УДК 51-37:544.127

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ РЕНТГЕНОСПЕКТРАЛЬНЫХ ДАННЫХ ДЛЯ СМЕСИ СОЕДИНЕНИЙ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

© 2021 г. А. С. Алгасов[b, *], С. А. Гуда[a, b], А. А. Гуда[a, **], Ю. В. Русалёв[a], А. В. Солдатов[a]

[a] *Международный исследовательский институт интеллектуальных материалов, Южный федеральный университет, Ростов-на-Дону, 344090 Россия*

[b] *Институт математики, механики и компьютерных наук, Южный федеральный университет, Ростов-на-Дону, 344090 Россия*

**e-mail: alexander.algasov@gmail.com*

***e-mail: guda@sfedu.ru*

Поступила в редакцию 27.03.2020 г.
После доработки 20.05.2020 г.
Принята к публикации 25.05.2020 г.

На основе алгоритмов машинного обучения разработан метод определения структурных параметров компонентов смеси по данным рентгеновских спектров поглощения. Строится база данных спектров для всевозможных предполагаемых деформаций структуры каждого компонента. Метод машинного обучения, реализованный в программном комплексе PyFitIt, позволяет быстро рассчитать спектр для деформаций структур из рассматриваемого семейства и оптимизировать структурные параметры смеси путем подгонки теоретического спектра под экспериментальный. Возможности метода рассмотрены на примере анализа изменений структурных характеристик и концентраций компонентов смеси для бис-диоксоленового комплекса кобальта с функционализированными иминопиридиновыми лигандами при его валентно-таутомерной интерконверсии в зависимости от температуры.

Ключевые слова: анализ компонентов смеси, PyFitIt, машинное обучение, валентно-таутомерная интерконверсия.

DOI: 10.31857/S1028096021050034

ВВЕДЕНИЕ

Для определения структурных параметров компонентов смеси и их концентраций на основе анализа околопороговой тонкой структуры рентгеновских спектров поглощения (XANES – X-ray Absorption Near Edge Structure) существует несколько подходов. Исторически первым был подход моделирования смеси линейной комбинацией заданных спектров и поиск соответствующих коэффициентов [1–3]. Когда неизвестно количество компонентов и их спектры, применяют анализ главных компонентов (PCA – Principal Component Analysis), фактор-анализ [4] и в последнее время метод анализа многомерных кривых путем среднеквадратичной вариации (MCR-ALS – Multivariate Curve Resolution using Alternating Least Squares) [5, 6].

Рассматриваемая в настоящей работе задача имеет свою специфику, которую нельзя учесть перечисленными методами анализа смесей. В случае, когда с большой вероятностью известно, к какому семейству принадлежат атомные структуры компонентов смеси, для определения каждого компонента можно воспользоваться классическим, зарекомендовавшим себя подходом – подбором геометрических параметров атомной структуры так, чтобы ее расчетный (теоретический) спектр был как можно ближе к экспериментальному. Первая программа MXAN, автоматизирующая данный процесс, появилась в 2001 г. [7–10]. Впоследствии многие другие программы, рассчитывающие спектр, включили в себя функции автоматического подбора параметров. Чтобы один раз посчитать значение оптимизируемой функции, необходимо рассчитать теоретический спектр, что требует значительного времени. Автоматическая оптимизация часто работает дольше недели и иногда приводит к физически маловероятным или даже некорректным структурам. Требуется ручное вмешательство человека в процесс оптимизации. Чтобы сделать удобной ручную оптимизацию, нужно обеспечить высокую скорость расчета спектра. В программе FitIt [11] применяют предварительный расчет множества спектров для набора геометрических параметров и их последующую интерполяцию. В настоящее время процедура аппроксимации спектров, реализован-

ная в программе FitIt, была значительно улучшена методами машинного обучения в приложении PyFitIt [12]. Функции PyFitIt были также расширены для решения задач определения структурных параметров на основе анализа рентгеновских спектров поглощения в случае смеси веществ. Наряду с интерфейсом пользователя, в котором можно менять атомные структуры компонентов смеси, добиваясь наилучшего совпадения с экспериментальными спектрами, полученными для разных температур, в PyFitIt также была встроена процедура автоматического подбора геометрических параметров компонентов смеси и их концентраций равномерно по всему интервалу температур.

