Supplementary information for the manuscript

Fingerprint Analysis of X-Ray Absorption Spectra With Machine-Learning Method Trained on Multielement Experimental Library

B. O. Protsenko[1], Y. Kakiuchi[2], S. A. Guda[1,3], D. Trummer[2], A. Zabilska[4], S. Shapovalova[1], A. V. Soldatov[1], O. V. Safonova*[4], C. Copéret*[2], A. A. Guda*[1]

[1] The Smart Materials Research Institute, Southern Federal University, 344090 Rostov-on-Don, Russia
[2] Department of Chemistry and Applied Biosciences, ETH Zürich, CH-8093 Zürich, Switzerland
[3] Institute of Mathematics, Mechanics and Computer Science, Southern Federal University, 344090 Rostov-on-Don, Russia
[4] Paul Scherrer Institute, 5232 Villigen, Switzerland

*corresponding authors: guda@sfedu.ru, olga.safonova@psi.ch, ccoperet@ethz.ch

Contents

1. Definition of spectral descriptors

X-ray absorption spectrum is a complex multidimensional object that consists of hundreds of points. Structural information is hidden in the spectral features that often overlap, and in general there are no well-defined rules to decipher the spectrum as tabulated for FTIR, NMR or cell parameters in XRD. Based on theoretical datasets, ML may help to understand the information contained in different regions of spectra. Torrisi et al. [1] applied random forest ML model to find relationship between polynomial descriptors of spectrum and material properties in terms of Bader charge and mean nearest neighbor distance materials. Martini et.al. applied principal component descriptors for the step-by-step fit of XAS spectrum refining first such distortions that affect the spectrum in a stronger way [2]. The descriptor-based approach was found to be useful for studying analytical relationships between spectral and structural descriptors even for amorphous materials [3]. Those combinations of descriptors that demonstrate the highest quality in cross-validation may be further applied for the speciation of local coordination of the catalyst in the reaction mixture as demonstrated for Ru molecular complexes [4]. Herein, we employ several sets of descriptors, based on different approaches to the featurization of spectra for machine learning and visualization.

Table S1. Combinations of descriptors applied in the main text. The energy intervals are given in the Cr K-edge photon energy units.

#	Name	Description
1	Spectrum	All points if spectrum in the energy interval [5980…6080] eV and values of the first derivative in the same energy interval
2	PCA	Three first PCA components evaluated from the merged library Cr+V in the energy range [5980…6080] eV of shifted library
3	t-SNE	Two components of the t-SNE dimensionality reduction procedure, applied to the spectra in range of [5980…6080] eV and few strong spectral features (*vide infra*)
4	Pre-edge	Pre-edge centre and area after the subtraction of baseline

On Table S2 one can see the full list of implemented spectral descriptors and their short description.

Table S2. Full list of descriptors of spectra constructed from selected spectral features or mathematical processing of the whole set of spectra aiming to improve sensitivity to a target property. The energy intervals are given in the Cr K-edge photon energy units.

Short Name	Full name	Details
1. Descriptors based on spectral features		
Spec	Spectrum	Spectrum itself in the form of the intensity vector in range of [5980…6080] eV.
$Edge_E$	Energy of the absorption edge	Calculated as an inflection point in the arctangent function that fits the whole spectrum.
$Edge_{Slope}$	Slope of the rising edge of the spectrum	The slope of the arctangent function in the inflection point.
PE_{area}	area under the pre-edge	The pre-edge region is fitted by a baseline and after subtraction the area is calculated. Pre-

		edge should not be mixed with shoulder for square planar complexes and transitions to the delocalized d-band[5].
PE_{center}	centroid energy of the pre-edge	Pre-edge centroid position and area after subtracting the baseline.
Value at E energy	Intensity of the selected spectrum point at the specific energy E	Impurity-based feature importance of the trained ExtraTrees model is used to select the most valuable spectrum points and use them for further analysis. Four values of spectra at specific energies: 5995, 6003, 6037, 6043 eV for shifted, common, energy scale.
1st maximum	Centroid energy and intensity of the main maximum	Centroid position (energy and intensity) for the XANES curve (not area in contract to pre-edge) in the region Intensity > 0.6 Energy < EFermi+20. Such definition is more stable than single maximum since the latter may be splitted into several small peaks.
2. Database related descriptors		
PCA	Projections on the principal components	Spectra are projected onto three first principal components, calculated in the region from 5980 to 6120 eV for whole dataset after the edge shift.
Scaled PCA	Normalized PCA components	Three PCA components from PCA analysis applied to scaled XANES spectra (at each point of spectrum subtracted average and normalized by dispersion over the whole database)
t-SNE	t-distributed stochastic neighbour embedding	Nonlinear projection of high-dimensional spectra onto n-dimensional (here we use n=2) space constrained by similarity between probability distributions for original and projected data. The Euclidean norm is used to estimate distance between objects. Six values from the two-dimensional t-SNE decompositions for: - five PCA components of spectra; - five PCA components for scaled spectra; - four points of spectra as specified above.
2.1 Descriptors related to the target-property		
BL	Best linear combination of points in terms of prediction quality for a given property	Linear support vector machine (SVR) is applied to selected points of spectra to a construct a one-dimensional embedding that is of a high quality for the prediction of metal charge state or coordination number. • linear combination from two values [first maximum E, 1st maximum intensity] for Formal charge

