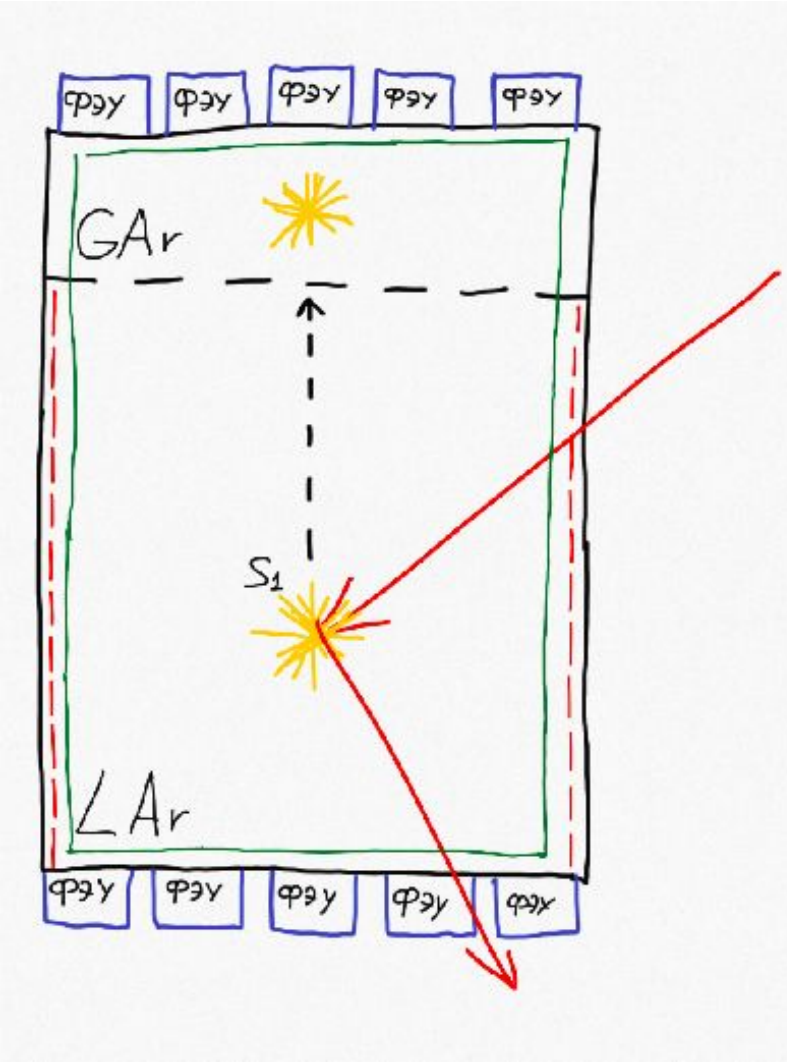
Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
Национальный исследовательский ядерный университет «МИФИ»

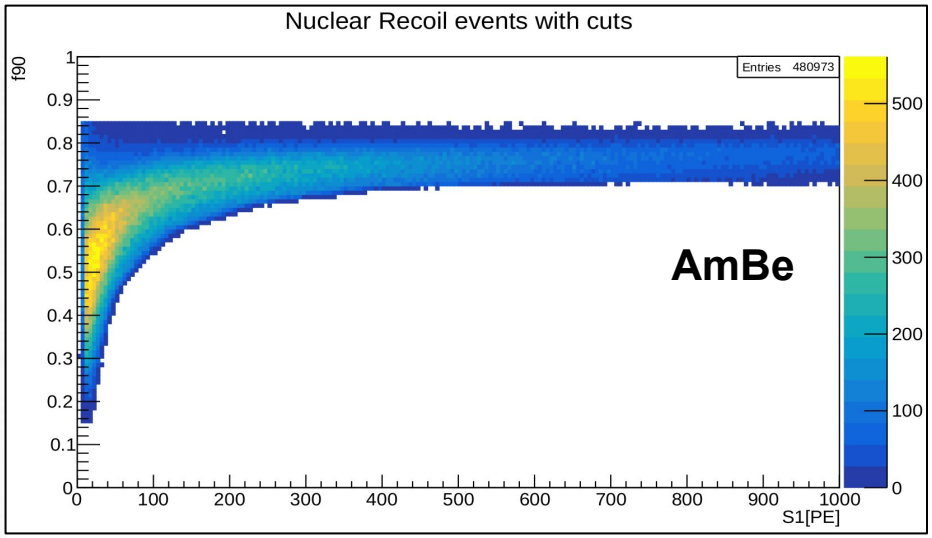
Отчет о научно-исследовательской работе на тему: Монте Карло моделирование и разделение по форме импульса низкоэнергетических событий эксперимента DarkSide-50

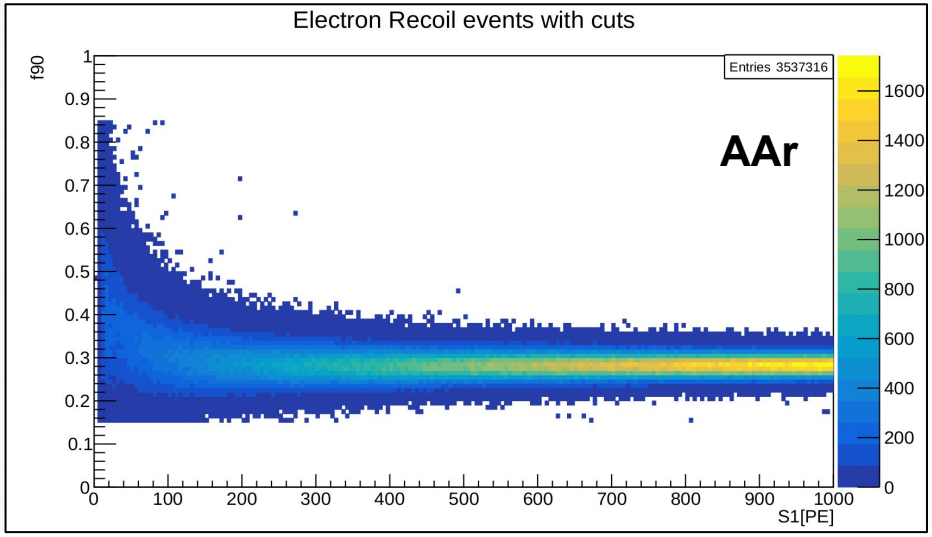
Выполнил:
Студент гр. М19-115
Ильясов А.И.
Научный руководитель:
к.ф.-м.н.
Гробов А.В.

