

УДК 538.91, 004.032.26

КОМПАКТНОЕ ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ О ЛОКАЛЬНОЙ СТРУКТУРЕ ВЕЩЕСТВА В МЕТОДАХ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧ ОБРАБОТКИ ДАННЫХ XANES-СПЕКТРОСКОПИИ

© 2024 г. И. А. Викленко^{а,*}, В. В. Срабионян^а, В. А. Дурыманов^а,
Я. Н. Гладченко-Джевелекис^а, В. Н. Раздоров^а, Л. А. Авакян^а, Л. А. Бугаев^а

^аЮжный федеральный университет, Ростов-на-Дону, 344090 Россия

*e-mail: viklenko@sfedu.ru

Поступила в редакцию 25.09.2023 г.

После доработки 22.11.2023 г.

Принята к публикации 22.11.2023 г.

В работе предложен метод представления данных о локальной атомной структуре в виде гистограмм парных радиальных функций распределения с разрешением по типам атомов. Метод используют для построения дескриптора структуры, необходимого для определения структуры вещества методами машинного обучения и искусственного интеллекта. Особенностью метода является использование одновременно двух наборов парных радиальных функций распределения: для пар всех типов атомов и для пар с выделенным поглощающим атомом. Разработанный подход апробирован на примере задачи определения ближнего окружения атомов серебра в центрах окраски в натриево-силикатных стеклах по данным спектров рентгеновской абсорбционной спектроскопии вблизи *K*-края поглощения Ag. Информативность предложенного дескриптора структуры продемонстрирована путем возможности воссоздания трехмерной модели структуры центра окраски серебра из соответствующих гистограмм парных расстояний. При использовании нескольких методов машинного обучения было показано, что предложенный дескриптор позволяет добиться качественного воспроизведения спектров XANES для центров окраски в стекле в рамках метода конечных разностей, что позволяет сократить время расчета спектров XANES на четыре порядка. Построенная модель машинного обучения позволяет установить фундаментальную связь между атомным строением центров окраски в стеклах и XANES-спектром серебра, необходимую для определения строения стекол.

Ключевые слова: XANES, машинное обучение, локальная атомная структура, центры окраски, градиентный бустинг, алгоритмы случайного леса, метод линейной регрессии.

DOI: 10.31857/S1028096024040045, **EDN:** GKASFX

ВВЕДЕНИЕ

Методам машинного обучения находят все большее применение в современной науке в целом и материаловедении в частности [1–5]. Одним из широко используемых и развивающихся методов получения информации об атомной структуре вещества является спектроскопия поглощения рентгеновского излучения в энергетическом интервале до ~50 эВ относительно края поглощения (XANES). Этот метод обладает высокой чувствительностью к локальному строению вещества вблизи поглощающего атома и позволяет

исследовать структуру даже в отсутствие дальнего порядка. Однако получение информации о трехмерной атомной структуре вещества по данным XANES является достаточно сложной и ресурсоемкой задачей, не всегда приводящей к успеху. Для решения этой задачи могут быть применены методы машинного обучения [6]. В основе этих методов лежит использование математических моделей, параметры которых предварительно подлагают настройке в процессе обучения. В области материаловедения наиболее распространенным способом такой настройки является обучение с учителем, требующее достаточный объем данных

о физических характеристиках и свойствах вещества. В настоящее время существуют достаточно крупные базы данных, содержащие информацию о структуре и свойствах различных материалов, такие как AFLOW [7], Materials Project [8], OMDb [9], OQMD [10], QM9 [11, 12], которые можно использовать в процессе настройки модели машинного обучения. Набор данных, предназначенный для обучения, должен быть как можно большего размера и максимально разнообразным для достижения требуемой обобщающей способности модели машинного обучения. К сожалению, при решении узкоспециализированных задач необходимый для обучения модели набор данных может отсутствовать в открытом доступе. В таком случае для улучшения обобщающей способности моделей машинного обучения прибегают к различным техникам для улучшения качества “обучающего” набора данных. Например, в работе [14, 15] применена выборка набора данных по методу латинского гиперкуба, для предположительного уменьшения общего размера набора данных без потери качества обучения на нем. Также, как правило, исследователи ограничиваются рассмотрением нерелаксированных групп атомов вместо более реалистичных, получаемых после выполнения атомной релаксации, а также используют информацию о зависимости входных данных и целевых значений для существенного увеличения объема синтетического набора данных, как это, например, было проведено в работе [6] при генерации теоретических спектров XANES путем смешивания сигналов, полученных от атомов, находящихся в различных неэквивалентных позициях (без учета реалистичности весовых множителей).

В результате использования методов машинного обучения для изучения структуры и свойств материалов возникает потребность в создании удобного представления входных данных о структуре вещества. В частности, в задачах спектроскопии рентгеновского поглощения используют информацию о локальном окружении атомов вблизи поглощающего атома. Для улучшения качества работы методов машинного обучения, а также для учета физики процесса, функция представления данных о структуре вещества должна быть инвариантной по отношению к преобразованиям системы координат с помощью поворота, отражения и сдвига структуры как единого целого, а также такое представление должно быть инвариантно к изменению нумерации атомов в системе. Таким образом, функция представления, называемая “дескриптор”, должна быть однозначно определена для любой рассматриваемой структуры материала и одновременно отражать

небольшие изменения в ней, а вычисление этой функции должно быть значительно быстрее прямого вычисления интересующего целевого значения (физического свойства) [15].

