УДК 519.6:539.1

ИНТЕРПРЕТИРУЕМОЕ ВОССТАНОВЛЕНИЕ НЕЙТРОННЫХ СПЕКТРОВ НЕЙРО-НЕЧЁТКОЙ МОДЕЛЬЮ С ПОЛНОЙ SHAP-РЕГУЛЯРИЗАЦИЕЙ1

Лебедев Александр Дмитриевич[1], Трофимов Юрий Владиславович[2], Кондрашова Елизавета Сергеевна[3], Лебедев Михаил Дмитриевич[4], Ильин Андрей Сергеевич[5], Чижов Константин Алексеевич[6]

[1]Студент;
Государственный университет «Дубна»;
Россия, 141980, Московская область, г. Дубна, ул. Университетская, 19;
e-mail: lad.24@uni-dubna.ru.

[2]Младший научный сотрудник;
Государственный университет «Дубна»;
Россия, 141980, Московская область, г. Дубна, ул. Университетская, 19;
Инженер-программист;
Объединённый институт ядерных исследований;
Россия, 141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
e-mail: ytrofimov@jinr.ru.

[3]Ассистент;
Государственный университет «Дубна»;
Россия, 141980, Московская область, г. Дубна, ул. Университетская, 19;
e-mail: kes.th@uni-dubna.ru.

[4]Студент;
Национальный исследовательский технологический университет «МИСИС»;
Россия, 119049, г. Москва, Ленинский пр-кт, 4, стр. 1.;
e-mail: lebedevmisha2003@gmail.ru.

[5]Студент;
Университет Иннополис;
Россия, 420500, Республика Татарстан, г. Иннополис, ул. Университетская, 1;
e-mail: a.ilin@innopolis.university.

[6]Младший научный сотрудник;
Государственный университет «Дубна»;
Россия, 141980, Московская область, г. Дубна, ул. Университетская, 19;
Инженер-программист;
Объединённый институт ядерных исследований;
Россия, 141980, Московская обл., г. Дубна, ул. Жолио-Кюри, 6;
e-mail: kchizhov@jinr.ru.

Работа посвящена восстановлению энергетического спектра нейтронов по откликам многошарового спектрометра Боннера. Эта задача относится к некорректно поставленным обратным задачам, поскольку малые ошибки в измеренных откликах могут заметно искажать восстановленный спектр. В качестве основного подхода предложен двухэтапный метод на основе многовыходной ANFIS-модели с полной SHAP-регуляризацией и тихоновским условием гладкости. SHAP-цель вычисляется в режиме exact SHAP по коалициям десяти входных измерений, а внутренняя важность определяется через градиент ошибки восстановления. Обучение, валидация и тестирование проводились на наборе из 375 нейтронных спектров с разбиением 225/75/75. В итоговом

Статья находится в открытом доступе и распространяется в соответствии с лицензией Creative Commons «Attribution» («Атрибуция») 4.0 Всемирная (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/deed.ru

прогоне достигнуто значение $R^2 = 0,84$, а проверка по нескольким начальным состояниям генератора случайных чисел подтвердила положительный AUC-разрыв внутренней важности.

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

Ключевые слова: обратная задача, нейтронный спектр, многошаровой спектрометр Боннера, ANFIS, SHAP-регуляризация, регуляризация Тихонова.

Для цитирования:

Интерпретируемое восстановление нейтронных спектров нейро-нечёткой моделью с полной SHAP-регуляризацией / А. Д. Лебедев, Ю. В. Трофимов, Е. С. Кондрашова [и др.] // Системный анализ в науке и образовании. 2026. № 3. С.34–47. EDN: EKOJHZ. URL: https://sanse.ru/index.php/sanse/article/view/734.

INTERPRETABLE NEUTRON SPECTRUM RECONSTRUCTION BY A NEURO-FUZZY MODEL WITH FULL SHAP REGULARIZATION

Lebedev Alexander D.[1], Trofimov Yury V.[2], Kondrashova Elizaveta S.[3], Lebedev Mikhail D.[4], Ilyin Andrey S.[5], Chizhov Konstantin A.[6]

[1]Student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: lad.24@uni-dubna.ru.

[2]Junior Researcher;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
Software Engineer;
Joint Institute for Nuclear Research;
6 Joliot-Curie Str., Dubna, Moscow region, 141980, Russia;
e-mail: ura_trofim@bk.ru.

[3] Assistant;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: kes.th@uni-dubna.ru.

[4]Student;
National University of Science and Technology MISIS;
4 Leninsky Prospekt, building 1, Moscow, 119049, Russia
e-mail: lebedevmisha2003@gmail.ru.

[5]Student;
Innopolis University;
1 Universitetskaya St., Innopolis, Republic of Tatarstan,, 420500, Russia;
e-mail: a.ilin@innopolis.university.

[6]Junior Researcher;
Dubna State University;
19 Universitetskaya St., Dubna, Moscow Region, 141980, Russia;
Software Engineer;
Joint Institute for Nuclear Research;
6 Joliot-Curie Str., Dubna, Moscow region, 141980, Russia;
e-mail: kchizhov@jinr.ru.

This paper addresses the reconstruction of the neutron energy spectrum from the responses of a Bonner multisphere spectrometer. This problem belongs to the class of ill-posed inverse problems, since small errors in the measured responses can noticeably distort the reconstructed spectrum. As the main approach, a two-stage method is proposed based on a multi-output ANFIS model with full SHAP regularization and Tikhonov

smoothness constraint. The SHAP target is computed in the exact SHAP mode over coalitions of ten input measurements, while internal importance is determined through the gradient of the reconstruction error. Training, validation, and testing were performed on a dataset of 375 neutron spectra split into 225/75/75 samples. In the final run, the model achieved $R^2 = 0.84$, and testing over several random generator initial states confirmed a positive AUC gap for internal importance.Keywords: inverse problem, neutron spectrum, Bonner sphere spectrometer, ANFIS, SHAP regularization, Tikhonov regularization.

