npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms

A. A. Guda[1]✉, S. A. Guda[1,2]✉, A. Martini[1,3]✉, A. N. Kravtsova[1], A. Algasov[1,2], A. Bugaev[1], S. P. Kubrin[4], L. V. Guda[1], P. Šot[5], J. A. van Bokhoven[5,6], C. Copéret[5] and A. V. Soldatov[1]

X-ray absorption near-edge structure (XANES) spectra are the fingerprint of the local atomic and electronic structures around the absorbing atom. However, the quantitative analysis of these spectra is not straightforward. Even with the most recent advances in this area, for a given spectrum, it is not clear a priori which structural parameters can be refined and how uncertainties should be estimated. Here, we present an alternative concept for the analysis of XANES spectra, which is based on machine learning algorithms and establishes the relationship between intuitive descriptors of spectra, such as edge position, intensities, positions, and curvatures of minima and maxima on the one hand, and those related to the local atomic and electronic structure which are the coordination numbers, bond distances and angles and oxidation state on the other hand. This approach overcoms the problem of the systematic difference between theoretical and experimental spectra. Furthermore, the numerical relations can be expressed in analytical formulas providing a simple and fast tool to extract structural parameters based on the spectral shape. The methodology was successfully applied to experimental data for the multicomponent Fe:SiO_2 system and reference iron compounds, demonstrating the high prediction quality for both the theoretical validation sets and experimental data.

npj Computational Materials (2021)7:203 ; https://doi.org/10.1038/s41524-021-00664-9

INTRODUCTION

X-ray absorption spectroscopy is widely employed to probe the local atomic and electronic structure around the absorbing atom[1–3]. The X-ray absorption near-edge structure (XANES), spanning a region of 50–200 eV above the absorption edge, contains information about the structural descriptors involving the bond distances and angles, type of ligand surrounding, oxidation state, which affect the spectral descriptors; edge position, shapes and positions of spectral maxima and minima. An experienced researcher can, for example, distinguish the pure metallic state from metal oxide compound, or discriminate between tetrahedral and octahedral surroundings based on a qualitative inspection of the related spectral features.

Figure 1 shows a series of typical experimental Cu *K*-edge XANES spectra for different copper compounds. The pre-edge feature A originates from the transition to the spatially localized 3*d* states. The pre-edge shape depends on the number of electrons in the *d*-shell[4], its intensity is proportional to the amount of 3*d*–4*p* hybridization[5], while its energy position can be employed to realize the calibration of the 3d metal oxidation state[6]. The sharp shoulder B on the rising edge is indicative of a linear or square planar geometry with a lower energy of empty 4*p* orbitals perpendicular to the chemical bonds[7]. A similar shoulder appears in the spectra of metals. K-edge XANES of metals with an *fcc* structure is further characterized by the splitting of the main peak into M1 and M2 features. Intensities of M2 and M3 are sensitive to the scattering from the second coordination shell[8] similar to the feature D in molecular covalent complexes, and their reduction can be therefore used to probe the nanosized effects[9]. Positions of M4 and further high-energy maxima relate to the interatomic distances via the semi-empirical Natoli's rule[10]. The absorption edge position depends on the oxidation state[11] and also interatomic distances[12]. The intensity of the white line C is higher in spectra of metal complexes with the octahedral coordination. Planar complexes are characterized by energy splitting of this peak[13]. Characteristic spectral features can be further established for the *K*-edges of light atoms, $L_{2,3}$ edges for 3*d* metals with strong multiplet splitting, or $L_{2,3}$ spectra for 4*d* metals possessing a characteristic white line.

For a data scientist, the above-mentioned spectral features are recognized as descriptors, and the relationships between spectral descriptors and the structural ones (coordination number, geometry, bond distances, angles…) can be established, for example, by using machine learning (ML) algorithms. Using all points of a spectrum as descriptors, Zheng et al.[14] managed to classify the atomic coordination environments via random forest models. The convolutional neural network was applied to predict Cu–Cu coordination numbers (CN)[15] and to evaluate several CNs for platinum nanoparticles to refine their sizes and shapes[16]. Rankine et al.[17] demonstrated the ability of a deep neural network to predict a XANES spectrum from geometric information about the local environment around the absorbing atom. To achieve better performance of ML, the dimensionality of both spectral and structural descriptors should be reduced. For example, 3 *N* atomic coordinates for *N* atoms can be converted into radial and angle distribution functions[18] or into generalized radial distribution functions[19]. Alternatively, geometry parameters can be filtered in terms of their importance for XANES variation[20]. The multiple points of a spectrum can also be reduced to only a few descriptors. Commonly used spectral descriptors are the

[1]The Smart Materials Research Institute, Southern Federal University, 344090 Sladkova 178/24, Rostov-on-Don, Russia. [2]Institute of mathematics, mechanics and computer science, Southern Federal University, 344090 Milchakova 8a, Rostov-on-Don, Russia. [3]Department of Chemistry, INSTM Reference Center and NIS and CrisDi Interdepartmental Centers, University of Torino, Via P. Giuria 7, I-10125 Torino, Italy. [4]Research Institute of Physics, Southern Federal University, 194, Stachki Ave., Rostov-on-Don 344090, Russia. [5]Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 1-5, 8093 Zurich, Switzerland. [6]Paul Scherrer Institut, Villigen 5232, Switzerland. ✉email: guda@sfedu.ru; gudasergey@gmail.com; andrea.martini@unito.it

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

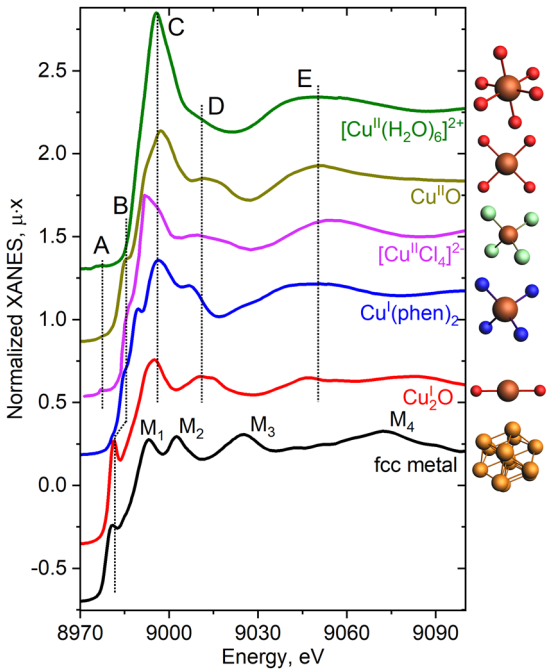

Fig. 1 Characteristic spectral features for different structural motifs. Cu *K*-edge XANES spectra for different oxidation states and local coordinations of the copper atom (from bottom to top): Cu^0 in *fcc* metal, linear Cu^I in Cu_2O, pseudo-tetrahedral Cu^I and Cu^{II} in Cu $(phen)_2$ and $[CuCl_4]^{2-}$ complexes, square planar Cu^{II} in CuO, pseudo-octahedral Cu^{II} in $[Cu(H_2O)_6]^{2+}$.

pre-edge centroid and the pre-edge area which are, for example, combined to analyze the Fe oxidation state and coordination number[6]. Carbone et al. demonstrated that the principal components calculated from a series of theoretical spectra can be used to realize the classification of four-, five-, and six-coordinated metal environments[21] and of type of functional groups[22]. Recently, Torrisi et al.[23] demonstrated the concept of constructing descriptors from a polynomial fit of equidistant energy intervals of the spectrum.

The present work aims to extend these approaches of identifying descriptors of the spectral features, such as positions of the absorption edge, minima, and maxima, their amplitudes, and curvatures. We present a step-by-step procedure to prepare a training set, evaluate the descriptors, train the machine learning algorithm, apply cross-validation, and finally analyze the experimental data. The variable structural parameters are introduced and the problem of classification of the calculated spectra in terms of these parameters is addressed. The analytical formulas establishing the relations between the spectral features and the structural parameters are then derived. Finally, we validate the approach for a set of experimental spectra belonging to oxides, silicates, geological samples (tektites, impactites), and amorphous glasses as well as silica-supported Fe single-site catalysts prepared via surface organometallic chemistry[24–26].

RESULTS AND DISCUSSION

Descriptors of spectrum

In general, the theoretical XANES spectrum contains ~100 energy points. A common approach to improve the efficiency of ML algorithms is to reduce the dimensionality of such object by extracting only informative features, notably the spectral descriptors[23]. Table 1 and Fig. 2 describe a set of descriptors evaluated for each spectrum: edge position (feature A), white line position and intensity (feature B), first pit (minimum) position and intensity

Table 1. Descriptors of spectra with their short notation and details on the evaluation.

No.	Short notation	Descriptor	Comment
1	PC_1	Projections on the first three principal components of the dataset	Principal component analysis is performed for the whole dataset of spectra consisting of calculations for two-, three-, four-, five- and sixfold coordinated Fe, together for Fe^{2+} and Fe^{3+}.
2	PC_2		
3	PC_3		
4	$Edge_E$	Edge energy	Center and slope of the arctangent function which fits the whole spectrum.
5	$Edge_{slope}$	Edge slope	
6	WL_{int}	White line intensity	Polynomial fit for the region of the first maximum and first minimum in the spectrum. For better stability of the fit, the spectra were convoluted with a Lorentzian of 5 eV width.
7	WL_E	White line center	
8	WL_{curv}	White line curvature	
9	Pit_E	Pit energy	
10	Pit_{int}	Pit intensity	
11	Pit_{curv}	Pit curvature	
12	WL-Pit_{slope}	White line - Pit slope	The slope of the line connects the maximum of the white line and the minimum of the first pit.
13	rPC_1	Projections on the relative principle components	Same as 1–3 but singular value decomposition is performed for a data set of spectra aligned according to their edge energy position.
14	rPC_2		
15	rPC_3		

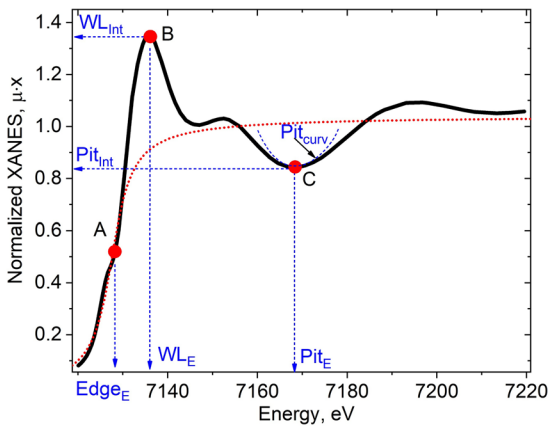

Fig. 2 A set of suggested descriptors for a XANES spectrum. Edge energy and edge slope are evaluated in point A, white line intensity, curvature, and position are evaluated in B, pit intensity, curvature, and position are evaluated in C. Arctangent function (red dotted line) is used to determine the edge position and its slope.

