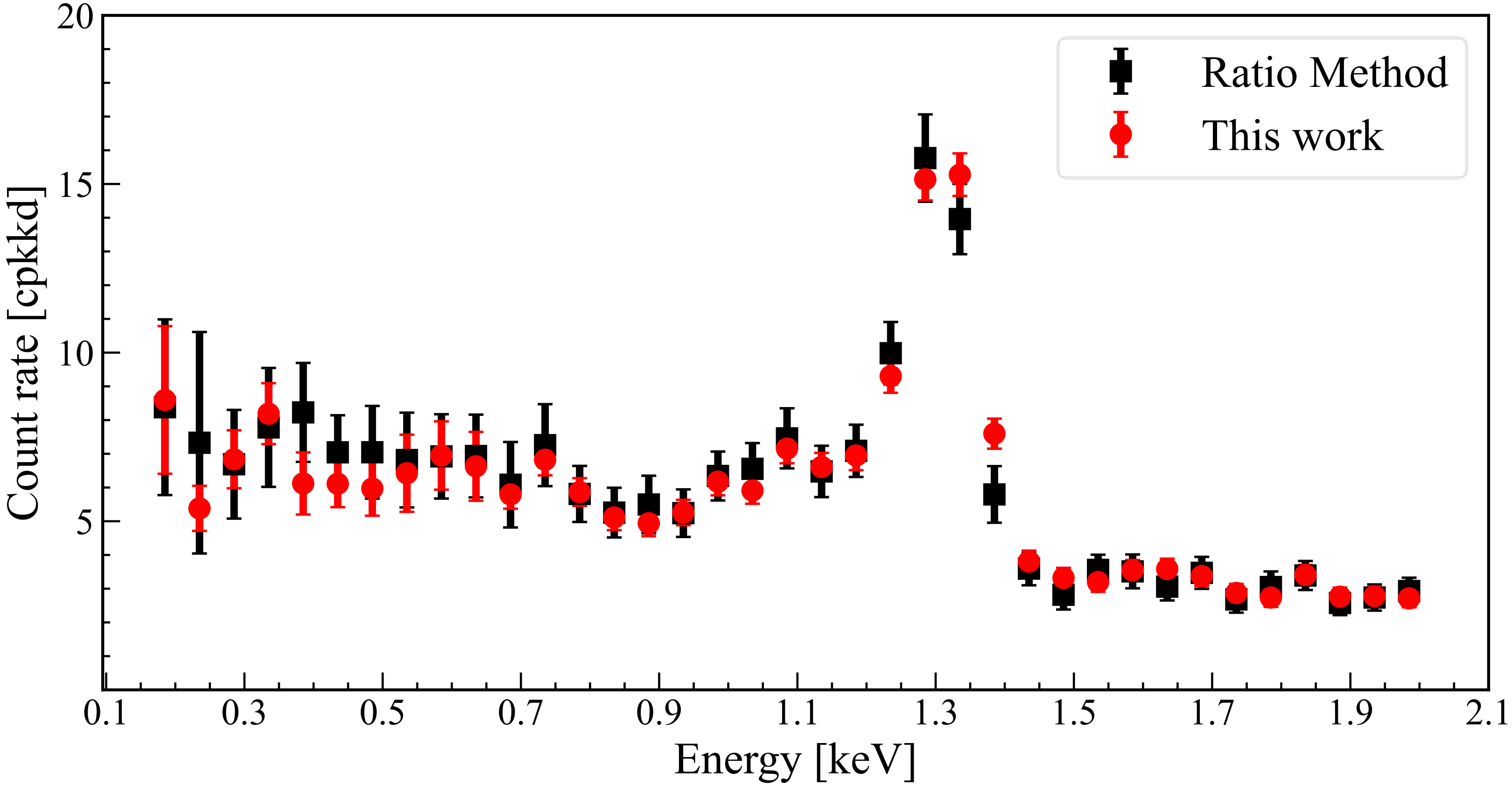




Results: On the Physics data

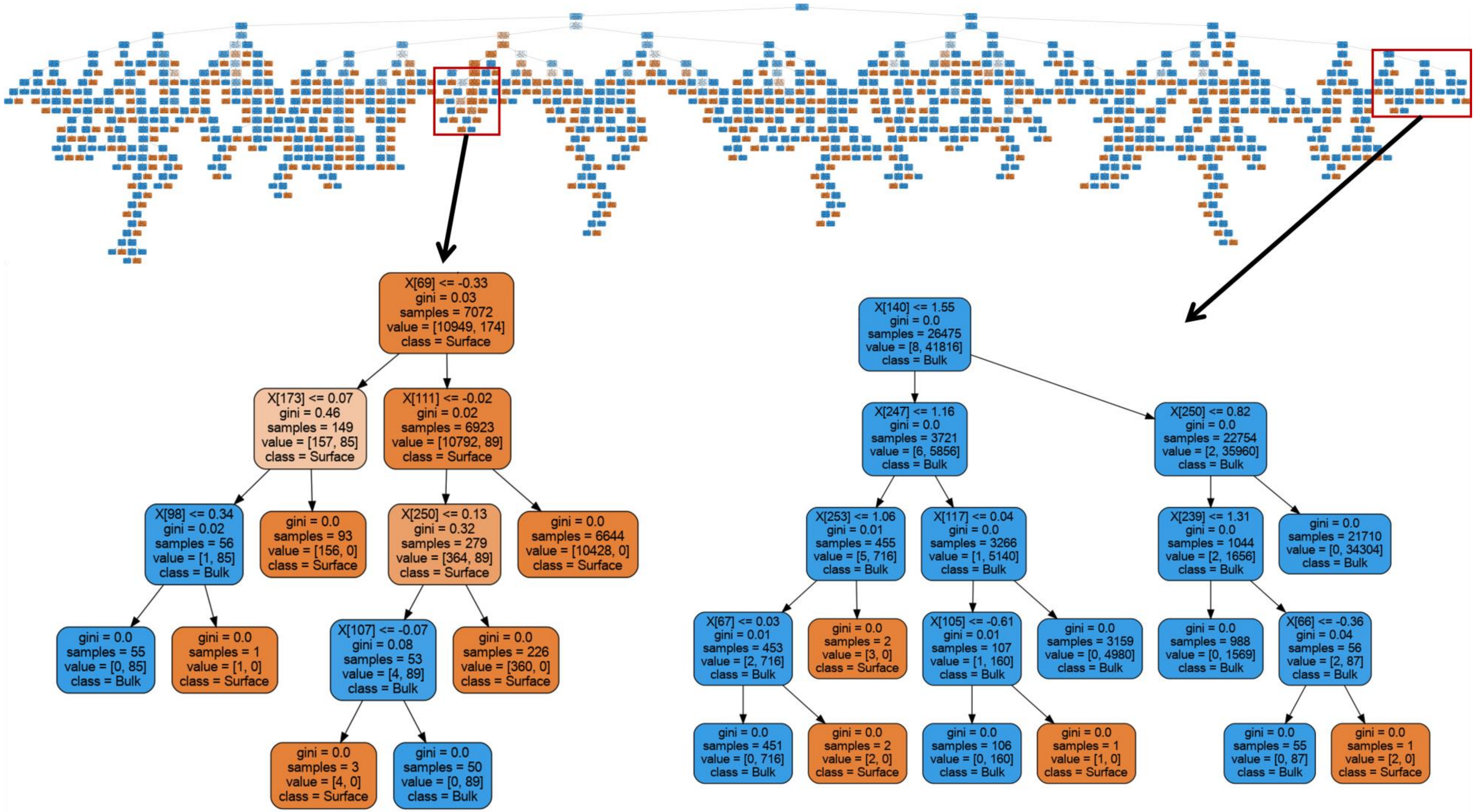
✓ **Consistent with the 'Ratio' method. Achieve lower uncertainties**

- **More robust than the risetime-fitting algorithm. More information of waveforms being used**
- **A better BE count correction scheme**

