Генерация синтетических спектров комбинационного рассеяния с использованием метода главных компонент

© Е.С. Прихожденко

Московский физико-технический институт (национальный исследовательский университет), Институт биофизики будущего, Лаборатория медицинского оборудования в области *in vitro* диагностики,
Долгопрудный, Московская обл., Россия

e-mail: prikhozhdenko.es@mipt.ru

Поступила в редакцию 21.11.2025 г.
В окончательной редакции 17.12.2025 г.
Принята к публикации 17.12.2025 г.

Предложен подход к генерации синтетических спектров комбинационного рассеяния с использованием метода главных компонент. В качестве исходных данных используются спектры комбинационного рассеяния, полученные для жировой ткани до и после воздействия липазы. Предлагаемые данные характеризуются схожестью спектров, поскольку исследуемые материалы обладают схожей химической структурой: жировая ткань представляет собой молекулы триглицеридов. После воздействия липазы образуется смесь из ди- и моноглицеридов, а также свободных жирных кислот. Предложенный в работе подход к генерации спектральных данных заключается в следующем: (1) применение метода главных компонент к набору реальных спектральных данных, (2) вычисление среднего значения и стандартного отклонения счетов выбранного количества главных компонент, (3) генерация нового набора счетов на основе среднего значения и стандартного отклонения, (4) генерация спектров по нагрузкам выбранного количества главных компонент и новому набору счетов. Для подтверждения схожести исходных и синтетических данных производилось обучение модели классификации на исходных данных с последующей оценкой точности на синтетических данных. Предложенный подход позволил синтезировать данные спектроскопии комбинационного рассеяния со средней точностью $(90.6 \pm 0.5)\%$.

Ключевые слова: генерация спектров комбинационного рассеяния, метод главных компонент, машинное обучение, случайный лес.

DOI: 10.61011/OS.2026.05.63344.53-25

Введение

Спектроскопия комбинационного рассеяния (КР) света — мощный инструмент, который широко используется в аналитической химии, биомедицине и материаловедении, позволяя получить подробную информацию о молекулярном составе и структуре вещества [1,2]. Однако применение методов машинного обучения для анализа спектров КР часто сталкивается с проблемой нехватки экспериментальных данных. Это связано с тем, что сбор данных требует значительных временных и финансовых затрат [3,4]. Недостаток данных может привести к статистической ненадежности моделей, их переобучению и снижению точности прогнозов.

Чтобы преодолеть дефицит данных в спектроскопии КР, был разработан целый ряд методов, которые можно разделить на две категории: простые детерминированные и сложные генеративные. К традиционным подходам относятся элементарные преобразования, такие как добавление гауссова шума к исходному спектру для имитации инструментальной погрешности [5]. Также используются аффинные преобразования, такие как смещение базовой линии, изменение наклона и интенсивности, чтобы учесть изменчивость условий эксперимента [6]. Хотя эти методы легки в реализации и не требуют больших вычислительных ресурсов, они имеют существенный недостаток: генерируемые данные представляют собой лишь небольшие вариации исходных образцов и не способны значительно расширить область определения модели, так как не создают принципиально новых спектральных особенностей.

Революция в области генерации данных была достигнута благодаря использованию глубокого обучения, особенно генеративно-состязательных сетей (GAN) [7] и вариационных автоэнкодеров (VAE) [8]. Эти модели способны адаптироваться к сложным и многообразным распределениям исходных данных и создавать высококачественные синтетические спектры, которые визуально неотличимы от реальных [9,10]. Однако, несмотря на впечатляющие результаты, у этих подходов есть существенные недостатки. Они требуют значительных вычислительных мощностей и больших объемов данных для обучения, что делает их не всегда доступными. Кроме того, они сложны в настройке и требуют особых методов для стабилизации обучения [7].

Метод главных компонент используется для разложения многомерного набора данных на набор последовательных ортогональных компонент (нагрузок), которые объясняют максимальную долю дисперсии данных [11]. В спектроскопии КР метод главных компонент используется для понижения размерности данных, эффективно отделяя полезный сигнал (содержащийся в первых

главных компонентах) от шума (лежащего в последних компонентах) [12]. Также метод главных компонент используется для выявления паттернов в данных спектроскопии с акцентом на колебательные моды, наиболее характерные для анализируемого набора данных [13,14]. Данный подход позволяет провести кластеризацию спектров КР графически, визуализируя каждый отдельный спектр как точку в пространстве выбранных главных компонент: координатами в данном случае являются счета [15].

В настоящей работе представлен новый подход к генерации синтетических спектров КР, который основывается на методе главных компонент. Чтобы доказать эффективность этого метода, были использованы реальные спектры жировой ткани до ($n = 100$) и после ($n = 40$) ферментативного гидролиза под воздействием липазы. Эта модельная система обладает высокой спектральной схожестью, что связано с близкой химической структурой исследуемых материалов. Исходная жировая ткань состоит преимущественно из триглицеридов [16], а продукт ее расщепления представляет собой сложную смесь диглицеридов, моноглицеридов и свободных жирных кислот [17–19]. Таким образом, проверка предложенного метода генерации спектральных данных на сложных и похожих спектрах биоткани на примере жировой ткани до и после воздействия липазы представляет большой интерес.