The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation (theme No. 124112200072-2).

Keywords: inverse problem, neutron spectrum, Bonner multisphere spectrometer, ANFIS, SHAP regularization, Tikhonov regularization.

For citation:

Lebedev A. D., et al. Interpretable Neutron Spectrum Reconstruction by a Neuro-Fuzzy Model with Full SHAP Regularization. *System analysis in science and education*, 2026;(3):33–47 (in Russ). EDN: EKOJHZ. Available from: https://sanse.ru/index.php/sanse/article/view/734.

Введение

Многошаровой спектрометр Боннера, далее МСБ, широко применяется для спектрометрии нейтронов. Его преимущество связано с тем, что один и тот же детектор в сочетании с шарами-замедлителями разных диаметров позволяет получать информацию о широком диапазоне энергий [1, 2]. При каждом выбранном шаре измеряется скорость счёта, а набор таких скоростей образует вектор откликов спектрометра. Этот вектор связан с искомым спектром системой интегральных уравнений Фредгольма первого рода.

$$Q_j = \int_{E_{\min}}^{E_{\max}} R_j\,(E)\varphi(E)\,dE, \quad j = 1, \dots, M$$

где $R_j(E)$ обозначает функцию чувствительности j-го шара-замедлителя, Q_j является измеренным откликом, а $\varphi(E)$ задаёт энергетическое распределение нейтронов. После перехода к дискретной энергетической сетке задача принимает матричный вид.

$$\mathbf{A}\varphi = Q, \qquad \varphi \geq 0$$

Здесь $A \in \mathbb{R}\hat{}\{M \times N\}$ обозначает матрицу отклика спектрометра. Она связывает дискретизированный спектр с измеренными скоростями счёта, однако такая связь плохо обусловлена. Поэтому восстановление спектра оказывается чувствительным к шумам измерений: небольшое возмущение вектора Q может привести к заметному отклонению искомого φ. На практике это проявляется в локальных осцилляциях и отрицательных значениях, которые противоречат физическому смыслу нейтронного спектра.

Классические методы восстановления спектров опираются на регуляризацию Тихонова, усечённое сингулярное разложение, итерационные процедуры и принцип максимальной энтропии [3, 4]. Для задач МСБ также развивались регуляризационные схемы, в которых спектр аппроксимируется с использованием полиномов Лежандра и их сдвинутых вариантов [5, 6]. Эти подходы хорошо согласуются с математической природой обратной задачи, однако часто требуют ручного выбора регуляризационных параметров и не всегда дают удобное объяснение роли отдельных входных измерений. В последние годы для этой задачи всё чаще используются методы машинного обучения, включая нейронные сети, свёрточные архитектуры и байесовские нейронные сети [7–9]. Родственная прикладная линия для МСБ уже рассматривала случайный лес, градиентный бустинг и нейронную сеть на спектральных данных с десятью входными измерениями и шестьюдесятью энергетическими уровнями [10]. Отдельно для этой постановки изучалось восстановление спектров с помощью моделей глубокого обучения для табличных данных [11]. Настоящая работа развивает это направление в сторону нейро-нечёткой модели, где интерпретация входных измерений формируется во время обучения, а не только после получения прогноза.

Поэтому к модели реконструкции предъявляется более жёсткое требование, чем простое уменьшение ошибки прогноза. Она должна восстанавливать спектр с достаточной точностью,

сохранять гладкость по энергии и давать понятное распределение роли входных измерений. В настоящей работе эта задача решается через многовыходную *ANFIS*-модель, обучаемую в два этапа. На первом этапе параметры нейро-нечёткой системы инициализируются *BioANFIS*-процедурой с методом роя частиц. На втором этапе модель донастраивается на реальных измерениях, а функция потерь дополняется членами, которые согласуют внутреннюю важность с SHAP-целью и уменьшают нефизические изгибы спектра.

Методическая идея работы состоит в том, что объяснение модели включается в обучение как самостоятельный регуляризационный механизм. В отличие от однофакторного маскирования, используемая *SHAP*-цель строится по коалициям входных измерений. Поэтому внутренняя градиентная карта согласуется с распределением вкладов Шепли, а не с упрощённой оценкой удаления одного признака.

Цель работы состоит в построении и проверке двухэтапного метода восстановления нейтронных спектров по данным МСБ. В нём качество реконструкции оценивается вместе с гладкостью спектра, вкладом регуляризации, верностью карты важности и согласованием внутренней важности с exact *SHAP*-целью.

Табл.1. Основные обозначения, используемые в статье

Обозначение	Смысл
МСБ	многошаровой спектрометр Боннера
Q_j	измеренный отклик при j-м шаре-замедлителе
$R_j(E)$	функция чувствительности j-го шара-замедлителя
$\varphi(E)$	энергетическое распределение нейтронов
A	матрица отклика спектрометра после дискретизации
x, y	вектор входных измерений и дискретизированный спектр
B	размер мини-выборки, то есть число образцов в одном шаге обучения
m, d	число энергетических бинов и число входных измерений
p_θ	дифференцируемая внутренняя карта важности модели
q	SHAP-цель, построенная по exact SHAP-значениям
ε	малая положительная величина для защиты от деления на ноль
карта важности	нормированное распределение вклада входных измерений в прогноз модели

1. Постановка задачи

Пусть $x \in \mathbb{R}^n$ обозначает вектор входных измерений МСБ, а $y \in \mathbb{R}^m$ обозначает дискретизированный нейтронный спектр. В рассматриваемой постановке использованы n = 10 входных измерений и $m = 60$ энергетических бинов. В используемом датасете, содержащем 375 спектров, Q1 соответствует регистрации без шара-замедлителя, а Q2…Q10 шарам диаметром 2, 3, 5, 6, 8, 10, 12, 15 и 18 дюймов. Энергетическая сетка соответствует IAEA-совместимому разбиению на 60 бинов в диапазоне от 10^{-3} эВ до $3{,}98 \cdot 10^{8}$ эВ.