		• linear combination from four values of spectrum for target property CN
PLS	Partial least squares	A PLS model finds the linear combination of spectral points that explains the maximum variance of the target property: charge or coordination number. We construct PLS descriptors from the following datasets: - full spectra in the [5980…6080] eV energy interval correspondingly using coordination number as target property - descriptors of spectra except t-SNE, PLS, BL using metal valence as target property - descriptors of spectra except t-SNE, PLS, BL using coordination number as target property

For unknown spectra all database-related features are calculated by adding them to the dataset of library references and direct application of the methods, except BL and PLS components, where representations, constructed on the library of references, are used to project the unknown spectra.

PCA descriptors

In Principal Component Analysis (PCA) XANES dataset $\mathbf{X}$ is decomposed in the following way:

$$\mathbf{X} = \mathbf{U}_{(\mathrm{m}\times\mathrm{m})}\mathbf{\Sigma}_{(\mathrm{m}\times\mathrm{n})}\mathbf{V}^{\mathrm{T}}{}_{(\mathrm{n}\times\mathrm{n})} \quad (1)$$

where $\mathbf{X}$ can be considered as a matrix of dimensions $(m \times n)$, where *m* is the number of XANES energy points while *n* is the number of theoretical spectra constituting it, $\mathbf{U}$ and $\mathbf{V}$ are two square unitary matrices, $\mathbf{\Sigma}$ is a diagonal rectangular matrix while T denotes the transpose operator. The diagonal elements of $\mathbf{\Sigma}$ are referred as the dataset singular values, whose magnitudes are proportional on the amount of variance of the related component. Columns in matrix $\mathbf{U}$ have the dimensionality of a XANES spectrum and are referred as abstract mathematical components (i.e. they do not look line spectra, but their proper linear combination does). Matrix $\mathbf{\Sigma V}$ provides the weights which need to be employed to reconstruct each spectrum of $\mathbf{X}$ from $\mathbf{U}$. Considering equation (1), the *i*th XANES spectrum $\boldsymbol{\mu}_{\mathrm{i}}$ of $\mathbf{X}$, can be rewritten as:

$$\boldsymbol{\mu}_{\mathrm{i}}(\mathbf{E},\mathbf{p}) = \sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{h}_{\mathrm{ij}}(\mathbf{p})\mathbf{u}_{\mathrm{j}}(\mathbf{E}) \quad (2)$$

where $\mathbf{E} = (\mathrm{E}_1,\ldots\mathrm{E}_{\mathrm{m}})$ is the set of XANES energy points, $\mathbf{u}_{\mathrm{j}}$ represents the j$^{\mathrm{th}}$ column vector of $\mathbf{U}$ while h_{ij} is the fraction of the *j*th component in the *i*th spectrum provided by matrix $\mathbf{\Sigma V}$. We explicitly introduce in equation (2) dependence of $\boldsymbol{\mu}_{\mathrm{i}}$ and hence h_{ij} on structural parameters **p**, since the theoretical training set is obtained by structural deformations, i.e. variation of **p**. It follows that each coefficient h_{ij}, here named as XANES multipliers, can

