ISSN 1027-4510, Journal of Surface Investigation: X-ray, Synchrotron and Neutron Techniques, 2021, Vol. 15, No. 3, pp. 495–501. © Pleiades Publishing, Ltd., 2021.
Russian Text © The Author(s), 2021, published in Poverkhnost', 2021, No. 5, pp. 95–101.

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

A. S. Algasov[a, b, *], S. A. Guda[a], A. A. Guda[a, b, **], Yu. V. Rusalev[a], and A. V. Soldatov[a]

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

Received March 27, 2020; revised May 20, 2020; accepted May 25, 2020

Abstract—Based on machine-learning algorithms, a method is developed for determining the structural parameters of the components of a mixture 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 quick calculation of the spectrum for deformations of structures from the considered family and optimization of 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
DOI: 10.1134/S1027451021030034

INTRODUCTION

There are several approaches to determine the structural parameters of the components of a mixture and their concentrations based on the analysis of the near-threshold fine structure of the X-ray absorption spectra (XANES—X-ray Absorption Near Edge Structure). Historically, the first approach was to simulate a mixture by a linear combination of given spectra and the search for corresponding coefficients [1–3]. When the number of components and their spectra are unknown, Principal Component Analysis (PCA), factor analysis [4] and more recently, the method of multivariate curve analysis by means of root-mean-square variation (MCR-ALS—Multivariate Curve Resolution using Alternating Least Squares) [5, 6] are used.

The problem considered in this work has its own specifics, which cannot be taken into account by the listed methods of analyzing mixtures. In the case when it is most likely known to which family the atomic structures of the mixture components belong, to determine each component, one can use a classical, proven approach: selection of the geometric parameters of the atomic structure so that its calculated (theoretical) spectrum is as close as possible to the experimental one. The first MXAN program to automate this process appeared in 2001 [7–10]. Subsequently, many other programs for calculating spectra have included automatic-parameter-selection functions. In order to calculate the value of the optimized function once, it is necessary to calculate the theoretical spectrum, which takes a lot of time. Automatic optimization often takes longer than a week and sometimes results in physically unlikely or even incorrect structures. Manual intervention in the optimization process is required. To make manual optimization convenient, you need to provide a high speed of spectrum calculation. The FitIt program [11] uses the preliminary calculation of a set of spectra for a set of geometric parameters and their subsequent interpolation. The spectrum-approximation procedure implemented in the FitIt program has been significantly improved by machine-learning methods in the PyFitIt application [12]. PyFitIt functions have also been extended to solve problems of determining structural parameters based on analysis of the X-ray absorption spectra in the case of a mixture of substances. Along with the user interface, in which you can change the atomic structures of the mixture components, achieving the best match with the experimental spectra obtained for different temperatures, PyFitIt also has a built-in procedure for automatic selection of the geometric parameters of the mixture components and their concentrations uniformly over the entire temperature range.

The proposed approach for determining the parameters of mixture components is an alternative to

the existing algorithms. i.e., PCA, factor analysis, MCR-ALS. and can be used in conjunction with them to confirm the results obtained. This approach has several advantages. In particular, only one experimental spectrum is sufficient to analyze a mixture; the algorithm also works well in the case of spectra with high noise.

In this work, the advantages of the machine-learning method are demonstrated by the example of a study of valence-tautomeric interconversion in a cobalt complex $(diox)_2Co(imPy\text{-}TEMPO)$. Earlier [13], it was shown that this bis-dioxolene cobalt complex, including the functionalized imopyridine ligand, undergoes a spin transition in the temperature range of 200–300 K in the solid state. According to magnetic-susceptibility data, this interconversion is most likely caused by valence tautomerism. However, all attempts to study the structural changes associated with the observed transformation by the method of the X-ray diffraction of single crystals have not been successful, since the crystal is destroyed upon cooling below 220 K, and the features of the molecular structure of the low-temperature isomer remain unknown. The XANES spectra analysis method for the *K* absorption edge of Co for $(diox)_2Co(imPy\text{-}TEMPO)$ in a wide temperature range from 30 to 300 K using machine-learning algorithms makes it possible to determine the parameters of the local atomic structure of various isomers and their concentration depending on temperature.