Требуется построить отображение от вектора измерений к восстановленному спектру.

$$\hat{y} = f_\theta(x)$$

где f_θ является многовыходной ANFIS-моделью с параметрами θ. Перед обучением применяется нормировка по сумме входных сигналов. Для i-го образца вводятся величины

$$S_i = \sum_{j=1}^{n} x_{ij}\,, \qquad x_i{}' = \frac{x_i}{S_i}, \qquad y_i{}' = \frac{y_i}{S_i}$$

Если сумма входных сигналов оказывается нулевой, она заменяется малой положительной величиной ε. Такая нормировка уменьшает влияние образцов с большой абсолютной интенсивностью и переносит внимание модели на форму спектра. Обучение выполняется в нормированном пространстве, а при необходимости прогноз возвращается к исходному масштабу.

$$\hat{y}_i = S_i \hat{y}_i'$$

2. Модель реконструкции и схема обучения

В качестве оператора реконструкции используется *ANFIS*-модель типа *Sugeno* с R нечёткими правилами и гауссовскими функциями принадлежности [12]. Для правила r задаётся локальная линейная многовыходная модель.

$$z_r(x) = \mathbf{A}_r x + b_r, \qquad \mathbf{A}_r \in \mathbb{R}^{m\times n}, \quad b_r \in \mathbb{R}^m$$

Сила срабатывания правила определяется произведением функций принадлежности по входным измерениям, после чего выполняется нормировка весов.

$$w_r(x) = \prod_{j=1}^{n} \mu_{rj}\,(x_j), \qquad \overline{w}_r(x) = \frac{w_r(x)}{\sum_{s=1}^{R} w_s\,(x) + \varepsilon}$$

Итоговый прогноз строится как взвешенная сумма локальных моделей.

$$\hat{y} = f_\theta(x) = \sum_{r=1}^{R} \overline{w}_r\,(x) z_r(x)$$

В такой форме каждый энергетический бин описывается гладкой комбинацией локальных линейных зависимостей, а нелинейность задаётся системой нечётких правил. Используемая схема обучения отличается от обычной *ANFIS*-регрессии, где параметры часто подбираются локальным градиентным методом из случайной начальной точки.

На первом этапе применяется *BioANFIS*-подход с глобальной популяционной оптимизацией методом роя частиц [13]. Положение частицы в методе роя частиц соответствует полному набору параметров *ANFIS*. Для p-й частицы на итерации s шаг обновления записывается как

$$v_p^{(s+1)} = \omega v_p^{(s)} + c_1 r_1 \left(\theta_p^{best} - \theta_p^{(s)}\right) + c_2 r_2 \left(\theta_{glob}^{best} - \theta_p^{(s)}\right)$$

$$\theta_p^{(s+1)} = \theta_p^{(s)} + v_p^{(s+1)}$$

Здесь личное лучшее положение частицы задаёт индивидуальную составляющую поиска, а глобальное лучшее положение популяции задаёт коллективную составляющую. Коэффициент инерции управляет сохранением прежнего направления движения частицы, коэффициенты c_1 и c_2 задают относительный вклад индивидуального и коллективного поиска, а r_1 и r_2 являются случайными числами из интервала [0, 1]. В вычислительном эксперименте использовались 50 нечётких правил, гауссовские функции принадлежности, 100 частиц и 150 итераций метода роя частиц. Такая инициализация снижает зависимость модели от случайного старта и формирует более устойчивую начальную структуру правил.

На втором этапе параметры инициализируются значением $\theta(0)$ и донастраиваются на реальных измерениях в течение 300 эпох при размере мини-выборки B = 32 и скорости обучения 10^{-3}. Под мини-выборкой здесь понимается группа из B образцов, по которой вычисляется один шаг градиентного обновления. Основная ошибка восстановления имеет вид

$$L_{main} = \frac{1}{Bm} \sum_{i=1}^{B} \sum_{k=1}^{m} (\hat{y}_{ik} - y_{ik})^2$$

Полный функционал второй стадии объединяет ошибку реконструкции, *SHAP*-регуляризацию и тихоновское условие гладкости.

$$L = L_{task} + \gamma_t \tilde{s}_t L_{shap} + \lambda_{tik} L_{tik}$$

Здесь основной член функционала соответствует ошибке восстановления спектра, *SHAP*-штраф отвечает за согласование внутренней важности с SHAP-целью, а тихоновский член подавляет резкие нефизические изгибы по энергетической оси. Коэффициент γ_t задаёт расписание включения *SHAP*-регуляризации, адаптивный множитель регулирует её целевую долю в функции потерь, а λ_{tik} определяет силу тихоновского штрафа.

3. Полная SHAP-регуляризация важности входных измерений

В этой работе *SHAP*-регуляризация рассматривается как контур, встроенный в обучение ANFIS-модели. Он связывает дифференцируемую внутреннюю карту важности с целью, вычисленной по коалициям входных измерений [14, 15]. Такой расчёт учитывает вклад измерения в разных сочетаниях с остальными входами и тем самым соответствует принципу значений Шепли.

