ISSN 1027-4510, Journal of Surface Investigation: X-ray, Synchrotron and Neutron Techniques, 2021, Vol. 15, No. 5, pp. 934–938. © Pleiades Publishing, Ltd., 2021.
Russian Text © The Author(s), 2021, published in Poverkhnost', 2021, No. 9, pp. 49–54.

Search for Analytical Relations between X-Ray Absorption Spectra Descriptors and the Local Atomic Structure Using Machine Learning

S. A. Guda[a, b], A. S. Algasov[b], A. A. Guda[a, *], A. Martini[a], A. N. Kravtsova[a], A. L. Bugaev[a], L. V. Guda[a], and A. V. Soldatov[a]

[a] *Smart Material Research Institute, Southern Federal University, Rostov-on-Don, 344090 Russia*
[b] *Vorovich Institute of Mathematics, Mechanics and Computer Science, Southern Federal University, Rostov-on-Don, 344090 Russia*
**e-mail: guda@sfedu.ru*

Received January 19, 2021; revised February 18, 2021; accepted February 25, 2021

Abstract—In this paper, we develop a new technique for quantitative analysis of the near region of X-ray absorption spectra that is based on the extraction of spectrum descriptors and machine learning. The use of descriptors (edge position, intensity and curvature of minima and maxima, and tangent of the slope of the absorption edge) allows solution of the problem of systematic differences between theoretical calculations and experimental data, reducing the dimension of the problem and thereby improving the accuracy of machine-learning algorithms. We obtain analytical relations between the spectrum descriptors and the parameters of the local atomic structure of a substance, which extend the range of applicability of the empirical Natoli rule and analysis of the chemical shift of spectra to arbitrary classes of chemical compounds.

Keywords: spectrum descriptors, machine learning, Natoli rule, X-ray absorption spectroscopy
DOI: 10.1134/S1027451021050050

INTRODUCTION

X-ray absorption spectroscopy is an effective method for studying the local atomic and electronic structure around an absorbing atom [1, 2]. The energy region of incident photons in the 200-eV range beyond the absorption edge contains information about the structure descriptors: distances in the first coordination sphere, bond angles, the type of nearest neighbors, and the oxidation state. The structure descriptors affect the spectrum descriptors: the position of the absorption edge, the height of the white line, the position of the minima and maxima, and the splitting of the peaks. A spectroscopist can easily distinguish the absorption spectra of 3*d* and 4*d* metal oxides by their shape and also see the characteristic features of close-packed metal lattices. At the formation stage of theoretical models for calculating the absorption spectra, the semi-empirical Natoli rule [3], which relates the positions of the maxima in the absorption spectrum with the interatomic distances, was discovered. The chemical-shift rule, on the contrary, relates the oxidation state of the metal atom to the shift of the absorption edge [4].

Over the past few years, machine-learning algorithms for quantitative spectral analysis have been actively developed. For example, a random forest model trained at all points of the spectrum to classify the symmetry of the local environment of 3*d* metal atoms was used in [5]. A convolutional neural network was used to estimate the coordination numbers in the first coordination sphere of copper atoms of copper-oxide clusters [6] and to predict the first three coordination numbers of platinum nanoparticles in order to determine their shape and size [7]. A deep neural network can predict the functions of the radial distribution of atoms by the X-ray absorption spectrum and, conversely, the spectrum by the given structure descriptors [8]. To optimize the operation of machine-learning algorithms, the dimension of the data supplied to the input is reduced. For example, a Coulomb matrix can be constructed from $3N$ atomic coordinates for N atoms, and generalized radial or angular distribution functions can be calculated, taking into account the atomic masses [9, 10]. By analyzing the main components and the values of the norm, the structural parameters can be sorted according to their effect on the shape of the X-ray absorption near-edge structure (XANES) spectra [11]. Hundreds of absorption-coefficient values measured with a small energy step in one spectrum can be reduced to several descriptors. This approach is often used to analyze the pre-edge region of the spectra. The center of mass of the leading edge and its area are calculated after subtracting the background, which were studied in [12] to analyze the oxidation state and coordination numbers. In [13], it was demonstrated that the projections onto