EXPERIMENTAL

Following the approach implemented in the PyFitIt application [12], machine-learning models are built for each component of the mixture, which allow quick calculation of the spectrum for a given geometric structure. In an interactive application, the user can use sliders to change the geometric parameters of the structures of the mixture components and immediately see the resulting spectrum (Fig. 1), which makes it possible to conveniently adjust the theoretical spectrum to the experimental one in the manual mode or launch the automatic optimizer.

For this approach to work, the user of the program needs for each *i*th component of the mixture to determine the vector of the geometric parameters g_i and compose the function M_i, which creates an atomic structure for a given vector g_i, $i = 1, ..., n$; n is the number of components. Displacements of groups of atoms in some directions, rotations of parts of the structure, displacements of individual atoms or changes in the coordinates of all atoms of the structure can act as vector coordinates g_i. It should be borne in mind that a large number of geometric parameters will take a long time to build the training sample.

The obtained parameters of the atomic structure of the model compounds are then entered into a program for calculating the X-ray absorption spectra, for example, FDMNES [14, 15]. We denote the procedure for calculating the spectrum from a given atomic structure by S. The exact calculation of the spectrum is time-consuming and cannot be used interactively. Therefore, we use the machine-learning approximation. We construct a training sample by calculating the XANES spectra *i*th component for a set of vector values g_{ij}, $j = 1, ..., m_i$:

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

The calculated database of spectra makes it possible to construct an approximation A_i of the superposition $S \circ M_i$, to quickly obtain an approximation of the spectrum for a given arbitrary vector of geometric parameters g_i:

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

The final XANES spectrum is obtained by summing the approximate spectra multiplied by some weights $C_i(T)$, depending on the case of the considered cobalt complex on temperature T:

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

Point selection g_{ij} has a significant effect on the accuracy of the constructed approximation. Practice has shown that good quality results can be obtained from an improved method for selecting points based on the Latin hypercube [16] (IHS—Improved Hypercube Sampling). In contrast to the method for selecting points at the nodes of the coordinate grid, with the IHS approach to generating vectors g points with non-repeating coordinates which are evenly distributed in space are obtained. This allows a better approximation of the functions of several variables, provided that the function depends weakly on one or more variables.

Any regression reconstruction methods can be used as an approximation model. As shown in Fig. 1, PyFitIt allows a choice between the following methods: forest of trees with increased randomness [17], ridge regression (linear/quadratic) [18], and radial basis functions [19]. The latter method is interpolation and often gives the best results, although not always.

The final scheme for the approximation of spectra and the selection of geometric parameters is shown in Fig. 2. In addition to the graphical user interface, PyFitIt provides the ability to fully automatically select the geometric parameters of the atomic structures of components and the dependence of their concentrations on temperature. This approach optimizes the function $F(g_1, ..., g_n)$ structural parameters:

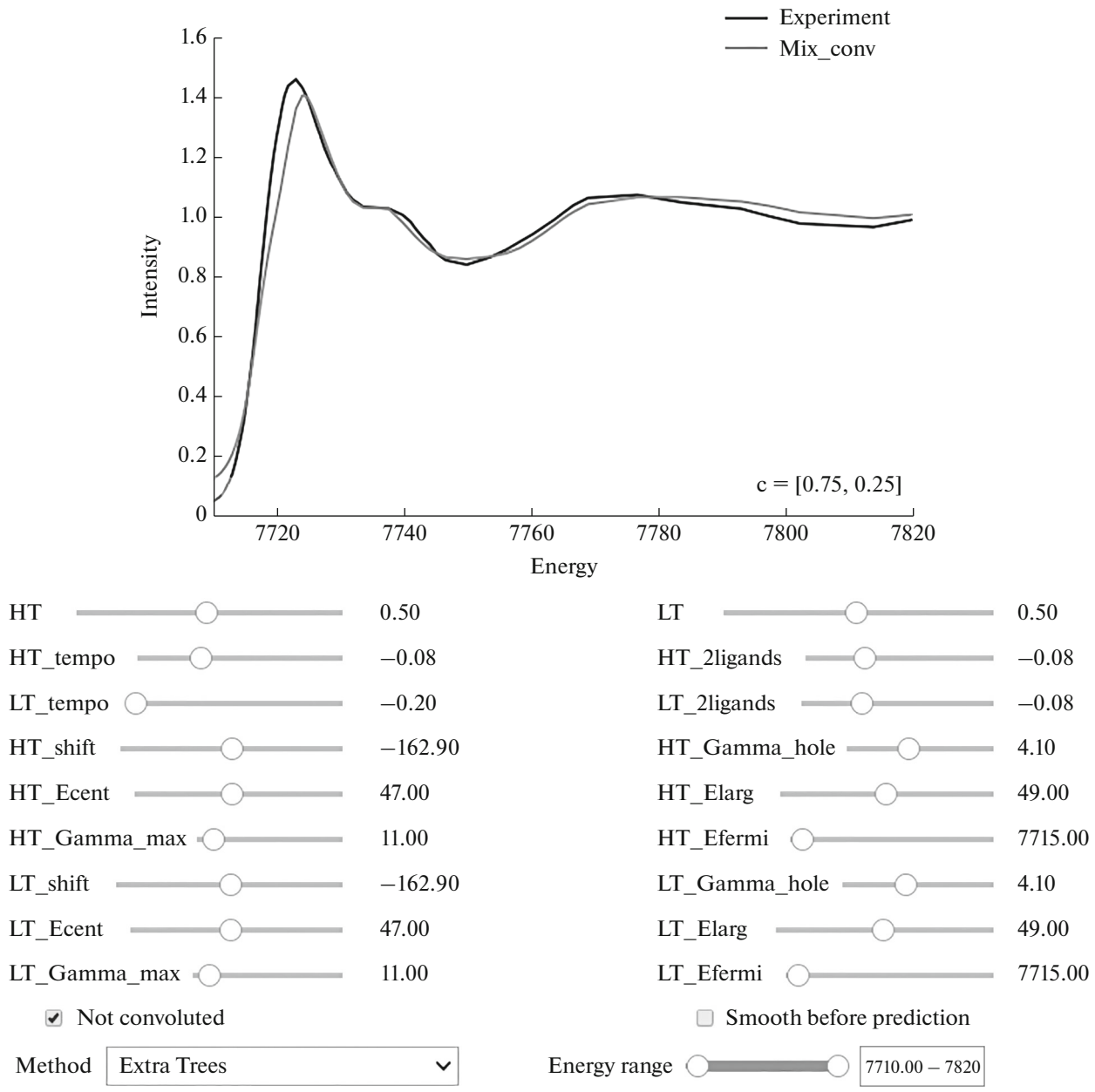

Fig. 1. PyFitIt interface with sliders based on Jupyter Notebook framework.

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

Since optimization can converge to a local minimum, PyFitIt makes several attempts to carry out minimization $F(g_1, ..., g_n)$ for various initial configurations of the atomic structure. The resulting dependences of the concentrations on temperature $C_i(T)$ in some cases may turn out to be nonsmooth. To eliminate this drawback, PyFitIt has the ability to set the temperature nodes for which the concentration is sought $C_i(T)$ and summation over T in (4). The concentrations at intermediate points are calculated by spline interpolation.

The XANES Spectra of the K-absorption edges of Co for the calculated structures $(diox)_2Co(imPy\text{-}TEMPO)$ at valence-tautomeric interconversion were calculated using the method of the finite difference of the total potential implemented in FDMNES code