Москва, 2020

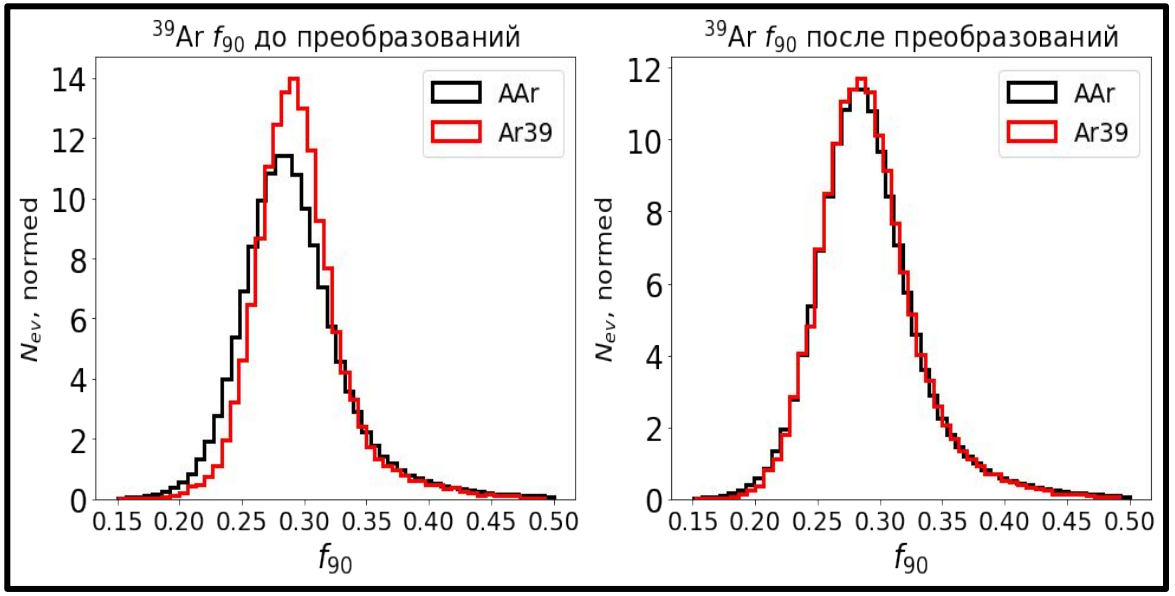
Эксперимент DarkSide-50

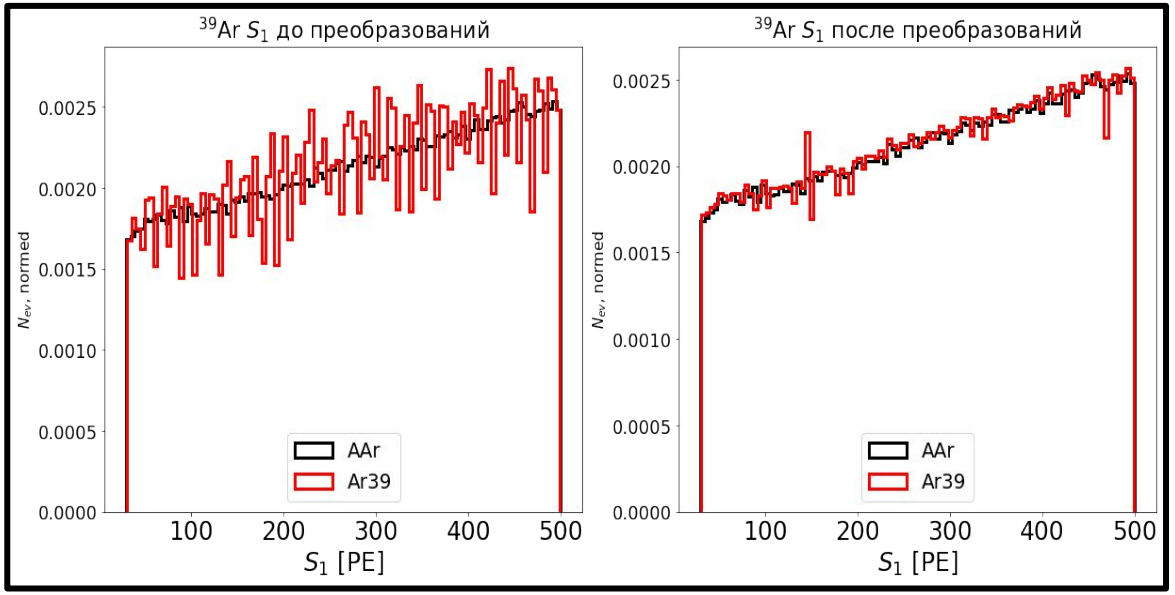






Данные Монте Карло моделирования ^{39}Ar





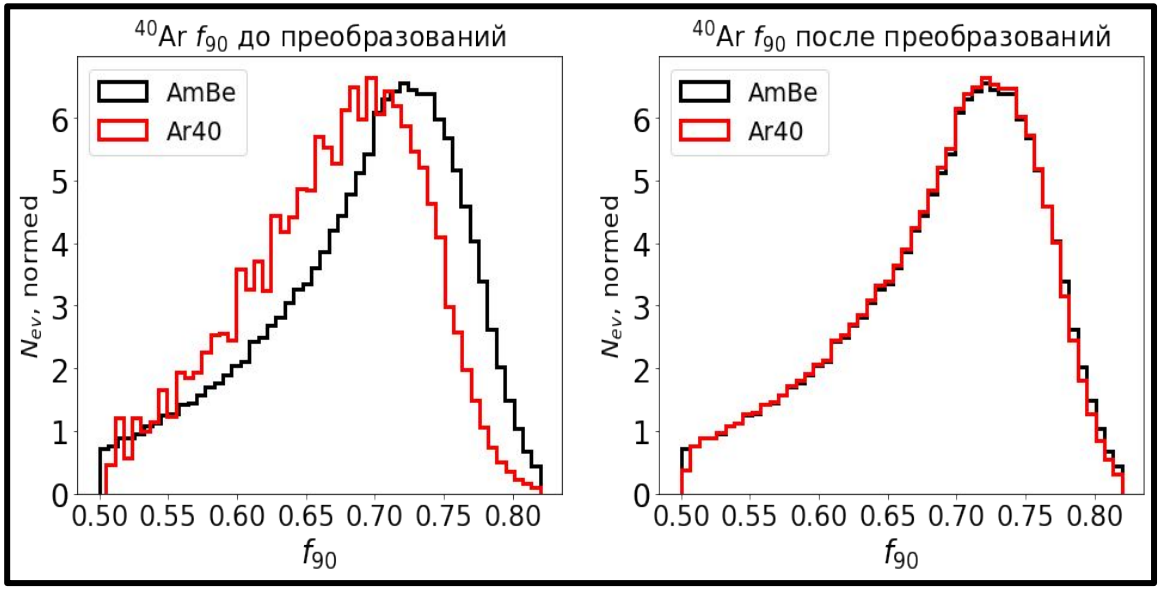
Функция распределения Хинкли:

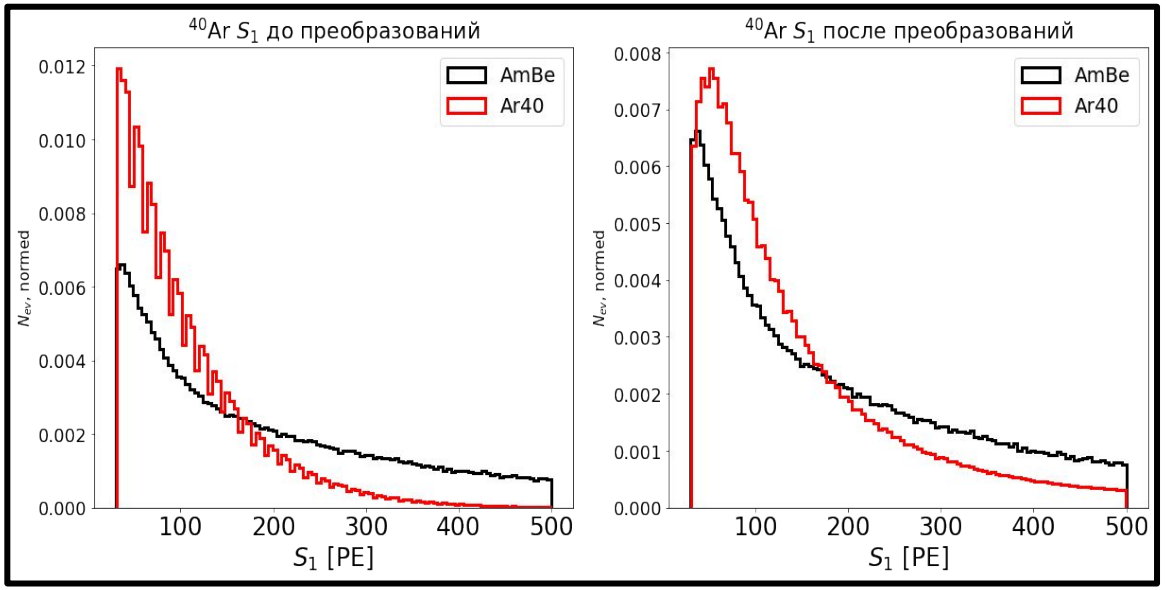
$$f(x) = \frac{\sigma_l^2 \mu_p x + \sigma_p^2 \mu_l (1-x)}{\sqrt{2\pi}(\sigma_l^2 x^2 + \sigma_p^2 (1-x)^2)^{3/2}} \times \exp\left[- \frac{(\mu_l x - \mu_p (1-x)^2)}{2(\sigma_l^2 x^2 + \sigma_p^2 (1-x)^2)} \right]$$

$$f_{90} = \frac{\int_0^{90ns} S_1 \, dt}{\int_0^{7\mu s} S_1 \, dt}$$

Statistic/ p-value	old data	new data
f_{90}	26.721/1.0	1.239/1.0
S_1	0.084/1.0	0.025/1.0
N событий	52 тыс.	607 тыс.

Данные Монте Карло моделирования ^{40}Ar





Функция распределения Хинкли:

$$f(x) = \frac{\sigma_l^2 \mu_p x + \sigma_p^2 \mu_l (1-x)}{\sqrt{2\pi}(\sigma_l^2 x^2 + \sigma_p^2 (1-x)^2)^{3/2}} \times \exp\left[- \frac{(\mu_l x - \mu_p (1-x)^2)}{2(\sigma_l^2 x^2 + \sigma_p^2 (1-x)^2)} \right]$$

$$f_{90} = \frac{\int_0^{90ns} S_1 \, dt}{\int_0^{7\mu s} S_1 \, dt}$$

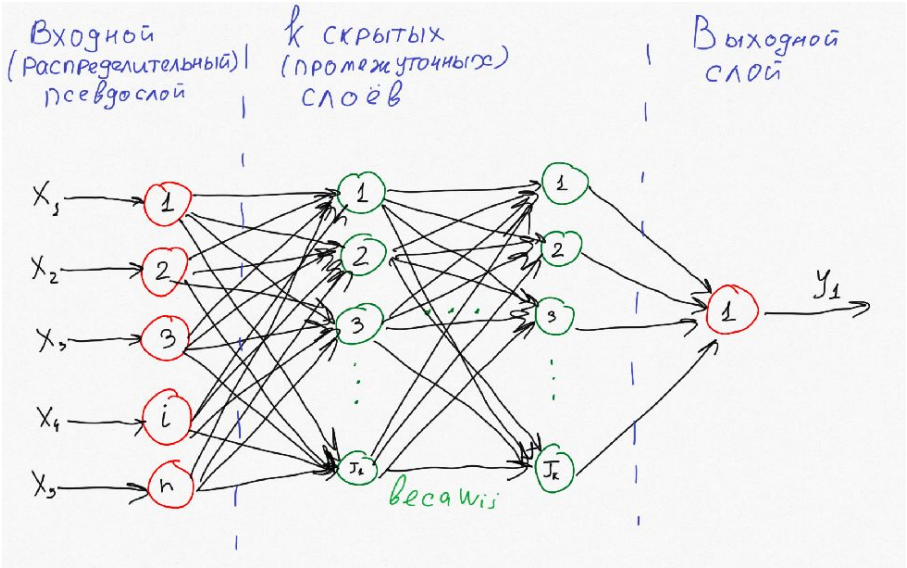
Statistic/ p-value	old data	new data
f_{90}	12.271/1.0	0.781/1.0
S_1	0.003/1.0	0.0003/1.0
N событий	1.3 млн.	1.3 млн.