Предлагаемая методология включает следующие этапы: 1) применение метода главных компонент к исходному набору спектров для выделения главных компонент (нагрузок) и соответствующих счетов, 2) статистический анализ вычисленных счетов с определением их средних значений и стандартных отклонений, 3) генерация новых синтетических счетов путем случайной выборки из нормального распределения с найденными параметрами, 4) восстановление полного синтетического спектра путем линейной комбинации нагрузок главных компонент и сгенерированных счетов. В работе представлены результаты сравнительного анализа, а также проведена оценка сходства между синтезированными и реальными наборами данных с использованием методов машинного обучения.

Методы

Материалы и методы

Липаза Б *Candida antarctica* была закуплена в Sigma Aldrich (Сент-Луис, Миссури, США). Исследование жировой ткани, взятой из подкожной жировой клетчатки *post-mortem*, было проведено с одобрения этического комитета ФГБОУ ВО „Саратовский государственный медицинский университет имени В.И. Разумовского“ (протокол № 1 от 03.09.2013).

Регистрация спектров КР

В ходе измерений применялся конфокальный микроскоп КР Renishaw inVia (Великобритания), оснащенный лазером с длиной волны 785 nm. Лазерный луч направлялся через объектив микроскопа 50× Leica N PLAN L с числовой апертурой 0.5. Для получения спектров КР в диапазоне от 150 до 3200 cm^{-1} использовался режим SynchroScan в программном обеспечении WiRE 4.2 для непрерывного сканирования.

Исследование необработанной жировой ткани (исходный образец) проводилось с применением лазерного излучения мощностью 3 mW. Количество спектров КР равнялось 100. После обработки жировой ткани липазой Б *Candida antarctica* (0.75 mg/ml, водный раствор) в течение 30 min при температуре 37°C образец представлял собой мицеллярный раствор жирных кислот в воде. Регистрация КР-спектров обработанного образца проводилась после его охлаждения до комнатной температуры. Мощность лазерного излучения составляла 15 mW, количество спектров — 40. Время накопления КР-сигнала для обоих типов образцов составляло 10 s.

Предварительная обработка данных

В результате применения инструмента *Subtract baseline*, входящего в состав программного обеспечения WiRE 4.2, из полученных спектров был удален полиномиальный фон, описываемый уравнением степени 11. Дальнейшая обработка данных и применение методов машинного обучения осуществлялись на языке программирования Python версии 3 в среде *Jupyter Notebook* с использованием библиотеки *scikit-learn (sklearn)* версии 1.7.2 [20]. К набору данных также был применен метод *sklearn.preprocessing.normalize* для нормализации данных с использованием l^2-нормы:

$$||x||_2 = \sqrt{\sum_{i=1}^{m} |x_i|^2},$$

где $||x||_2$ — делитель при использовании l^2-нормы, x_i — интенсивность КР при i-м волновом числе, m — количество волновых чисел в спектре ($m = 3047$ для рассматриваемого набора данных). Таким образом, для каждого спектра производилось деление интенсивностей на полученное значение делителя ($||x||_2$). В результате были вычислены нормированные значения интенсивностей.

Генерация спектров КР

Для реализации анализа главных компонент использовался метод *sklearn.decomposition.PCA* с варьированием параметра $n_{\text{components}}$ в диапазоне 5−130 с шагом 5. С помощью *fit_transform* проводилось преобразование исходного массива спектральных данных размером 140×3047 к массиву $140 \times n_{\text{components}}$. Данный

массив содержал искомые коэффициенты для линейной комбинации (счета).

Далее были рассчитаны средние значения и стандартные отклонения счетов. Затем с помощью библиотеки *numpy* и метода *random.normal* по среднему и стандартному отклонению были сгенерированы по 1000 счетов для спектров жировой ткани до и после воздействия липазы.

По сгенерированным счетам и спектрам главных компонент (нагрузкам) была произведена обратная трансформация методом *inverse_transform*. Таким образом были получены наборы синтетических спектров КР жировой ткани до и после воздействия липазы. Размерность сгенерированных массивов данных составила 1000×3047.

Оценка генерации с помощью модели классификации на основе случайного леса

Реальные данные были предварительно поделены на обучающую и тестовую выборки в соотношении 3:1 с сохранением пропорции между спектрами до и после воздействия липазы с помощью библиотеки *scikit-learn* (*sklearn*) методом *model_selection.train_test_split*. В качестве модели классификации использовалась модель *sklearn.ensemble. RandomForestClasssifier*(*max_depth*=3, $n_estimators = 10$) [21]. В качестве метрик были использованы точность (*sklearn.metrics.accuracy_score*), прецизионность (*sklearn.metrics.precision_score*), полнота (*sklearn.metrics.recall_score*) и матрица неточности (*sklearn.metrics.confusion_matrix*), благодаря которой были получены доля верно классифицированных синтетических наборов данных жировой ткани до и после воздействия липазы.