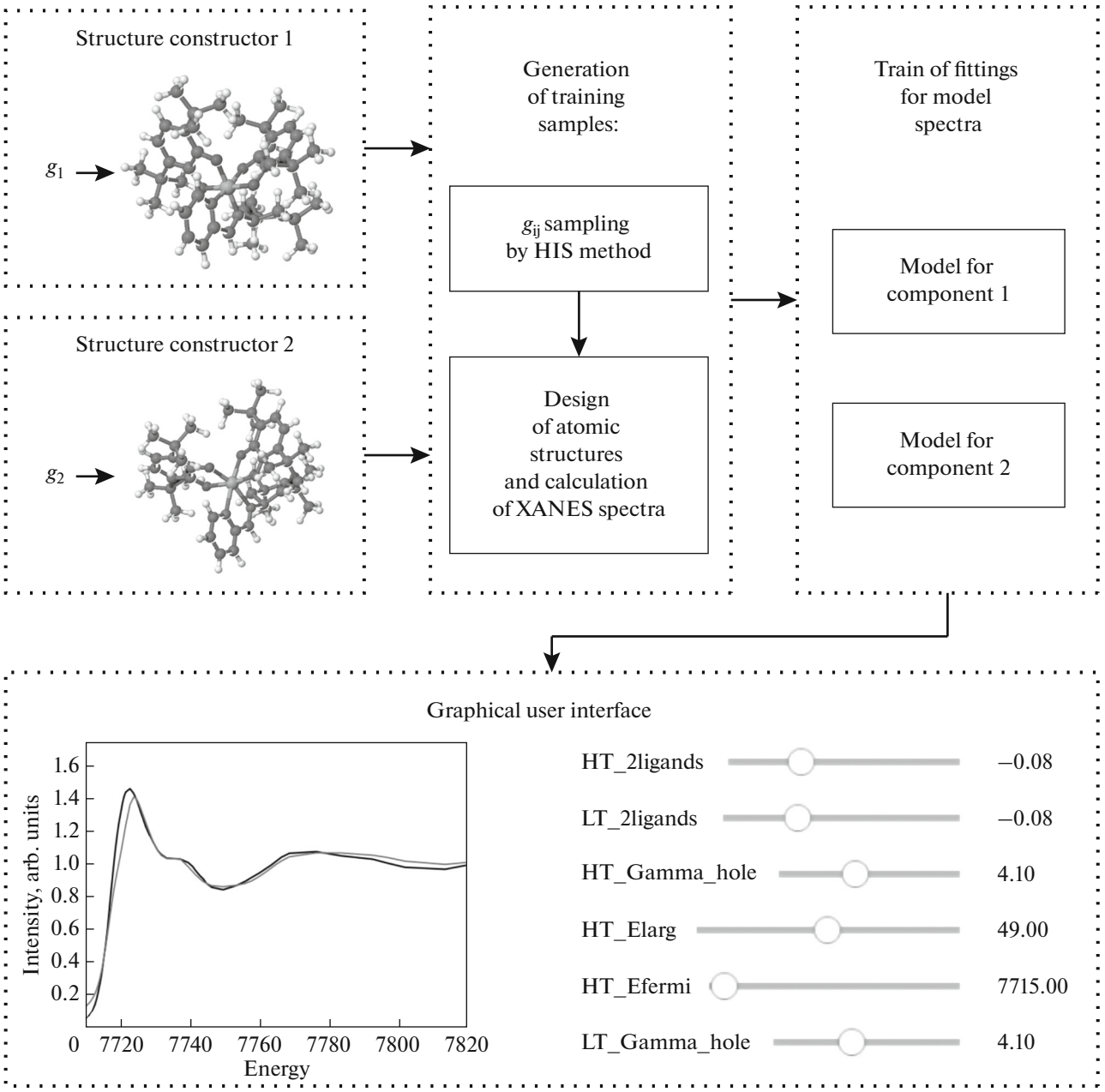

Fig. 2. Scheme for constructing approximation models for a mixture.

[14, 15]. We used a finite-difference grid with a distance of 0.2 Å between adjacent points inside a sphere with a radius of 6 Å around the absorbing cobalt atom. The theoretical spectra were additionally smoothed to take into account the broadening of the peaks due to the tunneling effect and instrumental errors (the arctangent was used to model the energy dependence of the Lorentzian width). Then the XANES spectra were fitted using PyFitIt software [12]. Starting from density functional theory of the optimized structure, a variation of two structural parameters in the complex was applied: the distance between the cobalt and nitrogen atoms of the imPy-TEMPO ligand and the distance between the cobalt atom and four oxygen atoms of the two diox-ligands. For each point in two-dimensional space of the structural parameters generated by the IHS algorithm, the XANES spectrum of the *K*-absorption edge of Co was calculated. Based on the obtained training sample, the spectra were then approximated at each point of two-dimensional space of the structural parameters by the method of radial basis functions.

