

На правах рукописи

Петров Никита Александрович

**Исследование состава космических лучей
в диапазоне энергий 1,5 – 100 ПэВ
на архивных данных эксперимента KASCADE
с использованием машинного обучения**

1.3.15. Физика атомных ядер и элементарных частиц,
физика высоких энергий

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2025

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте ядерных исследований Российской академии наук (ИЯИ РАН).

Научный руководитель:

Кузнецов Михаил Юрьевич, кандидат физико-математических наук, Федеральное государственное бюджетное учреждение науки Институт ядерных исследований Российской академии наук, лаборатория обработки больших данных в физике частиц и астрофизике, старший научный сотрудник.

Официальные оппоненты:

Быков Андрей Михайлович, доктор физико-математических наук, чл.-корр. РАН, профессор, Федеральное государственное бюджетное учреждение науки Физико-технический институт им. А.Ф.Иоффе Российской академии наук, отделение физики плазмы, атомной физики и астрофизики, руководитель отделения.

Деркач Денис Александрович, кандидат наук (PhD), Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики», Институт искусственного интеллекта и цифровых наук, директор института по прикладным исследованиям и разработкам.

Ведущая организация:

Федеральное государственное бюджетное образовательное учреждение высшего образования «Ярославский государственный университет им. П.Г.Демидова»

Защита состоится _____ г. в _____ часов на заседании диссертационного совета 24.1.163.01 на базе Федерального государственного бюджетного учреждения науки Института ядерных исследований Российской академии наук по адресу: 117312, Москва, проспект 60-летия Октября, 7А.

С диссертацией можно ознакомиться в библиотеке и на сайте ИЯИ РАН по адресу: <https://www.inr.ru>

Автореферат разослан _____

Ученый секретарь
диссертационного совета 24.1.163.01,
кандидат физико-математических наук

Демидов Сергей Владимирович

Общая характеристика работы

Актуальность темы

Космические лучи были открыты В. Гессом в 1912 г. в серии экспериментов на аэростатах по измерению ионизации в зависимости от высоты. В результате Гесс понял, что измеряемое им излучение связано не с землёй, а с космосом. Спустя более чем сто лет мы знаем, что космические лучи состоят по большей части из высокоэнергетичных атомных ядер и элементарных частиц, однако остаётся большое количество фундаментальных вопросов связанных с ними, например, вопросы об источниках космических лучей и механизмах их ускорения. Исследование энергетического спектра космических лучей и его массового состава может дать важную информацию для ответа на эти вопросы.

Энергетический спектр космических лучей, охватывающий диапазон энергий от ~ 10 ГэВ до 10^{11} ГэВ, в первом приближении может быть описан простым степенным законом, однако при более внимательном рассмотрении в нём на сегодняшний день обнаруживается ряд особенностей, таких, как отклонение от простого степенного закона для космических лучей с магнитной жёсткостью около 300 ГВ; так называемое «колени» в области $\sim 4 \cdot 10^6$ ГэВ; второе «колени» в области $\sim 3,2 \cdot 10^8$ ГэВ; «лодыжка» $10^{9,7}$ ГэВ и, наконец, обрезание спектра при $\sim 4 \cdot 10^{10}$ ГэВ. Причины возникновения этих особенностей являются предметом дискуссии, в которой одно из значимых мест занимает вопрос об интерпретации колена.

В области колена ~ 4 ПэВ энергетический спектр космических лучей резко меняет наклон, эта особенность спектра была впервые обнаружена в 1958 г. в эксперименте ШАЛ МГУ и впоследствии подтверждалась в работах. Разрабатываемые теории об источниках и механизмах ускорения космических лучей в качестве наблюдаемого эффекта предполагают различные позиции колена, его форму, химический состав. Часть теорий связывает колени в энергетическом спектре с максимальной энергией, которую могут достичь космические лучи в источниках, другие объясняют происхождение колена высвобождением космических лучей из нашей галактики. В обоих этих случаях положение излома в спектрах индивидуальных массовых компонент прямо зависит от заряда первичной частицы. В то же время существуют более экзотичные теории, предполагающие новую физику и зависимость энергии колена от массы первичной частицы.

Степень разработанности темы

Впервые систематическое изменение массового состава в области колена с увеличением энергии наблюдалось в эксперименте KASCADE, так-

же массовый состав изучался в работах TibetAS γ , KASCADE-Grande, IceTop, TALE, MAKET-ANI, Tunka-133, TAIGA, а также недавно опубликованном анализе LHAASO. И хотя другие эксперименты в целом подтверждают выводы KASCADE, между ними существуют разногласия, поэтому необходимо продолжать эксперименты и улучшать точность измерений.

На сегодняшний день прямое изучение космических лучей с энергиями порядка и выше ПэВ из-за их низкой интенсивности оказывается недоступным, и исследования проводятся только посредством регистрации вторичных частиц широких атмосферных ливней (ШАЛов). ШАЛы были открыты П. Оже в 1939 г. и представляют собой каскады элементарных частиц и атомных ядер, образующихся в результате взаимодействия первичного космического луча с атомами атмосферы Земли. В таком эксперименте наземные детекторные станции регистрируют различные компоненты ШАЛа и из этих данных делаются выводы о первичной частице. В области энергий 1–100 ПэВ продолжают работать эксперименты TALE, Tunka, IceTop, LHAASO, Ковер.

Основная сложность такого рода экспериментов связана с тем, что требуется точное моделирование развития ШАЛа методами Монте-Карло. Для такого моделирования необходимо детальное знание электромагнитных и адронных взаимодействий, однако в случае генераторов адронных взаимодействий при высоких энергиях, которые ответственны за начальное развитие ливня, продолжает сохраняться расхождение в наблюдаемых величинах для различных независимых генераторов. В настоящее время активно развиваются генераторы QGSJet, EPOS и Sibyll, и ситуация в целом улучшается, поэтому важно не только анализировать новые данные экспериментальных установок, но и проводить переобработку архивных измерений с учётом новых версий генераторов.

Для повторного анализа массового состава при энергиях 1–100 ПэВ хорошо подходит эксперимент KASCADE, поскольку за время работы с 1996 г. по 2013 г. им был собран большой объём качественных экспериментальных данных, который впоследствии был сохранён, задокументирован и выложен в открытый доступ вместе с Монте-Карло моделированием, основанным на современных версиях генераторов адронных взаимодействий. Кроме того, с момента публикации последних результатов KASCADE также получили развитие методы анализа данных с использованием машинного обучения, которые могут более эффективно использовать информацию из эксперимента. Применительно к анализу массового состава космических лучей, машинное

обучение использовалось ещё в эксперименте TibetAS γ , однако распространение получило сейчас и продолжает доказывать свою полезность.

Целью данной работы является исследование массового состава космических лучей в диапазоне энергий 1,5–100 ПэВ по архивным данным эксперимента KASCADE.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. разработать методику восстановления энергетических спектров индивидуальных массовых компонент с точностью, превышающей оригинальные результаты KASCADE;
2. разработать, обучить и провалидировать классификатор, реконструирующий массовый тип первичной частицы;
3. провести процедуру анфолдинга реконструированного типа и энергии первичных частиц, учесть систематические неопределённости;
4. проанализировать поведение спектров в зависимости от энергии, проверить спектры на предмет коленоподобных структур и изломов.