Результаты и дискуссия

Описание исходного набора данных

Для демонстрации подхода к генерации данных в спектроскопии КР был выбран набор ранее полученных спектров, соответствующих жировой ткани до ($n = 100$) и после ($n = 40$) воздействия липазы [22]. Таким образом, исходный набор данных представляет собой таблицу размерностью 140×3047, где 140 соответствует количеству отдельных спектров, а 3047 — длине каждого спектра. Усредненные нормированные спектры, а также разность спектров представлены на рис. 1, *a*. Спектры КР приведенных образцов обладают малыми различиями, что в свою очередь делает их отличными кандидатами для отработки предлагаемого принципа генерации данных. В качестве основного инструмента был предложен метод главных компонент [15]. При его применении оказалось, что 100% дисперсии данных объясняются 139 главными компонентами. Нагрузки, соответствующие первым трем главным компонентам, представлены на рис. 1, *b*. Эти компоненты объясняют соответственно 53.3, 5.1 и 4.2% дисперсии данных. При реализации метода главных компонент происходит трансформация исходных данных размером 140×3047 в массив $140\times$ количество главных компонент. В полученном таким образом массиве каждый образец характеризуется своим набором счетов в линейной комбинации вычисленных главных компонент, что позволяет определить и графически изобразить средние счета для жировой ткани до и после воздействия липазы (рис. 1, *c*). Таким образом, с точки зрения метода главных компонент спектры исследуемых образцов обладают видимыми различиями.

Помимо прямой трансформации — уменьшения размерности исходных данных в приведенном случае с 3047 волновых чисел до 139 главных компонент — можно провести обратную операцию. Таким образом, если использовать метод главных компонент на наборе данных спектроскопии КР, получить характерные нагрузки и распределение счетов при них, то на их основе можно сгенерировать дополнительный набор счетов и провести обратную трансформацию в синтетические спектры КР.

Создание синтетического набора данных с помощью анализа главных компонент

Для генерации счетов метода главных компонент в настоящей работе было использовано допущение, что счета главных компонент, соответствующие определенному классу образцов (жировой ткани до и после воздействия липазы), характеризуются нормальным распределением. Таким образом, с помощью библиотеки *numpy* и метода $numpy.random.normal$, указав средние и стандартные отклонения счетов, можно получить дополнительные счета для линейной комбинации нагрузок главных компонент с последующим восстановлением до синтетических спектров КР. На процесс генерации влияют два параметра: $random_state$ или $numpy$, $seed$ — любое целое число, позволяющее воспроизводимо генерировать случайные величины при работе на языке *Python3* (важно иметь постоянное значение во всем используемом коде); $n_{\text{components}}$ — количество главных компонент.

Для жировой ткани до и после воздействия липазы было сгенерировано по 1000 спектров КР при разном количестве главных компонент: 5–130 с шагом 5. Результаты генерации спектров для $n_{\text{components}} = 130$ приведены на рис. 2. По графикам разности между усредненными спектрами КР реальных и синтетических данных (кривые *3* на рис. 2, *a, c*) видно, что в случае использования спектров жировой ткани до воздействия липазы (рис. 2, *a*) генерация позволила получить достаточно близкие к реальным данным результаты: график разности при его увеличении в 100 раз представлял собой случайный шум, не содержащий никаких отдельных КР-полос. Схожесть синтетического набора данных для жировой ткани после воздействия липазы с его реальным прототипом