В настоящей работе предложен метод представления данных о локальной атомной структуре (дескрипторов) в виде гистограмм парных радиальных функций распределения атомов (ПРФРА), построенных с учетом различия типов атомов. Метод применен для решения задачи определения структуры ближнего окружения атомов серебра в центрах окраски в натриево-силикатных стеклах по данным спектров рентгеновской абсорбционной спектроскопии вблизи K -края поглощения серебра. Известно, что спектры XANES являются крайне чувствительными к локальному окружению атомов вблизи поглощающего атома [16–18]. Поэтому используемые дескрипторы должны однозначным образом кодировать информацию о структуре, а также иметь возможность указания выделенного поглощающего атома.

МЕТОД ГИСТОГРАММ

Описание метода

В настоящей работе применен метод представления данных о локальной атомной структуре вещества в виде гистограмм парных радиальных функций распределений атомов с учетом типов атомов. Были использованы наборы ПРФРА двух видов. Первый вид соответствует общепринятому выбору ПРФРА, при котором усреднены все парные расстояния между выбранными типами атомов. Так, для случая системы из атомов двух сортов A и B будут сформированы четыре ПРФРА: $A-A$, $A-B$, $B-A$ и $B-B$ (рис. 1а). Дополнительно мы рассматривали ПРФРА, построенные для одного выделенного центрального атома A^* , интерпретируемого как атом, поглощающий рентгеновский фотон. В случае присутствия в системе атомов двух сортов это приводит к добавлению еще двух ПРФРА: A^*-A и A^*-B (рис. 1б, 1в). Для этого выбранного центрального атома A^* в дальнейшем строили вклад от него в XANES спектр, поэтому усреднение по всем возможным позициям A не проводили.

Гистограммы радиального распределения строили в диапазоне возможных радиусов от $R_{\min}$, обусловленного минимально возможным расстоянием между парой атомов, до $R_{\max}$, характеризующего размеры области, существенной для формирования спектра поглощения, протяженность которой ограничивается конечностью длины свободного пробега электрона, выбитого рентгеновским фотоном. Гистограммы парных

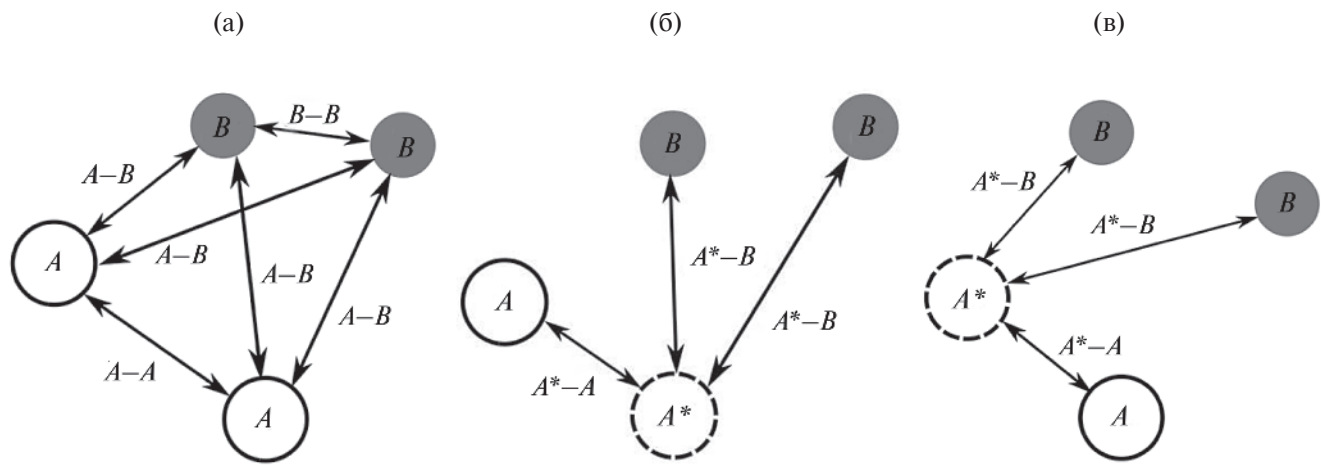


Рис. 1. Иллюстрация пар атомов типов *A* и *B*, учитываемых при построении ПРФА, при отсутствии (а) и наличии (б, в) выделенного поглощающего рентгеновский фотон атома *A**.

расстояний для каждой возможной пары атомов строили в диапазоне от $R_{\min}$ до $2R_{\max}$ (максимальное допустимое межатомное расстояние внутри рассматриваемого объема) с заранее выбранным шагом ΔR . Общее количество гистограмм N_{hist} зависит только от количества уникальных атомных типов в структуре:

$$N_{\text{hist}} = 2N_{\text{atomtype}} + C^2_{\text{Natomtype}},$$

где N_{atomtype} — число типов атомов в системе; $C^2_{\text{Natomtype}}$, число сочетаний из N_{atomtype} по 2. Полученные гистограммы являются инвариантными к изменению нумерации атомов в системе, однако для унификации используемого метода представлений данных необходимо дополнительно определить порядок конкатенации гистограмм для формирования входного набора данных. В настоящей работе использовали представление данных в виде вектора, а соответствующие гистограммы конкатенировали в порядке увеличения атомных номеров в рассматриваемых подсистемах.