Научная новизна

1. Впервые была разработана и применена к данным эксперимента KASCADE методика по восстановлению спектров пяти массовых компонент (протонной, гелиевой, углеродной, кремниевой, железной) с использованием моделей машинного обучения.
2. Определены энергетические спектры пяти массовых компонент в диапазоне 1,5–100 ПэВ с точностью превышающей исследования KASCADE и других экспериментов.
3. С наибольшей на текущий момент статистической значимостью подтверждена коленоподобная структура протонной и гелиевой массовых компонент, а также впервые получено указание на излом в железном спектре в исследуемом диапазоне энергий.

Теоретическая и практическая значимость

- Независимое измерение энергетических спектров массовых компонент в области энергий 1,5–100 ПэВ необходимо для проверки теоретических моделей, объясняющих происхождение и механизмы ускорения галактических космических лучей.
- Использование методов машинного обучения для анализа данных подобных экспериментов открывает путь к увеличению точности результатов, благодаря более эффективному использованию информации.

- Повторный анализ открытых архивных экспериментальных данных с использованием новых методов демонстрирует важность сохранения и публикации данных и других экспериментов.

Методология и методы исследования

В основе диссертации лежат методы машинного обучения (нейронные сети, случайный лес) для решения задачи классификации массового типа первичной частицы, методы Монте-Карло, методы анфолдинга, а также методы статистической обработки данных.

Основные положения, выносимые на защиту

1. Разработанная методика с использованием методов машинного обучения позволяет определить энергетические спектры пяти индивидуальных массовых компонент (протоны, гелий, углерод, кремний, железо) в диапазоне энергий 1,5–100 ПэВ на архивных данных KASCADE.
2. Систематическая неопределённость энергетических спектров пяти индивидуальных массовых компонент (протоны, гелий, углерод, кремний, железо) космических лучей в диапазоне энергий 1,5–100 ПэВ на архивных данных KASCADE, полученных с помощью разработанной методики, ниже соответствующей в оригинальных работах KASCADE и в современных экспериментах (IceTop).
3. Получена оценка неопределённостей в спектрах индивидуальных массовых компонент, связанная с использованием трёх различных современных генераторов адронных взаимодействий: QGSJet-II.04, EPOS-LHC, Sibyll 2.3c.
4. Энергетический спектр протонной компоненты демонстрирует избыток (~ 10 раз при энергии 10 ПэВ), а спектры остальных компонент — недостаток относительно оригинальных результатов KASCADE.
5. Колоноподобное поведение в спектрах протонной и гелиевой массовых компонент наблюдается со значимостью более 5σ и 3σ при энергиях $\sim 4,4$ ПэВ и ~ 11 ПэВ соответственно. Указание на излом в спектре железной компоненты обнаружено при энергиях $\sim 4,5$ ПэВ.

Степень достоверности и апробация результатов

Достоверность полученных результатов обеспечивается с помощью различных проверок и тестов, проведённых в данной работе, сравнением полученных результатов с другими экспериментами, а также публикацией результатов данного исследования в рецензируемых научных журналах и обсуждениями на научных конференциях.

Основные результаты работы докладывались на конференциях:

1. Workshop on Machine Learning for Cosmic-Ray Air Showers (Delaware, USA, 2022);
2. The 27th European Cosmic Ray Symposium (ECRS, Nijmegen, Netherlands, 2022);
3. International Conference on Particle Physics and Cosmology (Yerevan, Armenia, 2023);
4. 38-я Всероссийская конференция по космическим лучам (Москва, Россия, 2024).

Результаты также демонстрировались и обсуждались на семинаре лаборатории обработки больших данных ИЯИ РАН 17.01.2024, экспериментальном семинаре ИЯФ СО РАН 01.03.2024; были представлены на конкурсе молодых учёных ИЯФ СО РАН 14.05.2024.

Публикации

По теме диссертации опубликовано 3 работы, из них 2 в научных журналах из списка ВАК [1; 2] а также 1 работа в трудах конференций [3]. Все 3 публикации проиндексированы в базе данных Scopus.

Личный вклад автора

Личный вклад автора в получение научных результатов, лежащих в основе диссертации является определяющим. Автором выполнены все этапы по предварительной обработке данных и обучению свёрточной нейронной сети в качестве классификатора типа первичной частицы, разработаны и проведены необходимые проверки по валидации классификатора. Автор провёл анфолдинг и получил полный и индивидуальные энергетические спектры массовых компонент в диапазоне энергий от 1,5 ПэВ до 100 ПэВ с учётом оценки систематических неопределённостей. А также, автором выполнена процедура по поиску коленоподобных структур и изломов в полученных энергетических спектрах и интерпретация результатов.

Авторский вклад в статье [1] заключается в следующем: автором проведена предварительная подготовка данных для обучения классификатора типа первичной частицы; предложена и обучена свёрточная нейронная сеть, проведены исследования по оптимизации параметров архитектуры сети; реализован метод случайных смесей для демонстрации и сравнения качества работы различных классификаторов. Автор исследовал влияние входных параметров на результат работы свёрточной нейронной сети, провёл ряд тестов с целью проверки корректности её работы, реализовал сравнение качества результатов реконструкции типа первичной частицы, полученных на Монте-Карло моделировании с использованием различных генераторов адронных взаимодействий. Автор предложил и реализовал процедуру анфолдинга энер-

гетических спектров отдельных массовых компонент, провёл систематизацию, анализ и оценку вклада источников систематических неопределённостей, а также продемонстрировал прямое сравнение результатов с последними оригинальными результатами эксперимента KASCADE на 20 % экспериментальных данных.

В статье [2] авторский вклад состоит в следующем: автором была применена методика, описанная в статье [1], к закрытой части (оставшиеся 80 %) экспериментальных данных, проведена оценка неопределённостей, связанных с использованием различных генераторов адронных взаимодействий, реализовано представление полученных результатов в виде зависимости среднего логарифма массы от энергии первичной частицы. Автор исследовал поведение энергетических спектров отдельных компонент на предмет коллоидных структур и изломов, сравнил полученные результаты с другими экспериментами, а также провёл интерпретацию полученных результатов.

Содержание работы

Во введении обосновывается актуальность и значимость данной работы в контексте существующих исследований, формулируются цели и ставятся задачи. Описывается объём и структура диссертации.

Первая глава посвящена описанию эксперимента KASCADE. Эксперимент KASCADE, располагавшийся в Карлсруэ, Германия, был предназначен для изучения первичного состава космических лучей в энергетическом диапазоне 500 ТэВ–100 ПэВ посредством регистрации широких атмосферных ливней (ШАЛов). Из различных компонент эксперимента, на протяжении всего его времени работы с 1996 г. по 2013 г. данные набирались на KASCADE Аггау. Данные с этой компоненты эксперимента и используются в текущем исследовании.

Общая схема KASCADE Аггау и его детекторных станций представлена на Рисунках 1 и 2. KASCADE Аггау состоял из 252 детекторных станций, расположенных в узлах прямоугольной сетки покрывающей площадь $200 \times 200 \text{ м}^2$. Станции внутренних четырёх кластеров (типа II) были оборудованы слоем жидкого сцинтиллятора для определения электромагнитной компоненты ШАЛа, а также ФЭУ для регистрации времени высвечивания. Остальные станции (типа I) дополнительно оборудованы пластиковым сцинтиллятором для определения мюонной компоненты под специальным защитным слоем.