be considered as the projection of $\boldsymbol{\mu}_\mathrm{i}$ over $\mathbf{u}_\mathrm{j}$. Because the dataset components $\mathbf{u}_\mathrm{j}$ are common for every spectrum in $\mathbf{X}$, the dependence of XANES of parameters **p** resides in its multipliers h_{ij}. The latter, in this way, constitute a new class of descriptors. For a given experimental spectrum $\boldsymbol{\mu}^{\mathrm{exp}}$ PCA descriptors are calculated in the following way. The mean of the dataset is subtracted from the experimental spectrum first. Then the scalar product among the mean-corrected spectrum $\tilde{\boldsymbol{\mu}}^{\mathrm{exp}}$ and each of the selected columns ($\mathbf{u}_\mathrm{i}$) of $\mathbf{U}$ is calculated: $\mathrm{h}_\mathrm{i}^{\mathrm{exp}} = \tilde{\boldsymbol{\mu}}^{\mathrm{exp}} \cdot \mathbf{u}_\mathrm{i}$. The first multiplier $\mathrm{h}_{\mathrm{i}=1}^{\mathrm{exp}}$ (first PCA-descriptor) will be the most intense while the subsequent PCA-descriptors will be sorted in descending order of intensity of XANES variation.

t-SNE descriptors

The t-Distributed Stochastic Neighbor Embedding (t-SNE), contrary to the PCA, represents a popular nonlinear dimensionality reduction technique, routinely used to visualize complex multidimensional data. One of the simple heuristics behind the t-SNE method is that it nonlinearly "projects" data points to a lower dimensional space, keeping the "proximity" relation of data points: ones, that are "neighbors" in the original space, remain neighbors in the projected space, whilst distant points remain distant compared to others after the dimensionality reduction [6].

In t-SNE, original datapoints **X** (XAS spectra or its features) are embedded in the space **Y** (t-SNE descriptors) by optimizing the Kullback–Leibler divergence:

$$C = KL(\mathbf{P}||\mathbf{Q}) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3}$$

between probability distributions of the data points in the original **X** and embedding **Y** spaces, **P** and **Q** respectively. These distributions, in turn, are represented by joint probabilities p_{ij} and $q_{ij,}$ obtained by symmetrization

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2}, \;\; q_{ij} = \frac{q_{j|i} + q_{i|j}}{2} \tag{4}$$

of probability densities of the dataset points in **X** and **Y**, modelled by Gaussian and Student's t-distributions

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp\left(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2\right)}, \; q_{ij} = \frac{(1+||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k \neq l} (1+||\mathbf{y}_k - \mathbf{y}_l||^2)^{-1}} \tag{5}$$

where values σ_i are usually set such that each Gaussian kernel fits *k* nearest neighbors within one standard deviation of the probability density[7]. The *C* optimization method depends on the specific implementations of the algorithm and additional assumptions, usually introduced to speed up the calculation (for example, interpolation of Barnes-Hut t-SNE). In PyFitIt we mostly use python openTSNE implementation[7]. This method is affected by the curse of dimensionality. Therefore, we apply t-SNE not to the whole spectra but to their first five principal components derived from PCA analysis or to four selected points in the spectrum selected according to their importance

PLS descriptors

When we use ML algorithm to predict label based on spectra data, we train it on the matrix in which a row is a spectrum and a column is all the spectra values at some energy point. Due to spectrum continuity the neighbor columns of the training matrix are approximately collinear. It results in troubles when training some ML models, in particular linear regression. Dimensionality reduction by PCA can improve the linear regression training process. PCA keeps the features with the most variance, but primary components may be irrelevant to predict the target. Partial Least Squares is a generalization of the PCA. While choosing the primary components PLS regression takes into account target feature and returns directions most correlated with the target feature. We use the PLS-regression implementation from sklearn library. The sklearn.cross_decomposition.PLSRegression class was constructed for three data configurations:

- full spectra in the [5460…5560] eV and [5980…6080] eV energy intervals for V and Cr correspondingly using coordination number as target property. The energy intervals are given in the Cr and V K-edge photon energy units before alignment.
- descriptors of spectra except t-SNE, PLS, BL using metal valence as target property
- descriptors of spectra except t-SNE, PLS, BL using coordination number as target property

At first for each configuration the PLS-regression model with two components was fitted using data with known target properties. Then we apply its transform method for the full dataset. Thus, we obtain 6 pls-descriptors.

Feature comparison and selection

Supervised ML algorithm can be trained on any combination of descriptors and provides expected quality of the prediction via cross-validation procedure. In this section we demonstrate methodology to select their optimal combinations for a given target property. The best quality, in terms of accuracy and reliability of the prediction, can be obtained using the most complete set of noncorrelated descriptors. To find such uncorrelated pairs, we use the concept of mutual information (MI)[8], which estimates mutual dependence of two descriptors by notions of information theory[9]. A non-negative value of a MI score is proportional to the amount of information one variable reveals about another. Figure S1 shows the matrix of mutual information values for all pairs of descriptors, calculated as implemented for regression task in scikit-learn.