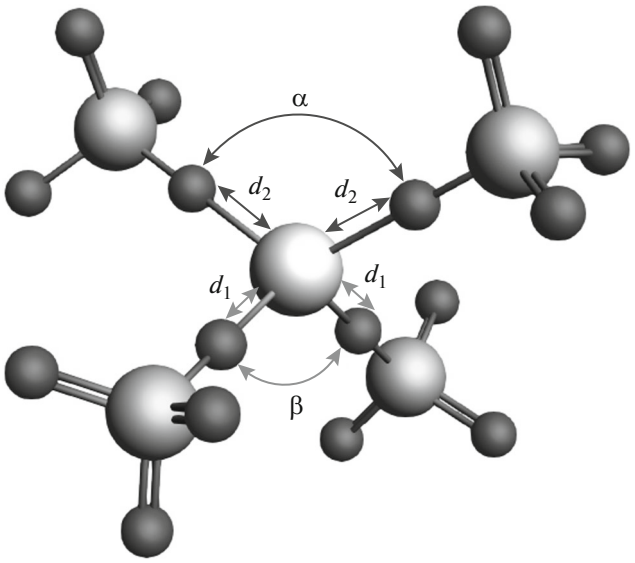

Fig. 1. Cluster $Fe(SiO_4)_N$ for $N = 4$ and variable structural parameters used to calculate the training sample.

the principal components of the training set of theoretical spectra can be used to classify the four-, five-, and six-coordinated environments of a 3*d* metal atom, as well as the type of functional groups for light atoms [14]. The construction of descriptors based on the approximation of parts of the spectrum using polynomials of various orders was recently demonstrated in [15].

This study is aimed at expanding the methodology for calculating the spectrum descriptors for training machine-learning algorithms and then obtaining analytical relations between spectrum descriptors and structural parameters.

EXPERIMENTAL

The X-ray absorption spectra were calculated using the finite difference and total potential method [16] implemented in FDMNES software [17]. The photoelectron wave function was calculated on a three-dimensional grid of points in a sphere with a radius of 6 Å around an absorbing atom. The distance between grid points was 0.25 Å. To take into account the finite lifetime of the photoelectron, the theoretical spectra were processed using the convolution operation for line broadening. The dependence of the width of the Lorentzian kernel for convolution was approximated using the arctangent function.

Figure 1 shows an example of a four-coordinated cluster for calculating the X-ray absorption spectrum behind the *K*-edge of iron (central atom). Modeling was performed for structures of $Fe(SiO_4)_N$ silicates with coordination numbers (CN) of $N = 2$–6. For each coordination number, the interatomic distances d_1 and d_2 were varied in the range of 1.8–2.2 Å and the bond angles O–Fe–O α, β were varied in the range of 70°–110°. In the space of structural parameters, points for calculating the spectra were selected by the improved Latin hypercube (IHS) method in the amount of 700 points. 3500 X-ray absorption spectra, which were used to train a machine-learning algorithm based on radial basis functions with a linear kernel, were calculated for all CNs. Thus, in terms of the machine-learning domain, the set of calculated spectra constituted a training sample.

RESULTS AND DISCUSSION

The calculated spectra for CN = 4 are shown in Fig. 2. A change in the interatomic distances leads both to a shift in the absorption edge and to a change in the positions of the minima and maxima in the spectrum. Typically, a theoretical XANES spectrum contains about 100 energy points. A common approach to increasing the efficiency of machine-learning algorithms is to reduce the dimension of such an object by extracting only informative features, i.e., corresponding spectral descriptors [15]. Figure 3 shows a set of descriptors calculated for each individual spectrum: position of the absorption edge, position and intensity of the main maximum, position and intensity of the main minimum, and curvature of the minima and maxima. To estimate the position of the absorption edge, the entire spectrum was approximated by the arctan function, the parameters of the arctangensoid were also used as descriptors: the position of the center of the arctangensoid and the slope in the center. For stable calculation of the descriptors, additional "blurring" of the absorption spectra by 5 eV was carried out before calculating the curvature and fitting by the arctangensoid.