Для расчёта *SHAP*-цели многовыходной прогноз *ANFIS* сводится к скалярному функционалу полезности. В основной конфигурации используется среднее значение нормированного спектра по 60 энергетическим бинам. Этот функционал не является дозовой характеристикой и служит только техническим способом построить единую карту важности. Его выбор проверялся отдельно: вместо среднего выхода использовались сумма и *L*2-функционал, а базовый вектор задавался средним, медианой и нулевым входом. Табл. 4 показывает, что ключевые показатели в этой сетке меняются несущественно.

$$s_i = \frac{1}{m}\sum_{k=1}^{m} \hat{y}_{ik}$$

Внутренняя карта модели строится через градиент, по которому регуляризационный сигнал может менять параметры. В финальной версии используется режим с учётом ошибки восстановления: вклад входного измерения определяется модулем градиента основной ошибки по этому измерению с учётом масштаба входа. Такая карта связана непосредственно с задачей реконструкции и поэтому лучше подходит для проверки функциональной значимости.

Локальные вклады усредняются по мини-выборке и нормируются к распределению pθ с единичной суммой. В этой форме карта пригодна для согласования с SHAP-распределением q и одновременно остаётся связанной с ошибкой восстановления. Это устраняет слабость раннего варианта, где важность строилась только по среднему выходу модели.

$$p_{\theta,j} \propto \mathbb{E}_{x\in B}\left[\left|\frac{\partial L_{task}}{\partial x_j}\right| \cdot |x_j|\right], \qquad \sum_{j=1}^{d} p_{\theta,j} = 1$$

Цель q вычисляется в режиме *exact SHAP*. Пусть N обозначает полный набор из d входных измерений, а b является базовым вектором. В основной конфигурации b равен среднему нормированных входов на обучающей выборке; в проверке чувствительности также использовались медианный и нулевой варианты. Для подмножества S запись $x^{\wedge}(S)$ означает объект, в котором измерения из S сохранены, а остальные заменены базовыми значениями. Функция $v(S)$ задаёт выбранный скалярный функционал модели на мини-выборке.

$$v(S) = \mathbb{E}_{x\in B}\left[\frac{1}{m}\sum_{k=1}^{m} f_{\theta,k}\left(x^{(S)}\right)\right]$$

Для j-го входного измерения полное SHAP-значение записывается как взвешенная сумма предельных вкладов по всем коалициям остальных измерений. Здесь N_j обозначает множество N без j-го измерения.

$$\phi_j = \sum_{S\subseteq N_j} \frac{|S|!\,(d-|S|-1)!}{d!}\left[v(S\cup\{j\}) - v(S)\right]$$

В рассматриваемом эксперименте $d = 10$, поэтому строгий перебор коалиций остаётся вычислительно допустимым. После вычисления значений Шепли строится нормированная цель q.

$$q_j = \frac{|\phi_j|}{\sum_{\ell=1}^{d}|\phi_\ell| + \varepsilon}$$

В реализации q вычисляется вне вычислительного графа. Поэтому обратное распространение не проходит через процедуру расчёта значений Шепли. Градиент идёт через основной член ошибки, внутреннюю карту $p\theta$ и регуляризационные компоненты, а q на данном шаге выступает фиксированной целевой мерой.

Первый компонент *SHAP*-регуляризации отвечает за согласованность карт $p\theta$ и q. Он объединяет среднеквадратичное отклонение и дивергенцию Йенсена-Шеннона [16].

$$L_{consistency} = MSE(p_\theta, q) + \lambda_{JS} JS(p_\theta \| q), \qquad \lambda_{JS} = 0.5$$

Второй компонент отвечает за разреженность распределения важности. Для МСБ это означает, что модель не должна искусственно размазывать вклад по всем измерениям, если для реконструкции конкретной формы спектра существует ограниченное ядро наиболее информативных шаров-замедлителей. Разреженность задаётся через нормированную энтропию и коэффициент Джини [17].

$$H(p_\theta) = -\frac{1}{\log d} \sum_j p_{\theta,j} \log\left(p_{\theta,j} + \varepsilon\right), \qquad G(p_\theta) = \frac{2 \sum_j j\, p_{\theta,(j)}}{d \sum_j p_{\theta,(j)} + \varepsilon} - \frac{d+1}{d}$$

$$L_sparsity = H(p\theta) + \max(0, G_0 - G(p\theta))^2$$

Третий компонент отвечает за верность локальной градиентной аппроксимации. Он проверяет, насколько сумма градиентных вкладов объясняет изменение среднего отклика модели относительно базового состояния.

$$L_{faithfulness} = \frac{1}{B} \sum_{i=1}^{B} \left[\left(s_i - s_i^{(0)}\right) - \sum_j g_{ij}\, x_{ij} \right]^2$$

Четвёртый компонент отвечает за стабильность карты важности внутри мини-выборки. Он штрафует резкое расхождение индивидуальных локальных карт с их средним профилем.

$$L_{stability} = \frac{1}{Bd} \sum_{i=1}^{B} \sum_{j=1}^{d} \left(u_{ij} - \overline{u}_j\right)^2, \qquad \overline{u}_j = \frac{1}{B} \sum_i u_{ij}$$

Итоговый штраф объединяет четыре компоненты и ранговую добавку, усиливающую разницу между ведущей и нижней группами измерений.

$$L_{shap} = w_c L_{cons} + w_s L_{spar} + w_f L_{faith} + w_{st} L_{stab} + w_r L_{rank}$$

Вес каждого компонента пересчитывается по обратной величине соответствующего штрафа и затем нормируется. Это снижает влияние разных численных масштабов. В финальной схеме регуляризация включается после начального прогрева модели: сначала *ANFIS* осваивает основную задачу восстановления, затем начинает действовать *SHAP*-контур. Лучшее состояние выбиралось по валидационному *AUC*-разрыву теста удаления при ограничении на качество по взвешенному коэффициенту детерминации.