Предлагаемый подход определения параметров компонентов смеси является альтернативой существующим алгоритмам – PCA, фактор-анализу, MCR-ALS – и может быть использован совместно с ними для подтверждения полученных результатов. Этот подход имеет ряд преимуществ. В частности, для анализа смеси достаточно всего одного экспериментального спектра, также алгоритм хорошо работает в случае спектров с высоким шумом.

В настоящей работе преимущества метода машинного обучения продемонстрированы на примере исследования валентно-таутомерной интерконверсии в комплексе кобальта $(diox)_2Co(imPy\text{-}TEMPO)$. Ранее [13] было показано, что этот бис-диоксоленовый кобальтовый комплекс, включающий функционализированый иминопиридиновый лиганд, претерпевает спиновый переход в интервале температур 200–300 K в твердом состоянии. Согласно данным магнитной восприимчивости, такая интерконверсия, скорее всего, вызвана валентной таутомерией. Однако все попытки изучить структурные изменения, связанные с наблюдаемым превращением, методом рентгеновской дифракции монокристаллов не увенчались успехом, поскольку кристалл разрушается при его охлаждении ниже 220 K, и остаются неизвестными особенности молекулярной структуры низкотемпературного изомера. Метод анализа спектров XANES за K-краем поглощения Co для $(diox)_2Co(imPy\text{-}TEMPO)$ в широком температурном интервале от 30 до 300 K с использованием алгоритмов машинного обучения позволил определить параметры локальной атомной структуры различных изомеров и их концентрации в зависимости от температуры.

МЕТОДЫ

Следуя подходу, реализованному в приложении PyFitIt [12], для каждого компонента смеси строят модели машинного обучения, позволяющие быстро рассчитывать спектр для заданной геометрической структуры. В интерактивном приложении пользователь с помощью слайдеров меняет параметры геометрии структур компонентов смеси и тут же видит результирующий спектр (рис. 1), что позволяет удобно проводить подгонку теоретического спектра под экспериментальный в ручном режиме или запустить автоматический оптимизатор.

Для работы данного подхода пользователю программы необходимо для каждого предположительного i-го компонента смеси определить вектор геометрических параметров g_i и составить функцию M_i, которая создает атомную структуру по заданному вектору g_i, $i = 1, ..., n$, n – число компонентов. В качестве координат вектора g_i могут выступать смещения групп атомов в некоторых направлениях, повороты частей структуры, смещения отдельных атомов или изменение координат всех атомов структуры. Необходимо учитывать, что большое количество геометрических параметров потребует долгого времени при построении обучающей выборки.

Полученные параметры атомной структуры модельных соединений затем вводят в программу расчета спектров рентгеновского поглощения, например, FDMNES [14, 15]. Обозначим процедуру расчета спектра по заданной атомной структуре через S. Точное вычисление спектра требует много времени, что не позволяет использовать его в интерактивном режиме. Поэтому используем аппроксимацию методами машинного обучения. Построим обучающую выборку, рассчитав спектры XANES i-го компонента для набора значений вектора $g_{ij}, j = 1, ..., m_i$:

$$\begin{gathered} \mathrm{XANES}_{ij} = S(M_i(g_{ij})), \\ j = 1, \ldots, m_i, \quad i = 1, \ldots, n. \end{gathered} \tag{1}$$

Рассчитанная база данных спектров позволяет построить аппроксимацию A_i суперпозиции $S \circ M_i$, для быстрого получения приближения спектра по заданному произвольному вектору геометрических параметров g_i:

$$\mathrm{ApproxXANES}_i = A_i(g_i), \quad i = 1, \ldots, n. \tag{2}$$

Итоговый спектр XANES получаем суммированием приближенных спектров, умноженных на некоторые веса $C_i(T)$, зависящие в случае рассматриваемого комплекса кобальта от температуры T:

$$\mathrm{XANES}(T) = \sum_{i=1}^{n} C_i(T)\mathrm{Approx\ XANES}_i. \tag{3}$$

Выбор точек g_{ij} оказывает существенное влияние на точность построенной аппроксимации. Практика показала, что хорошее качество дает улучшенный метод выбора точек на основе латинского гиперкуба [16] (IHS – Improved Hyper-