The experimental absorption spectra of cobalt for the *K* edges were measured at the structural materials science station of the Kurchatov synchrotron radiation

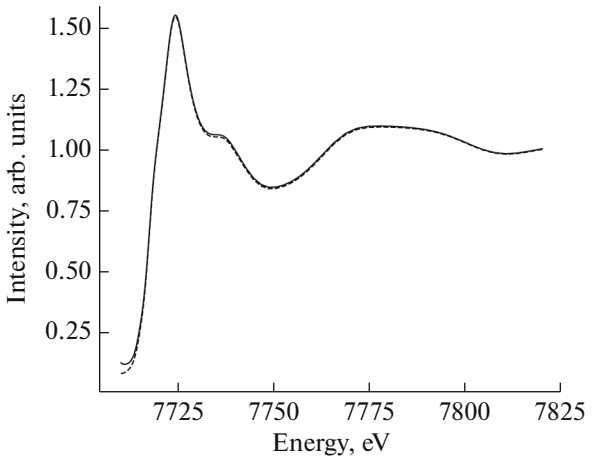

Fig. 3. Quality of spectrum approximation. The worst case of approximation (dashed line) of the theoretical spectrum (solid line) is presented, the largest approximation error is obtained when training the model for the remaining spectra.

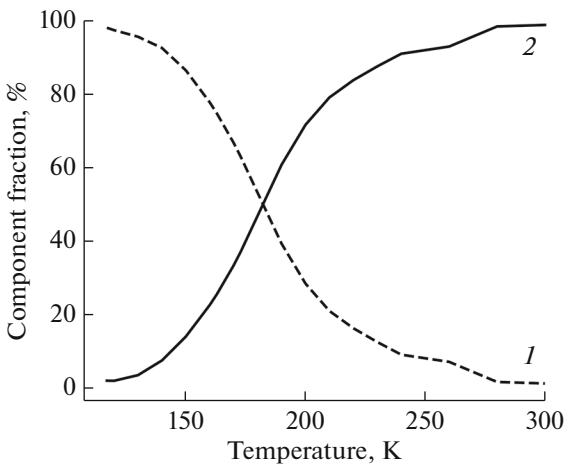

Fig. 4. Dependence of the component concentrations on temperature. The result is obtained with pure substances for $T = 117$ (*1*) and 300 K (*2*).

source using a Si(111) monochromator and transmission geometry.

RESULTS AND DISCUSSION

To interpolate the X-ray absorption spectra for the *K*-absorption edge of cobalt of the $(diox)_2Co(imPy\text{-}TEMPO)$ complex in the training sample, the best turned out to be the radial basis functions. To check the quality of the approximation, ten-block cross-validation was applied: the training sample was divided into ten blocks, each of which was used in turn as a test sample, while all the others were used as the training sample. As a result, the average relative error in the interpolation of the spectrum turned out to be 2.3% with respect to the error of approximation by the average spectrum, which is a fairly good indicator. For clarity, Fig. 3 shows the worst case of approximation: the theoretical spectrum, for which the largest approximation error was obtained when training the model from with the remaining spectra.

Two-component fit

The first attempt at modeling the experiment was a fit with two components obtained from the same family of structures with different geometric parameters. As a result of the process of fitting the experimental spectra for each individual temperature, a graph of the concentrations of the components was obtained (Fig. 4). Even without imposing artificial restrictions on the concentration of components for extreme temperatures, we obtained a result with pure substances for $T = 117$ and 300 K. The structural parameters of the component at 117 K: the average distance between the cobalt atom and two nitrogen atoms of the imPy-TEMPO ligand is 2.08 Å, (Co–N = 2.1080, 2.1443 Å at 240 K [13]); the average distance between the cobalt atom and four oxygen atoms of two dioxy ligands is 2.03 Å (Co–O = 2.0062–2.0659 Å at 240 K [13]). For the component at $T = 300$ K the average distance between the cobalt atom and two nitrogen atoms of the imPy-TEMPO ligand is 2.18 Å (Co–N = 2.1226, 2.1586 Å [13]), the average distance between cobalt and four oxygen atoms of two diox ligands is 2.03 Å (Co–O = 2.0150–2.0747 Å [13]). The spectra of the two components, which were used to fit the series of spectra, and the corresponding experimental spectra are shown in Fig. 5. The temperature dependence of the quality of fit is shown in Fig. 6 (solid line). For extreme temperatures, the *R* factor is slightly larger than for the temperatures at which the concentrations of both components are nonzero. The reason for this, apparently, is the broadening of the spectrum of the mixture in the case of several components.