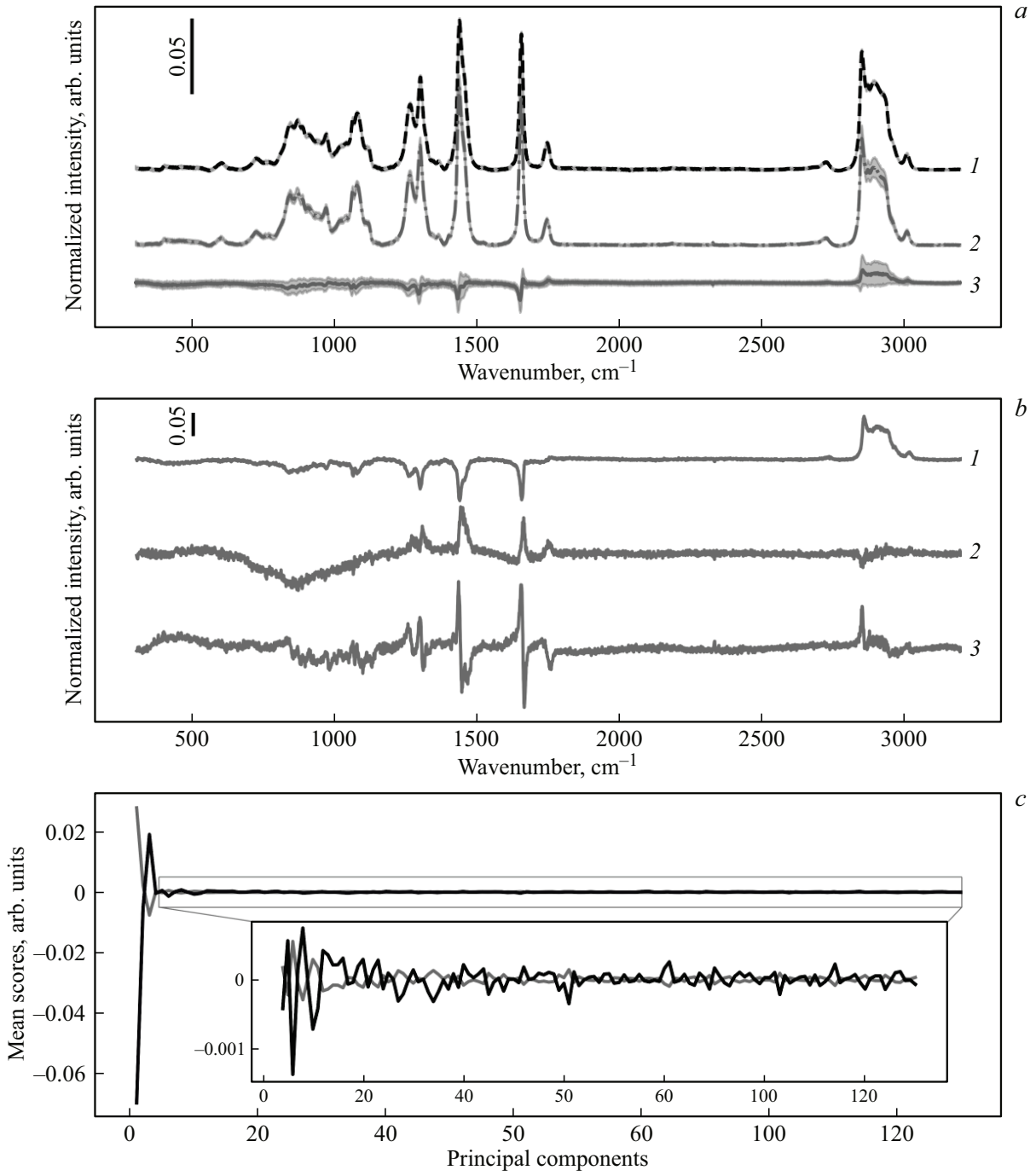

Рис. 1. (*a*) Усредненные нормализованные спектры КР жировой ткани: *1* — до воздействия липазы ($n = 100$), *2* — после ($n = 40$) воздействия липазы, *3* — спектр разности, закрашенная область характеризует стандартное отклонение нормированных интенсивностей. (*b*) Нагрузки первых трех главных компонент исследуемого набора данных. (*c*) Средние счета главных компонент для спектров жировой ткани до (серая кривая) и после (черная кривая) воздействия липазы, на вставке — график для главных компонент, начиная с четвертой.

менее выражена, спектр разности между реальными и сгенерированными данными при его увеличении в 100 раз обладает рядом КР-полос: в частности 1456 cm^{-1} (колебание связей CH_2), 1657 $см^{-1}$ (растяжение связи C=C), а также в области 2800–3000 cm^{-1} (растяжение C–H) [23,24].

Возможной причиной данного эффекта могут служить большие стандартные отклонения по сравнению с образцом жировой ткани до воздействия липазы (спектр *2*, рис. 1, *a*). Также стоит отметить простоту предложенного подхода и отсутствие необходимости проведения дополнительных шагов генерации спектров, в частности