Выбор представления данных в виде гистограмм обусловлен помимо всего прочего удобством их интерпретации. То есть, из гистограммы можем получить исчерпывающую информацию о типах атомов, их взаимном расположении. В настоящей работе использовали шаг $\Delta R = 0.1 \text{ \AA}$ для парных гистограмм, и $0.05 \text{ \AA}$ — для радиальных, что обусловлено, с одной стороны, чувствительностью спектров XANES к изменению координат атомов в ближайшем окружении относительно поглощающего атома, а с другой стороны — простотой настройки и обучения моделей машинного обучения.

В случае центра окраски в силикатном стекле имеются атомы трех типов — Ag, Si и O, — и один из атомов Ag является поглощающим. В качестве

примера рассмотрим центр окраски компонентного состава AgSi_2O_4 , структура которого приведена на вставке на рис. 2, а соответствующие координаты — в табл. 1. Панели на рис. 2 показывают построенные гистограммы ПРФА. Видно, что гистограммы парных расстояний с атомами серебра пусты, ввиду того, что поглощающий атом — Ag, удален из структуры центра окраски в процессе построения гистограмм. Очевидно, что в случае присутствия в структуре поглощающего атома Ag содержащаяся в них информация дублировала бы информацию из функций радиального распределения. Тем не менее наличие этих гистограмм, заполненных нулевыми значениями интенсивности, необходимо для обеспечения возможности рассмотрения систем, имеющих более одного атома того же типа, что и поглощающий (решение задачи без потери общности).

*Информативность используемого представления.
Обратная задача*

Для проверки информативности предложенного дескриптора была предпринята попытка

Таблица 1. Координаты атомов центра окраски с компонентным составом AgSi_2O_4

Тип атома	<i>X</i> , Å	<i>Y</i> , Å	<i>Z</i> , Å
Ag	0.00	0.00	0.00
O	0.94	1.27	0.87
O	0.26	1.49	−1.06
O	−0.85	−0.09	2.16
O	−1.46	−1.81	1.82
Si	−0.85	1.44	2.57
Si	0.94	1.27	0.87

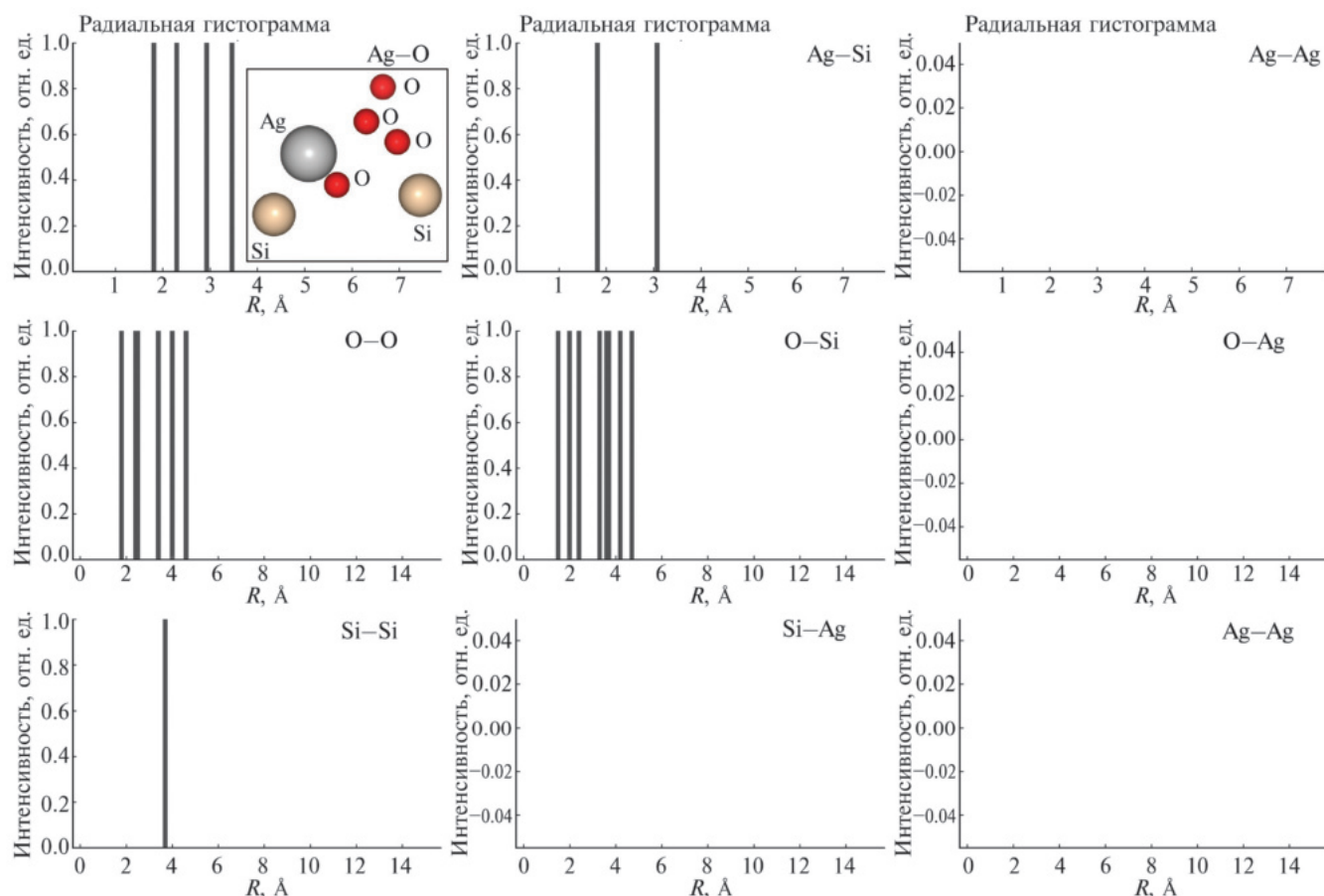


Рис. 2. Представление информации о локальном атомном окружении центра окраски с применением гистограмм радиального распределения относительно поглощающего атома (верхний ряд) и гистограмм всевозможных парных расстояний. На вставке приведена соответствующая структура центра окраски.