One Piece Fit

The methods described in the work allow the implementation of another type of modeling of the changes occurring in the experiment. We will fit the experimental spectra with one component that continuously changes its structure depending on temperature. The results of this fit are shown in Fig. 7. The parameters obtained for the extreme temperatures of 117 and 300 K agree with the parameters of the corresponding component in a multicomponent fit. For intermediate values, in order to reproduce the experimentally observed broadening of the spectral bands, the method chooses asymmetric ligand displacements, which indirectly simulates the coexistence of two phases. The graph of the dependence of the *R* factor versus temperature (Fig. 6, dashed line) is similar to the graph in the case of multicomponent fitting,

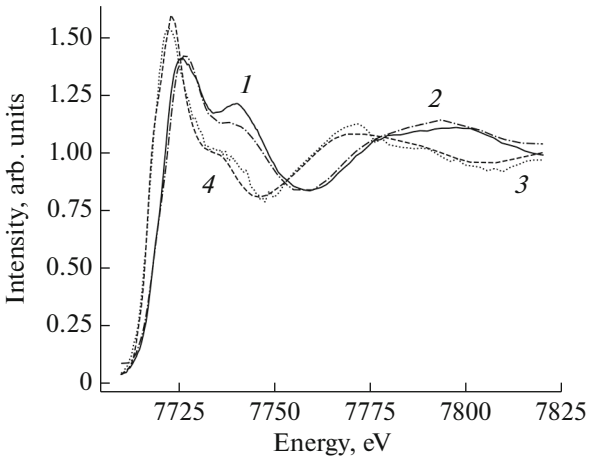

Fig. 5. Experimental (*1*, *3*) and theoretical (*2*, *4*) spectra for extreme temperatures 117 (*1*, *2*) and 300 K (*3*, *4*).

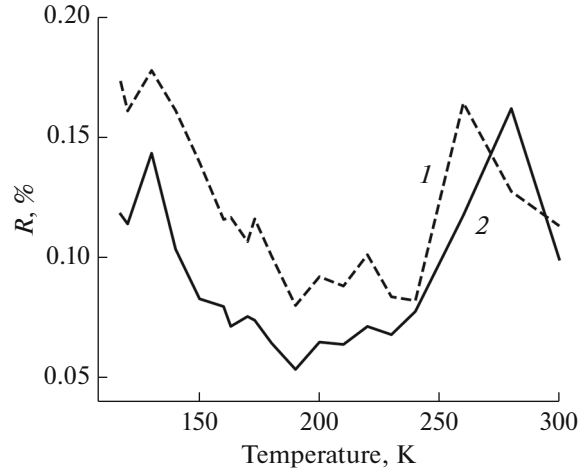

Fig. 6. Quality of spectrum fit via the deformation of one structure (one-component fit) (*1*) and using the superposition of two deformable structures (*2*) depending on the sample temperature.

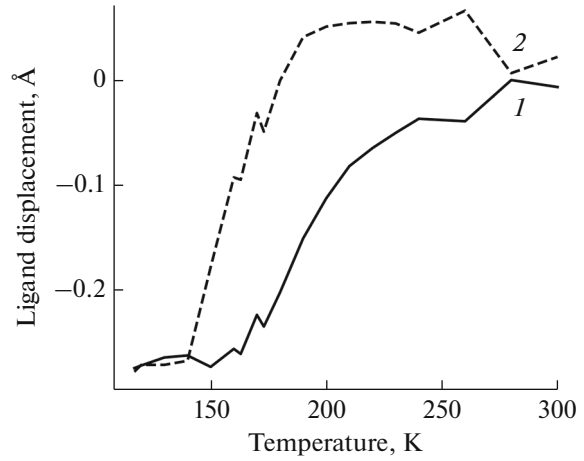

Fig. 7. Temperature dependence upon fitting with one component of ligand displacements: dioxolene pairs (*1*) and one imopyridine (*2*).

only it is located higher. This suggests that the results of two-component fit are in better agreement with experiment.