В данной работе используются экспериментальные данные и моделирование Монте-Карло KASCADE, предоставленные проектом KCDC. ШАЛы

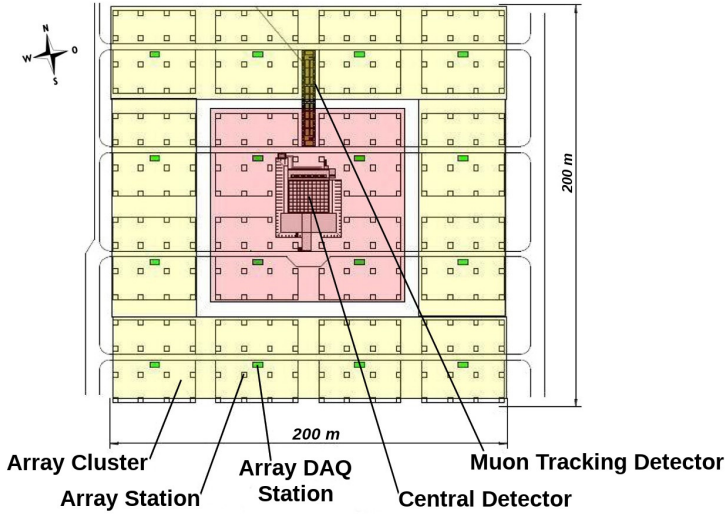


Рис. 1 — Схема KASCADE Array. Станции типа I (зелёный цвет) включают два класса детекторов (e/γ и μ) разделённых защитным слоем для одновременной регистрации электромагнитной и мюонной компоненты ШАЛов. Станции типа II (красный), расположенные в центральной области, содержат только e/γ детекторы

моделируются программой CORSIKA, отклик детектора симулируется пакетом GEANT3. Для событий моделирования и эксперимента реконструируется набор наблюдаемых параметров: интегральное число электронов (N_e) и мюонов (N_μ) в ливне, позиция ядра ливня (x, y), зенитный (θ) и азимутальный (ϕ) углы прихода первичной частицы, энергия первичной частицы (E), а также возраст ливня (s). В дополнение к откликам отдельных станций (энергосвечения e/γ , μ и временная метка попадания первой заряженной частицы) эти параметры формируют событие. Пример события из Монте-Карло моделирования протонов показан на Рисунке 3.

К используемым данным применяются критерии отбора рекомендованные KASCADE, а именно $\theta < 18^\circ$, $\sqrt{x^2 + y^2} < 91 \text{ м}$, $\log_{10} N_e > 4,8$, $\log_{10} N_\mu > 3,6$, $0,2 < s < 1,48$. Исследование массового состава космических лучей проводится, начиная с $10^{6,15} \text{ ГэВ}$, поскольку с учётом критериев отбора в этой области достигается полная эффективность триггера, а также начинает выходить на единицу эффективность критериев отбора. Объём экспериментальных данных составил $3,5 \cdot 10^6$ событий после применения критериев отбора в изучаемом диапазоне энергий, полная экспозиция $\mathcal{E}$

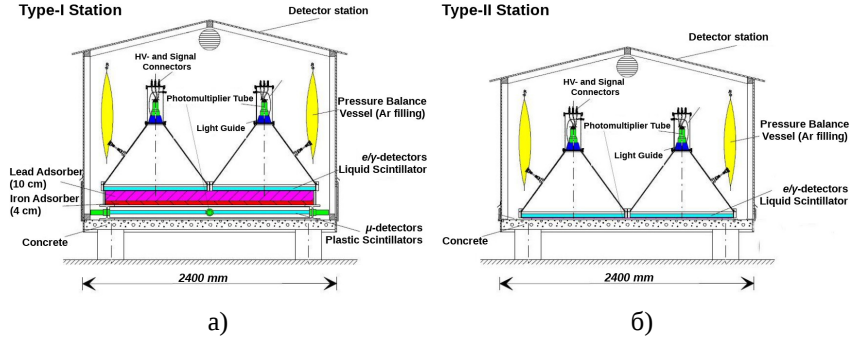


Рис. 2 — Схемы детекторных станций KASCADE Array типа I (а) и типа II (б)

Монте-Карло событие p : $\log_{10}(E/\text{эВ}) = 15.51$, $\theta = 20.78^\circ$, $\phi = 244.1^\circ$,
 $\log_{10} N_e = 5.32$, $\log_{10} N_\mu = 4.51$, $x = 15.3$ м, $y = 27.3$ м, $s = 0.88$

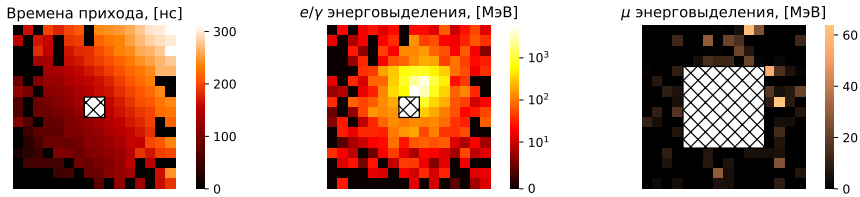


Рис. 3 — Пример события KASCADE из Монте-Карло моделирования протонов. Продемонстрированы реконструированные параметры события, а также времена прихода (слева), энерговыведения электромагнитной (центр) и мюонной компоненты (справа) со станций

составила

$$\mathcal{E} = (\pi \cdot \sin^2 18^\circ) \left(\pi \cdot (91 \text{ м})^2 \right) \cdot T_{lt},$$

где $T_{lt} = 4,2 \cdot 10^8$ с — полное время работы экспериментальной установки.

Также экспериментальный набор данных разбивался случайным образом на закрытую (80 %) и открытую (20 %) части данных. На открытой части проводились проверки методики восстановления спектров индивидуальных массовых компонент, описываемая во второй главе, а закрытая часть использовалась для получения финальных результатов, описываемых в третьей главе данной работы. Кроме того, для исследования влияния генераторов адронных взаимодействий высоких энергий использовались наборы Монте-Карло, полученные с использованием современных (post-LHC) гене-

раторов QGSJet-II.04 ($1,3 \cdot 10^5$), EPOS-LHC ($6,7 \cdot 10^4$), Sibyll 2.3c ($6,8 \cdot 10^4$), а также дополнительный набор Монте-Карло моделирования с помощью pre-LHC генератора QGSJet-II.02 ($3,7 \cdot 10^4$) для прямого сравнения с оригинальными результатами KASCADE, в скобках показан объём соответствующих наборов данных после применения критериев отбора в изучаемом диапазоне энергий. Все наборы Монте-Карло содержали события пяти исследуемых типов первичных частиц, спектральный индекс в Монте-Карло соответствовал $dN/dE \sim E^{-2}$ и перевешивался к $E^{-2.7}$ в ходе анализа.

Во **второй главе** подробно описывается разработанная процедура по восстановлению индивидуальных энергетических спектров пяти массовых компонент в диапазоне энергий 1,5–100 ПэВ.