In the early 1980s, Natoli formulated an empirical rule [3], which established the relation between the positions of the peaks in the XANES spectrum and interatomic distances for structures with similar symmetry. This rule is applicable, for example, to metals whose crystal lattices belong to the same space group (for example, body-centered cubic (BCC) Nb and Mo) or to structures that undergo volumetric expansion, for example, palladium upon hydrogen absorption [1, 18, 19]. Another example of the empirical analysis of X-ray absorption spectra is shown in [20]. The authors derived the analytical relation between the positions of the maxima in the absorption spectra behind the L_3 edge of uranyl complexes and the distances between uranium and oxygen atoms in the first coordination sphere. The considered examples are based on a limited number of spectrum descriptors and cannot be extended to a larger number of structural parameters. The question about the range of applicability of the technique for other metal complexes also remains open. In this paper, we describe a methodology for obtaining analytical relations between any set of spectral descriptors and structural parameters using a machine-learning algorithm. In

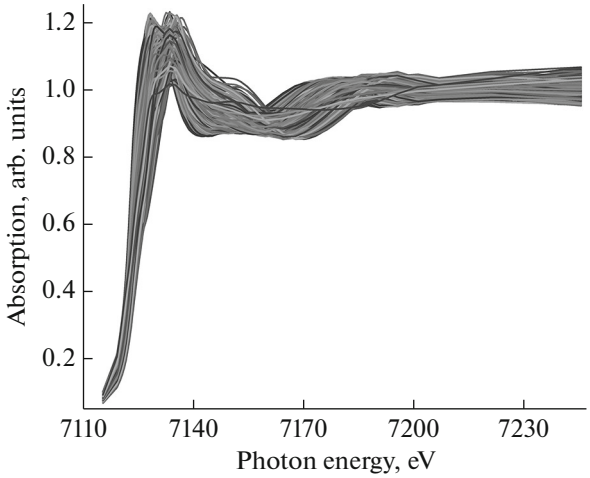

Fig. 2. Theoretical XANES spectra calculated using the finite difference method beyond the Fe *K* edge for the $Fe(SiO_4)_4$ cluster.

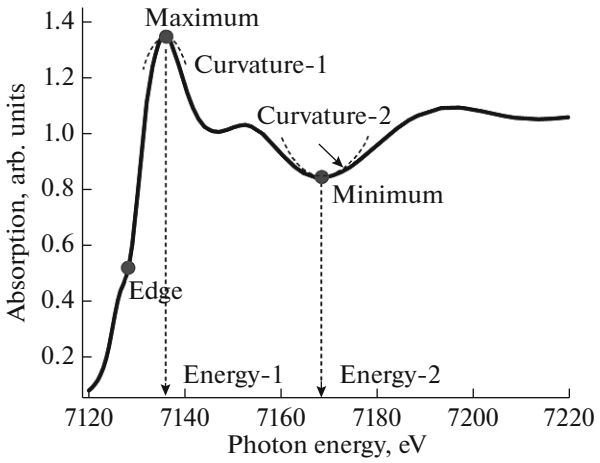

Fig. 3. Set of descriptors used for XANES analysis of the Fe *K* edge of the $Fe(SiO_4)_4$ cluster. The dots mark the parts of the spectrum, in which the descriptors are calculated.

general, the analytical relation between the known parameters $x_1...x_n$ and the target variable y is determined using linear regression:

$$y = a_1x_1 + a_2x_2 + ... + a_nx_n. \quad (1)$$

More complex relations are based on higher-order polynomials with a cross product of the parameters $x_1...x_n$. We are interested in simple analytical solutions with good approximation quality. The target criterion in this case is the absence of large a_i coefficients and the minimal possible number of nonzero coefficients. For the problem of integer relations, such optimization is achieved by using special orthogonalization algorithms (for example, [21] or [22]). In the case of rational coefficients, we use the properties of the elastic network algorithm [23] in combination with semiempirical reasoning. To construct analytical dependences, we restrict ourselves to a polynomial of the second degree with parameters $x_1...x_n$ and their pairwise products:

$$y = \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j. \quad (2)$$

At the first step, the values of all descriptors were normalized so that the mean over the sample was zero and the standard deviations of the values for all sets of descriptors were the same. Only under this condition is it possible to compare values with different dimensions, for example, the position of the absorption edges and the curvature of the minima and maxima. For the approximation, the elastic network method was used, which includes the LASSO algorithm [24] and ridge regression. In the case of a group of strongly correlated variables, the LASSO algorithm tends to select one variable from the group and ignore others, i.e., the rational selection of features is performed. If the linear formula returned by the elastic network is too complex, one can try to simplify it by making the model less accurate. To do this, we sort the coefficients (w_i, w_{ij}) returned by the elastic network algorithm in terms of their absolute values and try to construct a linear model based on the subsets of functions with the largest absolute coefficients. Analysis was performed for subsets of each size: 1, 2, 3, ..., and, for all of them, the value of the approximation quality was estimated. Table 1 shows the selected analytical relations between the descriptors of the spectra and the structural parameters of noncrystalline compounds of iron silicates.

Analytical relations between descriptors can be obtained for any number of spectral characteristics and structural parameters. Although, in general, machine-learning algorithms operate as a black box, Table 1 provides a visual interpretation of the best descriptor combinations. For example, a prediction quality of up to 93% can be achieved for interatomic distances if the energy positions of the edge and the first maximum and minimum are taken into account. It is important to note that, unlike the Natoli empirical rule, the accuracy estimate given in this study applies to all structures with a coordination number of CN = 2–6 and large variations in the structural parameters.

The intensity of the main maximum changes with shortening of the Fe–O bond length, so this descriptor can help distinguish shifts associated with the oxidation state or changes in the volume. The formulas for the CN depend on the curvature of the main maximum, which is consistent with the general behavior of EXAFS oscillations, the amplitude of which is proportional to the CN. As the dependences in Table 1 show, the position of the absorption edge should also be used for a reliable estimate of the CN.

The change in the oxidation state of the iron atom was also modeled by applying an energy shift to each

Table 1. Analytical relations between the descriptors of the spectra and the structural parameters of the compounds of noncrystalline iron silicates (the parameter R^2 score was chosen for estimation of the quality of approximation)

No.	Descriptor	Analytical formula	Approximation quality
		Structure descriptors	
1	CN(N)	$0.59\text{Edge}_E - 0.56\text{Max}_{\text{curv}}$	0.85
2		$0.58\text{Edge}_E - 0.43\text{Max}_{\text{curv}} + 0.22\text{Max}_{\text{int}}$	0.88
3	Average distance Fe–O, (Dist)	$-0.77(\text{Min}_E - \text{Max}_E) - 0.40\text{Max}_{\text{curv}}$	0.88
4		$-0.29\text{Max}_E + 0.54\text{Edge}_E - 0.94\text{Min}_E$	0.93
5	Standard deviation Dist from mean (Dev)	$+0.99\text{Edge}_E - 0.19\text{Edge}_{\text{slope}} - 1.12\text{Max}_{\text{curv}} - 0.42\text{Max}_E + 0.38\text{Max}_{\text{int}} - 0.94\text{Max} - \text{Min}_{\text{slope}} + 0.37\text{Min}_{\text{curv}} + 1.49\text{Min}_{\text{int}}$	0.66
		Spectrum descriptors	
1	$\text{Edge}_{\text{slope}}$	$-0.26N + 0.36N^2 + 0.80\text{Dist} - 0.36$	0.84
2	Max_{curv}	$-0.51\text{Dist} - 0.69N$	0.78
3		$-0.82N - 0.55\text{Dist} + 0.35\text{Dev}$	0.88
4	Min_E	$0.34N - 0.90\text{Dist}$	0.89
5	Min_{int}	$-0.96N + 0.63\text{Dev}$	0.84