воспроизвести с его помощью трехмерную структуру центра окраски. Для упрощения поставленной задачи можно разделить всю систему на набор подсистем, состоящих из атомов только одного типа. Рассмотрим одну из получившихся подсистем, содержащую атомы типа *A*. Из гистограммы радиального распределения мы можем легко получить число атомов в данной подсистеме, которое является суммой всех значений в гистограмме. Также очевидно, что число атомов в подсистеме можно получить из соответствующей гистограммы парных расстояний по формуле:

$$N_{\text{atoms}} = \frac{1 + \sqrt{1 + 8N_{\text{bonds}}}}{2},$$

где N_{bonds} — число межатомных связей в рассматриваемой подсистеме, которое определяется как сумма всех значений в соответствующей гистограмме парных расстояний.

Исходя из полученной информации, мы разместили необходимое количество атомов заданного

типа на расстояниях, определенных из соответствующей гистограммы радиального распределения. После этого приступили к решению задачи оптимизации гистограммы парных расстояний при сохранении условия нахождения атомов на правильных расстояниях от начала системы координат, соответствующую задачу можно просто решить в сферической системе координат путем варьирования углов Θ и ϕ (рис. 3). Очевидно, что восстановленная структура определена с точностью до произвольного поворота/отражения всей системы как единого целого.

Аналогичным образом восстанавливают структуру остальных подсистем, после чего проводят их “стыковку”. Все подсистемы имеют общий центр (поглощающий атом в нуле координат), поэтому мы можем осуществлять их попарное объединение, для этого фиксируем одну из подсистем, например, *A*, а вторую подсистему вращаем как единое целое относительно начала системы координат (рис. 4), решая задачу оптимизации функции невязки для гистограммы парных расстояний

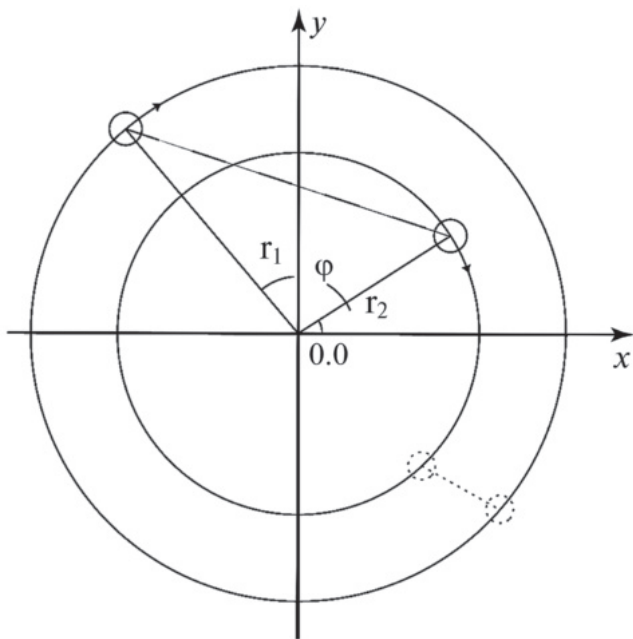


Рис. 3. Схематичное представление процедуры оптимизации структуры подсистемы $A-A$. Точечным пунктиром показано правильное расположение атомов типа A , сплошным черным цветом – исходные параметры системы.

атомов типов $A-B$. Данную процедуру повторяем для всех оставшихся подсистем.

Очевидно, что ввиду дискретного характера гистограмм, структура подсистем может быть воспроизведена только с ограниченной точностью, зависящей от выбранного шага ΔR . С учетом обозначенных особенностей используемого дескриптора задача оптимизации непосредственно множества гистограмм сводится к задаче минимизации расстояния между двумя точками в многомерном пространстве значений гистограмм, представляемых разреженными векторами (в которых данные представлены фрагментарно), не чувствительными к малому изменению координат в системе (чувствительность определяется шагом гистограммы), что усложняет процесс оптимизации. Для реконструкции трехмерной структуры системы целесообразно перейти к оптимизации векторов парных расстояний, компоненты которых формируются положениями столбцов в гистограммах, а общее количество компонент определяется суммой всех значений соответствующей гистограммы. На рис. 5 показано качество описания исходной гистограммы, построенной для центра окраски (рис. 1), восстановленного с помощью обсуждаемого дескриптора. Соответствующая восстановленная структура центра окраски также приведена на рис. 5, а ее координаты – в табл. 2.