В первую очередь проводилась разработка пособытийного классификатора типа первичной частицы (по пяти классам: протоны, гелий, углерод, кремний, железо); в частности были обучены модели машинного обучения, основанные на случайном лесе, многослойном перцептроне, свёрточной нейронной сети. Матрицы смешивания (confusion matrices), демонстрирующие в своих строках наборы вероятностей отнести первичную частицу к реконструированному типу для соответствующего истинного типа, проиллюстрированы на Рисунке 4 для трёх моделей, указанных выше, обученных и протестированных на Монте-Карло с генератором QGSJet-II.04.

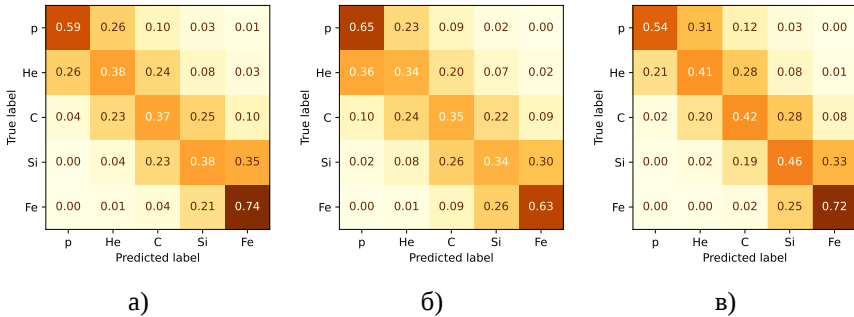


Рис. 4 — Матрицы смешивания для моделей машинного обучения: (а) случайный лес, (б) многослойный перцептрон, (в) свёрточная нейронная сеть, обученных и протестированных на наборе моделирования Монте-Карло с генератором QGSJet-II.04

Среди использованных моделей машинного обучения наилучшее качество показала свёрточная нейронная сеть (Convolutional Neural Network —

CNN), поэтому именно она использовалась в дальнейшем при разработке методики. Архитектура CNN вдохновлена классической свёрточной моделью LeNet-5 и содержала $\sim 30\,000$ обучаемых параметров. На вход этой нейронной сети подавались энерговыделения e/γ , μ со станций KASCADE в виде двухканального изображения 16×16 , а также реконструированные параметры N_e , N_μ , θ , s .

Далее текст посвящён детальному исследованию поведения, свойств и качества предсказаний используемой свёрточной нейронной сети. Сначала исследовалась важность отдельных входных параметров, для того чтобы понять, из каких признаков модель машинного обучения извлекает дискриминирующую информацию и насколько они физичны. С этой целью были обучены несколько версий CNN с различным набором входных параметров и их результаты сравнивались между собой. Выяснено, что наиболее важными характеристиками данных для нейронной сети являются матрицы с e/γ и μ энерговыделениями со станций, а также реконструированный зенитный угол θ . Из важных особенностей было отмечено, что e/γ и μ энерговыделения взаимозаменяемы с реконструированными параметрами N_e и N_μ соответственно, т. е. нейронная сеть умеет извлекать эти параметры из «сырых» энерговыделений. Также отмечено, что классификатор ведёт себя физическим образом, поскольку ему необходимо использование именно соотношение между e/γ и μ энерговыделениями, без любого одного из них качество результатов существенно деградирует, именно соотношение между электромагнитной и мюонной компонентой даёт наибольшую информацию о типе первичной частицы в ШАЛе, поскольку в зависимости от типа частицы меняется средняя высота начала развития ШАЛа, и соответственно пропорции между электромагнитной и мюонной компонентами во время регистрации. Ещё одной особенностью— влияние параметра θ , без использования которого матрица смешивания классификатора также деградирует. Этот параметр не является взаимозаменяемым с другими, использованными в нейронной сети, поскольку извлекается из временных меток первых попаданий заряженных частиц на детекторные станции, что может сообщать важную информацию о толщине атмосферы, в которой распространялся атмосферный ливень и, таким образом, улучшать качество классификации.

Также проверялась зависимость качества классификации от реконструированной энергии и зенитного угла первичной частицы, поскольку не интерпретируемые с точки зрения физики результаты могут помочь выявить возможные ошибки в процессе разработки и обучения модели машинного обучения. В результате была проверена и показана независимость предсказанных свёрточной нейронной сетью долей первичных частиц в зависимости

от реконструированного зенитного угла. Кроме того, было определено, что качество восстановления существенно не деградирует для стандартного отбора по зенитному углу ($\theta < 18^\circ$) и расширенного ($\theta < 30^\circ$) (среднее снижение диагональных элементов матрицы смешивания менее 1 %), что может быть полезно в задачах, где необходима дополнительная статистика. А показанные на Рисунке 5 матрицы смешивания CNN в зависимости от интервала по реконструированной энергии первичной частицы демонстрируют улучшение качества классификации с увеличением энергии. Это можно объяснить

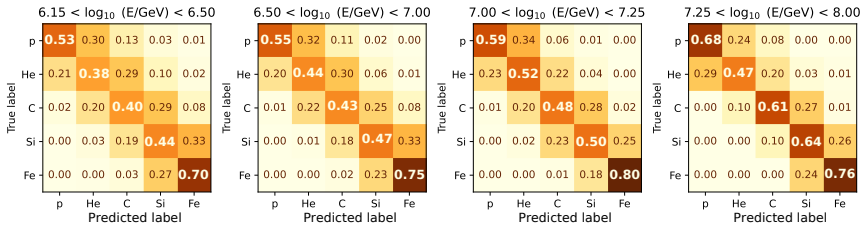


Рис. 5 — Матрицы смешивания CNN для различных диапазонов по реконструированной энергии. Нейронная сеть обучалась и тестировалась на Монте-Карло с генератором QGSJet-II.04

тем, что при более высоких энергиях первичной частицы, высота, на которой достигается максимальное развитие ШАЛа, становится ниже и ШАЛы покрывают большую площадь, тем самым, энерговыделения регистрируются в большем количестве станций, как следствие, сигнатура таких ливней чётче и возможность их покомпонентного разделения выше.

Поскольку, как было отмечено, качество классификации меняется с энергией, то при последующем анфолдинге результатов, необходимо использовать матрицы отклика для соответствующего интервала по энергии. Однако из-за того, что в Монте-Карло заложен спектр $dN/dE \sim E^{-2}$, то статистика на больших энергиях резко падает, поэтому было решено для анфолдинга использовать объединённые данные из энергетических интервалов $\log_{10}(E/\text{ГэВ})$: 6,15—6,25, 6,25—6,5, 6,5—7, 7—8. В этих диапазонах матрицы ведут себя стабильно и, таким образом, объединение не окажет влияние на анфолдинг, что также видно из Рисунка 6, который иллюстрирует зависимость диагональных элементов от энергии первичной частицы.