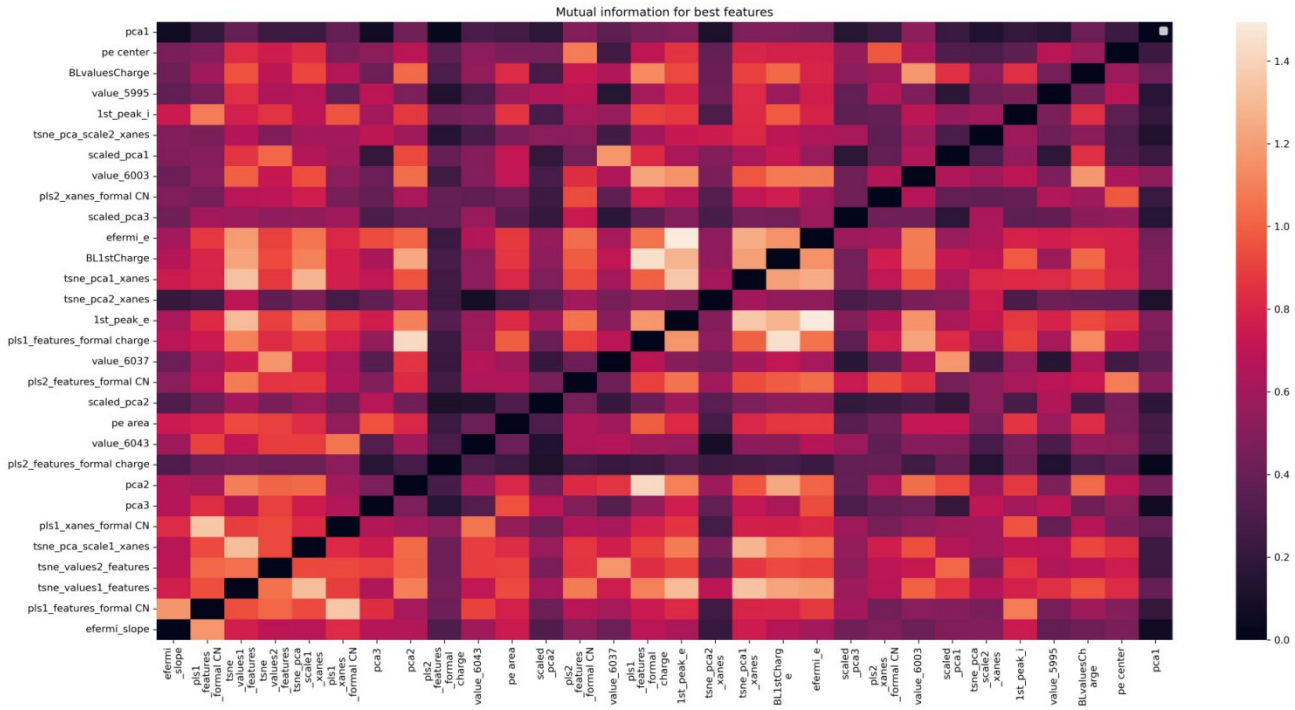


Figure S1. The matrix of mutual information scores for pairs descriptors for the task of formal CN prediction in Cr database. Features are sorted in the order of their quality, estimated as MI between the feature itself and the target variable (formal CN). The values of diagonal elements of the matrix, i.e., MI between pairs of identical descriptors, are set to zero for the sake representation, since they show the highest MI scores.

Although it is known that nonmetric algorithms, including classification and regression trees, can handle the curse of dimensionality, in case of a big amount of mutually dependent features (features with high MI), tree-based ensemble algorithms can form many trees with identical nodes, in other words, they tend to pick subsets of features with the same information since they choose random subsets of features from all given. That leads to a lower quality of the algorithm, especially in case of “small datasets”: ones that have number of samples comparable to the number of features. Thus, to train an accurate and robust algorithm, and obtain reliable prediction, a small collection of good and independent features is needed.