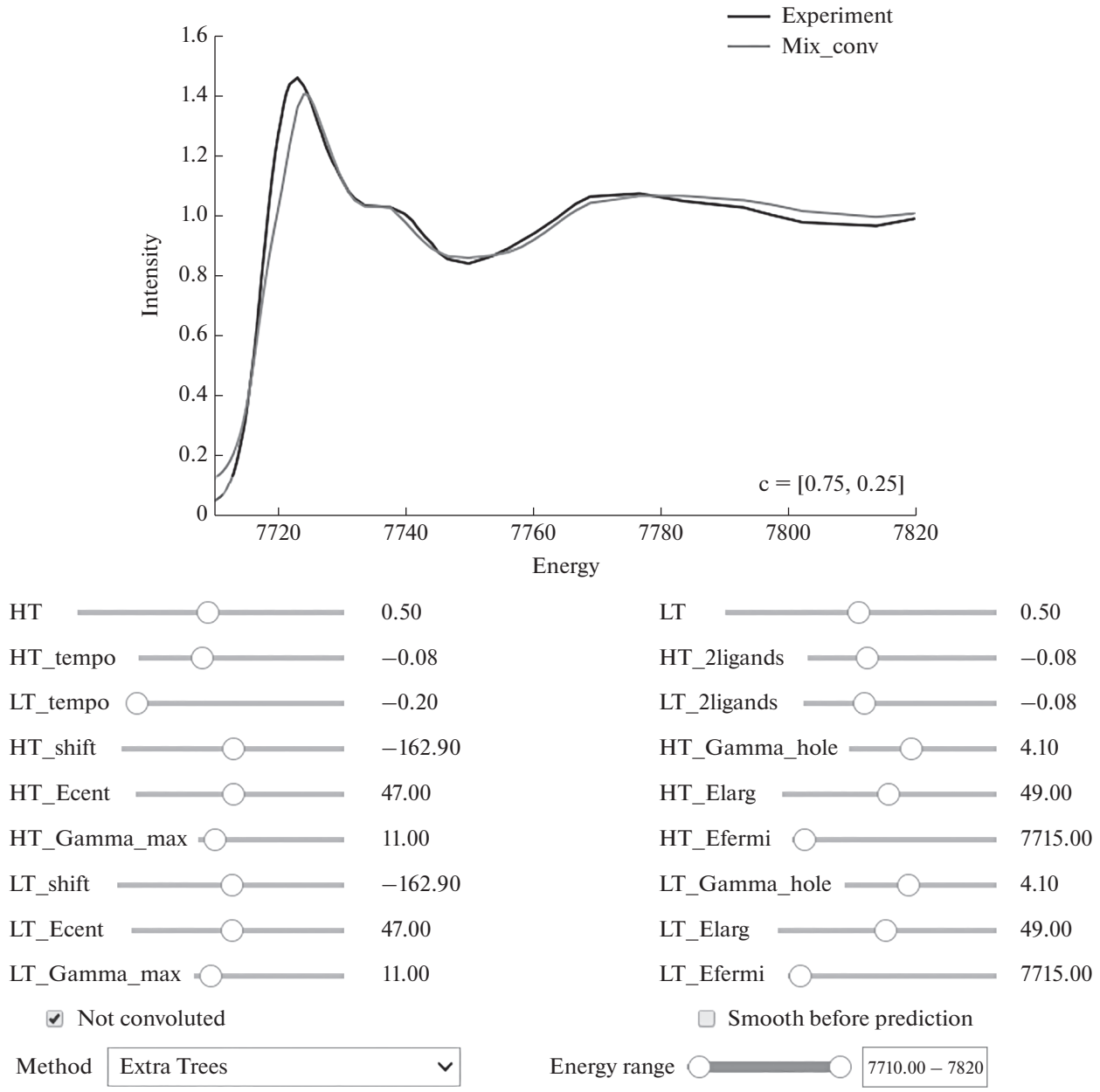

Рис. 1. Интерфейс PyFitIt со слайдерами на платформе Jupyter Notebook.

cube Sampling). В отличие от метода выбора точек в узлах координатной сетки, при IHS-подходе к генерации векторов g получают точки с неповторяющимися координатами и равномерно распределенными в пространстве. Это позволяет лучше аппроксимировать функции нескольких переменных при условии, что от одной или нескольких переменных функция зависит слабо.