В Монте-Карло моделировании в каждом событии все детекторные станции KASCADE Агга стабильно работают, однако в данных доля экспериментальных заходов с неработающими станциями была оценена в $\sim 50\%$,

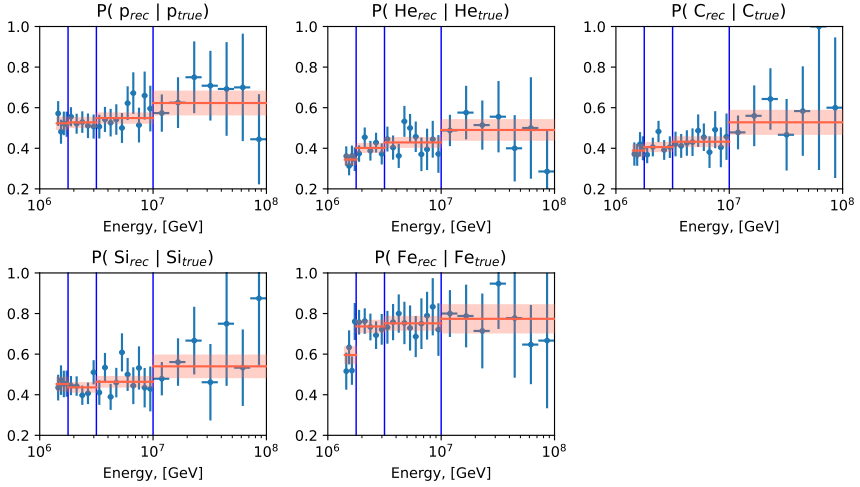


Рис. 6 — Зависимость диагональных элементов матриц смешивания от реконструированной энергии для CNN, обученной и тестируемой с генератором адронных взаимодействий QGSJet-II.04. Синие вертикальные линии обозначают диапазоны, в которых Монте-Карло объединяется для последующей процедуры анфолдинга. Красные горизонтальные линии показывают средние значения данных матричных элементов в ограниченных диапазонах энергий вместе с их неопределённостями

поэтому проводилось исследование влияния неработающих детекторных станций на результат работы используемого классификатора. Чтобы оценить влияние таких событий на качество классификации, необходимо симитировать неработающие детекторные станции в Монте-Карло моделировании («испортить» моделирование) и проверить работу стандартного классификатора на таких данных. Поскольку испортить просто только энерговыделения, зануляя случайным образом часть из них, как показано на Рисунке 7 (но не реконструированные параметры, которые поставляются KASCADE), то для более корректной оценки влияния необходимо сравнивать матрицы смешивания полученные для классификатора, обученного только на энерговыделениях. Средняя деградация диагональных элементов матриц смешивания CNN, обученной на энерговыделениях стандартного Монте-Карло, и протестированной на стандартном и «испорченном» наборе Монте-Карло составила $\sim 2\%$, и далее пересчитывается в систематическую неопределённость.

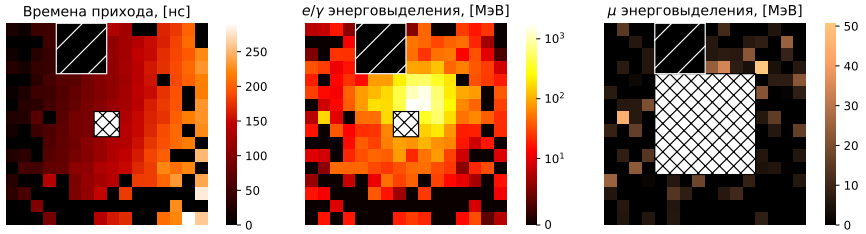


Рис. 7 — Пример «испорченного» события из Монте-Карло. Область зануления энергосвечений показана квадратом с белой штриховкой

Далее исследовалось влияние на результат работы классификатора трёх современных (post-LHC) генераторов адронных взаимодействий высоких энергий: QGSJet-II.04, EPOS-LHC, Sibyll 2.3c. Для этого три версии CNN, обученные на Монте-Карло моделировании с описанными выше генераторами, проверялись на тестовых частях каждого из трёх наборов Монте-Карло. В результате было отмечено, что качество классификации для одинаковых генераторов в обучении и проверке похожее, наибольшие значения вероятностей матрицы смешивания достигают в диагональных элементах, какие-либо выбросы или артефакты в матрицах отсутствуют, максимальное смешивание происходит для соседних массовых типов. Sibyll 2.3c выглядит в сравнении с остальными более «железным» генератором, поскольку классификатор, обученный на нём, систематически определяет компоненты в остальных наборах Монте-Карло более тяжёлыми, чем они заявлены, по аналогии, EPOS-LHC выглядит более «протонным». Поскольку для перекрёстных матриц смешивания заметны различия между генераторами, то их влияние на результат необходимо учитывать, речь об этом пойдёт в дальнейшем.

Затем были построены энергетические спектры индивидуальных массовых компонент на открытой части экспериментальных данных согласно используемой CNN без поправок на матрицу смешивания по следующей формуле:

$$\frac{dJ_j}{dE}(E_i) = \frac{N_{ij}}{\mathcal{E} \Delta E_i},$$

где $\frac{dJ_j}{dE}(E_i)$ — дифференциальный поток первичных частиц типа j (протоны, гелий, углерод, кремний, железо) в i -м энергетическом интервале, N_{ij} — количество первичных частиц типа j в энергетическом интервале i , полученное согласно CNN, $\mathcal{E}$ — экспозиция, ΔE_i — ширина i -го интервала по энергии. Спектры индивидуальных массовых компонент (с дополнительным фактором $E^{2.5}$ для лучшего визуального восприятия), полученных данным спосо-

бом, для CNN, обученной на Монте-Карло с генератором QGSJet-II.04, показаны на Рисунке 8.

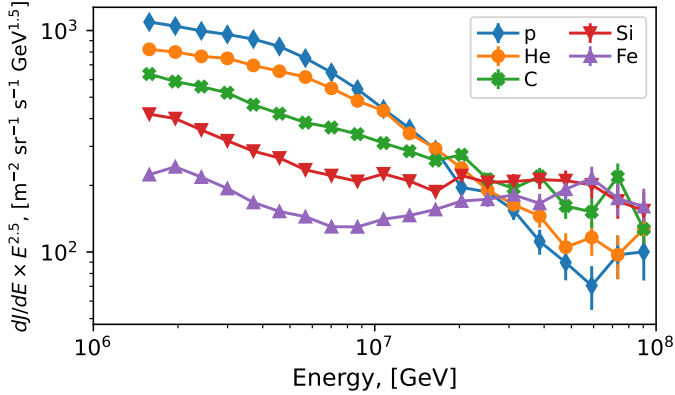


Рис. 8 — Спектры индивидуальных массовых компонент, полученные прямым предсказанием CNN, в зависимости от реконструированной энергии. CNN классификатор обучен на Монте-Карло с генератором QGSJet-II.04. Вертикальными линиями показаны статистические погрешности

Было отмечено, что спектры оказались достаточно гладкими в зависимости от реконструированной энергии, хотя CNN не получает информацию об этой переменной. Качественное поведение демонстрирует уменьшение долей массовых компонент вместе с увеличением их массового числа на низких энергиях и кардинальное изменение порядка на энергиях выше ~ 20 ПэВ. Также отмечается, что необходим анфолдинг для более точного определения спектров индивидуальных компонент.

Процедура анфолдинга, целью которой является определение оригинального распределения по наблюдаемому, может быть записана следующим образом

$$\nu_i = \sum_{j=1}^M R_{ij} \mu_j, i = 1, \dots, N, \quad (1)$$

ν_i — вектор ожидаемого числа событий в наблюдаемом распределении, μ_i — вектор ожидаемого числа событий в оригинальном распределении, N/M — число бинов в наблюдаемом / оригинальном распределении, R_{ij} — матрица отклика.