CONCLUSIONS

The paper describes a method for determining the structural parameters of the mixture components from the analysis of X-ray absorption spectra using machine-learning methods. Based on the PyFitIt software package, an application was created that allows one to calculate the structural parameters of mixture components from a given set of experimental spectra. This approach is an alternative to already known methods: analysis of the main components, factor analysis, MCR-ALS, since it there is the ability to select the parameters of the mixture from a single experimental spectrum. The developed method was used to determine the structural parameters of the mixture components and the variation in their concentrations during the temperature valence-tautomeric interconversion in the cobalt complex $(diox)_2Co(imPy\text{-}TEMPO)$.

FUNDING

This work was supported by the Council for Grants of the President of the Russian Federation for Young Scientists (Grant no. MK-2730.2019.2).

REFERENCES

1. M. A. Soldatov, A. Martini, A. L. Bugaev, et al., Polyhedron **155**, 232 (2018).
2. A. I. Frenkel, O. Kleifeld, S. R. Wasserman, and I. Sagi, J. Chem. Phys. **116**, 9449 (2002).
3. A. Piovano, G. Agostini, A. I. Frenkel, et al., J. Phys. Chem. C **115**, 1311 (2011).
4. M. Fernández-García, C. Márquez-Alvarez, and G. L. Haller, J. Phys. Chem. **99**, 12565 (1995).
5. J. Jaumot, A. de Juan, and R. Tauler, Chemom. Intell. Lab. Syst. **140**, 1 (2015).
6. J. Jaumot, R. Gargallo, A. de Juan, and R. Tauler, Chemom. Intell. Lab. Syst. **76**, 101 (2005).
7. S. Della Longa, A. Arcovito, M. Girasole, et al., Phys. Rev. Lett. **87**, 155501 (2001).
8. M. Benfatto, A. Congiu-Castellano, A. Daniele, and S. Della Longa, J. Synchrotron Radiat. **8**, 267 (2001).
9. M. Benfatto, LongaS. Della, and C. R. Natoli, J. Synchrotron Radiat. **10**, 51 (2003).
10. K. Hayakawa, K. Hatada, P. D'Angelo, et al., J. Am. Chem. Soc. **126**, 15618 (2004).
11. G. Smolentsev and A. V. Soldatov, Comput. Mat. Sci. **39**, 569 (2007).
12. A. Martini, S. A. Guda, A. A. Guda, et al., Comput. Phys. Commun. **250**, 107064 (2020). https://doi.org/10.1016/j.cpc.2019.107064

13. A. A. Zolotukhin, M. P. Bubnov, A. V. Arapova, et al., Inorg. Chem. **56**, 14751 (2017). https://doi.org/10.1021/acs.inorgchem.7b02597
14. O. Bunau and Y. Joly, J. Phys.: Condens. Matter **21**, 345501 (2009).
15. S. A. Guda, A. A. Guda, M. A. Soldatov, et al., J. Chem. Theory Comput. **11**, 4512 (2015).
16. B. K. Beachkofski and R. V. Grandhi, "Improved Distributed Hypercube Sampling," in *Proceedings of the 43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, Denver, 2002*. https://doi.org/10.2514/6.2002-1274
17. P. Geurts, D. Ernst, and L. Wehenkel, Mach. Learn. **63**, 3 (2006).
18. A. V. Tikhonov, Dokl. Akad. Nauk SSSR **151** (3), 501 (1963).
19. G. E. Fasshauer, *Meshfree Approximation Methods with Matlab* (World Sci., Singapore, 2007). https://doi.org/10.1142/6437