Ранговая добавка нужна, чтобы ведущие по q измерения имели большую внутреннюю важность, чем нижняя группа. Пусть T_+ обозначает наиболее значимые измерения, а T_- — наименее значимые. Через $\bar{p}_+$ и $\bar{p}_-$ далее обозначаются средние внутренние важности этих групп. Размеры групп задаются конфигурацией эксперимента; в тесте удаления дополнительно анализируются значения $k = 1 \ldots 4$.

$$L_rank = max(0, \delta - (\bar{p}_+ - \bar{p}_-))$$

Эта добавка делает согласование более направленным. Карта должна быть близка к q как распределение и одновременно сохранять правильный порядок сильных и слабых измерений. В ходе обучения величина рангового штрафа сохранялась в истории, что позволяло учитывать её при выборе состояния модели.

4. Условие гладкости спектра

Для подавления нефизических локальных осцилляций используется тихоновская регуляризация второго порядка в логарифмической шкале энергии. Пусть $\xi_k = log\, E_k$, где k нумерует энергетические бины. Тогда для i-го предсказанного спектра вводится разностный оператор кривизны.

$$\left(D_2^{\log} \hat{y}_i\right)_k = \frac{\hat{y}_{i,k+2} - \hat{y}_{i,k+1}}{\Delta\xi_{k+1}} - \frac{\hat{y}_{i,k+1} - \hat{y}_{i,k}}{\Delta\xi_k}, \qquad \Delta\xi_k = \xi_{k+1} - \xi_k$$

Индекс *i* относится к образцу мини-выборки, а индекс *k* относится к энергетическому бину. Этот оператор штрафует локальную кривизну спектра с учётом неравномерности энергетической сетки. Соответствующий член функционала имеет вид

$$L_{tik} = \frac{1}{B(m-2)} \sum_{i=1}^{B} \sum_{k=1}^{m-2} \left[\left(D_2^{\log} \hat{y}_i \right)_k \right]^2$$

В полном функционале этот штраф входит с коэффициентом λ_{tik}. Он не задаёт спектр жёстко, но уменьшает резкие изгибы, которые обычно возникают при решении некорректно поставленной обратной задачи.

5. Вычислительный эксперимент

Эксперименты проводились на наборе, описанном в разделе постановки задачи. Данные делились на обучающую, валидационную и тестовую части 225/75/75. Для проверки функциональной важности использовался единый протокол: перестановочное маскирование входов, k = 1…4 и 20 случайных повторов. После основной проверки были выполнены серия по нескольким начальным состояниям, отдельная оценка внутренней карты, контрольные классические развёртки и анализ чувствительности *SHAP*-семантики.

Во всех этих проверках согласование оценивалось по той же карте, которая входит в регуляризационный член. В финальной версии она строилась по градиенту основной ошибки по входам, поэтому обучение, выбор состояния и тест удаления использовали одну и ту же логику важности.

Табл. 2. Итоговые метрики практической проверки режима полной SHAP-регуляризации

R^2 weighted	Доля регуляризации	AUC-разрыв	Ведущие/случайные	Косинус согласования
0,838	0,089	0,016	1,934	0,654

6. Результаты

Метрики в табл. 2 показывают, что регуляризационный контур является содержательной частью обучения. Его вклад заметен, но он не вытесняет основную задачу восстановления спектра.

Контрольный блок развёртки был добавлен, чтобы сравнение не ограничивалось базовой *ANFIS*. В него вошли линейная тихоновская схема, её неотрицательный вариант, а также приближённые реализации идей *MAXED* и *GRAVEL* [3, 4]. Для каждого метода использовалась единая сетка λ, три начальных состояния 42, 43 и 44, а лучший вариант выбирался по максимуму $R^2_{weighted}$. Подробности приведены в табл. 3. В этой конфигурации все четыре схемы дали результаты, заметно уступающие основной *ANFIS*-модели.

Табл. 3. Контрольный блок классических и энтропийных приближённых методов развёртки

Метод	Схема и ограничения	Параметры протокола	R^2weighted, среднее ± std
Тихонов, линейная схема	$(R^{\mathrm{T}}R + \lambda L^{\mathrm{T}}L)^{-1}R^{\mathrm{T}}x$; неотрицательность не задаётся, сумма нормируется	λ-сетка 10^{-5}…10^{-1}; *D*2; прямое решение; *best* по R^2*weighted*	-0,140 ± 0,013
Тихонов с *NNLS*	$[R; \sqrt{\lambda L}]y \approx [x; 0], y \geq 0$, сумма нормируется	λ-сетка 10^{-5}…10^{-1}; D2; *scipy.optimize.nnls*; *best* по R^2*weighted*	-0,135 ± 0,032
MAXED-like	$1/2\|\|Ry - x\|\|^2 + \lambda KL(y\|\|prior),\ y \geq \varepsilon$, сумма нормируется	λ-сетка 10^{-5}…10^{-1}; *mirror descent*; 220 итераций; *lr* = 0,08	-0,208 ± 0,015

GRAVEL-like	мультипликативная коррекция $y \geq \varepsilon$, при $\lambda > 0$ добавляется демпфирование $L^T L$	λ-сетка $10^{-5}…10^{-1}$; 220 итераций; *eta* = 0,8; фиксированное число итераций	-0,363 ± 0,03

Чтобы проверить, не является ли карта важности артефактом конкретного выбора базового состояния или скаляризации многовыходного прогноза, была проведена отдельная проверка чувствительности *SHAP*-семантики. В ней сравнивались три способа задания базового вектора: среднее значение признаков, медиана и нулевой вектор. Для функции полезности проверялись средний выход, суммарный выход и L2-функционал. Во всех девяти сочетаниях значения ключевых метрик менялись в очень узких пределах, что показано в табл. 4. Эта проверка показывает, что в выбранной области настроек результат не зависит от одного произвольного определения *SHAP*-цели.