(feature C), the curvature of the white line, projections on the principal components (further called PC descriptors). The arctangent function (red dotted line) was used to fit the whole spectrum and the position of its center and slope were taken as the values of edge position ($Edge_E$) and slope ($Edge_{slope}$). For some deformations in the local geometry, the white line in the calculated spectra can consist of several close maxima. For a monotonic variation of descriptors across the training set we performed an additional convolution (5 eV Lorentzian width) of the spectral regions near extrema B and C before evaluating the curvature, amplitude, and energy position of these features.

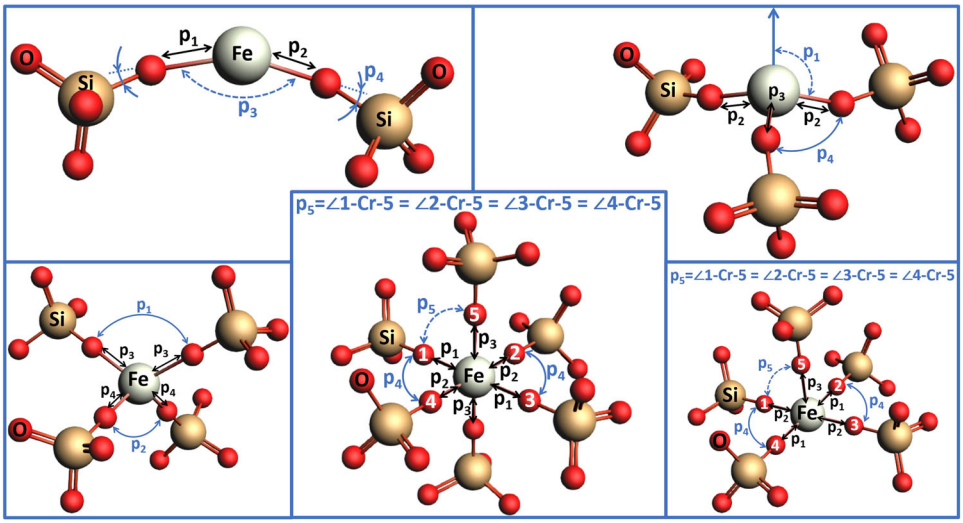

Fig. 3 The structures and their deformations applied for constructing the training set. $Fe(SiO_4)_{CN}$ clusters constructed for coordination numbers CN = 2, 3 …6. The variable structural deformations p1...p5 are applied to each structure and reproduce variety of iron local geometries in the amorphous silica.

The same convolution was applied to the experimental spectra before the evaluation of their descriptors.

Principal component analysis (PCA) was applied to the whole data set of theoretical spectra. Based on the singular value decomposition (SVD, see details in Supplementary Methods Section) three first principal components were evaluated. Each spectrum was projected on these components and the projections were used as the descriptors of the spectrum. We also applied SVD analysis to the data set where all edge positions (point A in Fig. 2) were aligned. In such a way, another set of three principal components was used to calculate projections of every theoretical spectrum. We call these projections relative PC descriptors (rPC).

If compared with the descriptors based on the curvature of fixed energy intervals for spectrum, as shown in the work Torrisi et al.[23] the search of minima and maxima along with edge characteristics relies on physically motivated features of the spectrum. As we show below the good prediction quality can be achieved by using just 2 or 3 such descriptors. The stable definition of the extrema may be tricky for flattened spectra or $L_{2,3}$ edges with rich multiplet splitting. Such systems may require additional spectral descriptors such as total variance, centers of mass and areas, fitted peak profiles, etc. These descriptors are beyond the scope of the present work but are included in the supplementary software.

Relationship between spectral features and structure

3*d* metal complexes can be found in a wide range of CNs and local symmetries around the metal center. The type of ligands determines the interatomic distances and symmetry for the given oxidation and spin state of the metal. Valuable catalysts or geological materials can contain iron ions in a silica matrix where oxygen coordination provides both Fe^{2+} and Fe^{3+} oxidation states along with several possible CNs. To address the problem of quantitative iron speciation, we calculated the training set consisting of spectra for $Fe(SiO_4)_{CN}$ complexes for CN = 2–6. The first shell distances and bond angles were varied for every CN using the improved Latin hypercube sampling (IHS) resulting in 3000 spectra calculated for all CN. A chemical shift was then applied for each spectrum to simulate absorption from Fe^{2+} and Fe^{3+} sites for every deformation. Figure 3 shows the clusters used for simulations and the variable structural parameters ***p***. The ranges of their variation are listed in Table 2.

Table 2. The ranges of variation for structural parameters p_1, p_2, p_3, p_4.

CN	Deformation	Range
2	p_1	1.8–2.3 Å
	p_2	
	p_3	120°–180°
	p_4	0°–70°
3	p_1	80°–135°
	p_4	60°–120°
	p_2	1.8–2.3 Å
	p_3	
4	p_1	65°–180°
	p_2	
	p_3	1.8–2.3 Å
	p_4	
5	p_1	1.8–2.3 Å
	p_2	
	p_3	
	p_4	60°–90°
	p_5	90°–120°
6	p_1	1.8–2.3 Å
	p_2	
	p_3	
	p_4	60°–90°
	p_5	90°–120°

The calculated spectra for Fe^{2+} are shown in Fig. 4. The library of spectra for Fe^{3+} contains the same entries but shifted according to the 1*s* core level energy difference evaluated within an accurate molecular orbital approach (see Methods section). Therefore, the total number of spectra in the training set was 6000, i.e. twice more than shown in Fig. 4.

Each spectrum in the training set is characterized by several descriptors (Table 1). The whole training set can be projected on a 2D plot for the selected pair of descriptors. Figure 5 compares the distribution of points in the training set over different 2D maps where each point is colored according to its structural parameters.

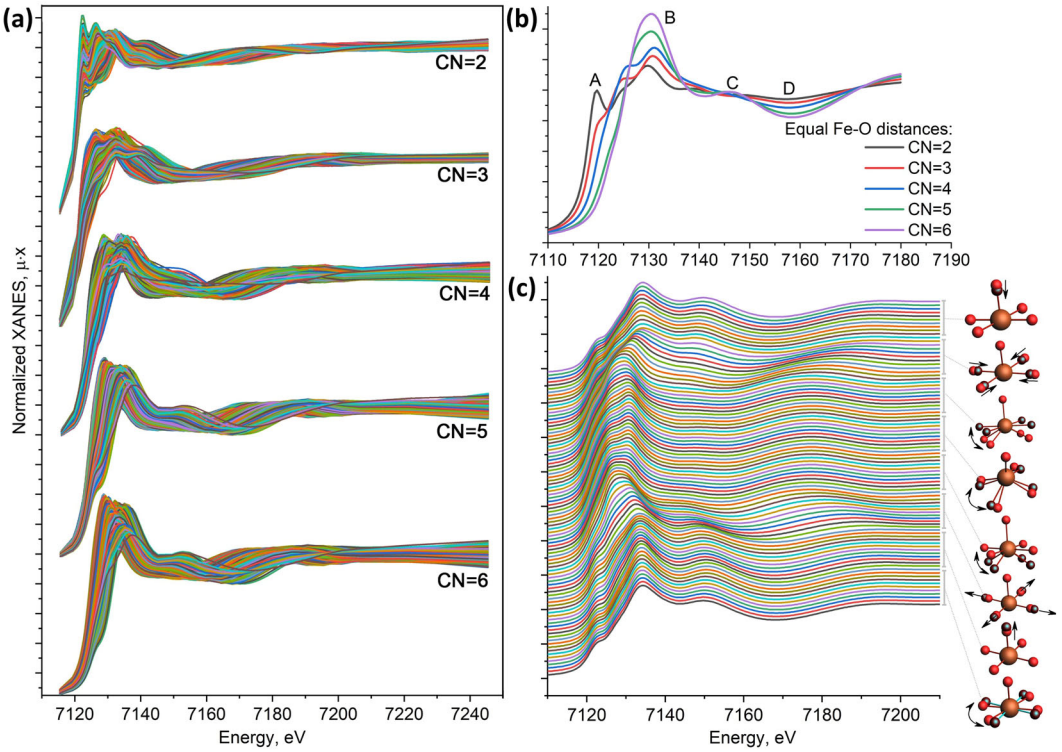

Fig. 4 Visualization of the theoretical training set and the trends in variations of XANES spectra upon studied deformations. a Six hundreds of Fe *K*-edge XANES spectra calculated for each coordination number by varying structural parameters (Fig. 3). **b** Comparison between shapes of spectra upon variation of coordination numbers while all Fe-O bond lengths fixed to 2.1 Å. **c** Sensitivity of the spectrum to variations of bending angles and interatomic distance for a five-coordinated model (only first coordination shell is shown for simplicity).