The notation "Dist" is used for the average Fe–O distance in the first coordination shell. "Dev" is used for the standard deviation of the Fe–O distances from the mean, a parameter that measures the disorder in the first coordination sphere. Edge_E is the position of the absorption edge, $\text{Edge}_{\text{slope}}$ is the tangent of the slope of the absorption edge, Max_{curv} is the curvature of the spectrum at the point of the main maximum, MinE is the energy of the main minimum, and Min_{int} is the intensity of the spectrum at the point of the minimum. Before constructing analytical relations, the training sample descriptors were normalized to the zero mean and single standard deviation.

spectrum in the sample. For the classification of spectra according to the charge state, the position of the absorption edge is of prime importance. However, the use of this descriptor alone provides a prediction accuracy of worse than 60%. The chemical shift of the entire spectrum can be misinterpreted due to a shift of the edge at changing distances. This effect is partially compensated by taking into account the intensity descriptor of the main maximum (Max_{int}), which, together with the position of the absorption edge, leads to correct classification in 75% of cases. A further increase in the prediction accuracy is possible by applying restrictions on the range of possible distances in the Fe^{2+} and Fe^{3+} structures.

The second part of Table 1 (spectrum descriptors) reflects the inverse dependence of the XANES-spectrum features if we consider the geometric parameters. For example, the tangent of the slope of the absorption edge depends on the average distances and the coordination number. The curvature of the white line correlates with disorder in the first iron coordination sphere (Dev). A wider spread of distances leads to broadening of the main maximum. The position of the first minimum (Min_E) is a rather important characteristic in the spectrum, although it is less often analyzed in comparison with the positions of the maxima. This feature is almost 90% due to the CN and the Fe–O distance. Its intensity is determined by the CN and the scatter of bond lengths in the first coordination sphere.

CONCLUSIONS

In this paper, a new approach to the quantitative analysis of X-ray absorption spectra using machine-learning algorithms is considered. For the function $\mu(E)$, the key features, namely, the energy position of the edge, minima, maxima, the intensity of the main maximum and minimum, the curvature of the function at the extrema, and the slope of the absorption edge, are selected. Using the radial-basis-function algorithm, combinations of descriptors that provide the best accuracy in predicting the structural parameters around the absorbing atom were selected. Typically, a machine-learning algorithm operates as a black box for researchers. The operation of a universal method for constructing analytical relations between spectrum descriptors and structural parameters, such as coordination numbers, interatomic distances, oxidation state, and standard deviations of interatomic distances, is shown.

FUNDING

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

REFERENCES

1. A. L. Bugaev, A. A. Guda, K. A. Lomachenko, V. V. Srabionyan, L. A. Bugaev, A. V. Soldatov,

C. Lamberti, V. P. Dmitriev, and J. A. Van Bokhoven, J. Phys. Chem. C **118**, 10416 (2014). https://doi.org/10.1021/jp500734p