Классификаторы

Многослойный перцептрон

Преимущества:
- Лёгок в понимании
- Легко настраивается
- Мало **гиперпараметров**

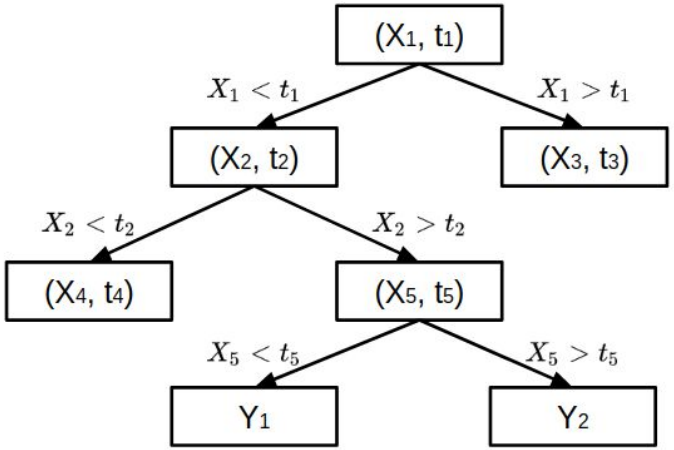
Недостатки:
- Легко **переобучается**
- Требует большую **статистику**
- **Долго** работает



Градиентный бустинг над деревьями решений

Преимущества:
- **Быстро** работает
- Работает хорошо с малой статистикой
- Хорошо **защищён от переобучения** с помощью гиперпараметров

Недостатки:
- **Трудно** настраивается
- Тяжелее в **понимании**



Оценка классификатора и используемые данные

Оценка

Метрики:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

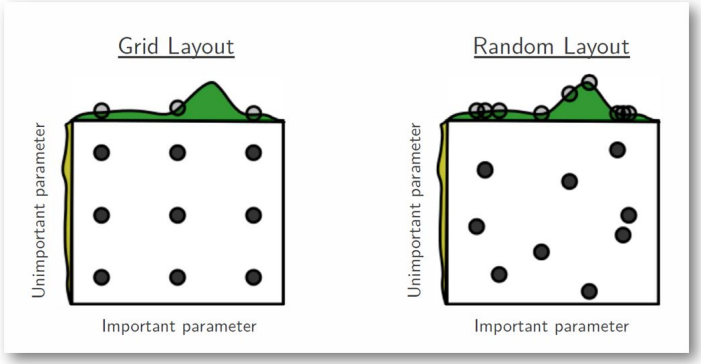
$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Параметры, необходимые для вычисления метрик		Истинные метки	
		1	0
Предсказание классификатора	1	TP	FP
	0	FN	TN

ROC-кривая - Зависимость ***recall*** от ***background rejection rate***:

$$BRR = \frac{TN}{FP + TN}$$

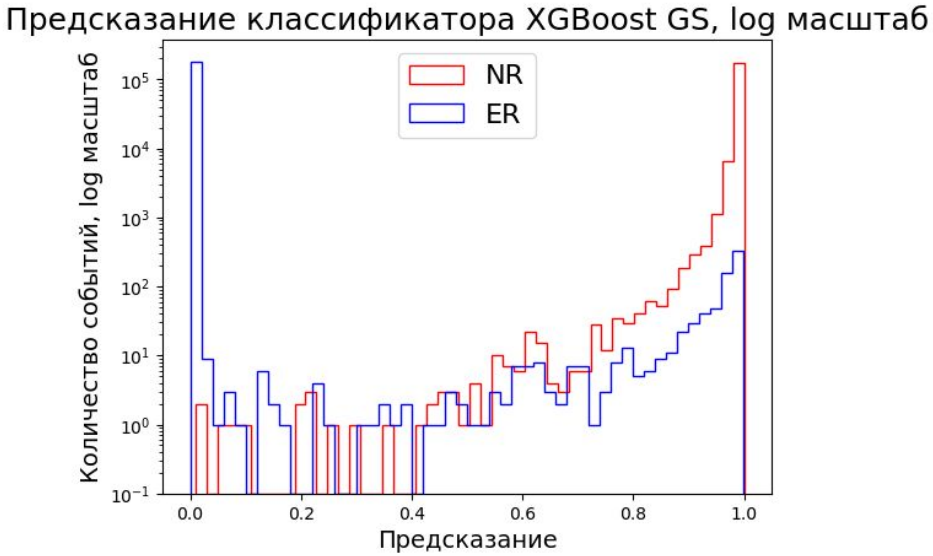
- Поиск параметров по сетке
- Рандомизированный поиск параметров

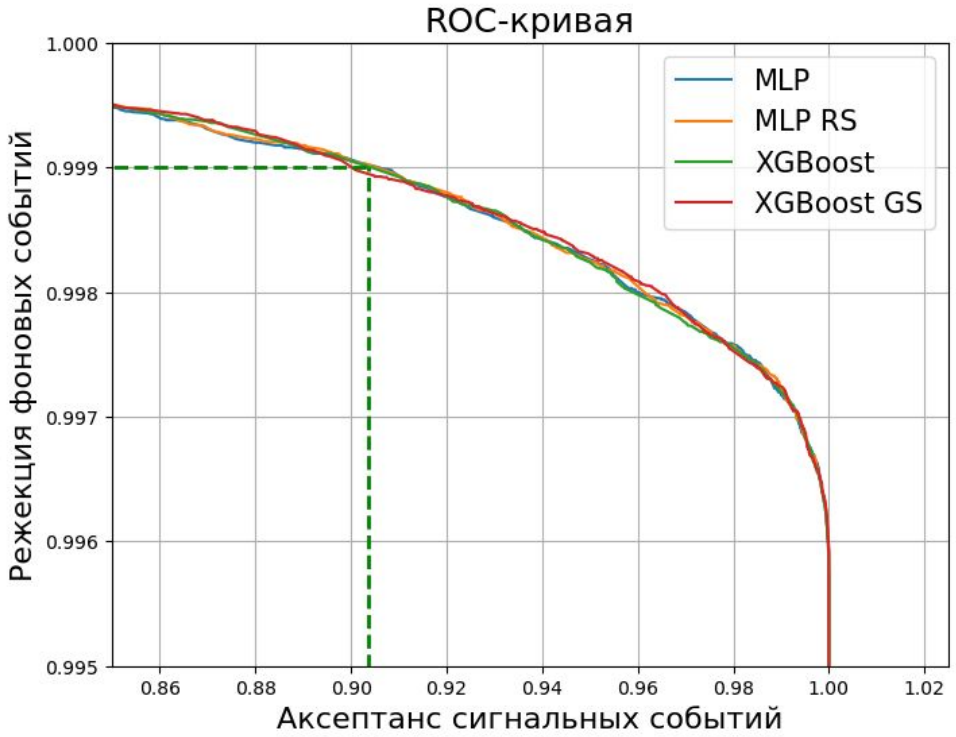


Используемые переменные:

- $f_{30}, f_{60}, f_{90}, f_{200}, S_1$
- $log(f_{30}, f_{60}, f_{90}, f_{200}, S_1)$
- Попарные произведения
- log от попарных произведений

Результаты



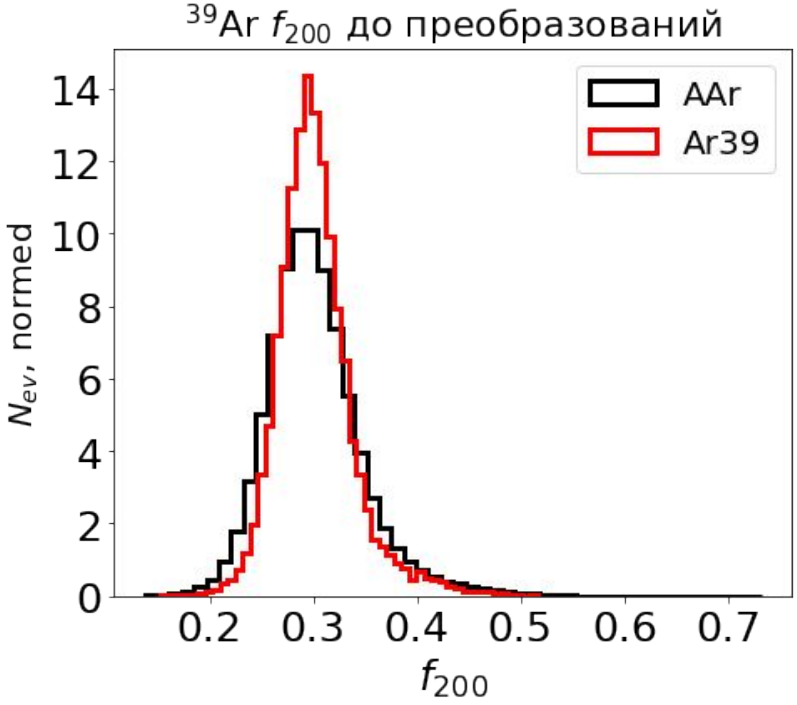


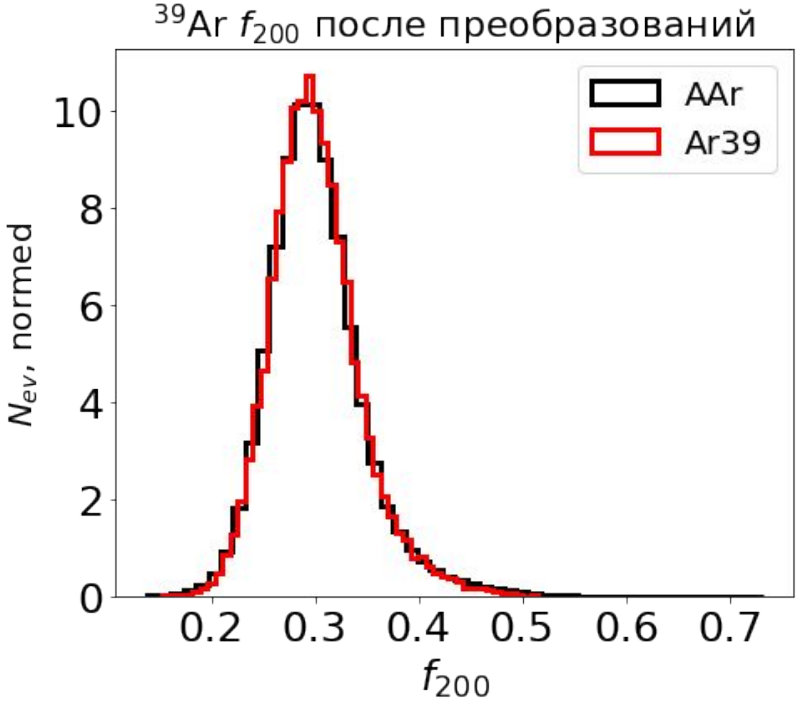
*Систематические неопределенности не учтены