Табл. 4. Чувствительность SHAP-семантики к выбору базового вектора и функционала полезности

Показатель	**Минимум**	**Максимум**	**Диапазон**
$R^2_{weighted}$	0,808787	0,808879	0,000092
AUC-разрыв	0,064011	0,064067	0,000056
Ведущие/случайные	1,029898	1,030638	0,000740

Форма средних восстановленных спектров показана на рис. 1. Средняя кривая режима *SHAP* + Тихонов близка к средней форме истинного спектра, а нижняя панель показывает области систематического смещения.



Рис. 1. Средние восстановленные спектры и отклонения моделей от среднего истинного спектра

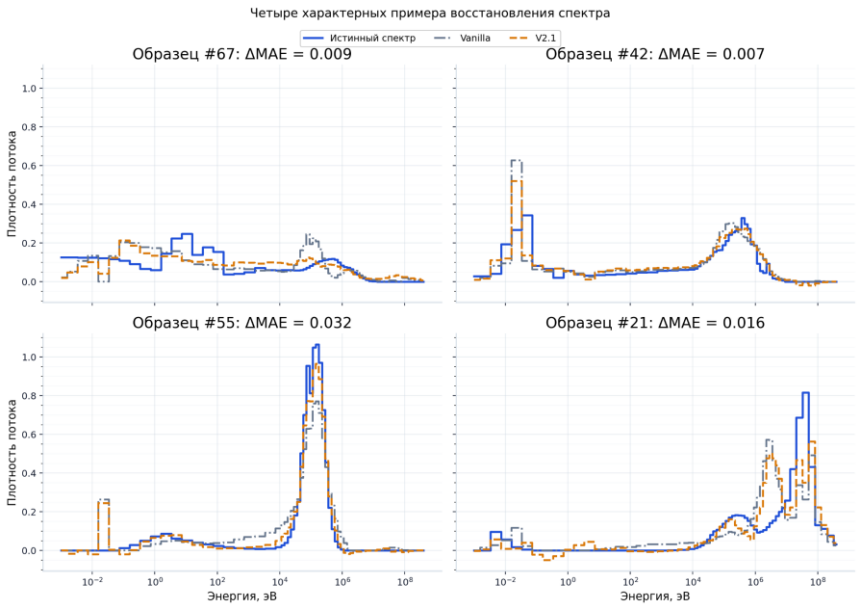


Рис. 2. Четыре характерных примера восстановления спектров из тестового набора

Нормированная карта важности входных измерений приведена на рис. 3. Её смысл задаётся согласованием внутренней градиентной важности с *SHAP*-целью, построенной по коалициям измерений. Наибольший вклад дают $Q3$, $Q5$ и $Q6$, причём суммарная важность трёх ведущих измерений равна 0,61. Такая концентрация не означает, что остальные измерения можно отбросить без потерь. Она показывает, что модель выделяет устойчивое ядро наиболее информативных откликов МСБ.



Рис. 3. Нормированная SHAP-важность входных измерений для режима полной SHAP-регуляризации

Для проверки верности объяснения использовался тест удаления. Входные измерения маскировались в трёх порядках: от наиболее важных, случайно и от наименее важных. Если карта согласована с прогнозом, удаление ведущих измерений должно сильнее увеличивать ошибку восстановления.

Табл. 5. Проверка верности и согласования карты важности в итоговом прогоне

Показатель	Значение	Критерий	Смысл
AUC-разрыв	0,016	> 0	удаление ведущих измерений влияет сильнее нижней группы
ведущие/случайные	1,934	> 1	эффект удаления ведущих почти в два раза выше случайного

косинус согласования	0,654	≥ 0,65	внутренняя важность согласована с *SHAP*-целью

Табл. 5 подтверждает ожидаемую иерархию: ведущие измерения влияют на восстановление сильнее нижней группы, а эффект их удаления почти вдвое превышает случайный порядок. Косинус согласования проходит заданный практический порог и показывает связь внутренней карты с *SHAP*-целью.

Устойчивость этого вывода проверялась на трёх начальных состояниях генератора случайных чисел. Модель выбиралась по валидационному *AUC*-разрыву, а сравнение с базовой *ANFIS* выполнялось по тому же протоколу удаления: перестановочное маскирование, k = 1…4 и 20 случайных повторов. Результаты приведены в табл. 6.

Табл. 6. Результаты

Seed	**Режим**	**AUC-разрыв**	**Ведущие/случайные**	$R^2_{weighted}$
42	*SHAP*-важность	0,011257	2,0873	0,834009
42	внутренняя важность	0,003591	0,7792	—
42	базовая *ANFIS*	0,002262	1,0540	0,825896
43	*SHAP*-важность	0,014151	2,9187	0,821660
43	внутренняя важность	-0,000217	0,0032	—
43	базовая *ANFIS*	-0,008041	0,2494	0,838027
44	*SHAP*-важность	0,009870	1,8592	0,707201
44	внутренняя важность	0,001657	0,3965	—
44	базовая *ANFIS*	-0,001019	0,5571	0,689649

После этого была выполнена дополнительная проверка для исправленной внутренней важности. В этой версии внутренняя карта строилась по градиенту основной ошибки восстановления, а не по среднему выходу модели. Цель проверки состояла в том, чтобы установить, становится ли внутренняя важность функционально валидной в тесте удаления, где удаление ведущих измерений должно ухудшать восстановление сильнее, чем удаление нижней группы.