2. J. A. Van Bokhoven and C. Lamberti, *X-Ray Absorption and X-Ray Emission Spectroscopy: Theory and Applications* (Wiley, New York, 2016).
3. C. R. Natoli, in *EXAFS and Near Edge Structure III*, Ed. by K. O. Hodgson, B. Hedman, and J. E. Penner-Hahn (Springer, Berlin, 1984), **Vol. 2**, p. 38.
4. J. García, G. Subías, and J. Blasco, in *X-Ray Absorption and X-Ray Emission Spectroscopy: Theory and Applications*, Ed. by J. A. van Bokhoven and C. Lamberti (Wiley, Chichester, 2016), p. 459.
5. C. Zheng, C. Chen, Y. Chen, and S. P. Ong, Patterns **1**, 100013 (2020). https://doi.org/10.1016/j.patter.2020.100013
6. Y. Liu, N. Marcella, J. Timoshenko, A. Halder, B. Yang, L. Kolipaka, M. J. Pellin, S. Seifert, S. Vajda, P. Liu, and A. I. Frenkel, J. Chem. Phys. **151**, 164201 (2019). https://doi.org/10.1063/1.5126597
7. J. Timoshenko, D. Y. Lu, Y. W. Lin, and A. I. Frenkel, J. Phys. Chem. Lett. **8**, 5091 (2017). https://doi.org/10.1021/acs.jpclett.7b02364
8. C. D. Rankine, M. M. M. Madkhali, and T. J. Penfold, J. Phys. Chem. A **124**, 4263 (2020). https://doi.org/10.1021/acs.jpca.0c03723
9. A. Martini, S. A. Guda, A. A. Guda, G. Smolentsev, A. Algasov, O. Usoltsev, M. A. Soldatov, A. Bugaev, Y. Rusalev, C. Lamberti, and A. V. Soldatov, Comput. Phys. Commun. **250**, 107064 (2019). https://doi.org/10.1016/j.cpc.2019.107064
10. J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, NPJ Comput. Mater. **5**, 83 (2019). https://doi.org/10.1038/s41524-019-0221-0
11. O. Trejo, A. L. Dadlani, F. De La Paz, S. Acharya, R. Kravec, D. Nordlund, R. Sarangi, F. B. Prinz, J. Torgersen, and N. P. Dasgupta, Chem. Mater. **31**, 8937 (2019). https://doi.org/10.1021/acs.chemmater.9b03025
12. M. Wilke, F. Farges, P. E. Petit, G. E. Brown, and F. Martin, Am. Mineral. **86**, 714 (2001). https://doi.org/10.2138/am-2001-5-612
13. M. R. Carbone, S. Yoo, M. Topsakal, and D. Y. Lu, Phys. Rev. Mater. **3**, 033604 (2019). https://doi.org/10.1103/PhysRevMaterials.3.033604
14. M. R. Carbone, M. Topsakal, D. Y. Lu, and S. Yoo, Phys. Rev. Lett. **124**, 156401 (2020). https://doi.org/10.1103/PhysRevLett.124.156401
15. S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram, and L. Hung, NPJ Comput. Mater. **6**, 109 (2020). https://doi.org/10.1038/s41524-020-00376-6
16. Y. Joly, Phys. Rev. **63**, 125120. https://doi.org/10.1103/PhysRevB.63.125120
17. S. A. Guda, A. A. Guda, M. A. Soldatov, K. A. Lomachenko, A. L. Bugaev, C. Lamberti, W. Gawelda, C. Bressler, G. Smolentsev, A. V. Soldatov, and Y. Joly, J. Chem. Theory Comput. **11**, 4512 (2015). https://doi.org/10.1021/acs.jctc.5b00327
18. A. L. Bugaev, V. V. Srabionyan, A. V. Soldatov, L. A. Bugaev, and J. A. van Bokhoven, J. Phys.: Conf. Ser. **430**, 012028 (2013). https://doi.org/10.1088/1742-6596/430/1/012028
19. A. L. Bugaev, A. A. Guda, K. A. Lomachenko, A. Lazzarini, V. V. Srabionyan, J. G. Vitillo, A. Piovano, E. Groppo, L. A. Bugaev, A. V. Soldatov, V. P. Dmitriev, R. Pellegrini, J. A. van Bokhoven, and C. Lamberti, J. Phys.: Conf. Ser. **712**, 012032 (2016). https://doi.org/10.1088/1742-6596/712/1/012032
20. L. J. Zhang, J. Zhou, J. Y. Zhang, J. Su, S. Zhang, N. Chen, Y. P. Jia, J. Li, Y. Wang, and J. Q. Wang, J. Synchrotron Radiat. **23**, 758 (2016). https://doi.org/10.1107/S1600577516001910
21. D. H. Bailey, Comput. Sci. Eng. **2**, 24 (2000). https://doi.org/10.1109/5992.814653
22. D. H. Bailey, J. Borwein, N. Calkin, R. Luke, R. Girgensohn, and V. Moll, *Experimental Mathematics in Action* (CRC, Boca Raton, 2007).
23. H. Zou and T. Hastie, J. R. Stat. Soc. B **67**, 301 (2005). https://doi.org/10.1111/J.1467-9868.2005.00503.X
24. R. Tibshirani, J. R. Stat. Soc. B **58**, 267 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Translated by A. Ivanov