В качестве модели аппроксимации допустимо использовать любые методы восстановления регрессии. Представленное на рис. 1 приложение PyFitIt позволяет осуществлять выбор между следующими методами: лес деревьев с повышенной случайностью [17], гребневая регрессия (линейная/квадратичная) [18], радиальные базисные функции [19]. Последний метод является интерполяцией и часто дает самые хорошие результаты, хотя и не всегда.

Итоговая схема аппроксимации спектров и подбора геометрических параметров приведена на рис. 2. Помимо графического интерфейса пользователя PyFitIt предоставляет возможность полностью автоматического подбора геометрических параметров атомных структур компонентов и зависимости их концентраций от температуры. Этот подход оптимизирует функцию $F(g_1, ..., g_n)$ структурных параметров:

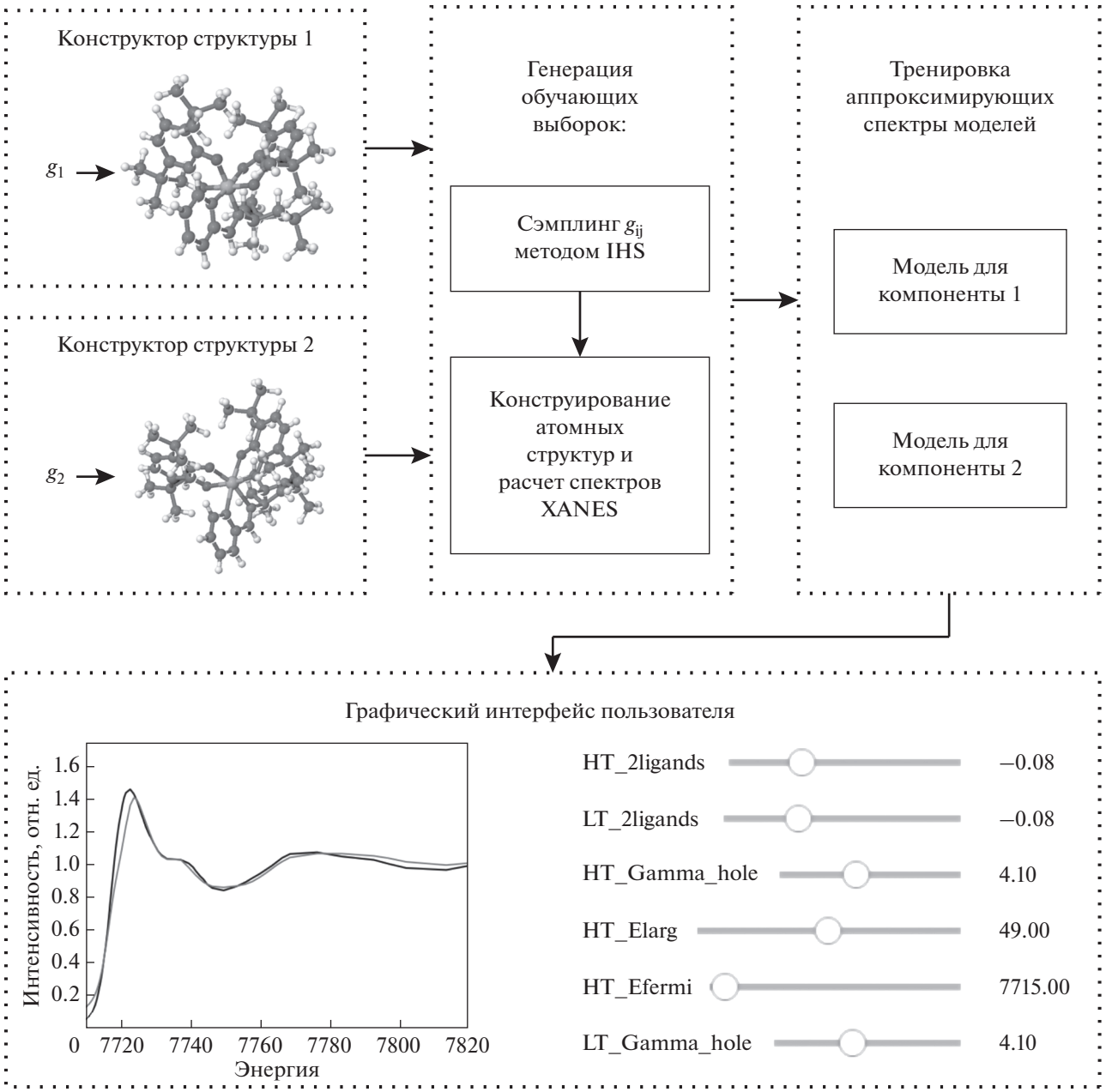

Рис. 2. Схема построения моделей аппроксимации для смеси.