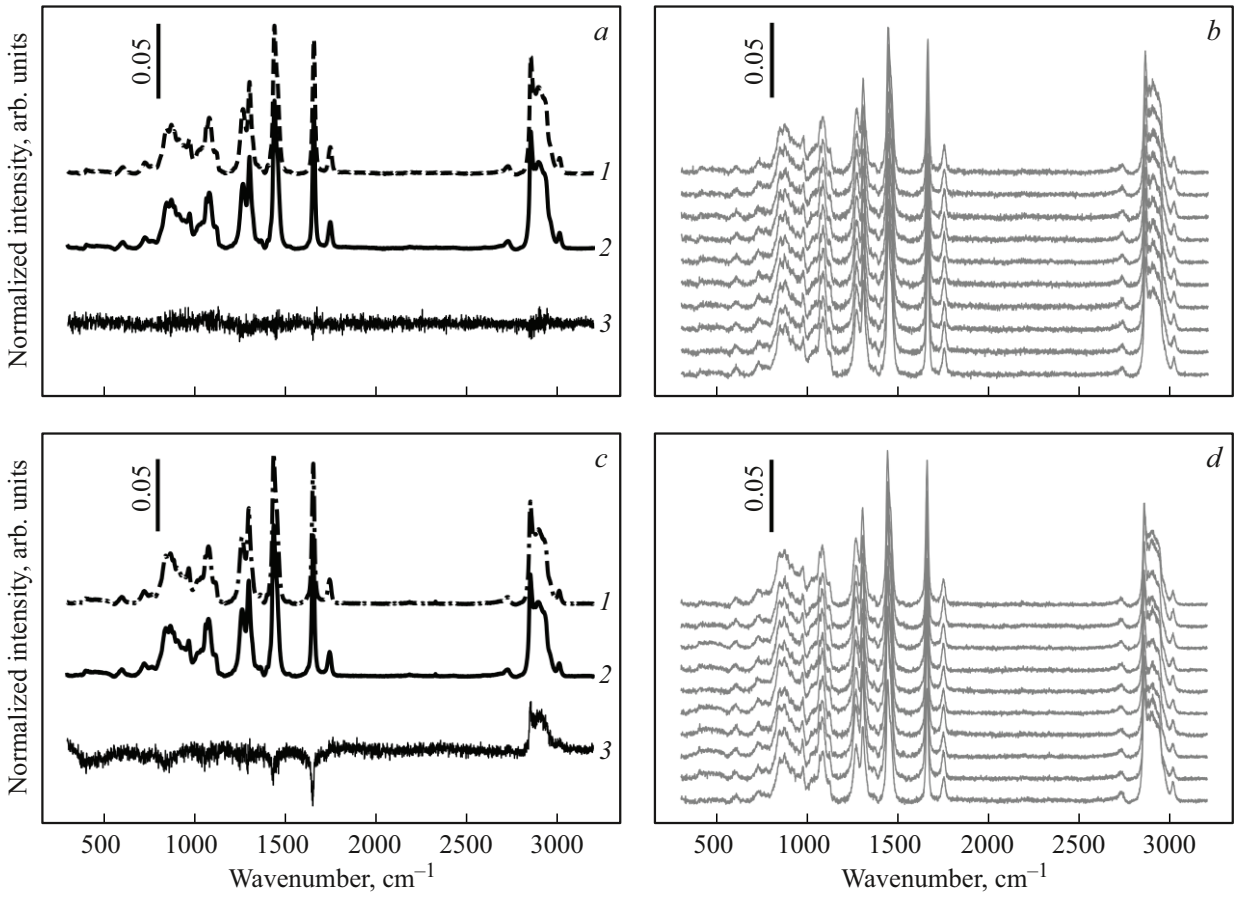

Рис. 2. Результаты генерации спектров КР с использованием метода главных компонент ($n_{\text{components}} = 130$) для образцов жировой ткани до (a, b) и после (c, d) воздействия липазы: (a, c) *1* — усредненный спектр КР реальных данных; *2* — усредненный спектр КР синтетических данных ($n = 1000$); *3* — разность усредненных спектров реальных и синтетических данных, увеличенная в 100 раз для наглядности. (b, d) Десять случайных спектров из набора синтетических данных.

внесения случайного шума. Так, при использовании 130 компонент для процесса генерации получаемые спектры КР характеризуются заметным уровнем шума (рис. 2, b, d). Более наглядно влияние количества главных компонент на результат генерации спектра представлен на рис. 3. На данном рисунке показана генерация одного спектра КР при разном количестве главных компонент: 10–130 с шагом 10. Особенно заметно увеличение шума в синтезированных спектрах с увеличением количества главных компонент на примере генерации спектров жировой ткани до воздействия липазы (рис. 3, a). На рис. 4, a приведены нагрузки для некоторых главных компонент. Шум на представленных графиках появляется, начиная со второй главной компоненты. Начиная с 20-й компоненты, график нагрузки содержит преимущественно шум. Доля объясненной дисперсии данных в зависимости от количества главных компонент приведена на рис. 4, b.

Поскольку каждая последующая компонента отвечает за меньшую долю объясненной дисперсии в данных, их влияние на процесс генерации спектров КР неравномерное (рис. 4, b). Важно также понимать, каким процентом потерь в данных можно пренебречь в генерации. Так, для анализируемого набора реальных данных при $n_{\text{components}} = 105$ доля потерь составляет менее 5%. Если необходима большая точность — большая доля сохраненной информации о наборе данных — при $n_{\text{components}} = 120$ доля потерь меньше 2.5%, а при $n_{\text{components}} = 130$ потери составляют менее 1%. Если количество компонент не является важным критерием, а более информативной является доля сохраненной дисперсии данных, то его можно указать непосредственно при использовании модуля $decomposition.PCA$ библиотеки $sci-kit\ learn$ на языке Python3. Так, при $n_{\text{components}} = 0.95$ будет выбрано число главных компонент такое, чтобы доля объясненной дисперсии превысила 95%.

Оценка схожести наборов данных с помощью модели классификации

Для оценки схожести синтетического и реального наборов данных использовалась модель классификации на основе случайного леса [21,25]. Эта модель относится