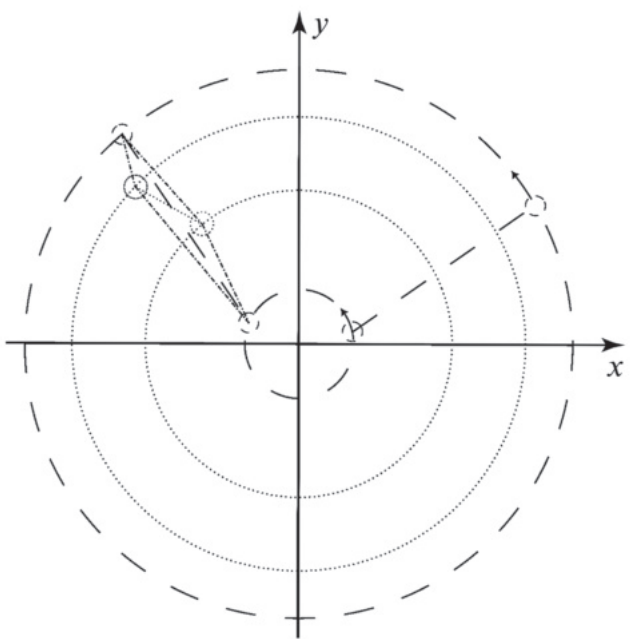


Рис. 4. Схематичное представление процедуры оптимизации относительного положения подсистем A и B . Точечным пунктиром показано положение подсистемы A , пунктиром положение подсистемы B , штрих-пунктиром – ожидаемое положения подсистемы B относительно подсистемы A .

Таким образом, было показано, что представление информации о локальной атомной структуре вещества в виде набора гистограмм парных расстояний и радиального распределения является информативным, и позволяет восстановить исходную структуру системы по соответствующему дескриптору.

ПРИМЕРЫ ОПИСАНИЯ
ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Для проверки применимости описанного дескриптора к решению задачи прямого расчета

Таблица 2. Координаты атомов восстановленного центра окраски с компонентным составом AgSi_2O_4

Тип атома	$X, \text{\AA}$	$Y, \text{\AA}$	$Z, \text{\AA}$
Ag	0.00	0.00	0.00
O	−1.01	2.84	1.66
O	0.43	1.64	1.61
O	1.43	−0.41	0.92
O	−0.24	−1.74	2.31
Si	1.93	1.30	1.93
Si	0.97	−0.18	−1.47

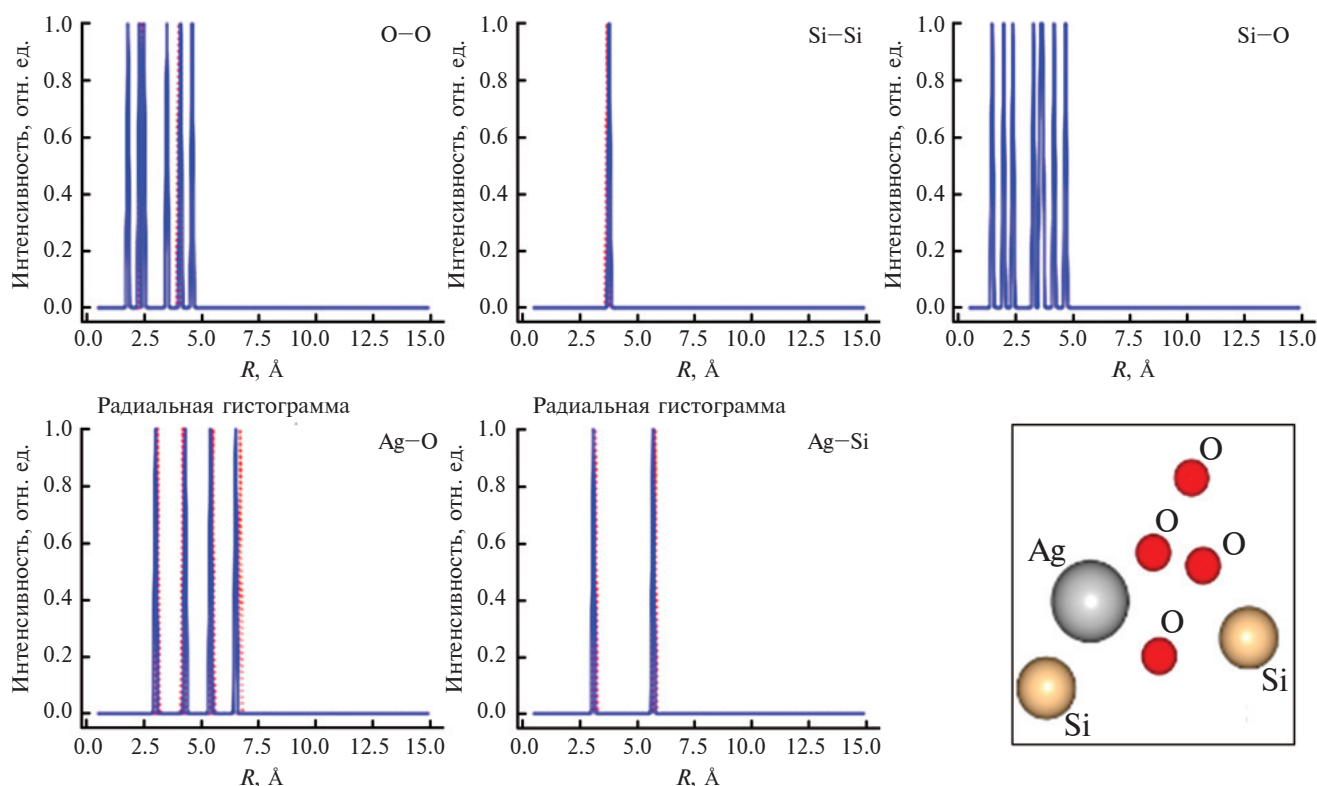


Рис. 5. Сопоставление гистограмм парных расстояний (верхний ряд) и гистограмм радиального распределения относительно поглощающего атома в центре окраски: исходных (пунктиром) и восстановленных (сплошными линиями); на вставке изображена трехмерная восстановленная структура центра окраски.