$$F(g_1,\ldots,g_n)=\sum_{T}\min_{\{C_1\ldots C_n\}}\left|\mathrm{TheorXANES}-\sum_{i=1}^{n}C_i\mathrm{ApproxXANES}_i\right|. \quad (4)$$

Так как оптимизация может сойтись к локальному минимуму, то PyFitIt делает несколько попыток запуска минимизации $F(g_1, ..., g_n)$ для различных начальных конфигураций атомной структуры. Получаемые зависимости концентраций от температуры $C_i(T)$ в некоторых случаях могут оказаться негладкими. Для устранения этого недостатка в PyFitIt имеется возможность задать узловые точки по температуре, для которых происходит поиск концентраций $C_i(T)$ и суммирование по T в (4). Концентрации в промежуточных точках вычисляются путем интерполяции сплайном.

Спектры XANES *K*-краев поглощения Co для предполагаемых структур $(diox)_2Co(imPy\text{-}TEMPO)$ при его валентно-таутомерной интерконверсии были рассчитаны с помощью метода конечных разностей полного потенциала, реализованного в коде FDMNES [14, 15]. Была использована конечно-разностная сетка с расстоянием 0.2 Å между соседними точками внутри сферы с радиусом 6 Å вокруг поглощающего атома кобальта. Теоретические спектры были дополнительно сглажены для учета уширения пиков из-за эффекта туннелирования и инструментальных погрешностей (для моделирования энергетической зависимости ширины лоренциана использован арктангенс). Затем был проведена подгонка спектров XANES с помощью программного обеспечения PyFitIt [12]. Начиная с теории функционала плотности оптимизированной структуры была применена вариация двух структурных параметров в комплексе: расстояния между атомами кобальта и азота лиганда imPy-TEMPO и расстояния между атомом кобальта и четырьмя атомами кислорода двух диокс-лигандов. Для каждой точки в двумерном пространстве структурных параметров, сгенерированной алгоритмом IHS, был рассчитан спектр XANES *K*-края поглощения Co. На основе полученной обучающей выборки спектры затем аппроксимировали в каждой точке двумерного пространства структурных параметров методом радиальных базисных функций.

Экспериментальные спектры поглощения кобальта за *K*-краем были измерены на станции структурного материаловедения Курчатовского источника синхротронного излучения при использовании монохроматора Si(111) и геометрии на прохождение.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Для интерполяции спектров рентгеновского поглощения за *K*-краем поглощения кобальта комплекса $(diox)_2Co(imPy\text{-}TEMPO)$ по обучающей выборке наилучшими оказались радиальные базисные функции. Для проверки качества аппроксимации была применена десятиблочная кросс-валидация (перекрестная проверка): обучающая выборка была разбита на десять блоков, каждый из которых по очереди использовали в качестве проверочной выборки, тогда как все остальные – в качестве тренировочной выборки. Средняя относительная погрешность интерполяции спектра в результате оказалась равной 2.3% по отношению к ошибке аппроксимации средним спектром, что является достаточно хорошим показателем. Для наглядности на рис. 3 представлен наихудший случай аппроксимации – теоретический спектр, для которого получена самая большая ошибка аппроксимации при обучении модели по остальным спектрам.