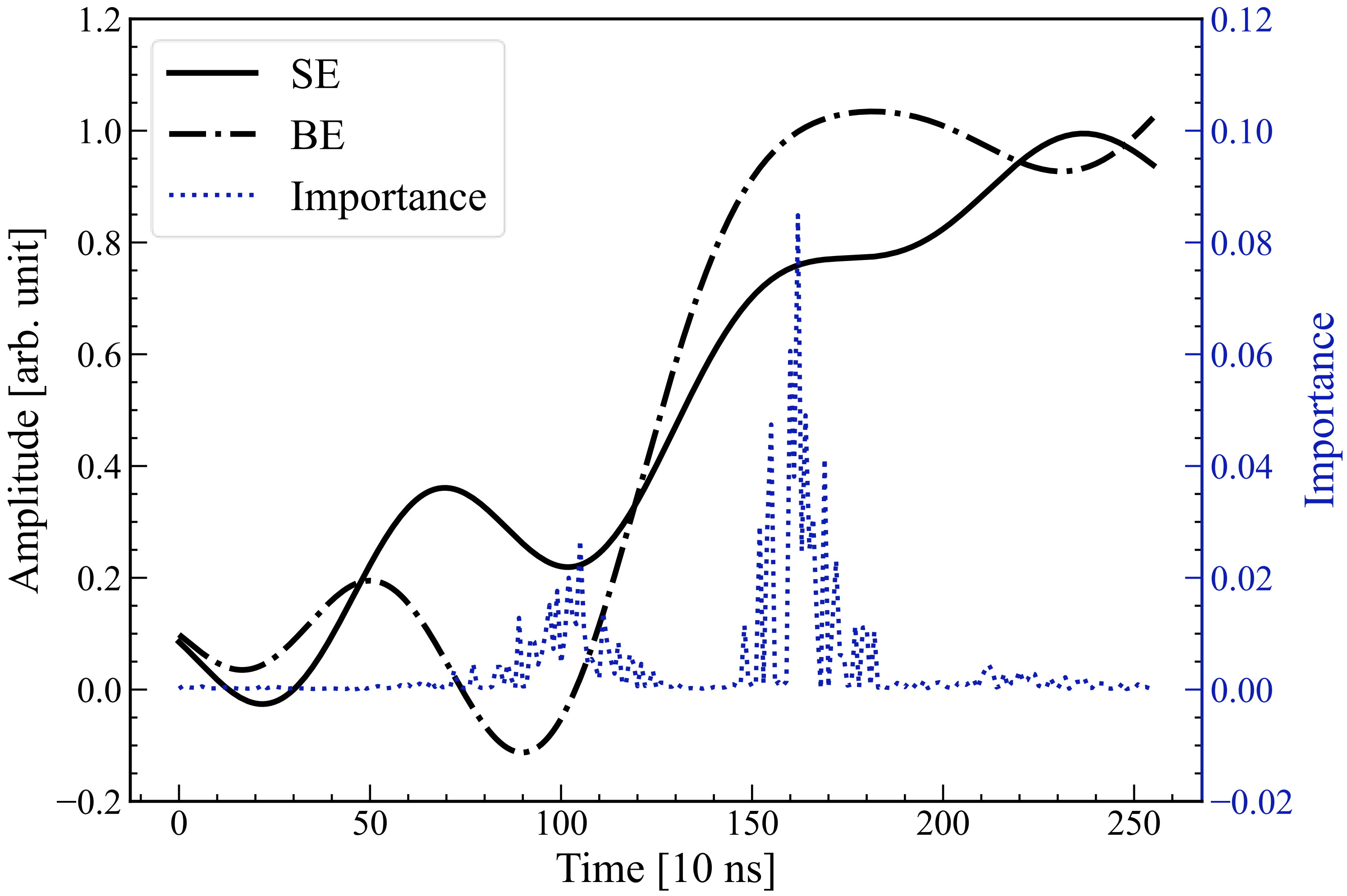


Results: Model interpretation

The structures of an optimized tree



Results: Model interpretation



5. Conclusions

Conclusions

Tsinghua University

- ✓ **Feature extraction algorithm + Efficiency correction scheme. Both outperform the traditional methods**
- ✓ **Verify the model with two datasets**
- ✓ **Apply onto the physics data. Consistent with the 'Ratio' method**
- ✓ **The machine learning algorithm is interpreted**

Thanks for your attention!

中国锦屏地下实验室
China Jinping Underground Laboratory
清华大学 · 雅砻江流域水电开发有限公司

中国暗物质实验
China Dark matter EXperiment