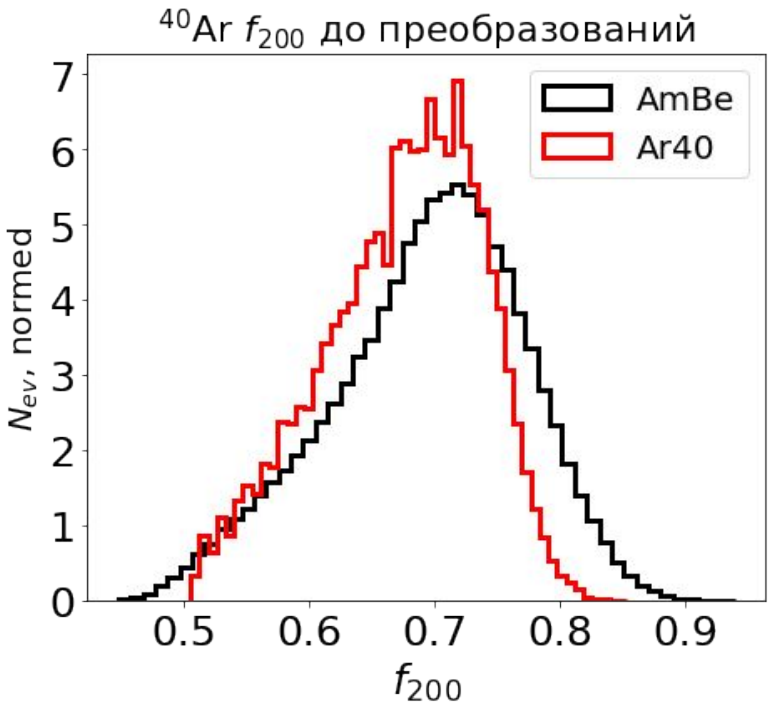
Заключение

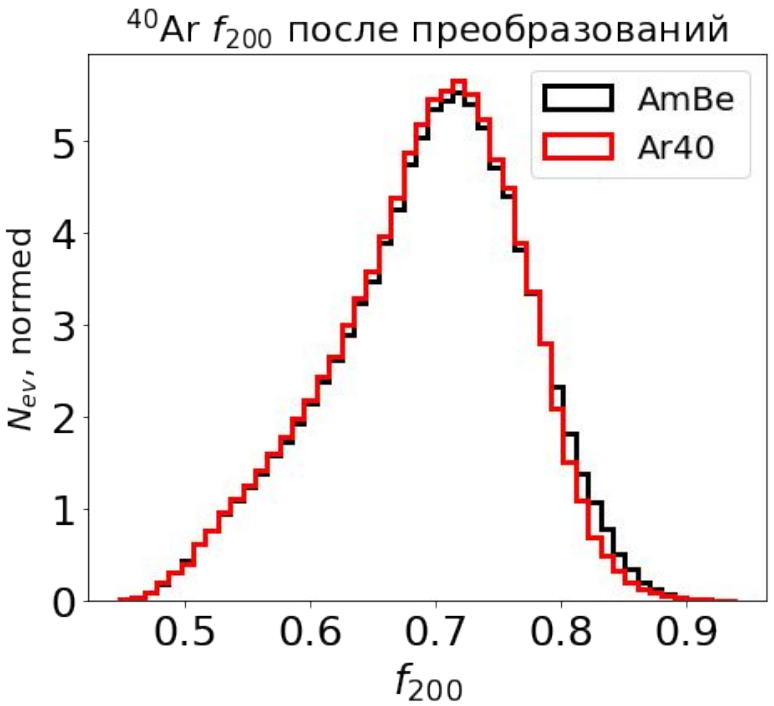
- Применение функции Хинкли уменьшает общую ошибку распределения, что позволяет применять данные Монте Карло моделирования для обучения моделей машинного обучения;
- ★ Из малого количества основных параметров (5) было получено большое количество пользовательских (33) параметров;
- ★ Все 4 модели имеют схожую эффективность. Каждая из них может быть выбрана в соответствии с поставленной задачей;
- ★ При пороговом значении в 99% режекции фоновых событий имеем 100% аксептанс сигнальных событий;
- ★ При пороговом значении 99.9% режекции фоновых событий имеем ~90% принятия сигнальных событий.