From a mathematical point of view, each color in Fig. 5 defines a class. If points for different classes are well separated on the 2D map, the chosen pair of descriptors is appropriate for the classification. For demonstration, we selected those pairs of the descriptors which separated points according to iron coordination number, Fe–O distances, or oxidation state. In particular, the best descriptors for discriminating different CN were the curvature of the white line (WL_{curv}), pit energy position (Pit_E), edge position ($Edge_E$). The average distances in the iron first coordination shell could be distinguished according to the energy positions of the pit (Pit_E) and edge ($Edge_E$). Projections on the principal components were able to separate structures with different distances and oxidation states, while pit energy and white line position (WL_E) distinguished between structures with different oxidation states.

Beyond the two-dimensional scatter plots, which are informative for the qualitative selection of good descriptors, the best quality of classification and the best choice of descriptors for ML algorithm was determined (Table 3) for combinations of 1, 2, 3, or 4 descriptors to predict CN, oxidation state, or distance in a pure compound (mixtures will be discussed further in section 2.4).

Two descriptors of spectra contain up to 95% of the information necessary for discrimination between Fe^{2+} and Fe^{3+}. Using the value of the edge energy alone provided 80% of the prediction quality. Considering the white line intensity in addition to the $Edge_E$ improved the quality to 90%. Other informative descriptors for the oxidation state were the energies of the main maximum and pit, the first principal components. Fe–O mean distance is uniquely characterized by the combination of edge and pit energies (95% quality). Projections on the second and third relative principal components, rPC_2 and rPC_3, were more important for this task than the first PC. Higher CNs are characterized by a sharp white line and a steep rising edge. Good quality of prediction for *CN* requires to use of at least three descriptors which include edge energy, slope, and curvature of the main maximum. The lowest accuracy in cross-validation analysis was observed for the standard deviation from the mean that measures the disorder in the first coordination shell. Four descriptors were necessary to reach the prediction quality equal to 90%.

The optimal choice of the descriptors in Table 3 does not guarantee their transferability to the experimental data and problem of the multicomponent system analysis. In Section 2.4, we address the quality of structural analysis by using descriptors in the training set composed of linear combinations of spectra.

Analytical relations between descriptors: beyond Natoli's rule

In the early eighties, Natoli formulated an empirical rule[10] that establishes dependence between peak positions in the XANES spectrum and interatomic distances for the structures with similar symmetry, which can be the case of metals within the same space group (e.g., *fcc* Cu and Ni, Supplementary Fig. 3) and to structures that undergo a volume expansion, such as palladium after hydrogen sorption[27]. In the latter case, we have previously observed that the relative intensities of the first two XANES maxima are proportional to the H/Pd ratio in palladium hydride samples[28,29]. Another example by Zhang et al.[30] provides an analytical relation between energy positions of maxima in U L_3-edge XANES spectra of uranyl complexes and distances between the uranium absorber and oxygen ligand atoms. Representing a useful tool for the analysis of XANES spectra, all these examples are limited to the usage of only one spectral descriptor and one descriptor of structure. In this section, we extend such methodology to derive the analytical relation between any set of spectral descriptors and structural parameters using machine learning algorithm. The common approach to find simple analytical

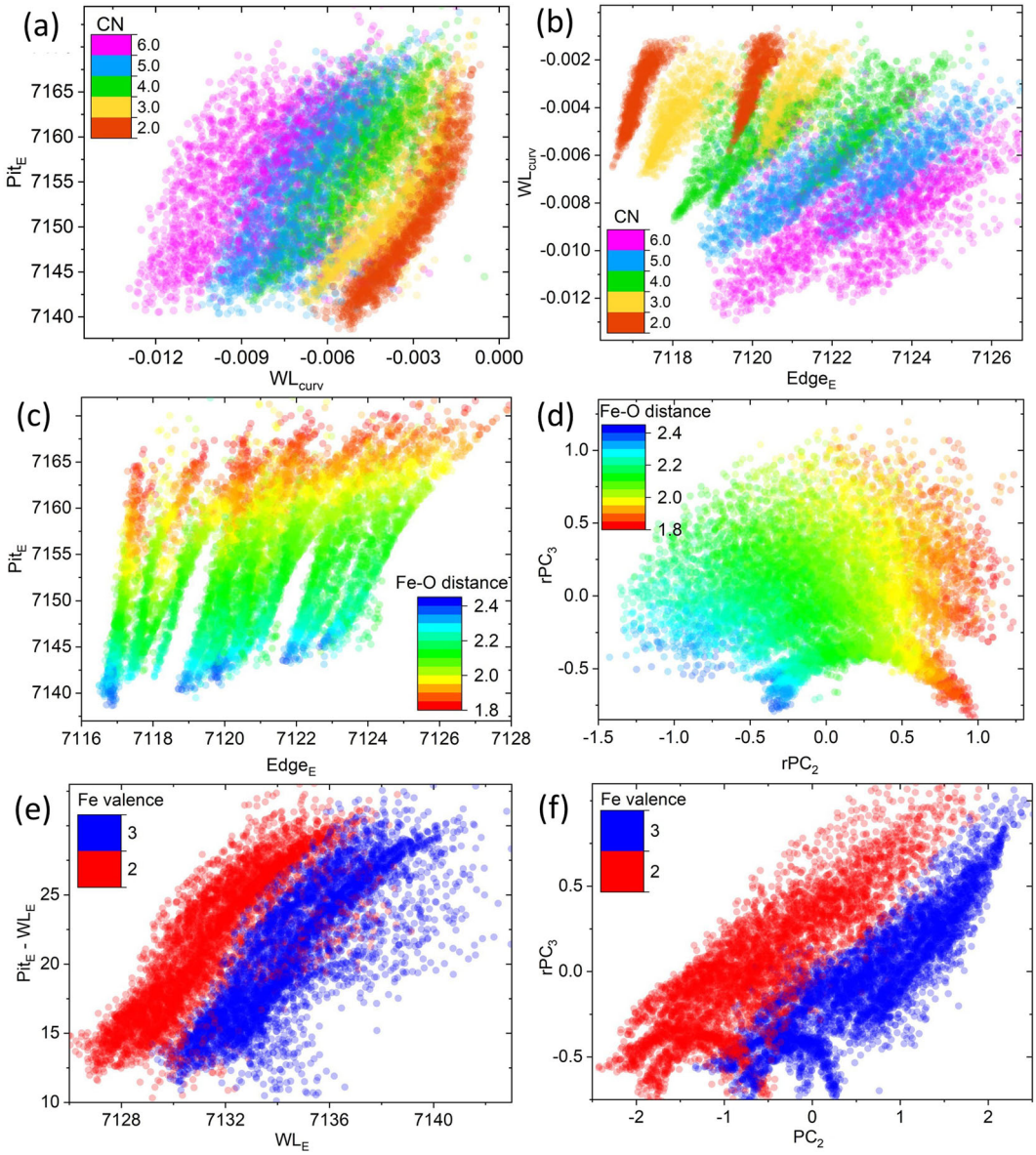

Fig. 5 Scatter plots for the selected pairs of descriptors. Each point corresponds to a single spectrum in the training set from Fig. 4a. The color reflects the CN values in **a**, **b**, the average Fe-O distance in **c**, **d**, and iron valence in **e**, **f**.

relations between known parameters $p_1 \ldots p_n$ and target variable y is the construction of linear regression:

$$y = w_1 \cdot p_1 + w_2 \cdot p_2 + \ldots + w_n \cdot p_n \quad (1)$$

More complex cases include pairwise and higher degree multiplications alongside parameters $p_1 \ldots p_n$. We are interested in pretty solutions with good approximation quality. The prettiness means the absence of large coefficients w_i and the smallest possible number of nonzero w_i. For the integer relations problem, the prettiness is achieved by applying special algorithms of integer orthogonalization (see e.g., [31] and §2.2 in ref. [32]). In a real-valued case, we use feature selecting properties of the Elastic Net algorithm[33] combined with some heuristics. For the theoretical data set, we restrict ourselves to the parameters $p_1 \ldots p_n$ and their pairwise multiplications, thus the Eq. (1) takes the form

$$y = \sum_{i=1}^{n} w_i \cdot p_i + \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \cdot p_i \cdot p_j \quad (2)$$

In the first step, the data were normalized to zero mean and unit standard deviation. We implemented the elastic net method that includes the LASSO[34] and ridge regression. In the case of the group of highly correlated variables, the LASSO algorithm tends to select one variable from a group and ignores the others, thus making feature selection. If the linear formula returned by Elastic Net is heavy, we try to simplify it at the expense of model precision. To do so, we sort the coefficients (w_i, w_{ij}) returned by Elastic Net by their absolute values and try to build a linear model based on subsets of features with the largest absolute coefficients. The analysis was performed for the subsets of each size: 1, 2, 3,…, and for all of them R^2-score was evaluated. Afterward, one can choose between pretty models with moderate quality or more complicated models with higher precision. Table 4 shows the selected analytical relations between descriptors of spectra and structural parameters.

Analytical relations between descriptors extend the qualitative classification of the 2d scatter plots. The obtained formulas explore dependencies between any number of spectral features

Table 3. The quality of structural parameters prediction by using selected good combinations of the descriptors of spectra from data set in Fig. 4a.