Табл. 7. Проверка внутренней важности p_θ на трёх начальных состояниях

Seed	*AUC*-разрыв по *SHAP*-важности *q*	*AUC*-разрыв по внутренней важности *pθ*(task_loss)	*AUC*-разрыв по важности базовой модели	Ведущие/случайные по *pθ*(*task_loss*)	Ведущие/случайные для базовой модели	R^2*weighted* модели
42	0,005222	0,018414	-0,001561	4,0035	0,4784	0,809471
43	0,010813	0,017359	-0,009120	3,6421	0,0770	0,829096
44	0,010167	0,010423	-0,001515	2,0330	0,4911	0,719495

Результаты табл. 7 показывают, что внутренняя карта выигрывает у базовой во всех трёх запусках. Отношение ведущие/случайные больше единицы для каждого начального состояния. Следовательно, внутренняя важность стала не только согласованной с *SHAP*-контуром, но и проверяемой через функциональный тест удаления.

7. Анализ устойчивости методом Монте-Карло

Дополнительная проверка устойчивости проводилась методом Монте-Карло. Входные измерения возмущались относительным гауссовским шумом, после чего оценивалась изменчивость восстановленного спектра. Рис. 4 показывает доверительный коридор при 10 % шуме: средний восстановленный спектр остаётся близким к базовому, а основная неопределённость сосредоточена около выраженного пика.

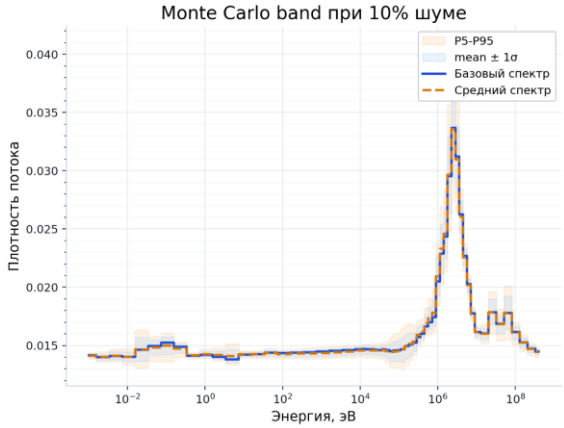


Рис. 4. Оценка устойчивости восстановленного спектра методом Монте-Карло

8. Обсуждение

Полученные результаты показывают, что полную *SHAP*-регуляризацию можно включить в обучение многовыходной нейро-нечёткой модели без перехода к однофакторной эвристике. Цель строится по коалициям входных измерений, а внутренняя карта согласуется с ней в процессе оптимизации. Поэтому объяснение не добавляется к готовой модели извне, а становится частью функции потерь.

Практическая проверка показывает, что регуляризационный член вносит измеримый вклад в обучение модели. Её средняя доля в функции потерь составила около 9 %, вклад *SHAP*-компоненты отличен от нуля, а косинус согласования достиг 0,654. После перехода к важности по градиенту ошибки тест удаления стал положительным и устойчивым по трём начальным состояниям.

Тест удаления делает интерпретационный вывод проверяемым. В основном прогоне получен *AUC*-разрыв 0,016, а серия по нескольким начальным состояниям показала выигрыш *SHAP*-регуляризованной модели над базовой *ANFIS*. Дополнительная проверка внутренней важности отвечает на более строгий вопрос: ведущие по $p\theta$ измерения также оказываются функционально значимыми, поскольку их удаление во всех трёх запусках сильнее ухудшало восстановление, чем удаление нижней группы.

Контрольный блок классических и энтропийных приближённых развёрток уточняет позиционирование работы. В выбранной воспроизводимой конфигурации эти схемы дали отрицательные значения $R^2_{weighted}$ и оказались существенно слабее *ANFIS*-линии. При этом *MAXED*-like и *GRAVEL-like* являются прокси-реализациями соответствующих идей, а не полными историческими пакетами. Поэтому данный блок закрывает минимальное контрольное сравнение на тех же данных, но не заменяет отдельное исследование с эталонными реализациями всех классических программ развёртки.

Проверка чувствительности *SHAP*-семантики снимает ещё один методический риск. В девяти сочетаниях базового вектора и функции полезности значения $R^2weighted$, *AUC*-разрыва и отношения ведущие/случайные изменялись только в третьем или четвёртом знаке. Это означает, что в данной постановке результат не определяется произвольным выбором скаляризации. Для более строгого вывода эту проверку следует расширить до нескольких начальных состояний.

Ограничения метода сохраняются. Полный *SHAP* требует перебора коалиций, поэтому вычислительная цена растёт с числом входов. В этой задаче $d = 10$, и точный расчёт остаётся практически допустимым; для более широких конфигураций потребуется аппроксимация или более редкое обновление цели. В статье также не проводится новая тренировка *MLP* или *CNN* Такой выбор

связан с фокусом работы: здесь проверяется нейро-нечёткая модель с внутренней картой важности, тогда как близкая ансамблевая и нейросетевая линия для восстановления спектров по данным МСБ уже рассматривалась ранее [10, 11]. Полученные результаты не закрывают вопрос о сравнении *ANFIS* со всеми возможными архитектурами. Дальнейшая работа должна включать перенос регуляризации на *MLP* и *CNN*, более длинные проверки по нескольким начальным состояниям, дозовые метрики, другие спектрометры и сравнение с эталонными реализациями классических программ развёртки. Идея согласования структурной интерпретируемости *ANFIS* и *SHAP*-объяснений также соответствует ранее предложенной архитектуре *GD-ANFIS*/*SHAP* для *XAI* 2.0 [18].

Заключение

Предложен двухэтапный метод восстановления нейтронных спектров по данным МСБ. В его основе лежит многовыходная нейро-нечёткая модель, которая сначала инициализируется биоинспирированной процедурой с методом роя частиц, а затем донастраивается на реальных измерениях. На втором этапе в обучение включаются полная *SHAP*-регуляризация *exact SHAP* и тихоновское условие гладкости.

В итоговом проверочном прогоне режим полной *SHAP*-регуляризации прошёл практический критерий состоятельности. Взвешенный коэффициент детерминации составил 0,838, средняя доля регуляризации в функции потерь равна 0,089, *AUC*-разрыв теста удаления составил 0,016, отношение ведущие/случайные равно 1,934, а косинус согласования достиг 0,654.