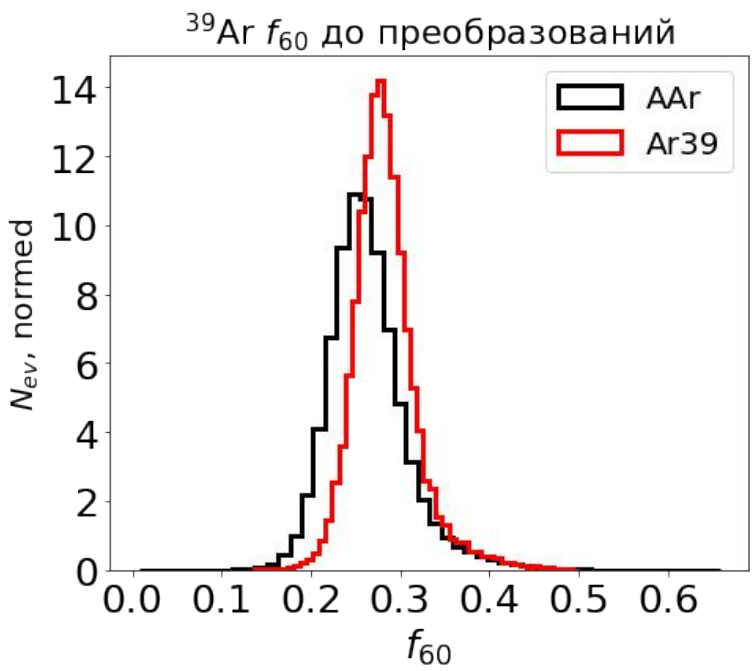
Backup slides

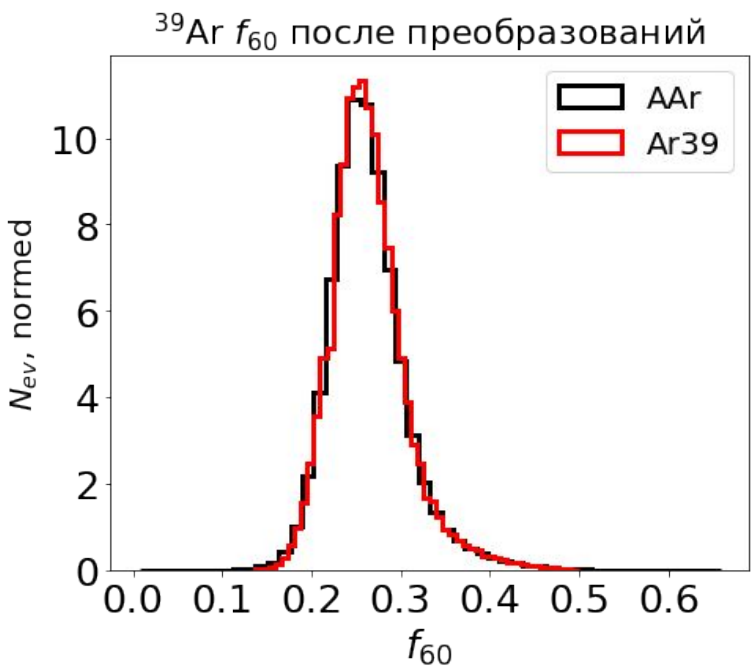

^{39}Ar f_{200} до преобразований
AAr
Ar39
N_{ev}, normed
f_{200}

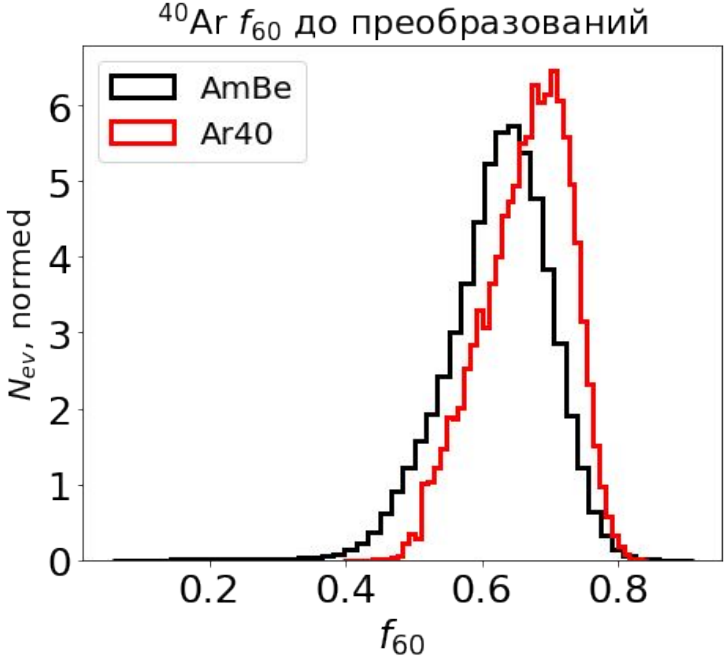

^{39}Ar f_{200} после преобразований
AAr
Ar39
N_{ev}, normed
f_{200}

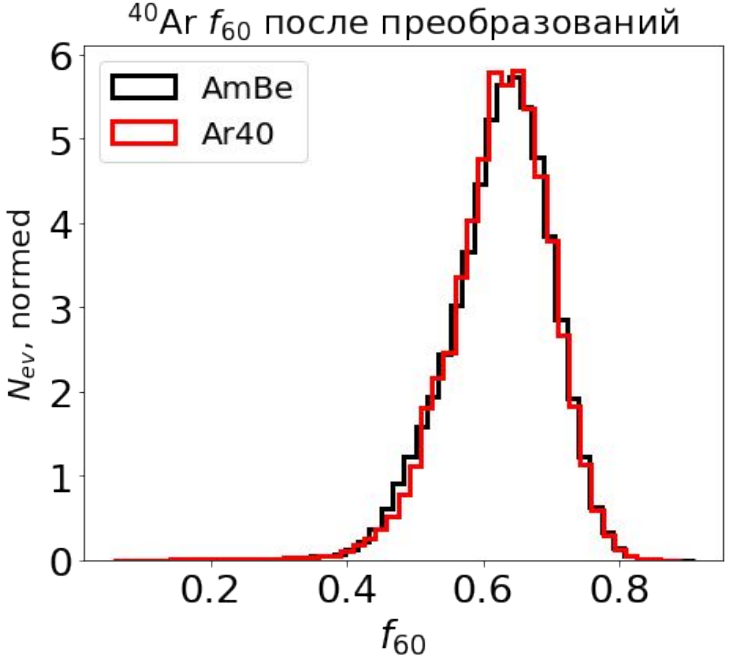

^{40}Ar f_{200} до преобразований
AmBe
Ar40
N_{ev}, normed
f_{200}

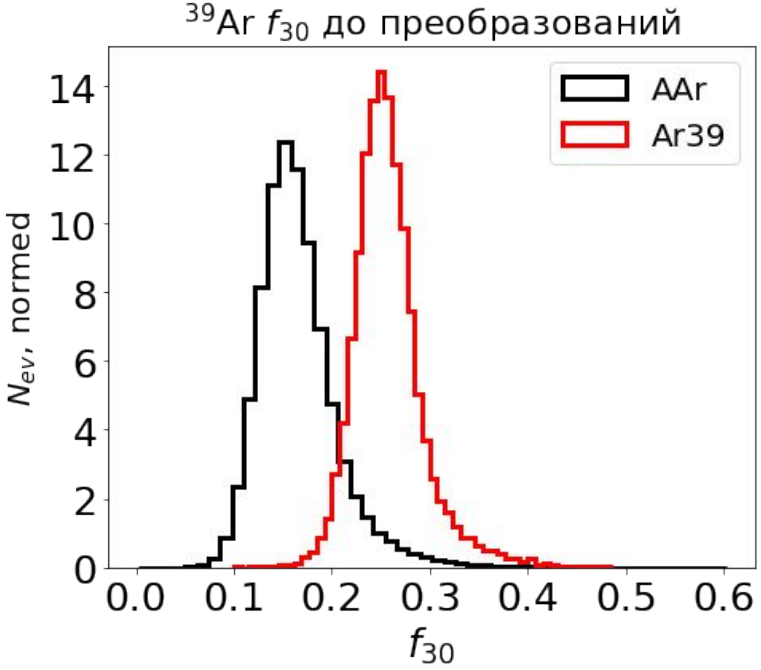

^{40}Ar f_{200} после преобразований
AmBe
Ar40
N_{ev}, normed
f_{200}

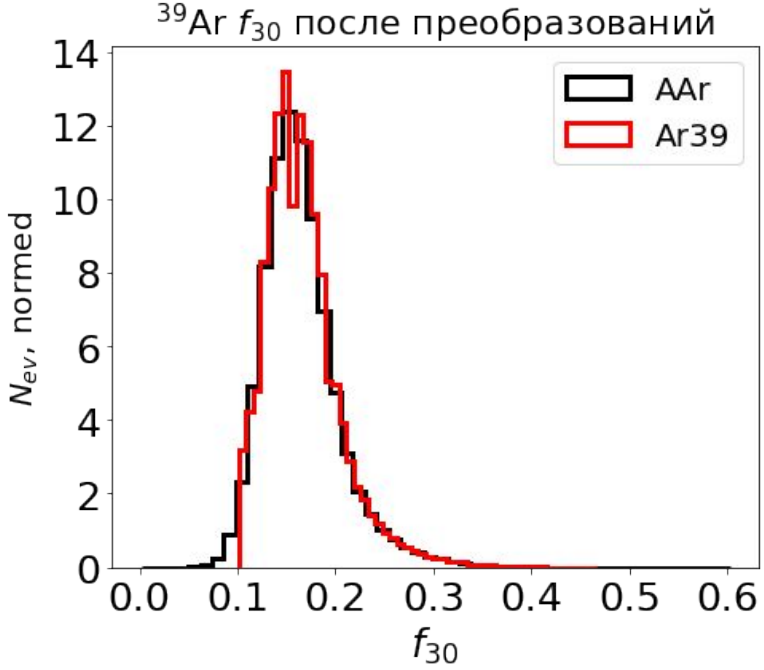

^{39}Ar f_{60} до преобразований
AAr
Ar39
N_{ev}, normed
f_{60}

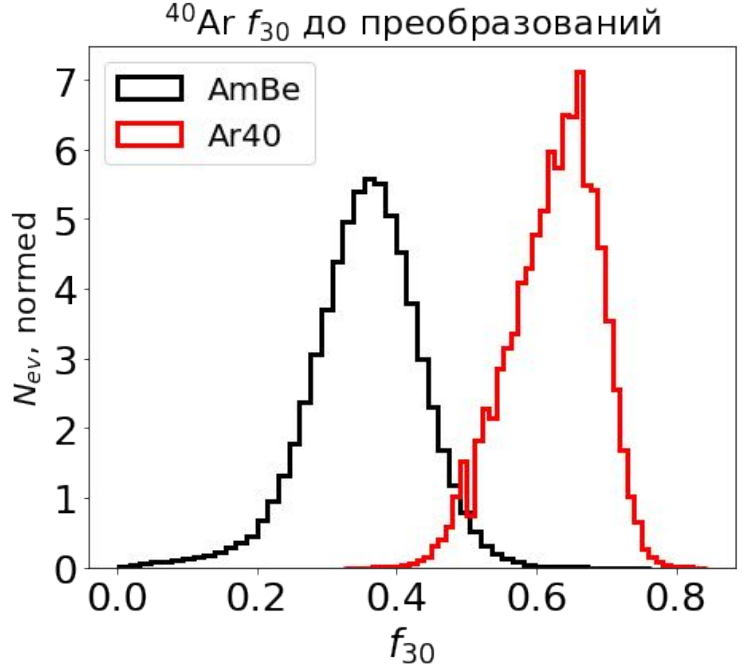

^{39}Ar f_{60} после преобразований
AAr
Ar39
N_{ev}, normed
f_{60}