спектров XANES по данным о трехмерной атомной структуре вещества нами было проведено обучение некоторых наиболее распространенных моделей машинного обучения, поддерживающих по умолчанию решение задачи многоцелевой регрессии. В частности, нами были рассмотрены методы на основе алгоритмов случайного леса (RandomForest), дополнительных деревьев (ExtraTrees), методов линейной регрессии с $L1$ и $L2$ регуляризацией, реализованных в библиотеке Sklearn [19], а также метод градиентного бустинга над решающими деревьями, реализованный в библиотеке CatBoost [20].

При построении набора данных для обучения рассматривали системы, компонентный состав которых гарантированно содержал атомы серебра, а также включал атомы кислорода и/или кремния. “Обучающий” набор данных включал в себя, как информацию о структуре, полученную из открытых баз данных Materials Project, AFLOW, так и синтетические структуры, полученные случайной генерацией или в процессе байесовской оптимизации. Результирующий набор содержал 1800 систем и соответствующих спектров XANES, посчитанных в программе FDMNES [21] в рамках метода конечных разностей. Весь набор данных

был поделен на тренировочную и проверочную части, проверочная часть составила 10% от всего объема данных. Оптимизация гиперпараметров моделей осуществлена методом поиска на сетке (GridSearchCV) с перекрестной проверкой на пяти подмножествах, оценка точности обученных моделей приведена на рис. 6.

Стоит отметить, что применение метода градиентного бустинга для задачи многоцелевой регрессии, реализованного в CatBoost, не только демонстрирует наилучшее описание данных по сравнению с другими методами машинного обучения, рассмотренными в рамках настоящей работы, но также позволяет добиться хорошей обобщающей способности модели машинного обучения, даже без существенной настройки гиперпараметров. Сопоставление некоторых случайно выбранных из многих спектров XANES, рассчитанных в программе FDMNES и полученных с применением модели машинного обучения на основе градиентного бустинга на деревьях решений, представлено на рис. 7. Из представленного сопоставления можно заключить, что качество работы модели машинного обучения практически не меняется при переходе от тренировочного набора данных, на которых проведена настройка

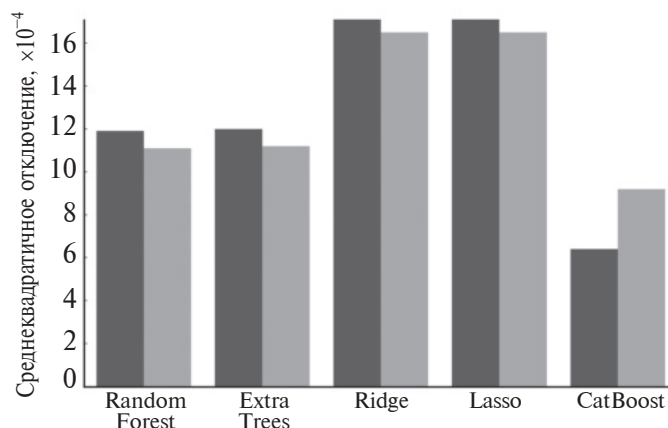


Рис. 6. Величина среднеквадратичной ошибки предсказания спектров XANES используемых моделей машинного обучения на тренировочном (темно-серым) и проверочном (светло-серым) подмножествах данных.