Structural parameter	Number of descriptors							
	1		2		3		4	
CN	Pit_{int}	0.55	$Edge_E$ WL_{curv}	0.85	$Edge_E$ WL_E rPC_3	0.95	$Edge_E$ $Edge_{slope}$ WL_E Pit_E	1.0
	WL_{int}	0.60						
	$Edge_E$	0.50	$Edge_E$ $Edge_{slope}$	0.80				
	WL_{curv}	0.55			$Edge_E$ $Edge_{slope}$ WL_{curv}	0.95		
	PC_3	0.50	WL_{curv} Pit_E	0.80				
	$(WL\text{-}Pit)_{slope}$	0.55						
Fe–O mean distance	Pit_E	0.80	rPC_2 rPC_3	0.95	PC_2 PC_3 rPC_3	0.95	$Edge_E$ WL_E Pit_{curv} Pit_E	1.0
	$Pit_E\text{-}WL_E$	0.80						
	$Edge_{slope}$	0.85	WL_{curv} rPC_3	0.95				
	WL_E	0.60			$Edge_E$ WL_E Pit_E	0.95		
	rPC_2	0.75	$Edge_E$ Pit_E	0.90				
	PC_3	0.60						
The standard deviation of distances from mean	Pit_{int}	0.20	WL_{curv} Pit_{int}	0.55	$Edge_E$ WL_{curv} Pit_{int}	0.85	$Edge_E$ WL_{curv} WL_{int} Pit_{int}	0.90
	Pit_{curv}	0.10						
	$Edge_E$	0.05	Pit_{int} rPC_2	0.50				
	rPC_3	0.05			PC_2 PC_3 Pit_{int}	0.80		
	rPC_2	0.05	PC_3 Pit_{int}	0.50				
	PC_3	0.05						
Fe valence	PC_1	0.95	PC_1 rPC_3	0.95	$Edge_E$ WL_{curv} $Pit_E\text{-}WL_E$	1.0	$Edge_E$ WL_E $Pit_E\text{-}WL_E$ Pit_{int}	1.0
	$Edge_E$	0.80						
	WL_E	0.70	$Edge_E$ WL_{int}	0.90				
	PC_2	0.80			PC_2 rPC_2 rPC_3	1.0		
	rPC_1	0.60	WL_E Pit_E	0.90				
	PC_3	0.60						

R^2 score and accuracy were used for regression and classification, respectively (1.0 means best accuracy, see Supplementary Equation 3 for definition).

and structural parameters. While, in general, ML algorithms work as a black box, Table 4 provides a geometrical interpretation of the best combinations of descriptors. For example, up to 90% prediction quality can be achieved for the interatomic distances if the energy positions of the edge, first maximum, and minimum are considered.

The accuracy of the quadratic analytical formulas for the oxidation state is above 80%. The quality of analysis could be improved if chemically relevant restrictions were imposed on the Fe–O distances for Fe^{2+} or Fe^{3+} ions in the training set. For better generalization, we assumed that the ranges of variations of structural parameters were equal for both the oxidation states. Therefore, the chemical shift of the whole spectrum can be misinterpreted by the edge shift upon distance variation. This effect is partially accounted for by the main maximum intensity (WL_{int}) descriptor that enters the formula. The intensity of the main maximum changes along with Fe–O bonds contraction therefore this descriptor can help to discriminate between shifts related to the oxidation state or volume changes. Formulas for *CN* depend on the curvature of the main maximum, which is consistent with the general behavior of EXAFS oscillations, whose amplitude is proportional to CN. One should note, however, that this conclusion should not be generalized to the structures with different types of bonds (e.g., metallic iron has larger CN, but the white line intensity is higher in the octahedral Fe-O oxide).

The second part of Table 4 interprets the features of the XANES spectrum in terms of geometry parameters. The slope of the edge depends on the average distances and coordination number. The curvature of the white line correlates with the disorder in the first coordination shell of iron, i.e., larger disorder makes the first maximum broader. The position of the first minimum is quite an important feature in the spectrum though it is less often analyzed as compared to the maxima. This feature is by almost 90% determined by the CN and Fe–O distance. Its intensity is determined by the CN and disorder in the first coordination shell.

Fitting a multi-component system

If the distribution of absorbing atoms in a material is heterogeneous, a linear combination of theoretical spectra with different oxidation states and coordination is required to describe the experimental spectrum. In this section, we extend the descriptor approach to the case of linear combinations and apply the descriptor analysis to the experimental Fe *K*-edge XANES data of iron oxide and iron silicate systems. The algorithm was applied to 56 experimental spectra of crystalline compounds[35], glasses[36], tektites, and impactites[37,38], as well as a single-site silica-supported Fe catalyst[24,39]. Figure 6 shows experimental spectra and Supplementary Tables 2–6 provide a description of each sample and results of ML-analysis. Iron coordination and oxidation state are heavily dependent on the conditions of synthesis. Studied samples are inherently heterogeneous systems. In particular, tektites are formed from molten high-speed ejecta during the early stages of impact crater formation[40]. Impactites have a more complex history of their formation and are the result of the melting of various types of rocks located at different depths in the Earth's crust. Iron in the amorphous silica structure has the potential to be a probe of impact rock formation conditions, such as pressure (P), temperature (T), oxygen fugacity[41,42].

Table 4. Analytical relations between descriptors of spectra and descriptors of structure.

No.	Descriptor	Analytical formula	R^2 score
Descriptors of structure			
1	CN	$-0.85 \cdot WL_{curv}$	0.7
		$-0.95 \cdot WL_{curv} + 0.36 \cdot Pit_E$	0.9
3	Average Fe–O distance, (R_{Fe-O})	$0.94 \cdot Edge_{slope}$	0.9
5		$-0.40 \cdot WL_{curv} - 0.77 \cdot (Pit_E - WL_E)$	0.9
6		$0.5 \cdot Edge_{slope} - 0.55 \cdot Pit_E + 0.2 \cdot Edge_E$	0.9
7	Fe oxidation state	$0.97 \cdot Edge_E + 0.52 \cdot Pit_{int}$	0.7
8		$1.11 \cdot Edge_E - 0.75 \cdot WL_{int} - 0.13 \cdot (Pit_E - WL_E)^2 + 0.13$	0.8
Descriptors of spectrum			
1	$Edge_{slope}$	$0.95 \cdot R_{Fe-O} - 0.15 \cdot Std + 0.09 \cdot CN^2 - 0.09$	0.9
2	WL_{curv}	$-0.35 \cdot R_{Fe-O} - 0.83 \cdot CN$	0.8
3		$-0.95 \cdot CN - 0.39 \cdot R_{Fe-O} + 0.32 \cdot Std$	0.9
3	WL-Pit_{slope}	$-0.47 \cdot Std + 0.31 \cdot R_{Fe-O} + 0.96 \cdot CN$	0.9
4	Pit_E	$-0.92 \cdot R_{Fe-O}$	0.8
5		$0.16 \cdot CN - 0.93 \cdot R_{Fe-O}$	0.9
6	Pit_{int}	$-0.97 \cdot CN + 0.67 \cdot Std$	0.8

Label "R_{Fe-O}" is used for the average Fe–O distances in the first coordination shell. "Std" is used for the standard deviation of Fe–O distances from mean, the parameter which measures disorder in the first coordination shell. Before constructing analytical dependencies, the descriptors of the training set were normalized to zero mean and unit standard deviation.

Figure 7 shows the steps required to apply the descriptor approach to the multicomponent system. We have constructed a database of linear combinations of theoretical spectra in the training set, using several random concentrations for every pair of spectra. The descriptors were then evaluated for the database. A cross-validation procedure was applied to different combinations of the descriptors to understand which combination works better for the mixtures.

The appropriate choice of the descriptors for the given structural property should provide good quality of analysis both for the theoretical validation set and a set of experimental references. Therefore, we have calculated the descriptors of the experimental and calculated spectra for the reference structures. The pairs of theoretical and experimental descriptors for the known structures can be used in step 2.2 to calibrate the descriptors in the theoretical training set for systematic energy shifts or intensity differences. The calibration step for intensities may be necessary when experimental spectra are measured in fluorescence mode and are flattened owing to self-absorption. In this work, we did not apply any calibration after the convolution of theoretical spectra. In step 2.4, the reference spectra are used for validation before predicting results for the unknown structures. Figure 8 shows selected scatter plots for pairs of descriptors that can discriminate efficiently between iron oxidation state, CN, and average Fe–O distance in the two-component mixture. While the classes were well separated in Fig. 5, their overlap occurs in Fig. 8 due to the linear combinations added to the training sample. We projected spectra of several references (hollow circles) on the two-dimensional scatter plots. Reference oxides and silicates have quite different structures from the entries in the training set, but descriptors Pit_E, WL_E, and WL_{curv} provided surprisingly good quality for their analysis. α-Fe_2O_3 and $NaFeSi_2O_6$ were properly projected to the region of 6-coordinated species. γ-Fe_2O_3 and Fe_3O_4 contain one-third of Fe ions in the tetrahedral positions and this point is projected to the region where 4-, 5-, and 6-coordinated points are overlapped (Fig. 8a). Fe_2SiO_4 reference has the longest Fe–O distances equal to 2.2 A and it is properly projected to the blue region of the plot in Fig. 8b, while γ-Fe_2O_3 has the shortest. In 8c, Fe_2O_3 and $NaFeSi_2O_6$ are assigned to Fe^{3+}, Fe_2SiO_4 to Fe^{2+}, whereas Fe_3O_4 contains a mixture of Fe^{2+} and Fe^{3+} sites.

The classes in the training set overlap when linear combinations of spectra are introduced along with the pure species. Figure 8a shows how CN classes are mixed if compared with Fig. 5a. The points with intermediate average valence are also become overlapped. In general, the prediction quality is 5–10% lower for the mixture if compared to the pure compound. The main difference was observed for the iron valence. For common structural parameters, the oxidation state affects only the energy position of spectra. Linear combination of spectra smears the localized distributions of Fe^{2+} and Fe^{3+} points in the scatter plots (compare Figs. 5e and 8c, respectively). Two descriptors can provide the quality of valence discrimination in the mixture up to 80% and the use of three or more descriptors is appreciated. The better choice should consider the joint analysis of descriptors from several spectral regions. Thus, in ref. [43], the multivariate approach was applied to XANES spectra to determine the iron redox state in silicate glasses. It was demonstrated that using the full spectral region from the pre-edge to the EXAFS provides more accurate results. Pre-edge descriptors alone can be applied to the charge state analysis as well. Wilke et.al. demonstrated for the Fe *K*-XANES[44] that the pre-edge contains information both about the oxidation state and coordination number. The method analyses the 2d scatter plot of the integrated pre-edge intensities versus the pre-edge centroid positions. The set of reference spectra was distributed in the localized regions attributed to the 4, 5, and 6-coordinated Fe ions in oxidation state Fe^{2+} and Fe^{3+}. The limitations of this methodology arise from the need for well-defined reference spectra since pre-edge XANES simulations are still difficult for real systems. However, no references were reported with the CNs below 4.