На первом шаге проводится анфолдинг по типу первичной частицы согласно формуле, записанной в терминах байесовских вероятностей,

$$P(A_i^{\mathcal{R}}|E_m^{\mathcal{R}}) = \sum_j P(A_i^{\mathcal{R}}|A_j^{\mathcal{T}}, E_m^{\mathcal{R}}) \cdot P(A_j^{\mathcal{T}}|E_m^{\mathcal{R}}), \quad (2)$$

где $A_i^{\mathcal{R}/\mathcal{T}}$ — i -й реконструированный/истинный тип первичной частицы, $E_m^{\mathcal{R}/\mathcal{T}}$ — m -й интервал реконструированной/истинной энергии первичной частицы. (Реконструированный тип первичной частицы — это тип, предсказанный CNN классификатором.) Это представление аналогично записанному в Формуле (1) и целью является определение $P(A_j^{\mathcal{T}}|E_m^{\mathcal{R}})$, причём матрица отклика R_{ij} — это матрица смешивания для классификатора CNN в данном случае с точностью до транспонирования, и, таким образом, извлекается из Монте-Карло.

На втором шаге проводится анфолдинг по энергии первичной частицы согласно формуле

$$P(E_m^{\mathcal{R}}) = \sum_j P(E_m^{\mathcal{R}}|E_j^{\mathcal{T}}) \cdot P(E_j^{\mathcal{T}}), \quad (3)$$

откуда определяется $P(E_j^{\mathcal{T}})$.

Таким образом, скомбинировав два результата, получаем следующее

$$\begin{aligned} \mu(A_i^{\mathcal{T}}, E_j^{\mathcal{T}}) &= \mu_{tot} P(A_i^{\mathcal{T}}, E_j^{\mathcal{T}}) = \\ &= \mu_{tot} P(A_i^{\mathcal{T}}|E_j^{\mathcal{T}}) P(E_j^{\mathcal{T}}) \rightarrow \mu_{tot} P(A_i^{\mathcal{T}}|E_j^{\mathcal{R}}) P(E_j^{\mathcal{T}}), \end{aligned} \quad (4)$$

где μ_{tot} — полное число событий, которое не меняется в процессе анфолдинга, поскольку в данном исследовании матрицы отклика фиксируют только перераспределение компонент и энергий. Так же, как видно из Формулы (4), производится замена $P(A_i^{\mathcal{T}}|E_j^{\mathcal{T}})$ на $P(A_i^{\mathcal{T}}|E_j^{\mathcal{R}})$, чтобы использовать результат анфолдинга по типу массовой компоненты. Эта замена является источником систематической неопределённости, и её влияние оценивается в соответствующем разделе.

Наконец, для учёта неединичной эффективности критериев отбора на двух нижних энергетических интервалах, проводилась коррекция

$$\mu^0(A_i^{\mathcal{T}}, E_j^{\mathcal{T}}) = \frac{\mu(A_i^{\mathcal{T}}, E_j^{\mathcal{T}})}{\epsilon(A_i^{\mathcal{T}}, E_j^{\mathcal{T}})}, \quad (5)$$

где μ^0 — скорректированное число событий, ϵ — эффективность регистрации событий конкретного массового типа в заданном энергетическом интервале (извлекается из соответствующего Монте-Карло).

Технически анфолдинг проводился байесовским итерационным методом в интервале энергий $10^{6,15} - 10^8$ ГэВ на 20 интервалах, эквидистантных в масштабе десятичного логарифма.

На следующем этапе проводилось изучение систематических погрешностей. Сначала рассматривались систематические неопределённости, связанные с использованием Монте-Карло с каким-либо конкретным генератором адронных взаимодействий. Такие неопределённости назывались «базовыми» для соответствующего генератора.

Среди источников таких погрешностей выделялись:

- неопределённости анфолдинга, включающие возможное смещение из-за регуляризации анфолдинга, погрешности элементов используемых матриц отклика, замену в процедуре анфолдинга (см. Формулу (4));
- погрешности Монте-Карло, включающие неопределённости связанные с заложенным массовым составом используемого Монте-Карло, из-за спектрального индекса Монте-Карло, из-за неработающих детекторных станций в экспериментальных данных.

В Таблице 1 указано сравнение оценок погрешностей различных источников для Монте-Карло с генератором QGSJet-II.02. Для массового анфол-

Таблица 1 — Список относительных значений систематических неопределённостей для Монте-Карло моделирования с генератором QGSJet-II.02

Тип	Масс. анф.	Эн. анф.
Регуляризация анфолдинга	1–11 %	0,1–13 %
Ограниченное Монте-Карло	6–20 %	2–15 %
Замена вероятностей	0–13 %	–
Энергетическое разрешение	–	8–10 %
Спектральный индекс	–	0,5–7 %
Неработающие детекторы	1–20 %	4 %
Суммарно: 13–30 %	6–24 %	11–20 %

динга приведён интервал относительных неопределённостей, усреднённый по всем типам первичных частиц. Суммарная неопределённость («базовая» для QGSJet-II.02) определялась как корень из суммы квадратов всех исследованных неопределённостей.

На Рисунке 9 представлено прямое сравнение полученных на открытой части экспериментальных данных индивидуальных энергетических спектров массовых компонент с использованием Монте-Карло моделирования с генератором QGSJet-II.02 и оригинальных результатов KASCADE.

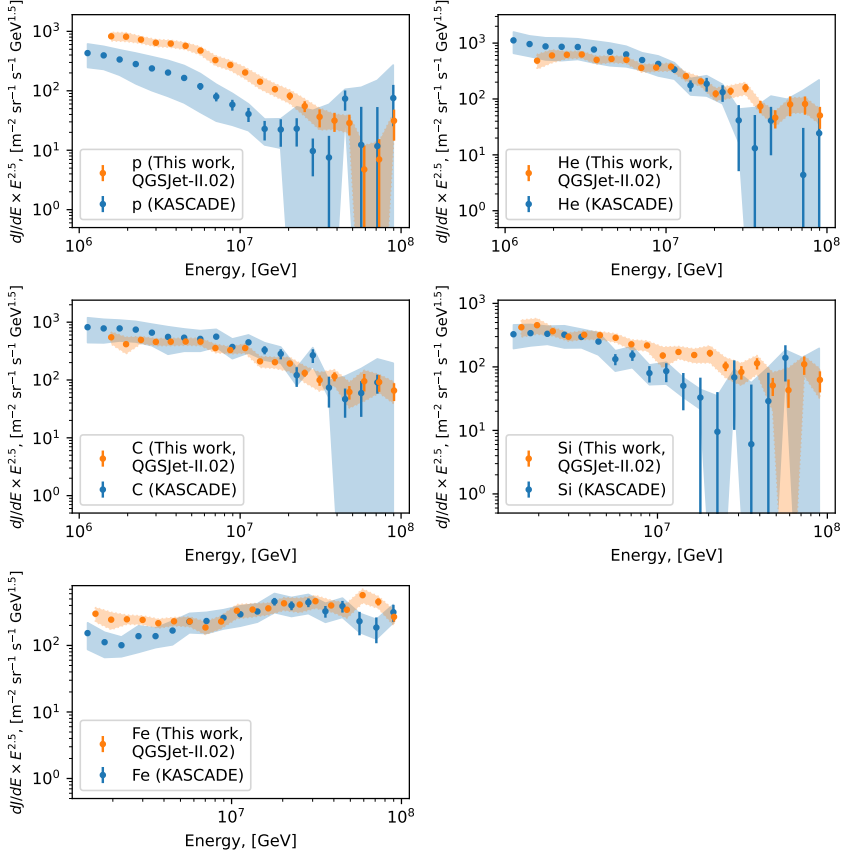


Рис. 9 — Сравнение индивидуальных спектров массовых компонент, полученных в текущей работе на открытой части экспериментальных данных (оранжевый), и оригинальных результатов KASCADE (синий). В обоих случаях результаты основаны на Монте-Карло с генератором QGSJet-II.02. Статистические ошибки изображены в виде вертикальных линий, систематические погрешности — в виде полосы соответствующих цветов