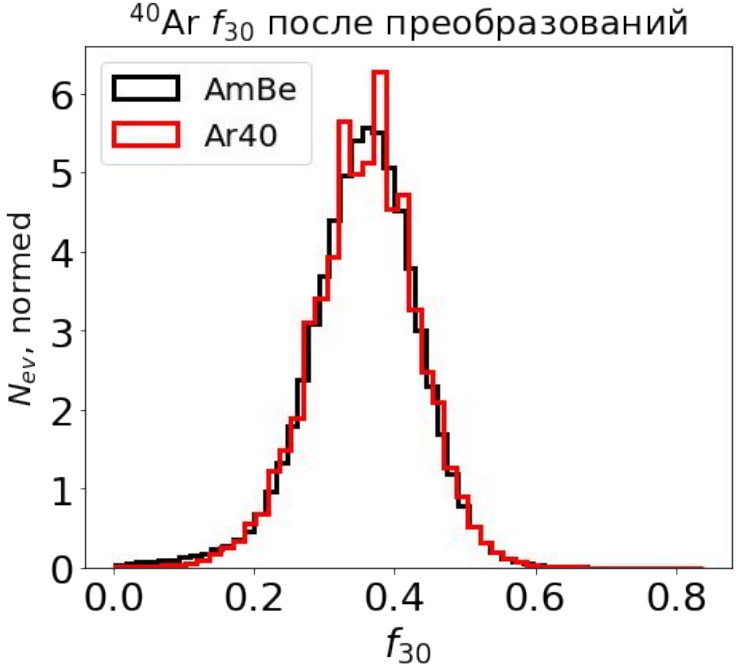

^{40}Ar f_{60} до преобразований
AmBe
Ar40
N_{ev}, normed
f_{60}


^{40}Ar f_{60} после преобразований
AmBe
Ar40
N_{ev}, normed
f_{60}

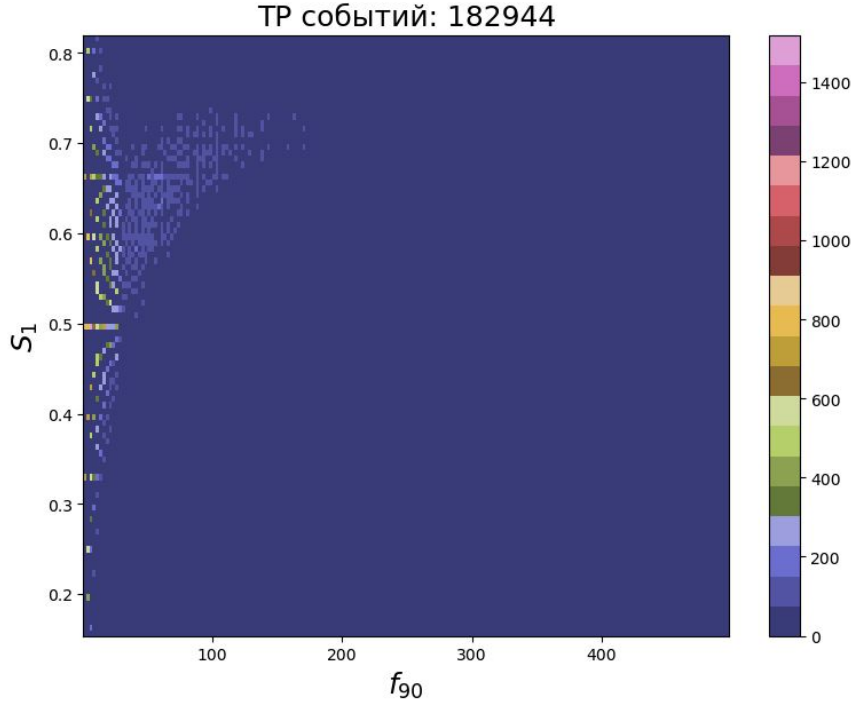

^{39}Ar f_{30} до преобразований
AAr
Ar39
N_{ev}, normed
f_{30}

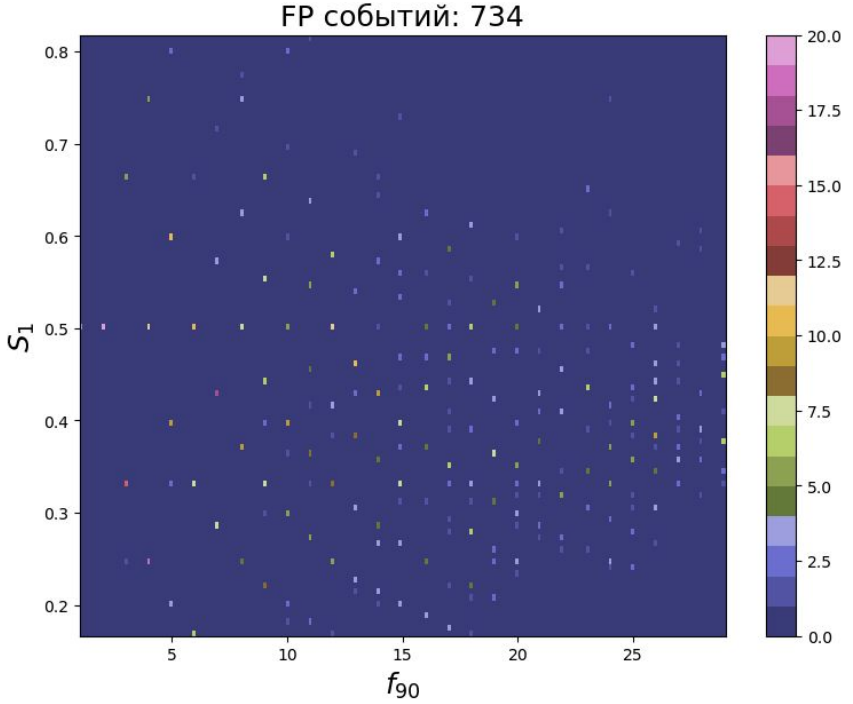

^{39}Ar f_{30} после преобразований
AAr
Ar39
N_{ev}, normed
f_{30}

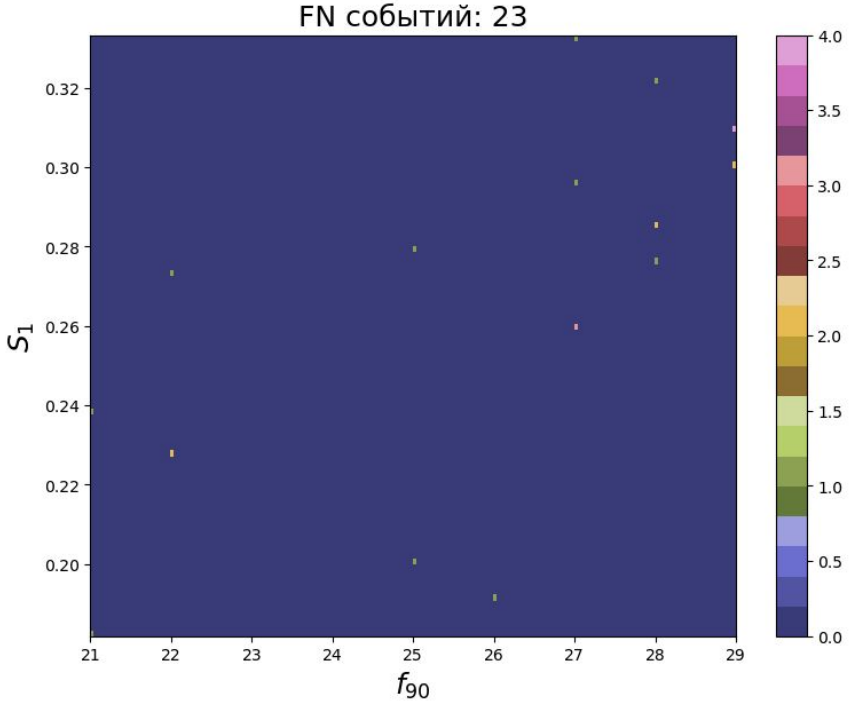

^{40}Ar f_{30} до преобразований
AmBe
Ar40
N_{ev}, normed
f_{30}

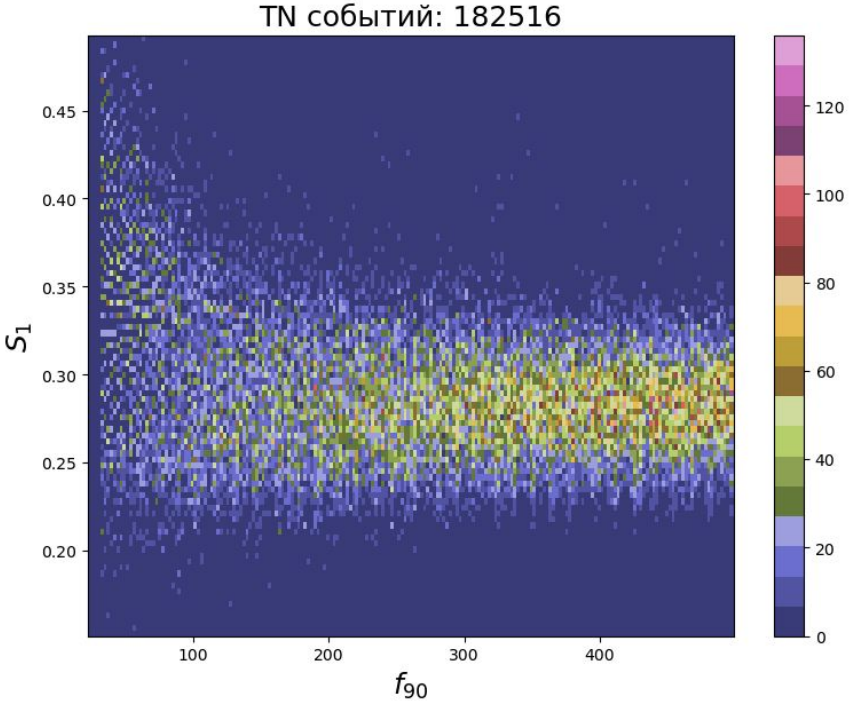

^{40}Ar f_{30} после преобразований
AmBe
Ar40
N_{ev}, normed
f_{30}

Алгоритм	TP	TN	FP	FN
XGBoost	182512	182941	738	26
XGBoost GS	182516	182944	734	23
MLP	182493	182960	757	7
MLP RS	182506	182952	744	15