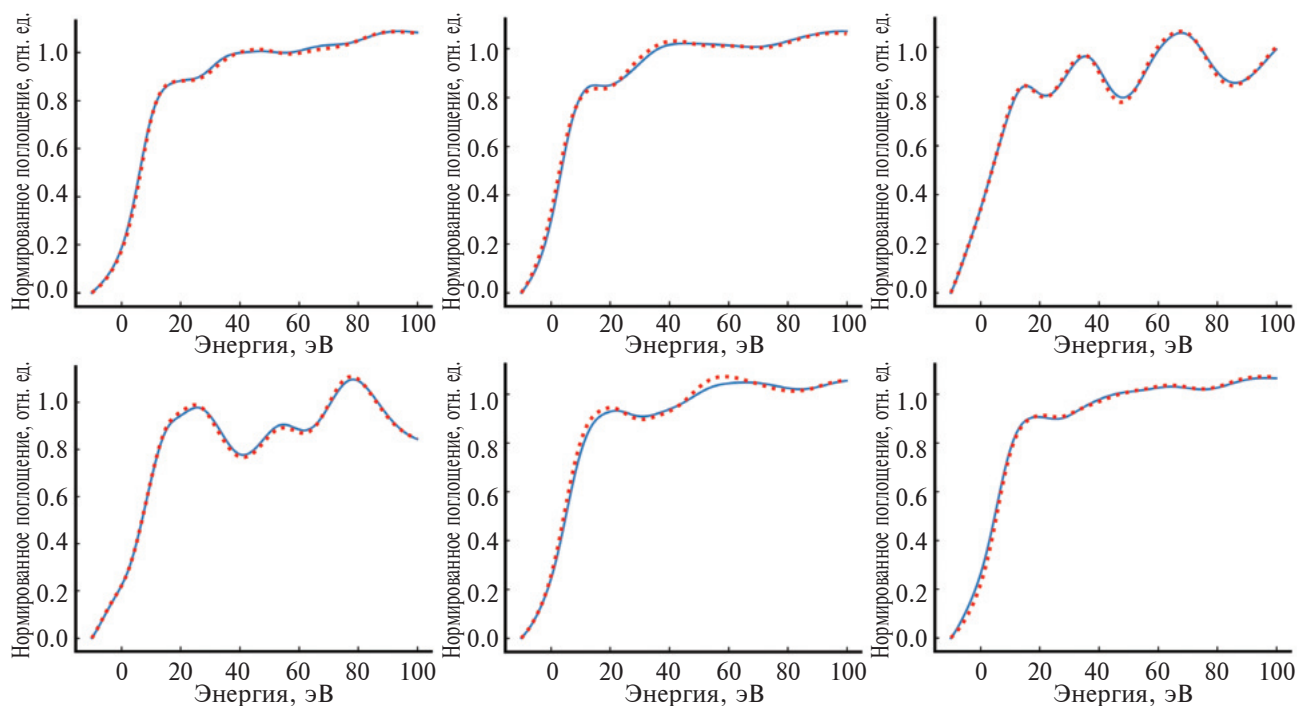


Рис. 7. Сопоставление спектров поглощения рентгеновского излучения вблизи K -края Ag, полученных с применением модели градиентного бустинга (сплошная линия), и спектров, рассчитанных в программе FDMNES (пунктирная линия), для тренировочного (сверху) и проверочного (снизу) подмножества данных.

параметров модели, к тестовой части данных, что указывает на отсутствие явного переобучения модели. Также стоит отметить, что рассматриваемая модель машинного обучения достаточно точно воспроизводит положения максимумов и их относительные интенсивности на спектрах XANES за краем поглощения (особенности в этой области очень чувствительны к изменениям в атомной структуре образца), а средняя невязка между рассчитанными спектрами составляет порядка $\sim 10^{-3}$.

Для дополнительной проверки обобщающей способности обученной модели был рассчитан спектр центра окраски в натриево-силикатном стекле, структура которого была определена в работе [22], не входящий в обучающих набор данных. Обученная модель на основе методов градиентного бустинга на деревьях решений воспроизводит все основные особенности спектра структуры центра окраски из работы [22], соответствующее качество описания спектра приведено на рис. 8.

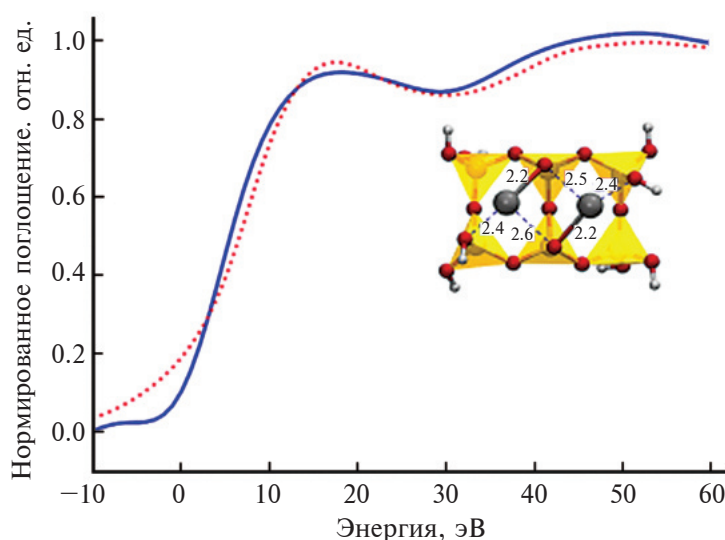


Рис. 8. Сопоставление спектров поглощения рентгеновского излучения вблизи K -края Ag, полученных с применением модели градиентного бустинга (сплошная линия), и спектра, рассчитанного в программе FMNESH (пунктирная линия), центра окраски в натриево-силикатной стеклянной матрице со структурой, полученной в работе [22] (показана на вставке).

ЗАКЛЮЧЕНИЕ

В рамках настоящей работы был предложен дескриптор для описания локальной атомной структуры вещества с выделенным атомом на основе представления данных в виде гистограмм всевозможных парных расстояний и гистограмм радиального распределения относительно выделенного (поглощающего) атома. Продемонстрирована информативность такого представления данных о структуре вещества, а также возможность интерпретации дескриптора для восстановления исходной трехмерной локальной атомной структуры вещества с точностью до шага гистограммы. Предложенный дескриптор позволяет добиться качественного описания (расхождение составляет $\sim 10^{-3}$) спектров XANES для центров окраски в стекле рассчитанных с помощью программы FDMNES, среднее время расчета одного спектра с применением FDMNES составляло ~ 30 мин, а время расчета одного спектра методами машинного обучения в среднем составило ~ 7 мс. Также применение предложенного дескриптора и моделей машинного обучения на основе градиентного бустинга на деревьях решений позволило качественно воспроизвести спектр структуры, описывающей экспериментальные данные центра окраски в натриево-силикатной стеклянной матрице, установленной в работе [22].

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при поддержке Российского научного фонда (грант № 23-21-00526).