Отмечается, что систематические неопределённости используемого метода выше чем в оригинальном исследовании KASCADE. Также видно заметное различие в спектрах индивидуальной компоненты протонов, где текущий метод демонстрирует заметное превышение относительно KASCADE, например, в около *пяти* раз при энергии 10 ПэВ, и, таким образом, в химическом составе до энергий ~ 5 ПэВ доминируют протоны. Остальные массовые компоненты согласуются с результатами KASCADE с учётом систематических неопределённостей.

В третьей главе проводится восстановление энергетических спектров пяти массовых компонент на закрытой части экспериментальных данных (80 %) с использованием Монте-Карло моделирования, основанном на современных (post-LHC) генераторах адронных взаимодействий QGSJet-II.04, EPOS-LHC, Sibyll 2.3c.

В качестве основного адронного генератора используется QGSJet-II.04, а остальные генераторы используются для оценки неопределённостей, связанной с их использованием. Восстановление массового состава производится аналогично согласно разработанной методике: обучение CNN на тренировочной части Монте-Карло с генератором QGSJet-II.04 для пособытийной реконструкции типа первичной частицы, полный анфолдинг полученных результатов по типу и энергии первичной частицы, определение систематических погрешностей.

Далее демонстрируется оценка неопределённостей, связанных с выбором среди современных (post-LHC) генераторов адронных взаимодействий, которые в данной работе называются теоретическими. Для этого в дополнение к проведённому анфолдингу (согласно QGSJet-II.04) выполняется полный анфолдинг для двух остальных наборов Монте-Карло (с генераторами EPOS-LHC и Sibyll 2.3c). В каждом случае определялась соответствующая полоса «базовых» неопределённостей, а теоретическая неопределённость индивидуальных спектров оценивалась как диапазон между минимальной и максимальной «базовой» неопределённостью для каждого энергетического интервала.

Полученные в результате спектры пяти индивидуальных массовых компонент в диапазоне энергий 1,5–100 ПэВ с учётом теоретических неопределённостей изображены на Рисунке 10. Качественное поведение спектров не отличается от полученных на QGSJet-II.02, на низких энергиях доли массовые компоненты убывают в порядке увеличения массового числа, изменение этого поведения происходит на энергиях 20–50 ПэВ. Однако в абсолютных значениях поток протонной компоненты в данном случае оказывает-

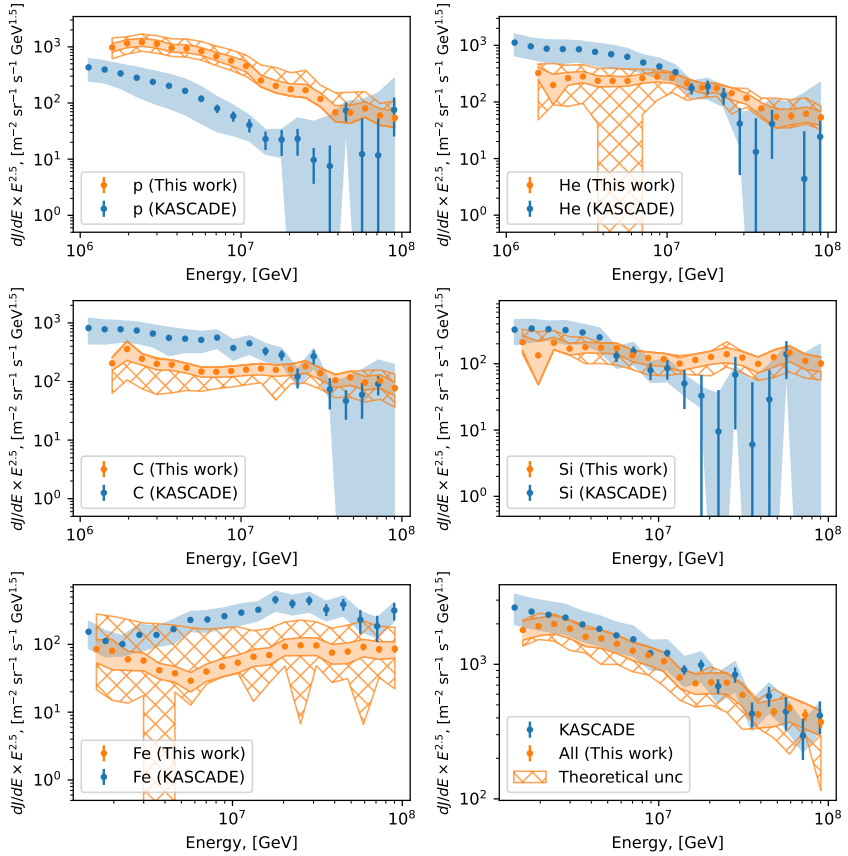


Рис. 10 — Энергетические спектры индивидуальных массовых компонент с учётом теоретической неопределённости, полученные в данном исследовании на закрытой части экспериментальных данных. Для иллюстрации продемонстрированы оригинальные спектры KASCADE (синим), полученные с Монте-Карло с генератором QGSJet-II.02. Статистические ошибки изображены в виде вертикальных линий, систематические погрешности — в виде полос соответствующих цветов, неопределённости с учётом теоретической обозначены штриховкой. Систематические погрешности спектров данной работы, показанные на рисунке, являются «базовыми» для генератора QGSJet-II.04

ся большим по сравнению с полученными выше результатами QGSJet-II.02, в ~ 2 раза при энергии 10 ПэВ.