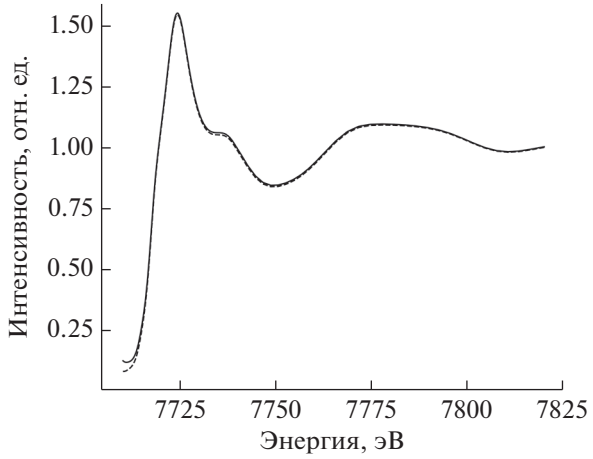

Рис. 3. Качество аппроксимации спектров. Представлен наихудший случай аппроксимации (штриховая линия) теоретического спектра (сплошная линия), получена самая большая ошибка аппроксимации при обучении модели по остальным спектрам.

Подгонка двумя компонентами

Первой попыткой моделирования эксперимента была подгонка двумя компонентами, полученными из одного семейства структур с различными геометрическими параметрами. В результате процесса подгонки под экспериментальные спектры для каждой отдельной температуры был получен график концентраций компонентов (рис. 4). Даже не накладывая искусственных ограничений на концентрации компонентов для крайних температур, получили результат с чистыми веществами для $T = 117$ и 300 K. Структурные параметры компонента при 117 K: среднее расстояние между атомом кобальтом и двумя атомами азота лиганда imPy-TEMPO равно 2.08 Å, (Co–N = = 2.1080, 2.1443 Å при 240 K [13]); среднее расстояние между атомом кобальта и четырьмя атомами кислорода двух диокси-лигандов 2.03 Å (Co–O = = 2.0062–2.0659 Å при 240 K [13]). Для компонента при $T = 300$ K среднее расстояние между атомом кобальта и двумя атомами азота лиганда imPy-TEMPO равно 2.18 Å (Co–N = 2.1226, 2.1586 Å [13]), среднее расстояние между кобальтом и четырьмя атомами кислорода двух диокс-лигандов 2.03 Å (Co–O = 2.0150–2.0747 Å [13]). Спектры двух компонентов, которые использовали при подгонке серии спектров, и соответствующие экспериментальные спектры приведены на рис. 5. Зависимость качества подгонки от температуры представлена на рис. 6 (сплошная линия). Для крайних температур *R*-фактор чуть больше, чем для температур, при которых концентрации обеих компонентов ненулевые. Причиной этого, видимо, является уширение спектра смеси в случае нескольких компонентов.

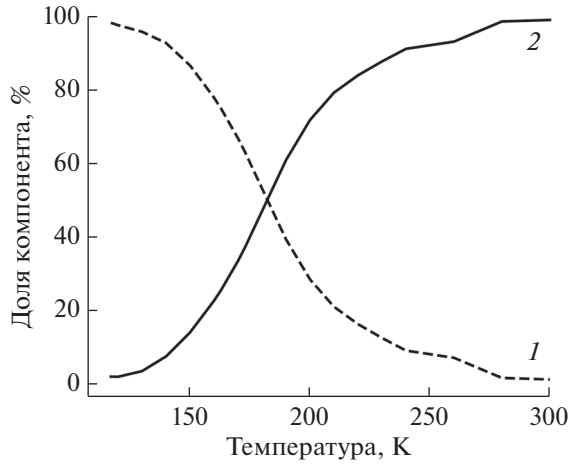

Рис. 4. Зависимость концентраций компонентов от температуры. Результат получен с чистыми веществами для $T = 117$ (*1*) и 300 K (*2*).

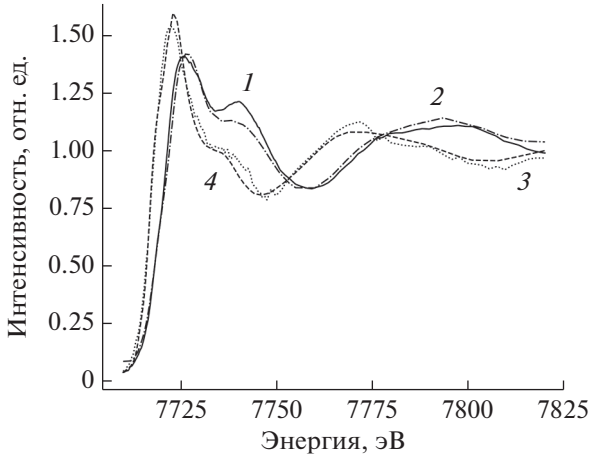

Рис. 5. Экспериментальные (*1*, *3*) и теоретические (*2*, *4*) спектры для крайних температур 117 (*1*, *2*) и 300 K (*3*, *4*).