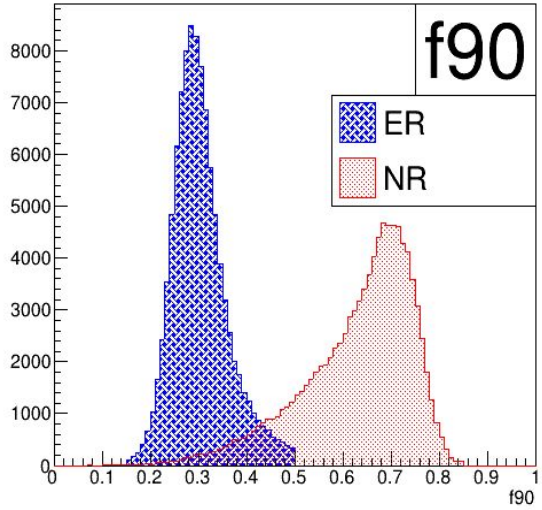

TP событий: 182944
S_1
f_{90}

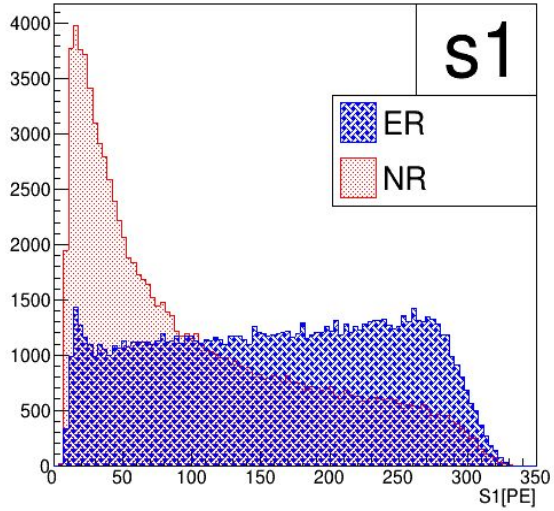

FP событий: 734
S_1
f_{90}

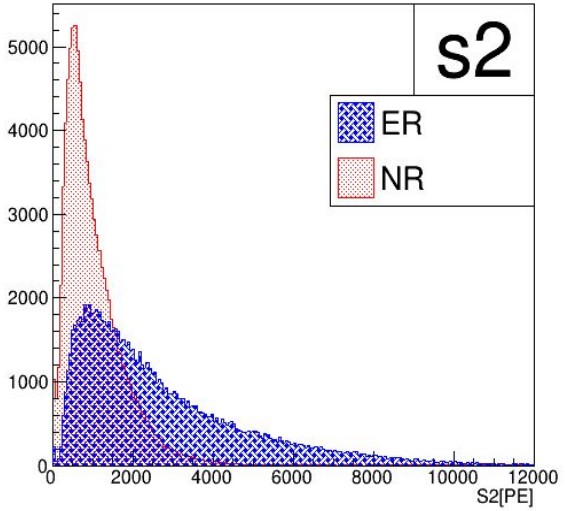

FN событий: 23
S_1
f_{90}

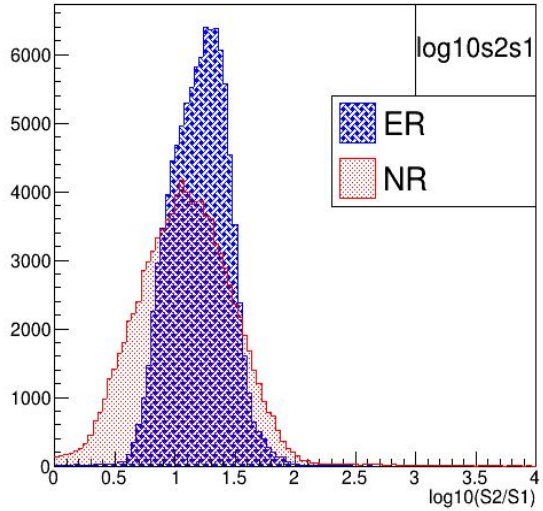

TN событий: 182516
S_1
f_{90}

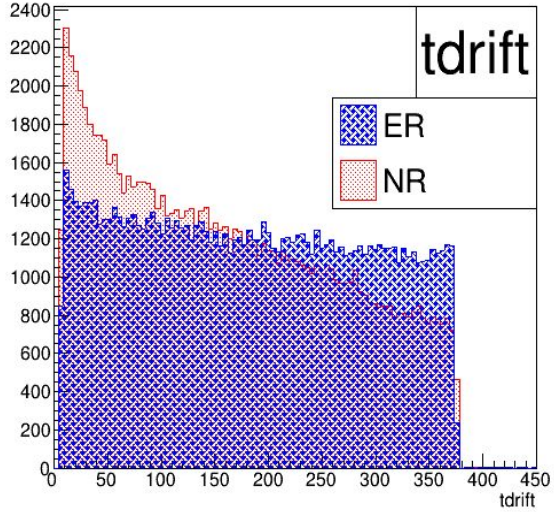
Используемые переменные

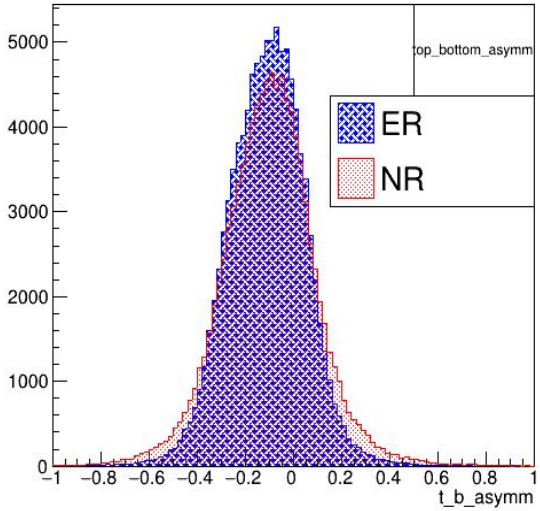


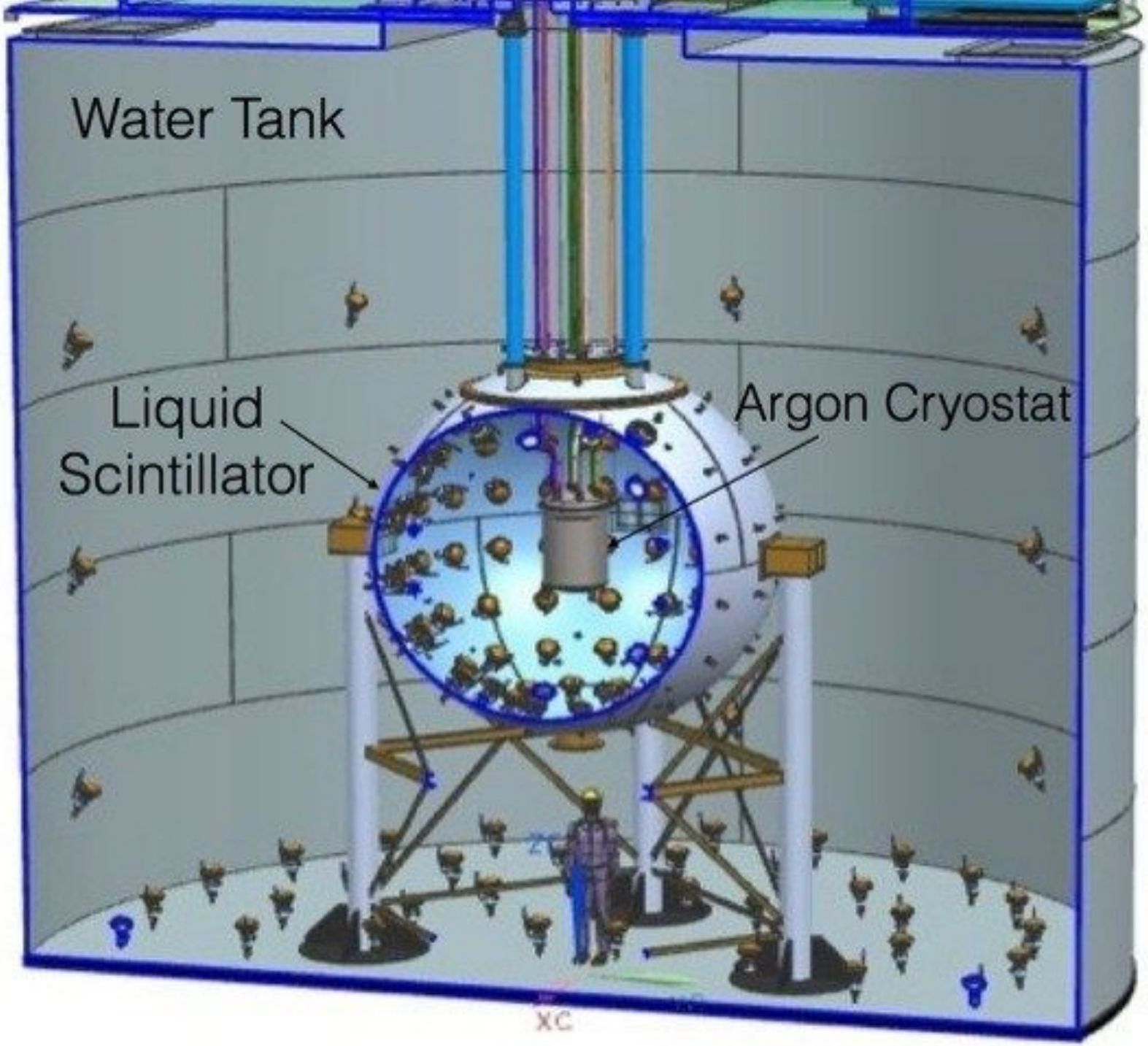




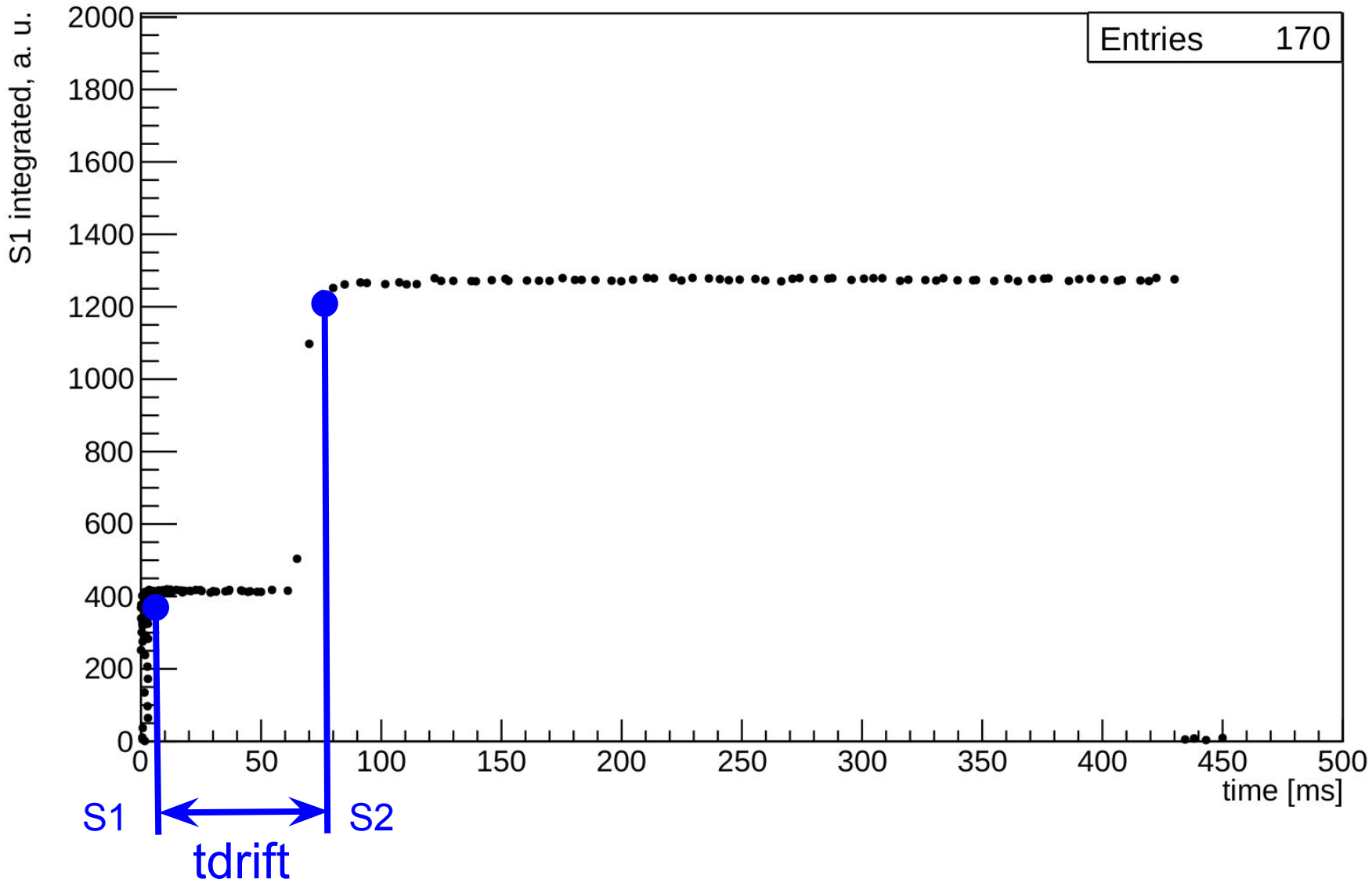


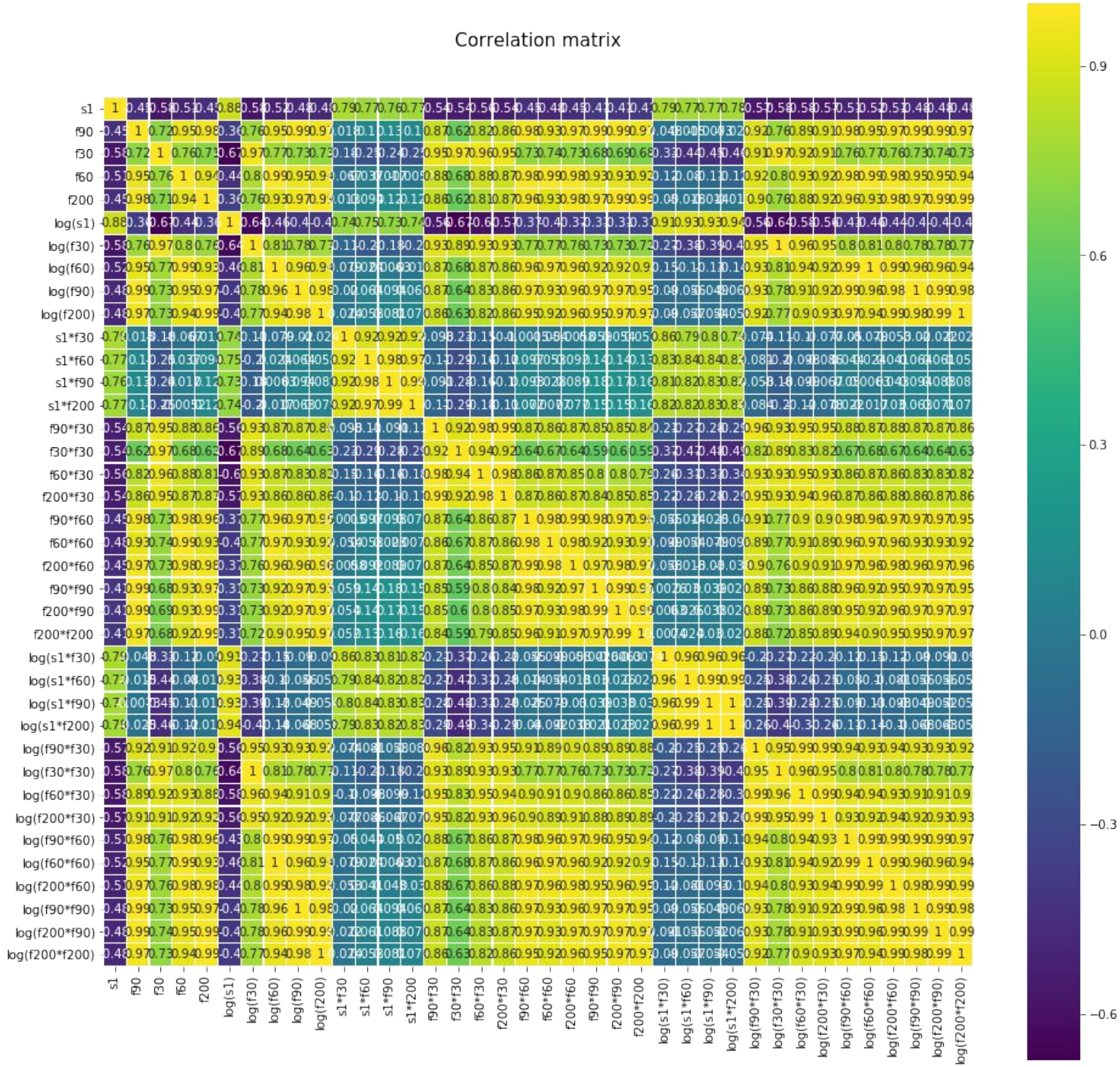





Water Tank
Liquid
Scintillator
Argon Cryostat

Используемые переменные

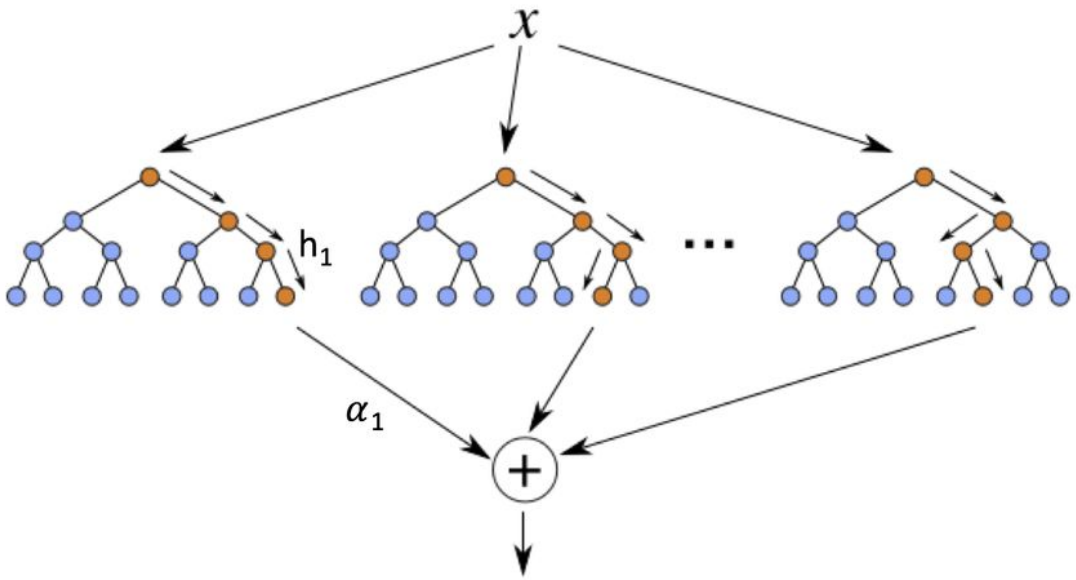



Correlation matrix
s1
f90
f30
f60
f200
log(s1)
log(f30)
log(f60)
log(f90)
log(f200)
s1*f30
s1*f60
s1*f90
s1*f200
f90*f30
f30*f30
f60*f30
f200*f30
f90*f60
f60*f60
f200*f60
f90*f90
f200*f90
f200*f200
log(s1*f30)
log(s1*f60)
log(s1*f90)
log(s1*f200)
log(f90*f30)
log(f30*f30)
log(f60*f30)