Tables 5 and 6 demonstrate the best combinations of descriptors in terms of their quality calculated over the whole theoretical database or set of experimental references. The best triples of descriptors are different for these two tasks. The fact can be understood due to statistical considerations. The area of variation of parameters in the theoretical training set is large and includes even chemically irrelevant species, e.g., Fe^{3+} with distances longer than in Fe^{2+}. In contrast, the range of structural parameters covered by experimental spectra is smaller and represents a subclass of the training set. The R^2 score quality is evaluated in the cross-validation procedure and depends on the size of the sample and its dimensions. Therefore Table 6 contains also the mean absolute error evaluated along with R^2 score for the experimental validation set. The triples of good descriptors for experimental analysis are listed in Table 6 for each structural parameter: [WL_E, Pit_E, rPC_2] for CN, [WL_{int}, Pit_{int}, rPC_2] for Fe–O distance, [$Edge_E$, WL_E, PC_3] for valence. Figure 9 reports the predicted structural parameters for reference experimental spectra compared with their actual values. Prediction for all experimental spectra can be found in Supplementary Tables 8–10. The mean absolute errors over the validation set were 0.1 for oxidation state, 0.4 for CNs, and 0.03 for distances. The largest errors of the ML algorithm were observed for crystalline compounds, which have a significantly different structure from entries in the training set. The latter was adapted for Fe:SiO_2 systems and contains silicon in the second coordination shell, while some reference minerals are composed of oxygen and iron/Al/CO in the nearest coordination

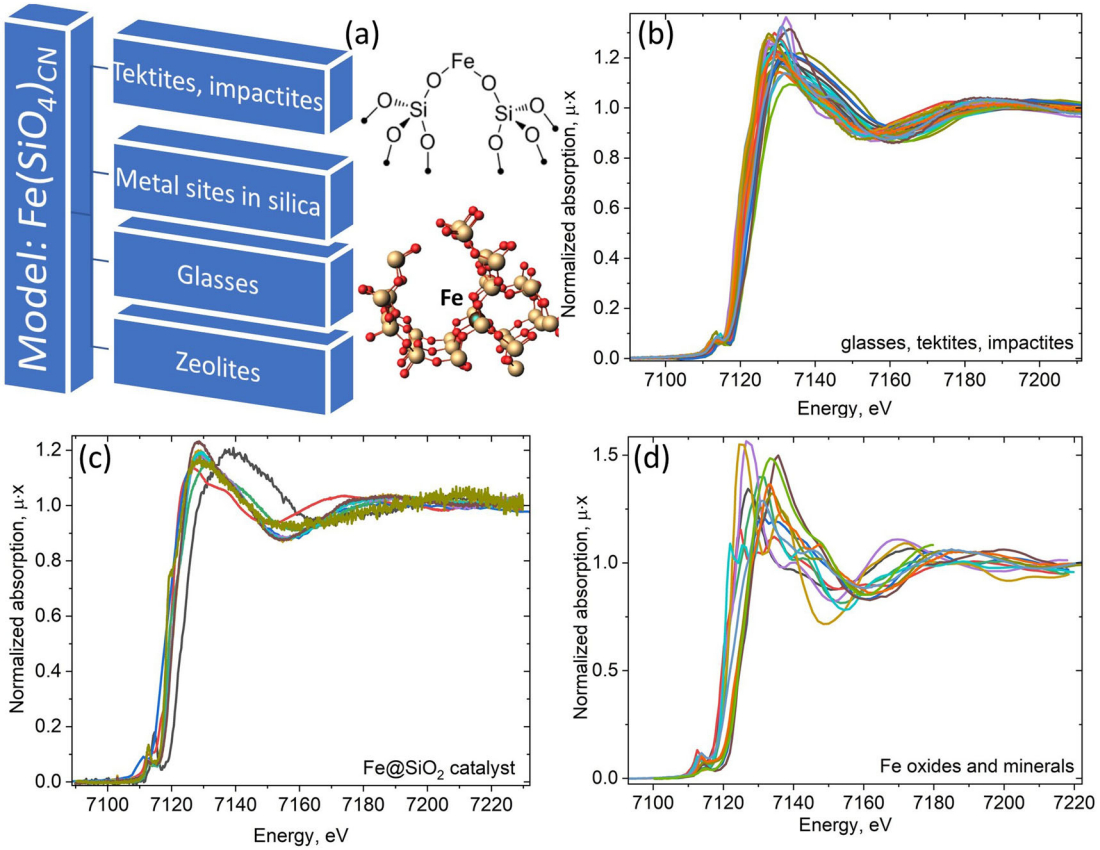

Fig. 6 Overview of the experimental validation dataset. The types of systems covered by the theoretical $Fe(SiO_4)_{CN}$ training set (**a**) and series of analyzed experimental Fe *K*-edge XANES spectra for (**b**) glasses, tektites, impactites, (**c**) single-site silica-supported Fe catalyst, (**d**) crystalline minerals. Only presented energy intervals of spectra were used for the analysis of Fe valence, Fe-O distances, and coordination numbers. See the complete list of studied samples and their description in Supplementary Tables 2–5.

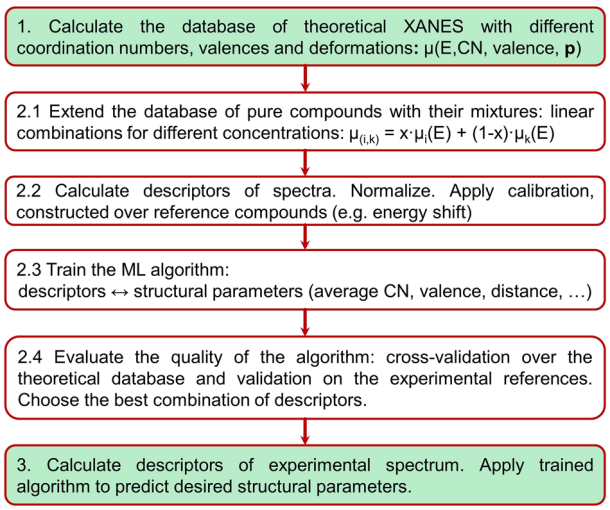

Fig. 7 The flowchart demonstrating how the descriptor analysis was applied to the mixture of spectra. Before training the algorithm, we normalize the descriptors to make them comparable (for a set of descriptors subtract the average and divide all entries by standard deviation over the set).

shells. The methodology can be directly applied to a new training set extended by ligands of different types. In this case, additional labels (e.g., atom types) should be added as descriptors of structure.

The obtained results for studied samples ("unknown" in Fig. 9) are in good agreement with other experimental methods. Mössbauer spectroscopy confirmed that the fraction of Fe^{3+} ions was larger in impactites (zhamanshinite, irghizite). A number of XAS-based[38,45] and non-XAS investigations have shown that iron oxidation state in tektites from different strew fields is about Fe^{2+} and generally $Fe^{3+}/\Sigma Fe$ ratio <0.15[42,46]. Iron in impact glasses can cover a wider range of Fe oxidation states[37,47,48] as compared with tektites, from purely Fe^{2+} to purely Fe^{3+}, and $Fe^{3+}/\Sigma Fe$ values are mainly within 0.25–0.59[42]. Fe–O distances are generally smaller for Fe^{3+} ions and we observed a similar trend for impactites as compared to tektites. Fe CNs in tektites is still a disputed issue. EXAFS studies have reported that mean Fe CNs in tektites are close to 4[45], whereas the coexistence of four and five-coordinated Fe was observed in[38]. Our estimations fall in the range CN = 3.5 ÷ 4.5, reproducing a similar trend as in EXAFS analysis. The absolute values of CN obtained from EXAFS analysis highly correlate with the Debye–Waller factor and can be affected also by self-absorption effects in the fluorescence regime of measurements (iron catalyst samples). Therefore, in the corresponding panel of Fig. 9, we omitted the expected values of CN to avoid confusion.