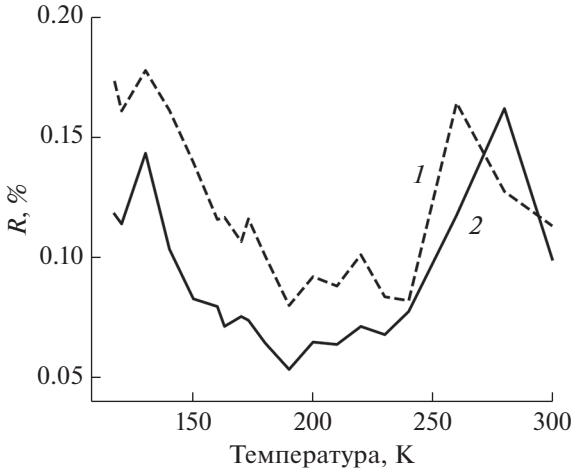

Рис. 6. Качество подгонки спектров с помощью деформации одной структуры (однокомпонентная подгонка) (*1*) и с помощью суперпозиции двух деформируемых структур (*2*) в зависимости от температуры образца.

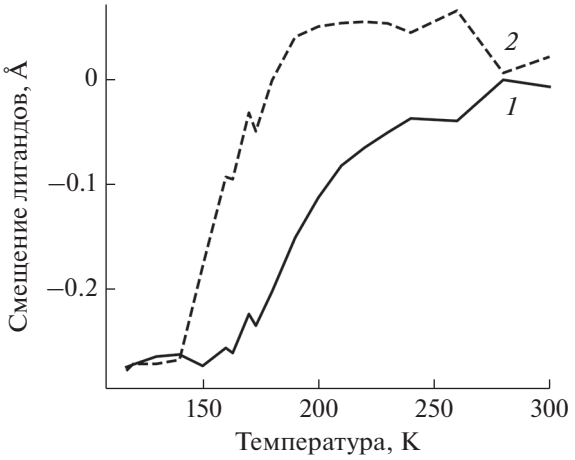

Рис. 7. Температурная зависимость при подгонке одним компонентом смещений лигандов: пары диоксоленовых (*1*) и одной иминопиридина (*2*).

Подгонка одним компонентом

Описываемые в работе методы позволяют реализовать другой тип моделирования происходящих в эксперименте изменений. Будем проводить подгонку экспериментальных спектров одним компонентом, непрерывно изменяющим свою структуру в зависимости от температуры. Результаты такой подгонки представлены на рис. 7. Полученные для крайних температур 117 и 300 K параметры согласуются с параметрами соответствующего компонента в многокомпонентной подгонке. Для промежуточных значений, чтобы воспроизвести наблюдаемые в эксперименте уширения спектральных полос, метод выбирает несимметричные смещения лигандов, что косвенно моделирует сосуществование двух фаз. График зависимости R-фактора от температуры (рис. 6, штриховая линия) похож на график в случае многокомпонентной подгонки, только располагается выше. Это говорит о том, что результаты двухкомпонентной подгонки лучше согласуются с экспериментом.

ЗАКЛЮЧЕНИЕ

В работе описан метод определения структурных параметров компонентов смеси из анализа рентгеновских спектров поглощения методами машинного обучения. На основе программного

комплекса PyFitIt было создано приложение, позволяющее вычислять структурные параметры компонентов смеси по заданному набору экспериментальных спектров. Данный подход является альтернативой уже известным методам – анализу основных компонентов, фактор-анализу, MCR-ALS, так как имеет возможность подбора параметров смеси по единственному экспериментальному спектру. Разработанный метод использован для определения структурных параметров компонентов смеси и вариации их концентраций при температурной валентно-таутомерной интерконверсии в комплексе кобальта $(diox)_2Co(imPy\text{-}TEMPO)$.

БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Совета по грантам Президента Российской Федерации молодых ученых (грант № МК-2730.2019.2).