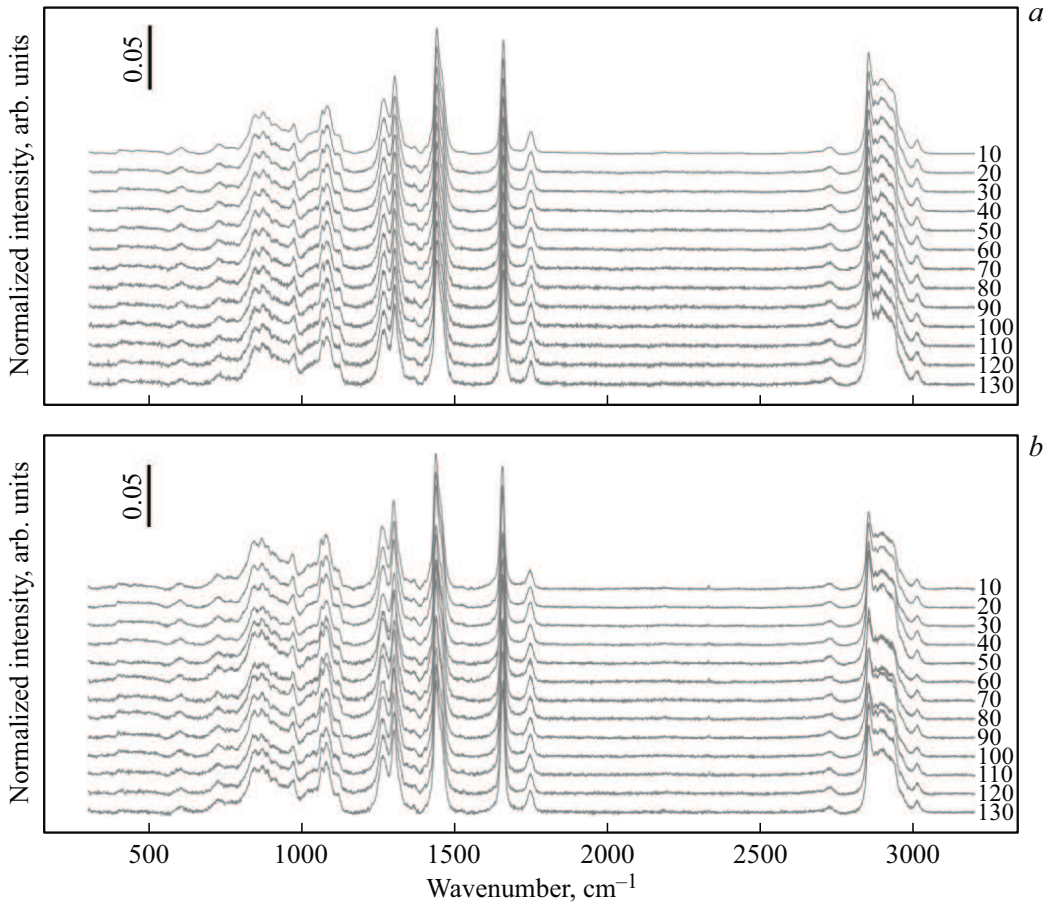

Рис. 3. Генерация спектра КР жировой ткани до (*a*) и после (*b*) воздействия липазы при диапазоне значений параметра $n_{\text{components}}$ 10–130.

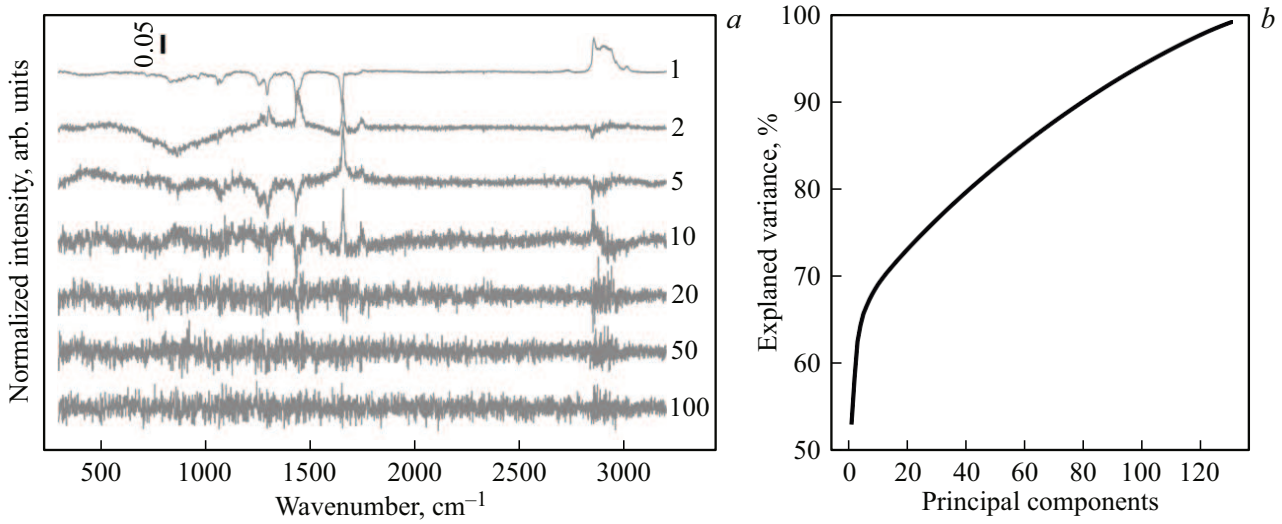

Рис. 4. (*a*) Графики нагрузки для указанных на рисунке главных компонент (1, 2, 5, 10, 20, 50, 100); (*b*) кумулятивный график доли объясненной дисперсии для 1–130 главных компонент.