Основной результат работы состоит в построении воспроизводимого контура полной SHAP-регуляризации, встроенной в обучение физической обратной задачи. Этот контур связывает реконструкцию спектра, гладкость решения, *SHAP*-цель и внутреннюю карту, построенную по градиенту ошибки восстановления. Дальнейшее развитие должно включать более длинные проверки по нескольким начальным состояниям, перенос регуляризации на *MLP* и *CNN*, сравнение с эталонными реализациями классических методов, дозовую валидацию и испытания на других конфигурациях МСБ.

Доступность кода и данных

Исходный код, конфигурации, протоколы вычислительных экспериментов и итоговые отчёты доступны в открытом репозитории проекта: *https://github.com/lebedeffson/bonner-shap-reconstruction*. В статье использовалась версия ветки main после фиксации внутренней важности с учётом ошибки восстановления, проверки чувствительности *SHAP*-семантики, контрольного блока классических приближённых развёрток и отчётов multi-seed проверок.

Список источников

1. Bramblett R. L., Ewing R. I., Bonner T. W. A new type of neutron spectrometer // Nuclear Instruments and Methods. 1960. Vol. 9, No. 1. P. 1–12. https://doi.org/10.1016/0029-554X(60)90043-4.
2. Compendium of Neutron Spectra and Detector Responses for Radiation Protection Purposes. Technical Reports Series No. 403. Vienna: International Atomic Energy Agency, 2001.
3. Tikhonov A. N., Goncharsky A. V., Stepanov V. V., Yagola A. G. Numerical Methods for the Solution of Ill-Posed Problems. Dordrecht: Kluwer Academic Publishers, 1995.
4. Reginatto M. Overview of spectral unfolding techniques and uncertainty estimation // Radiation Measurements. 2010. Vol. 45, No. 10. P. 1323–1329. https://doi.org/10.1016/j.radmeas.2010.06.016.
5. Chizhov K., Beskrovnaya L., Chizhov A. Neutron Spectra Unfolding from Bonner Spectrometer Readings by the Regularization Method Using the Legendre Polynomials // Physics of Particles and Nuclei. 2024. Vol. 55. № 3. P. 532–534. https://doi.org/10.1134/S1063779624030298
6. Chizhov K., Beskrovnaya L., Chizhov A. Neutron spectrum unfolding method based on shifted Legendre polynomials, its application to the IREN facility // Physics of Particles and Nuclei Letters. 2025. Vol. 22. № 2. P. 190. https://doi.org/10.1134/S154747712470239X

7. A neutron spectrum unfolding computer code based on artificial neural networks / J. M. Ortiz-Rodriguez, A. Reyes Alfaro, A. Reyes Haro [et al.] // Radiation Physics and Chemistry. 2014. Vol. 95. P. 428–431. https://doi.org/10.1016/j.radphyschem.2013.05.007.

8. Neutron spectrum unfolding using two architectures of convolutional neural networks / M. Bouhadida, A. Mazzi, M. Brovchenko [et al.] // Nuclear Engineering and Technology. 2023. Vol. 55, No. 6. P. 2276–2282. https://doi.org/10.1016/j.net.2023.03.025.

9. Bayesian neural networks for neutron spectrum unfolding in the EAST tokamak / B. Zhou, Z. Hu, M. Zhong [et al.] // IEEE Transactions on Instrumentation and Measurement. 2025. Vol. 74. Article No. 1007209. https://doi.org/10.1109/TIM.2025.3554325.

10. Белый А. А., Стариковская М. Д., Чижов К. А. Разработка веб-приложения для эксперимента по восстановлению спектра нейтронов с применением алгоритмов нейронный сетей // Системный анализ в науке и образовании. 2025. № 2. С. 49–57. EDN: XQZWTL. URL: https://sanse.ru/index.php/sanse/article/view/664.

11. Chizhov K. A., Bely A. A. Neutron Spectrum Unfolding Using Deep Learning Models for Tabular Data // Moscow Univ. Phys. 2025. Vol. 80. № S3. P. S987–S995. https://doi.org/10.3103/S0027134925702832.

12. Jang J.-S. R. ANFIS: adaptive-network-based fuzzy inference system // IEEE Transactions on Systems, Man, and Cybernetics. 1993. Vol. 23, No. 3. P. 665–685. https://doi.org/10.1109/21.256541.

13. Kennedy J. Particle swarm optimization / J. Kennedy, R. Eberhart // Proceedings of ICNN'95, International Conference on Neural Networks. 1995. Vol. 4. P. 1942–1948. https://doi.org/10.1109/ICNN.1995.488968.

14. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017. Vol. 30.

15. Shapley L. S. A value for n-person games // Contributions to the Theory of Games II / eds. H. W. Kuhn, A. W. Tucker. Princeton: Princeton University Press, 1953. P. 307–317. https://doi.org/10.1017/CBO9780511528446.003.

16. Lin J. Divergence measures based on the Shannon entropy // IEEE Transactions on Information Theory. 1991. Vol. 37, No. 1. P. 145–151. https://doi.org/10.1109/18.61115.

17. Gastwirth J. L. The estimation of the Lorenz curve and Gini index // The Review of Economics and Statistics. 1972. Vol. 54, No. 3. P. 306–316. https://doi.org/10.2307/1937992.

18. Verified Explainability Core: A GD–ANFIS/SHAP Hybrid Architecture for XAI 2.0 / Yu. V. Trofimov, A. D. Lebedev, A. S. Ilin, A. N. Averkin // Automatic Documentation and Mathematical Linguistics. 2025. Vol. 59. P. S469–S478. https://doi.org/10.3103/S0005105525701420.