Fig. 10 represents the formation process of the single-site Fe catalyst on silica. The analysis for this system implies that Fe remains at oxidation state +2 throughout the process consistent with Mossbauer analysis and magnetic characterization[24]. It also shows that after grafting of the molecular precursor, dimeric Fe(II) tris(*tert*-butoxy) siloxide on SiO_2 dehydroxylated at 1080 °C, the coordination number of Fe – CN(Fe) – remains close to 4 ($Fe@SiO_2$ 1), whereas it decreases to 3 after thermolysis at 1020 °C ($Fe@SiO_2$ 2) consistent with previously reported characterization data that

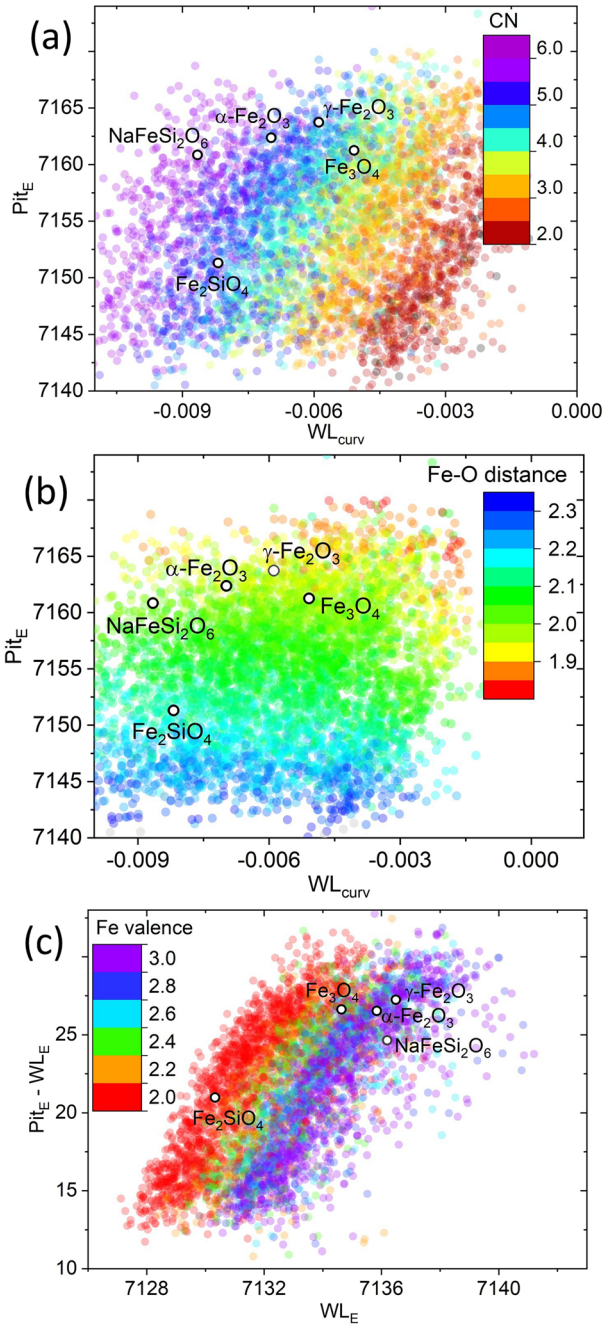

Fig. 8 Scatter plots for the selected pairs of descriptors for the library of mixtures. Descriptors were calculated for the theoretical database extended with the linear combinations of the spectra from Fig. 4. The experimental data for selected references (hollow circles) were projected onto each plot. Compare to the analogous Fig. 5 constructed for pure compounds.

Table 5. Descriptor performance over the database of linear combinations (mixtures) of theoretical spectra from the training set in Fig. 4a.

Structural parameter	Best combination of 3 descriptors	R^2 score
CN	$Edge_E$,Pit_E –WL_E,WL_{curv}	0.70
Average Fe–O distance	$Edge_E$, WL_E, Pit_E	0.90
Fe oxidation state	$Edge_E$, Pit_E, WL_{int}	0.80

R^2 score is obtained in a 10-fold cross-validation procedure.

Table 6. Descriptor performance over the database of experimental references.

Structural parameter	Best combination of 3 descriptors	R^2 score	Mean error
CN	WL_E, Pit_E, rPC_2	0.80	0.4
Average Fe–O distance	WL_{int}, Pit_{int}, rPC_2	0.85	0.03
Fe oxidation state	$Edge_E$, WL_E, PC_3	0.85	0.1

R^2 score and mean error are obtained by comparing the known experimental values and predicted values by the algorithm trained over the database of linear combinations of theoretical spectra. See Fig. 9 and Supplementary Tables 8–10 for details.

show a similar decrease of CN(Fe), albeit to a value of 2[24]. This confirms that thermal treatment leads to Fe(II) species with low coordination number, probably situated between 2 and 3. It is noteworthy the sample prepared at lower temperature both for the hydroxylation and thermolysis steps display Fe sites with a larger coordination number of 4.

As a concluding remark, we note that usually, the ML algorithms work as a "black box" for researchers since it is difficult to understand what structural information is contained in each part of the spectrum. We approach such understanding by using selected descriptors of the spectrum instead of individual points. The whole spectrum is substituted by several descriptors that intuitively characterize its shape, i.e. energy position of edge, minima, maxima, their intensities, and curvatures. Machine learning analysis established the rational choice of the combinations of descriptors providing the highest prediction accuracy for the structural parameters both for pure compounds and their mixtures. To visualize the spectrum-structure relations we use scatter plots and derive analytical dependencies between the descriptors of the spectrum and structural parameters.

Rational choice of descriptors isolates those features of spectra that are most sensitive to specific structural parameters, avoiding fitting the whole spectrum. The major problem of the practical application of ML methods for experimental data analysis arises from the systematic differences between theoretical calculations and measured data. This discrepancy can arise either from limitations of the theoretical approach or the experimental artefacts. The benefit of using descriptors over the full-spectrum stands in the possibility to correct the systematic differences by calibration on a dataset of theoretical and experimental spectra of reference compounds. However, as all methodologies based on supervised learning, our results are limited to the family of structures described by the training set. As an illustration, the algorithm was trained on Fe-O-Si system; it will thus fail for predicting proper parameters for metallic Fe or sulfide compounds that belong to very different types of materials. This certainly calls for expanding the training set in order to allow for distinguishing, for instance, the ligand types apart from coordination number or interatomic distances.

The further development of the approach is directed toward new ways of descriptor evaluation. A complete set of descriptors should provide the same amount of structural information as in a full spectrum. We foresee that a combination of descriptors from complementary experimental methods (nuclear magnetic resonance, electron paramagnetic resonance, X-ray diffraction, etc.) would significantly improve the quality of prediction.

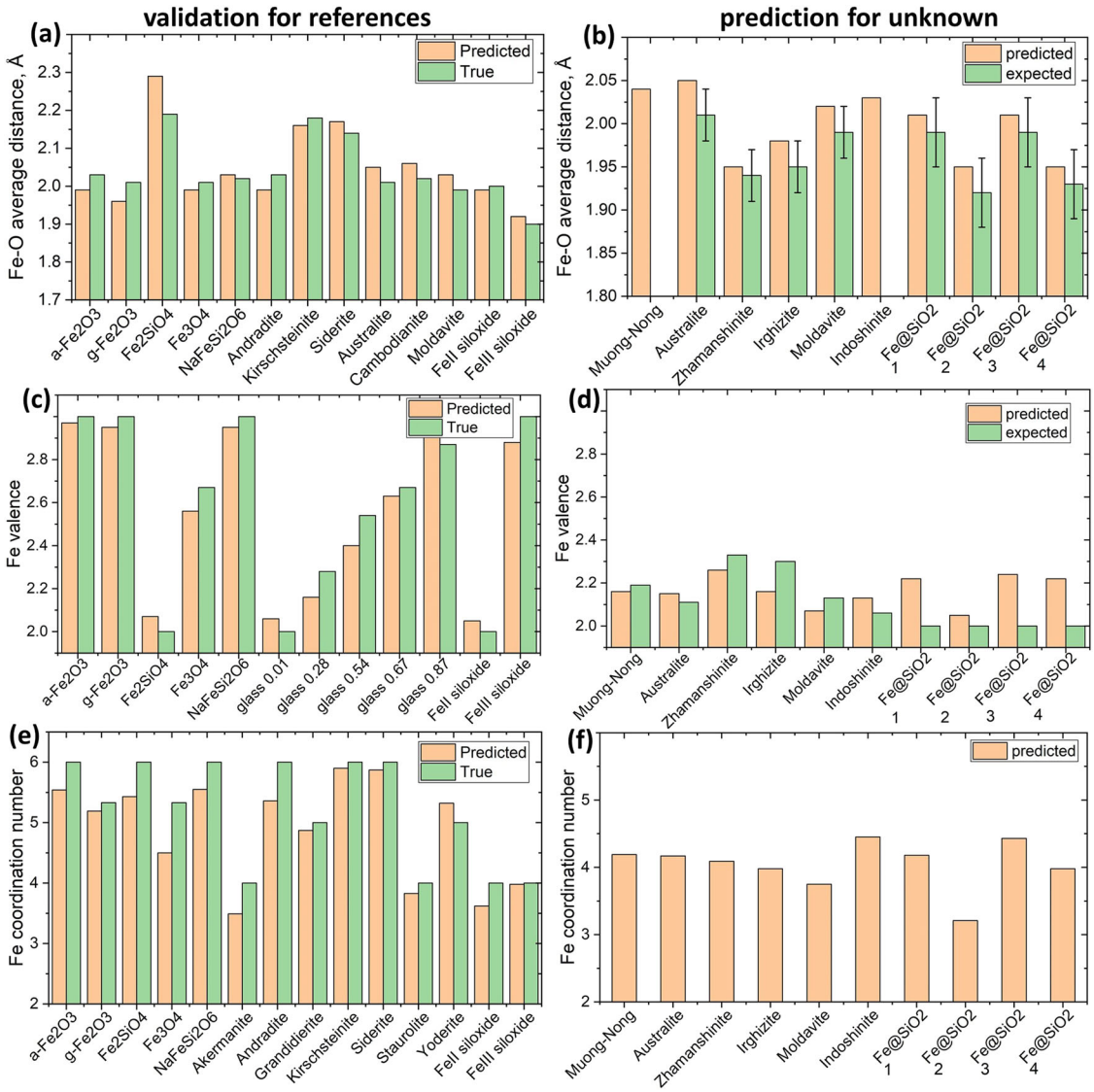

Fig. 9 Experimental validation of the Extra Trees algorithm trained over the theoretical data set. Prediction is based on three descriptors: [WL_{int}, Pit_{int}, rPC_2] for distances (**a**, **b**), [$Edge_E$, WL_E, PC_3] for iron valence (**c**, **d**) and [WL_E, Pit_E, rPC_2] for CN (**e**, **f**). The green bars in **a**, **c**, **e** are the values reported in the literature, and ones in **b**, **d** is the results of EXAFS and Mossbauer analysis performed independently by present authors. The error bars in **b** indicate the range of uncertainties provided by the EXAFS analysis. See also Supplementary Notes section for details.