к ансамблевым моделям машинного обучения, которая объединяет прогнозы нескольких отдельных деревьев принятия решений, работающих параллельно. Каждое отдельное дерево представляет собой иерархическую структуру, а итоговый прогноз ансамбля формируется мажоритарным голосованием. Основными параметрами

Доли верно классифицированных синтетических данных, точность, прецизионность и полнота классификации в зависимости от количества главных компонент

Количество главных компонент	Доля объясненной дисперсии в данных, %	Доля верно классифицированных синтетических спектров до воздействия липазы, %	Доля верно классифицированных синтетических спектров после воздействия липазы,%	Точность (accuracy), %	Прецизионность (precision), %	Полнота (recall), %
5	65.75	100.0	79.3	89.65	100.00	100.00
10	69.13	100.0	79.6	89.80	100.00	100.00
15	71.17	100.0	82.4	91.20	100.00	100.00
20	73.00	100.0	82.2	91.10	100.00	100.00
25	74.71	100.0	82.0	91.00	100.00	100.00
30	76.31	100.0	81.3	90.65	100.00	100.00
35	77.92	100.0	80.2	90.10	100.00	100.00
40	79.38	100.0	81.3	90.65	100.00	100.00
45	80.88	100.0	81.6	90.80	100.00	100.00
50	82.28	100.0	81.2	90.60	100.00	100.00
55	83.66	100.0	81.7	90.85	100.00	100.00
60	84.99	100.0	81.5	90.75	100.00	100.00
65	86.30	100.0	80.0	90.00	100.00	100.00
70	87.55	100.0	81.0	90.50	100.00	99.88
75	88.75	99.9	80.6	90.25	99.88	100.00
80	89.90	100.0	81.0	90.50	100.00	99.88
85	91.01	99.9	81.8	90.85	99.88	100.00
90	92.11	100.0	79.3	89.65	100.00	99.88
95	93.14	99.9	83.4	91.65	99.88	99.88
100	94.16	99.9	82.4	91.15	99.88	100.00
105	95.11	100.0	82.2	91.10	100.00	99.76
110	96.04	99.8	82.4	91.10	99.76	99.76
115	96.99	99.8	81.8	90.80	99.76	100.00
120	97.80	100.0	80.9	90.45	100.00	99.88
125	98.53	99.9	80.6	90.25	99.88	99.76
130	99.18	99.8	82.1	90.95	99.76	100.00

модели случайного леса являются количество использованных деревьев ($n_estimators$), а также „разветвленность“ каждого дерева (max_depth). В настоящей работе использовалась модель случайного леса со следующими параметрами: $n_estimators = 10$, $max_depth = 3$.

Для обучения модели использовались только реальные данные, которые были предварительно поделены на обучающую и тестовую выборки в соотношении 3 : 1. Точность прогнозов модели на тестовой выборке составила 100%. Полученная модель была затем использована для классификации синтетических наборов данных. Для оценки схожести использовались такие метрики, как точность — доля верно классифицированных синтетических данных жировой ткани до ($n = 1000$) и после ($n = 1000$) воздействия липазы, а также доли верно классифицированных синтетических данных по отдельности. Результаты приведены в таблице.

Вычисленная точность классификации синтетического набора данных лимитировалась спектрами КР, имитирующими сигнал жировой ткани после воздействия липазы. Средняя доля верно классифицируемых данных этой категории составила (81.3 ± 1.0)% вне зависимости от количества используемых для генерации главных компонент. Максимальная доля при этом достигла 83.4% при $n_{\text{components}} = 95$ и доле объясненной дисперсии 93.1%. Общая точность для двух наборов сгенерированных спектров КР составила 90.6 ± 0.5%.

Для представленного набора данных наблюдается слабая зависимость точности генерации спектров от количества главных компонент. Начиная с пятнадцати главных компонент, синтетические спектры КР жировой ткани после воздействия липазы были классифицированы верно с точностью более 80%. Важным параметром является не столько количество главных компонент, сколько доля объясненной дисперсии, а также наличие шума в синтетических спектрах. Таким образом, с точки зрения точности классификации достаточна доля объясненной дисперсии в 70% (количество компонент больше 15). Однако для лучшей имитации шума в данных предпочтительнее использовать долю объясненной дисперсии 90%, применяя *sklearn.decomposition.PCA* ($n_{\text{components}} = 0.9$) или выше, вплоть до 0.99.

Выводы

Предложен простой способ генерации данных спектроскопии КР на основе результатов реальных измерений и метода главных компонент. С помощью метода главных компонент набор реальных данных раскладывался на линейную комбинацию нагрузок с соответствующими счетами. На основе предположения о нормальном распределении полученных счетов (зная их среднее и стандартное отклонения) осуществлялась генерация новых случайных счетов. По полученному синтетическому набору и нагрузкам производилась обратная трансформация. Предложенный подход был апробирован на данных спектроскопии КР жировой ткани до и после воздействия липазы. Несмотря на материалы, схожие по набору химических связей, синтезированные наборы спектров имели отличия, что было продемонстрировано с помощью модели классификации на основе случайного леса. Средняя точность для синтетических данных жировой ткани до и после воздействия липазы составила $99.9 \pm 0.1\%$ и $81.3 \pm 1.0\%$ соответственно. Общая точность генерации составила $90.6 \pm 0.5\%$.