СПИСОК ЛИТЕРАТУРЫ

1. *Soldatov M.A., Martini A., Bugaev A.L. et al.* // Polyhedron. 2018. V. 155. P. 232.
2. *Frenkel A.I., Kleifeld O., Wasserman S.R., Sagi I.* // J. Chem. Phys. 2002. V. 116. P. 9449.
3. *Piovano A., Agostini G., Frenkel A.I. et al.* // J. Phys. Chem. C. 2011. V. 115. P. 1311.
4. *Fernandezgarcia M., Alvarez C.M., Haller G.L.* // J. Phys. Chem. 1995. V. 99. P. 12565.
5. *Jaumot J., de Juan A., Tauler R.* // Chemom. Intell. Lab. Syst. 2015. V. 140. P. 1.
6. *Jaumot J., Gargallo R., de Juan A., Tauler R.* // Chemom. Intell. Lab. Syst. 2005. V. 76. P. 101.
7. *Della Longa S., Arcovito A., Girasole M. et al.* // Phys. Rev. Lett. 2001. V. 87. P. 155501.
8. *Benfatto M., Congiu-Castellano A., Daniele A., Longa S.D.* // J. Synchr. Rad. 2001. V. 8. P. 267.
9. *Benfatto M., Della Longa S., Natoli C.R.* // J. Synchr. Rad. 2003. V. 10. P. 51.
10. *Hayakawa K., Hatada K., D'Angelo P. et al.* // J. Am. Chem. Soc. 2004. V. 126. P. 15618.
11. *Smolentsev G., Soldatov A.V.* // Comp. Matt. Sci. 2007. V. 39. № 3. P. 569.
12. *Martini A., Guda S.A., Guda A.A. et al.* // Comp. Phys. Commun. 2020. V. 250. P. 107064. https://doi.org/10.1016/j.cpc.2019.107064
13. *Zolotukhin A.A., Bubnov M.P., Arapova A.V. et al.* // Inorg. Chem. 2017. V. 56. P. 14751. https://doi.org/10.1021/acs.inorgchem.7b02597
14. *Bunau O., Joly Y.* // J. Phys.: Condens. Matt. 2009. V. 21. P. 345501.
15. *Guda S.A., Guda A.A., Soldatov M.A. et al.* // J. Chem. Theory Comp. 2015. V. 11. P. 4512.
16. *Beachkofski B.K., Grandhi R.V.* Improved Distributed Hypercube Sampling // 43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conf. Denver, Colorado, 2002. https://doi.org/10.2514/6.2002-1274
17. *Geurts P., Ernst D., Wehenkel L.* // Machine Learning. 2006. V. 63. P. 3.
18. *Тихонов А.В.* // ДАН СССР. 1963. Т. 151. № 3. С. 501.
19. *Fasshauer G.E.* Meshfree Approximation Methods with Matlab. World Scientific, 2007. 520 p. https://doi.org/10.1142/6437

Quantitative Analysis of X-Ray Spectral Data for a Mixture of Compounds Using Machine Learning Algorithms

A. S. Algasov[1, 2, *], S. A. Guda[1, 2], A. A. Guda[1, **], Yu. V. Rusalev[1], A. V. Soldatov[1]

[1] *International Research Institute of Intellectual Materials, Southern Federal University, Rostov-on-Don, 344090 Russia*

[2] *Institute of Mathematics, Mechanics and Computer Science, Southern Federal University, Rostov-on-Don, 344090 Russia*

**e-mail: alexander.algasov@gmail.com*

***e-mail: guda@sfedu.ru*

Based on machine learning algorithms, a method has been developed for determining the structural parameters of mixture components from X-ray absorption spectra. For each component, a database of spectra is constructed for all possible deformations of its structure. The machine learning method implemented in the PyFitIt software package allows quickly calculating the spectrum for deformations of structures from the considered family and optimizing the structural parameters of the mixture by fitting the theoretical spectrum to the experimental one. The capabilities of the method are examined by analyzing changes in the structural characteristics and concentrations of the components of the mixture for the bis-dioxolene complex of cobalt with functionalized iminopyridine ligands during its valence-tautomeric interconversion depending on temperature.

Keywords: mixture component analysis, PyFitIt, machine learning, valence tautomeric interconversion.