Для сравнения с результатами экспериментов IceTop (Sibyll 2.1), TALE (EPOS-LHC) и LHAASO (QGSJet-II.04) полученные спектры (QGSJet-II.04) были представлены в виде среднего логарифма массового числа ($\langle \ln A \rangle$), которые показаны на Рисунке 11. В скобках указаны генераторы, «базовые» систематические погрешности которых представлены на рисунке. Теоретические неопределённости в текущей работе определялись с использованием генераторов QGSJet-II.04, EPOS-LHC, Sibyll 2.3c в случае LHAASO — QGSJet-II.04, EPOS-LHC, Sibyll 2.3d. Результаты согла-

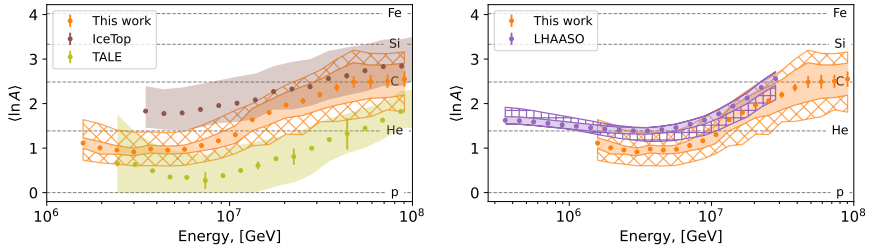


Рис. 11 — Сравнение среднего массового логарифма ($\langle \ln A \rangle$) в зависимости от энергии для текущего метода (оранжевый), результатов IceTop (коричневый), TALE (оливковый), и LHAASO (фиолетовый). Сплошной заливкой показаны «базовые» систематические погрешности соответствующих генераторов адронных взаимодействий, а штриховкой теоретические неопределённости

суются с учётом теоретических неопределённостей, причём верхняя граница теоретических неопределённостей в данной работе обусловлена результатами, полученными с генератором Sibyll 2.3c, а нижняя — с генератором EPOS-LHC.

Также проводился поиск коленоподобных структур в энергетических спектрах индивидуальных массовых компонент. Для этого проводилась аппроксимация спектров простой (PL, power law) и ломаной (BPL, broken power law) степенными функциями. с учётом «базовых» систематических погрешностей QGSJet-II.04. Статистическая значимость, насколько предпочтительнее гипотеза с изломом относительно гипотезы без излома, определялась из теста отношения правдоподобия. Среди индивидуальных спектров статистически значимое коленоподобное поведение удалось обнаружить для про-

тонной ($5,2\sigma$) и гелиевой ($3,9\sigma$) массовых компонент при энергиях $\sim 4,4$ ПэВ и ~ 11 ПэВ соответственно, которое является наиболее значимым среди других исследований на текущий момент. Кроме того, в спектре железной компоненты впервые показано указание ($2,4\sigma$) на излом при энергии $\sim 4,5$ ПэВ, которое может быть интерпретировано как аналог излома протонной компоненты при 166 ТэВ, продемонстрированном в эксперименте GRAPES-3. Описанные энергетические спектры, а также их аппроксимации представлены на Рисунке 12.

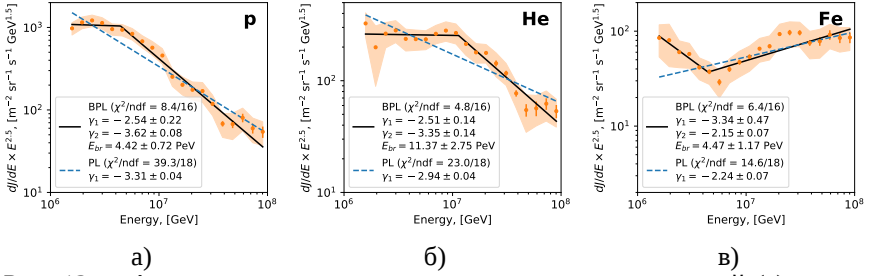


Рис. 12 — Аппроксимации энергетических спектров протонной (а), гелиевой (б) и железной (в) массовых компонент простой (PL, синий пунктир) и ломаной (BPL, чёрная линия) степенными функциями

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Разработана методика восстановления энергетических спектров пяти индивидуальных массовых компонент (протоны, гелий, углерод, кремний, железо) в диапазоне энергий 1,5–100 ПэВ на архивных данных эксперимента KASCADE в диапазоне энергий 1,5–100 ПэВ с использованием классификатора на основе машинного обучения для пособытийной реконструкции массовой компоненты первичной частицы, а также последовательного анфолдинга реконструированных массовой компоненты и энергии первичной частицы. При прямом сравнении продемонстрировано, что точность восстановления массового состава методикой, разработанной в данной работе, выше, чем в оригинальном исследовании KASCADE, а также в современных экспериментах (IceTop).
2. Получены энергетические спектры пяти массовых компонент в диапазоне энергий от 1,5 ПэВ до 100 ПэВ с использованием Монте-Карло моделирования, основанном на современных генераторах ад-

ронных взаимодействий QGSJet-II.04, EPOS-LHC, Sibyll 2.3c. Дополнительно результат продемонстрирован в виде среднего логарифма массового числа в зависимости от энергии первичной частицы. Отдельно учтены «теоретические» неопределённости, связанные с выбором генератора.

3. Продemonстрировано доминирующее поведение спектра протонной компоненты над остальными в диапазоне энергий 1,5–10 ПэВ (~ 10 раз при энергии 10 ПэВ), находящееся в противоречии с оригинальными результатами KASCADE, в которых доминирует гелиевая компонента.
4. Проведено сравнение полученных в данной работе результатов с последними результатами экспериментов IceTop, TALE и LHAASO. Показано согласие результатов с ними в рамках «теоретических» неопределённостей.
5. Исследовано поведение энергетических спектров индивидуальных массовых компонент. Впервые обнаружено статистически значимое коленоподобное поведение в случае протонной (более 5σ) и гелиевой (более 3σ) компонент при энергиях $\sim 4,4$ ПэВ и ~ 11 ПэВ соответственно. Впервые получено указание ($2,4\sigma$) на излом в железной компоненте при энергии $\sim 4,5$ ПэВ.

Список работ, опубликованных автором по теме диссертации

1. Methods of machine learning for the analysis of cosmic rays mass composition with the KASCADE experiment data / M. Y. Kuznetsov, N. A. Petrov, I. A. Plokhikh, V. V. Sotnikov. — Текст: электронный // Journal of Instrumentation. — 2024. — Янв. — Т. 19, № 01. — P01025. — URL: <https://doi.org/10.1088/1748-0221/19/01/P01025> ; Дата публикации: 25.01.2024.
2. Energy spectra of elemental groups of cosmic rays with the KASCADE experiment data and machine learning / M. Yu. Kuznetsov, N. A. Petrov, I. A. Plokhikh, V. V. Sotnikov. — Текст: электронный // Journal of Cosmology and Astroparticle Physics. — 2024. — Май. — Т. 2024, № 05. — С. 125. — URL: <https://doi.org/10.1088/1475-7516/2024/05/125> ; Дата публикации: 30.05.2024.
3. Towards mass composition study with KASCADE using deep neural networks / V. Sotnikov, M. Kuznetsov, N. Petrov, I. Plokhikh. — Текст: электронный // Proceedings of 27th European Cosmic Ray Symposium — PoS(ECRS). Т. 423. — 2023. — С. 092. — URL: <https://doi.org/10.22323/1.423.0092> ; Дата публикации: 14.12.2023.

Петров Никита Александрович
Исследование состава космических лучей
в диапазоне энергий 1,5 – 100 ПэВ
на архивных данных эксперимента KASCADE
с использованием машинного обучения

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Принято в печать 13.02.2025

Ф-т 60х84/16 Уч.-изд.л. 1,2 Зак. № 018/25 Тираж 80 экз. Бесплатно

Печать цифровая

Федеральное государственное бюджетное учреждение науки
Институт ядерных исследований Российской академии наук

Издательский отдел
117312, Москва, проспект 60-летия Октября, 7а