METHODS

XANES simulations and energy alignment

Fe K-edge XANES spectra were calculated utilizing the full potential finite difference method[49] implemented in the FDMNES software[50]. The photoelectron wave functions were evaluated on a grid of points in a 5.5 Å sphere around the absorbing atom with 0.2 Å interpoint distance. To account for the core-hole lifetime broadening and instrumental energy resolution, theoretical spectra were further convoluted using the arctangent function to model the energy dependence of the Lorentzian width.

For an accurate energy calibration of the spectra, the iron 1 s core level energy shifts between Fe^{2+} and Fe^{3+} oxidation states for each coordination number were estimated within the molecular orbital approach. The energy levels and the corresponding wave functions were calculated by density functional theory using the B3LYP exchange-correlation functional[51]. The largest available QZ4P basis set implemented in the ADF-2019 software[52,53] was used. For every coordination number in the range between two and six, we constructed a symmetric complex with Fe-O distances equal to 2 Å and evaluated transition matrix elements in the 50 eV energy interval both for the Fe^{2+} and Fe^{3+} oxidation states. The proper oxidation state was achieved by specifying the charge and spin state of the whole complex. After the convergence of the self-consistent procedure was achieved the charge states of iron atoms were confirmed by Mulliken charge analysis. Chemical shifts of the 1 *s* core levels were evaluated and applied to the spectrum calculated by the finite difference method. In this way, we simulated absorption from Fe^{2+} and Fe^{3+} sites for given values of structural parameters.

Machine learning algorithms

When we apply machine learning based on spectrum descriptors (calculate the quality of labels prediction, predict labels for experimental data) we use Extra Trees regressor or classifier models[54]. It consists of several randomly generated decision trees. A decision tree represents a flowchart of threshold conditions on parameters and divides the parameter space into non-intersecting rectangles, in each of which, for regression, the objective function μ(E, P) is approximated by a linear one using the least-squares method and for classification - probability table is calculated. The results obtained from several trees are averaged.

For XANES approximation (Supplementary Fig. 8) we use a supervised machine learning algorithm based on the Radial Basis Functions (RBF) that construct a continuous approximation of spectrum, μ(E), as a function of structural parameters $\boldsymbol{P} = (p_1, p_2, \ldots, p_k)$. The RBF method is a well-proven

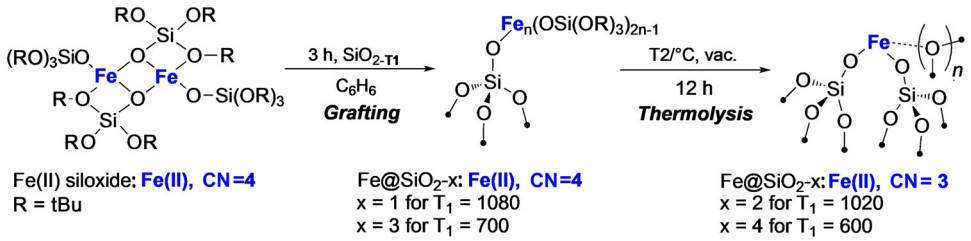

Fig. 10 The grafting and thermolysis process for highly dehydroxylated SiO_2. The iron coordination number and local symmetry change upon thermal treatment depending on the temperature and atmosphere.

mesh-free method[55–57]. The unknown function $\hat{\mu}(E, \boldsymbol{P})$ is represented in terms of a set of basis functions characterized by certain factors and polynomial terms as follows:

$$\hat{\mu}(E,\boldsymbol{P}) = \sum_{i=1}^{N} w_i(E) \cdot K(||\boldsymbol{P} - \boldsymbol{P}_i||) + \text{Polynomial}_E(\boldsymbol{P}) \quad (3)$$

where $K(r)$ is the radial basis function, $\text{Polynomial}_E(\boldsymbol{P})$ is a polynomial function of k-dimensional vector of structural parameters $\boldsymbol{P}$ with energy-dependent coefficients. The training set is composed of N calculated spectra. The points ($N = 600$ for each structure in Fig. 3) in the space of structural parameters $\boldsymbol{P}$ were chosen according to the IHS[58]. The unknown factors w_i and the polynomial coefficients are obtained by the ridge quadric regression method. Every basis function is a function of distance from the training set point $\boldsymbol{P}_i$. In our task, good results were obtained using linear basis functions and a second-order polynomial (see also Supplementary Table 1 for comparison with other ML methods).

It is important to define a proper norm in (1) to measure the distance between $\boldsymbol{P}$ and $\boldsymbol{P}_i$ for a good quality of the approximation. Structural parameters $p_1, p_2, \ldots, p_k$ have a different scale, e.g., interatomic distances and angles. Moreover, the variation of the target function, $\hat{\mu}(E, p_1, p_2, \ldots, p_k)$, greatly varies for different structural parameters. Spectrum changes caused by angle transformation are an order of magnitude less than caused by interatomic distance modification. That's why we estimate first the average partial variance of the target function ($\Delta_i\mu$) for each p_i and rescale structural parameters in the following way:

$$p_i' = p_i \frac{\Delta_i \mu}{maxp_i - minp_i} \quad (4)$$

The quality of approximation and prediction is calculated during 10-fold cross-validation. The training set, composed of spectra (the task of XANES approximation as a continuous function of structural parameters) or descriptors (the task of structural parameters prediction based on several spectral features) is divided randomly into 10 parts, nine of which are used for algorithm training and the tenth for validation. The quantitative measure of the quality is the R^2 score for the regression task and accuracy for the classification. Details of their evaluation are described in Supplementary Methods section, while supplementary Jupyter Notebook reports the steps necessary to repeat the calculations in the manuscript Fig. 10.

Section 2.4 of the main text deals with multicomponent systems. The algorithm training is then performed on the linear combinations instead of pure theoretical spectra. In total, more than 5000 pairs were constructed for randomized fractions of components with different CNs, valences, and Fe-O distances. The flowchart in Fig. 7 describes the details of the procedure for mixture analysis. We found the prediction quality may be improved for reference experimental data when sampling was performed according to the adaptive sampling scheme. Although the IHS scheme provides the uniform sampling over each structural parameter the adaptive sampling (or active learning)[59,60] chooses the points in the training sample to ensure a uniform variation of the XANES in the selected region of structural parameters. Both training sets are available as SI.

DATA AVAILABILITY

The data that support the findings of this study are available at the repository https://github.com/gudasergey/XANES_descriptors along with the source code.

CODE AVAILABILITY

The source code and executable Jupyter Notebook used to train the models and generate the figures in this publication are publicly available at the repository https://github.com/gudasergey/XANES_descriptors.

Received: 7 January 2021; Accepted: 6 November 2021;
Published online: 13 December 2021

REFERENCES

1. Calvin, S. *XAFS for Everyone*, (Taylor & Francis, 2013).
2. Henderson, G. S., de Groot, F. M. F. & Moulton, B. J. A. X-ray absorption near-edge structure (XANES) spectroscopy. *Rev. Mineral. Geochem.* **78**, 75–138 (2014).
3. Lamberti, C. & van Bokhoven, J. A. Introduction: historical perspective on XAS. In *X-Ray Absorption and X-Ray Emission Spectroscopy* 1–21 (John Wiley & Sons Ltd, 2016).
4. de Groot, F., Vanko, G. & Glatzel, P. The 1s x-ray absorption pre-edge structures in transition metal oxides. *J. Phys. Condens. Matter* **21**, 104207 (2009).
5. Westre, T. E. et al. A multiplet analysis of Fe K-edge 1s->3d pre-edge features of iron complexes. *J. Am. Chem. Soc.* **119**, 6297–6314 (1997).
6. Wilke, M., Farges, F., Petit, P. E., Brown, G. E. & Martin, F. Oxidation state and coordination of Fe in minerals: an FeK-XANES spectroscopic study. *Am. Mineral.* **86**, 714–730 (2001).
7. Zhang, R. Q. & McEwen, J. S. Local environment sensitivity of the Cu K-Edge XANES features in Cu-SSZ-13: analysis from first-principles. *J. Phys. Chem. Lett.* **9**, 3035–3042 (2018).
8. Oyanagi, H. et al. Small copper clusters studied by x-ray absorption near-edge structure. *J. Appl. Phys.* **111**, 084315 (2012).
9. Gombac, V. et al. CuOx-TiO2 photocatalysts for H-2 production from ethanol and glycerol solutions. *J. Phys. Chem. A* **114**, 3916–3925 (2010).
10. Natoli, C. R. Distance Dependence of Continuum and Bound State of Excitonic Resonances in X-ray absorption near-edge structure (XANES). In *EXAFS and Near Edge Structure III. Springer Proceedings in Physics*, **2**. 38–42 (Springer, 1984).
11. Arcon, I., Mirtic, B. & Kodre, A. Determination of valence states of chromium in calcium chromates by using X-ray absorption near-edge structure (XANES) spectroscopy. *J. Am. Chem. Soc.* **81**, 222–224 (1998).
12. Glatzel, P., Smolentsev, G. & Bunker, G. The electronic structure in 3d transition metal complexes: can we measure oxidation states? *J. Phys. Conf. Ser.* **190**, 012046 (2009).
13. Chaboy, J., Munoz-Paez, A., Carrera, F., Merkling, P. & Marcos, E. S. Ab initio x-ray absorption study of copper K-edge XANES spectra in Cu(II) compounds. *Phys. Rev. B* **71**, 134208 (2005).
14. Zheng, C., Chen, C., Chen, Y. & Ong, S. P. Random forest models for accurate identification of coordination environments from X-ray absorption near-edge. *Struct. Patterns* **1**, 100013 (2020).
15. Liu, Y. et al. Mapping XANES spectra on structural descriptors of copper oxide clusters using supervised machine learning. *J. Chem. Phys.* **151**, 164201 (2019).
16. Timoshenko, J., Lu, D. Y., Lin, Y. W. & Frenkel, A. I. Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles. *J. Phys. Chem. Lett.* **8**, 5091–5098 (2017).
17. Rankine, C. D., Madkhali, M. M. M. & Penfold, T. J. A deep neural network for the rapid prediction of X-ray absorption spectra. *J. Phys. Chem. A* **124**, 4263–4270 (2020).
18. Martini, A. et al. PyFitit: The software for quantitative analysis of XANES spectra using machine-learning algorithms. *Comput. Phys. Commun.* **250**, 107064 (2019).
19. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 83 (2019).