log(f200*f30)
log(f90*f60)
log(f60*f60)
log(f200*f60)
log(f90*f90)
log(f200*f90)
log(f200*f200)
0.9
0.6
0.3
0.0
−0.3
−0.6

Градиентный бустинг над деревьями решений

$$y \approx \hat{f}\,(x),$$

$$\hat{f}\,(x) = \arg\min_{f(x)} L(y, f(x))$$

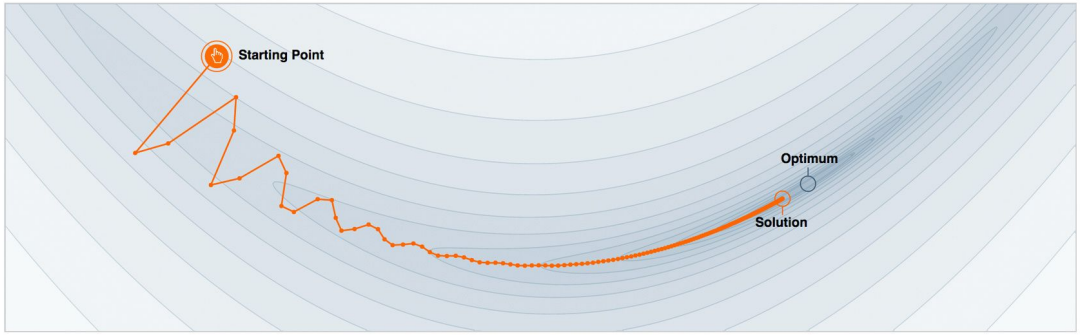


Где:

x - входные переменные,

y - целевые переменные,

L - функция потерь



Randomized Search/GridSearch

Randomized Search:

1) Задается набор значений для каждого интересующего параметра
2) Задается количество проверок **M**
3) Задается метрика проверки
4) Проверяется **M** классификаторов с параметрами, случайно взятыми из набора и выдаются параметры классификатора с наилучшим результатом по оцениваемой метрике

Grid Search:

1) Задается пространство гиперпараметров
2) Задается метрика проверки
3) Проверяются любые возможные комбинации параметров и выдаются параметры классификатора с наилучшим результатом по оцениваемой метрике