Финансирование работы

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (тема государственного задания FSMG-2025-0054).

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.

Список литературы

[1] G. Pezzotti. J. Raman Spectrosc., **52**, 2348.2443 (2021). DOI: 10.1002/jrs.6204
[2] D. Cialla-May, C. Krafft, P. Rösch, T. Deckert-Gaudig, T. Frosch, I.J. Jahn, S. Pahlow, C. Stiebing, T. Meyer-Zedler, T. Bocklitz, I. Schie, V. Deckert, J. Popp. Anal. Chem., **94**, 86–119 (2022). DOI: 10.1021/acs.analchem.1c03235
[3] J.S. Prell, J.T. O.Brien, E.R. Williams. J. Am. Soc. Mass Spectr., **21**, 800–809 (2010). DOI: 10.1016/j.jasms.2010.01.010
[4] L. Gao, R.T. Smith. J. Biophotonics, **8**, 441–456 (2015). DOI: 10.1002/jbio.201400051
[5] A. Conlin, E. Martin, A. Morris. Chemom. Intell. Lab. Syst., **44**, 161–173 (1998). DOI: 10.1016/S0169-7439(98)00071-9
[6] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø. Chemom. Intell. Lab. Syst., **118**, 62–69 (2012). DOI: 10.1016/j.chemolab.2012.07.010
[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Commun. ACM, **63**, 139–144 (2020). DOI: 10.1145/3422622
[8] D. Papadopoulos, V.D. Karalis. Appl. Sci., **13**, 8793 (2023). DOI: 10.3390/app13158793
[9] S. Yu, H. Li, X. Li, Y.V. Fu, F. Liu. Sci. Total Environ., **726**, 138477 (2020). DOI: 10.1016/j.scitotenv.2020.138477
[10] Y. Du, D. Han, S. Liu, X. Sun, B. Ning, T. Han, J. Wang, Z. Gao. Talanta, **237**, 122901 (2022). DOI: 10.1016/j.talanta.2021.122901
[11] A. Coccato, M.C. Caggiani. J. Raman Spectrosc., **55**, 125–147 (2024). DOI: 10.1002/jrs.6621
[12] J.E. Guimar.es, R. Nadas, W. Zhang, T. Endo, K. Watanabe, T. Taniguchi, R. Saito, Y. Miyata, A. Jorio. Phys. St. Solidi, 2500291 (2025). DOI: 10.1002/pssb.202500291
[13] C. Huang, X. Xu, J. Fu, D.-G. Yu, Y. Liu. Polymers, **14**, 3266 (2022). DOI: 10.3390/polym14163266
[14] N. Jin, Y. Song, R. Ma, J. Li, G. Li, D. Zhang. Anal. Chim. Acta, **1197**, 339519 (2022). DOI: 10.1016/j.aca.2022.339519
[15] A. Massei, N. Falco, D. Fissore. Eur. J. Pharm. Biopharm., **200**, 114342 (2024). DOI: 10.1016/j.ejpb.2024.114342
[16] D.R. Body. Prog. Lipid Res., **27**, 39–60 (1988). DOI: 10.1016/0163-7827(88)90004-5
[17] U.M. Lankage, S.A. Holt, S. Bridge, B. Cornell, C.G. Cranfield. ACS Appl. Mater. Interfaces, **15** (45), 52237–52243 (2023). DOI: 10.1021/acsami.3c11767
[18] J. Xuan, Z. Wang, Q. Xia, T. Luo, Q. Mao, Q. Sun, Z. Han, Y. Liu, S. Wei, S. Liu. Foods, **11**, 3664 (2022). DOI: 10.3390/foods11223664
[19] H. hai Wang, Q. Zhang, X. Yu, J. Liang, Y. Zhang, Y. Jiang, W. Su. Ind. Eng. Chem. Res., **62**, 15733–15751 (2023). DOI: 10.1021/acs.iecr.3c02132
[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. J. Mach. Learn. Res., **12**, 2825–2830 (2011).
[21] L. Breiman. Random Forests Mach. Learn., **45**, 5–32 (2001). DOI: 10.1023/A:1010933404324
[22] I.Y. Yanina, Y.I. Svenskaya, E.S. Prikhozhdenko, D.N. Bratashov, M.V. Lomova, D.A. Gorin, G.B. Sukhorukov, V.V. Tuchin. J. Biophotonics, e201800058 (2018). DOI: 10.1002/jbio.201800058
[23] F. Gao, D. Ben-Amotz, S. Zhou, Z. Yang, L. Han, X. Liu. LWT, **134**, 110105 (2020). DOI: 10.1016/j.lwt.2020.110105
[24] N.N. Yazgan Karacaglar, T. Bulat, I.H. Boyaci, A. Topcu. J. Food Drug Anal., **27**, 101–110 (2019). DOI: 10.1016/j.jfda.2018.06.008
[25] M. Poth, G. Magill, A. Filgertshofer, O. Popp, T. Großkopf. J. Raman Spectrosc., **53**, 1580–1591 (2022). DOI: 10.1002/jrs.6402