20. Trejo, O. et al. Elucidating the evolving atomic structure in atomic layer deposition reactions with in situ XANES and machine learning. *Chem. Mater.* **31**, 8937–8947 (2019).
21. Carbone, M. R., Yoo, S., Topsakal, M. & Lu, D. Y. Classification of local chemical environments from x-ray absorption spectra using supervised machine learning. *Phys. Rev. Mater.* **3**, 033604 (2019).
22. Carbone, M. R., Topsakal, M., Lu, D. Y. & Yoo, S. Machine-learning X-ray absorption spectra to quantitative accuracy. *Phys. Rev. Lett.* **124**, 156401 (2020).
23. Torrisi, S. B. et al. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *npj Comput. Mater.* **6**, 109 (2020).
24. Sot, P. et al. Non-oxidative methane coupling over silica versus silica-supported iron(II) single sites. *Chem. Eur. J.* **26**, 8012–8016 (2020).
25. Pak, C., Bell, A. T. & Tilley, T. D. Oxidative dehydrogenation of propane over vanadia-magnesia catalysts prepared by thermolysis of OV((OBu)-Bu-t)(3) in the presence of nanocrystalline MgO. *J. Catal.* **206**, 49–59 (2002).
26. Coperet, C. et al. Surface organometallic and coordination chemistry toward single-site heterogeneous catalysts: strategies, methods, structures, and activities. *Chem. Rev.* **116**, 323–421 (2016).
27. Bugaev, A. L. et al. Temperature- and pressure-dependent hydrogen concentration in supported PdHx nanoparticles by Pd K-edge X-ray absorption spectroscopy. *J. Phys. Chem. C.* **118**, 10416–10423 (2014).
28. Bugaev, A. L., Srabionyan, V. V., Soldatov, A. V., Bugaev, L. A. & van Bokhoven, J. A. The role of hydrogen in formation of Pd XANES in Pd-nanoparticles. *J. Phys. Conf. Ser.* **430**, 012028 (2013).
29. Bugaev, A. L. et al. Hydride phase formation in carbon supported palladium hydride nanoparticles by in situ EXAFS and XRD. *J. Phys. Conf. Ser.* **712**, 012032 (2016).
30. Zhang, L. J. et al. Extraction of local coordination structure in a low-concentration uranyl system by XANES. *J. Synchrotron Rad.* **23**, 758–768 (2016).
31. Bailey, D. H. Integer relation detection. *Comput. Sci. Eng.* **2**, 24–28 (2000).
32. Bailey, D. H. et al. *Experimental Mathematics in Action* (CRC Press, 2007).
33. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
34. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288 (1996).
35. Giuli, G., Paris, E., Pratesi, G., Koeberl, C. & Cipriani, C. Iron oxidation state in the Fe-rich layer and silica matrix of Libyan Desert Glass: a high-resolution XANES study. *Meteorit. Planet. Sci.* **38**, 1181–1186 (2003).
36. Berry, A. J., O'Neill, H. S., Jayasuriya, K. D., Campbell, S. J. & Foran, G. J. XANES calibrations for the oxidation state of iron in a silicate glass. *Am. Mineral.* **88**, 967–977 (2003).
37. Giuli, G., Eeckhout, S. G., Paris, E., Koeberl, C. & Pratesi, G. Iron oxidation state in impact glass from the K/T boundary at Beloc, Haiti, by high-resolution XANES spectroscopy. *Meteorit. Planet. Sci.* **40**, 1575–1580 (2005).
38. Wang, L. et al. Local structure of iron in tektites and natural glass: an insight through X-ray absorption fine structure spectroscopy. *J. Mineral. Petrol. Sci.* **108**, 288–294 (2013).
39. Holland, A. W. et al. New Fe/SiO2 materials prepared using diiron molecular precursors: synthesis, characterization and catalysis. *J. Catal.* **235**, 150–163 (2005).
40. Artemieva, N. High-velocity impact ejecta: tektites and martian meteorites. *In Catastrophic Events Caused by Cosmic Objects* 267–289 (Springer, Dordrecht, 2008).
41. Moretti, R. & Ottonello, G. Polymerization and disproportionation of iron and sulfur in silicate melts: insights from an optical basicity-based approach. *J. Non Cryst. Solids* **323**, 111–119 (2003).
42. Lukanin, O. A. & Kadik, A. A. Decompression mechanism of ferric iron reduction in tektite melts during their formation in the impact process. *Geochem. Int.* **45**, 857–881 (2007).
43. Dyar, M. D., McCanta, M., Breves, E., Carey, C. J. & Lanzirotti, A. Accurate predictions of iron redox state in silicate glasses: a multivariate approach using X-ray absorption spectroscopy. *Am. Mineral.* **101**, 744–747 (2016).
44. Wilke, M., Farges, F. O., Petit, P.-E., Brown, G. E. Jr. & Martin, F. O. Oxidation state and coordination of Fe in minerals: an Fe K-XANES spectroscopic study. *Am. Mineral.* **86**, 714–730 (2001).
45. Giuli, G., Pratesi, G., Cipriani, C. & Paris, E. Iron local structure in tektites and impact glasses by extended X-ray absorption fine structure and high-resolution X-ray absorption near-edge structure spectroscopy. *Geochim. Cosmochim. Acta* **66**, 4347–4353 (2002).
46. Giuli, G. Tektites and microtektites iron oxidation state and water content. *Rend. Lincei Sci. Fis. Nat.* **28**, 615–621 (2017).
47. Giuli, G., Eeckhout, S. G., Koeberl, C., Pratesi, G. & Paris, E. Yellow impact glass from the K/T boundary at Beloc (Haiti): XANES determination of the Fe oxidation state and implications for formation conditions. *Meteorit. Planet. Sci.* **43**, 981–986 (2008).
48. Kravtsova, A. N. et al. Iron oxidation state of impact glasses from the Zhamanshin crater studied by X-ray absorption spectroscopy. *Radiat. Phys. Chem.* **175**, 108097 (2020).
49. Joly, Y. X-ray absorption near-edge structure calculations beyond the muffin-tin approximation. *Phys. Rev. B* **63**, 125120 (2001).
50. Guda, S. A. et al. Optimized finite difference method for the full-potential XANES simulations: application to molecular adsorption geometries in mofs and metal-ligand intersystem crossing transients. *J. Chem. Theory Comput.* **11**, 4512–4521 (2015).
51. Reiher, M., Salomon, O. & Hess, B. A. Reparameterization of hybrid functionals based on energy differences of states of different multiplicity. *Theor. Chem. Acc.* **107**, 48–55 (2001).
52. Guerra, C. F., Snijders, J. G., te Velde, G. & Baerends, E. J. Towards an order-N DFT method. *Theor. Chem. Acc.* **99**, 391–403 (1998).
53. te Velde, G. et al. Chemistry with ADF. *J. Comput. Chem.* **22**, 931–967 (2001).
54. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
55. Fasshauer, G. E. *Meshfree Approximation Methods with Matlab*, **6** (WORLD SCIENTIFIC, 2007).
56. Myers, D. E. Smoothing and interpolation with radial basis functions. In *Boundary Element Technology Xiii: Incorporating Computational Methods and Testing for Engineering Integrity* **2**, 365–374 (WIT Press, 1999).
57. Wendland, H. Computational aspects of radial basis function approximation. In *Studies in Computational Mathematics*, Vol. **12**, 12231–256 (Elsevier, 2006).
58. Beachkofski, B. & Grandhi, R. Improved distributed hypercube sampling. In *43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference* (2002).
59. Fuhg, J. N., Fau, A. & Nackenhorst, U. State-of-the-art and comparative review of adaptive sampling methods for kriging. *Arch. Comput. Methods Eng.* **28**, 2689–2747 (2021).
60. Liu, H., Ong, Y.-S. & Cai, J. A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design. *Struct. Multidiscipl. Optim.* **57**, 393–416 (2018).

ACKNOWLEDGEMENTS

A. Guda acknowledges the financial support from the Russian Foundation for Basic Research (project number 20-32-70227) for the work on the multicomponent mixtures. A. Bugaev and A.V. Soldatov acknowledge the Russian Science Foundation grant #20-43-01015 for the financial support for the work on the spectral descriptors. Authors acknowledge D.D. Badyukov from Vernadsky Institute of Geochemistry and Analytical Chemistry of Russian Academy of Sciences for providing samples for analysis. P. Šot acknowledges the Shell Global Solutions International, B.V. for funding the work on the synthesis of Fe-containing catalyst, and European Synchrotron Research Facility for awarded beamtimes at beamlines ID26, BM25, and Swiss Light Source for the beamtime at SuperXAS beamline.

AUTHOR CONTRIBUTIONS

A.A.G., A.M., and S.A.G. contributed equally and developed the concept of the approach, selected a set of descriptors for spectra and wrote the manuscript. A.A. contributed to the open-source PyFitIt code development within Jupyter Notebooks interface. A.V.S. and A.B. designed the theoretical structures, calculated training sets, and performed calibration. A.N.K., L.V.G., and S.P.K. performed experimental characterization of references and studied tektites and impactites, interpreted results for these samples from machine learning analysis. P.Š., J.A.v.B., and C.C. performed the synthesis and measurements of Fe K-edge XANES for Fe@SiO_2 catalyst. A.V.S. supervised the project and provided guidance. All authors provided contributions to the manuscript and discussed the results.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00664-9.

Correspondence and requests for materials should be addressed to A. A. Guda, S. A. Guda or A. Martini.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.