Конфликт интересов. Авторы данной работы заявляют, что у них нет конфликта интересов

СПИСОК ЛИТЕРАТУРЫ

1. Seko A., Toyoura K., Muto S., Mizoguchi T., Broderick S. // MRS Bull. 2018. V. 43. № 9. P. 6905.
2. Хаметова Э.Ф., Бакиева О.Р. // Ученые записки Физического факультета МГУ. 2022. Т. 4. С. 2240703.
3. Гуда С.А., Алгасов А.С. Технологии машинного обучения для анализа геометрии молекул // Вестник Ростовского Государственного университета путей сообщения. 2019. Т. 2. Вып. 74. С. 84.
4. Орешко Е.И., Ерасов В.С., Сибяев И.Г., Луценко А.Н., Шершак П.В. // Авиационные материалы и технологии. 2022. Т. 4. Вып. 69. P. 132.
5. Bratchenko I.A., Artemyev D.N., Khristoforova Y.A., Bratchenko L.A. // Biomed. Opt. Express. 2019. V. 10. № 9. P. 4489.
6. Timoshenko J., Lu D., Lin Y., Frenkel A.I. // J. Phys. Chem. Lett. 2017. V. 8. № 20. P. 5091.
7. Curtarolo S., Setyawan W., Wang S., Xue J., Yang K., Taylor R.H., Nelson L.J., Hart G.L.W., Sanvito S., Buongiorno-Nardelli M., Mingo N., Levy O. // Comput. Mater. Sci. 2012. V. 58. P. 227.
8. Jain A., Ong S.P., Hautier G., Chen W., Richards W.D., Dacek S., Cholia S., Gunter D., Skinner D., Ceder G., Persson K.A. // APL Mater. 2013. V. 1. № 1. P. 011002.
9. Borysov S.S., Geilhufe R.M., Balatsky A.V. // PLoS One. 2017. V. 12. № 2. P. 0171501. <https://doi.org/10.1371/journal.pone.0171501>
10. Saal J.E., Kirklin S., Aykol M., Meredig B., Wolverton C. // JOM. 2013. V. 65. № 11. P. 1501.
11. Ruddigkeit L., van Deursen R., Blum L.C., Raymond J.-L. // J. Chem. Inf. Model. 2012. V. 52. № 11. P. 2864.

12. Ramakrishnan R., Dral P.O., Rupp M., von Lilienfeld O.A. // *Sci. Data*. 2014. V. 1. № 1. P. 140022.
13. Shields M.D., Zhang J. // *Reliability Engineering System Safety*. 2016. V. 148. P. 96.
<https://doi.org/10.1016/j.ress.2015.12.002>
14. Guda A.A., Guda S.A., Martini A., Bugaev A.L., Soldatov M.A., Soldatov A.V., Lamberti C. // *Radiat. Phys. Chem.* 2020. V. 175. P. 108430.
15. Himanen L., Jäger M.O.J., Morooka E. V., Federici Canova F., Ranawat Y.S., Gao D.Z., Rinke P., Foster A.S. // *Comput. Phys. Commun.* 2020. V. 247. P. 106949.
16. Vedrinskii R.V., Kraizman V.L. // *Uspekhi Fiz. Nauk*. 1988. V. 154. № 1. P. 172.
17. Koningsberger D.C., Prins R. *X-ray absorption: principles, applications, techniques of EXAFS, SEXAFS and XANES*. N.Y.: John Wiley and Sons Inc., 1987.
18. van Bokhoven J.A., Lamberti C. *X-Ray Absorption and X-Ray Emission Spectroscopy* // *X-Ray Absorption and X-Ray Emission Spectroscopy: Theory and Applications* / Ed. Van Bokhoven J.A., Lamberti C. Chichester, UK: John Wiley & Sons, Ltd, 2016. P. 1.
19. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. // *J. Mach. Learn. Res.* 2011. V. 12. P. 2825.
20. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: unbiased boosting with categorical features. 2017. arXiv:1706.09516
<https://doi.org/10.48550/arXiv.1706.09516>
21. Joly Y. // *Phys. Rev. B*. 2001. V. 63. № 12. P. 125120.
22. Srabionyan V.V., Avakyan L.A., Durymanov V.A., Rubanik D.S., Viklenko I.A., Skunova A.V., Bugaev L.A. // *J. Phys. Chem. Solids*. 2023. V. 179. P. 111412.

Compact Representation of the Local Atomic Structure of Matter for Machine Learning in XANES Spectroscopy Data Processing

I. A. Viklenko^{1,*}, V. V. Srabionyan¹, V. A. Durymanov¹, Ya. N. Gladchenko-Dzhevelekis¹,
V. N. Razdorov¹, L. A. Avakyan¹, L. A. Bugaev¹

¹*Southern Federal University, Rostov-on-Don, 344090 Russia*

^{*}*e-mail: viklenko@sfedu.ru*

A method for representing data on the local structure of atoms in the form of histograms of paired radial distribution functions is proposed. This method is used to construct a structure descriptor needed to determine the structure of materials using machine learning and artificial intelligence techniques. A special feature of the method is the use of two sets of paired radial distribution functions simultaneously: for pairs of all types of atoms and for pairs with a selected absorbing atom. The developed approach was tested on the problem of determining the local atomic structure of the environment of the silver color center in sodium silicate glasses using data from X-ray absorption near-edge structure for the Ag *K*-edge. The information content of the proposed structure descriptor is demonstrated by the ability to reconstruct the three-dimensional structure of a silver color center model from the corresponding pairwise distance histograms. Using several machine learning methods, it was shown that the proposed descriptor allows to achieve high-quality reproduction (mean square error $\sim 10^{-3}$) of X-ray absorption near-edge structure spectra for silver color centers in glass, which makes it possible to reduce the time for calculating X-ray absorption near-edge structure spectra by 4 orders of magnitude. The resulting machine learning model allows us to establish a fundamental connection between the atomic structure of silver color centers in glasses and the Ag X-ray absorption near-edge structure spectrum, which is necessary for determining the structure of glasses.

Keywords: X-ray absorption near-edge structure, machine learning, local atomic structure, color centers, gradient boosting, random forest algorithms, linear regression method.