For example, MI matrix, calculated for descriptors in the task of formal CN prediction of Cr database, (Figure 6) shows that the best feature (e_femi_slope) is highly correlated with pls1_features_formal_CN, which in turn is highly correlated with t-SNE-based features, pls1_xanes_formal_CN and pca2. On the other hand, MI scores between them and pls2_xanes_formal_charge and pls2_features_formal_charge are small. It is unclear what subset of these features will show the best performance. We found that the best accuracy is obtained for a combination of e_fermi_slope, pls1_features_formal_CN, tsne_values1_features and, highly independent from others, pls2_features_formal_charge, which was used further in CV procedure (*vide infra*), and same issues were addressed in the same manner also. Compilation of descriptors selected according to their independency as calculated by mutual information gives following sets of independent descriptors for separated libraries of edges:

- for Cr CN: efermi_slope', 'pls1_features_formal CN', 'tsne_values1_features', 'pls2_features_formal charge
- for Cr charge: BLvaluesCharge', 'pls1_features_formal charge', 'value_6003'
- for V CN: BLvaluesCN', 'pls2_features_formal charge', 'value_5468'
- for V charge: BL1stCharge', 'pls1_features_formal charge', 'pe center', 'efermi_e', 'value_5520', 'value_5485'

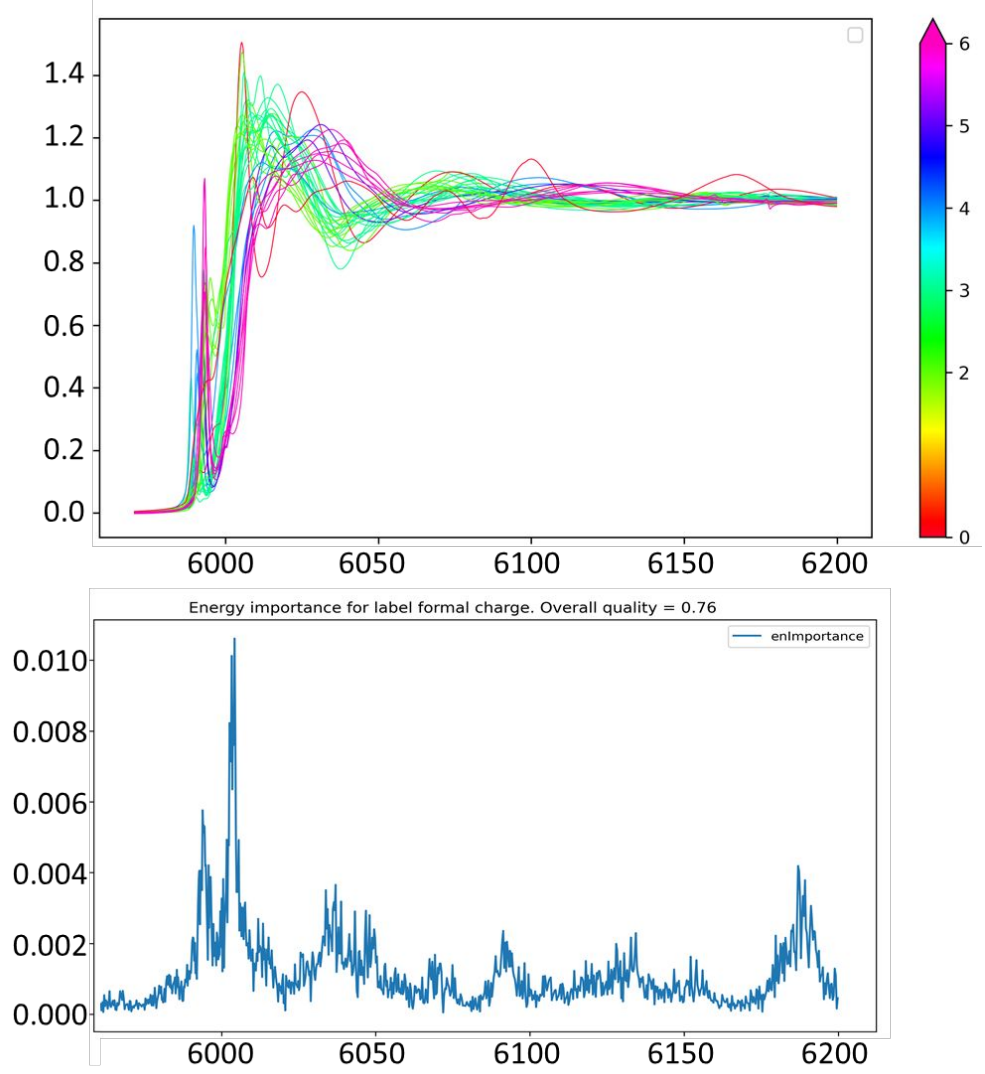


Figure S2. Dataset of Cr K-edge reference spectra and importance of individual points in the spectra for predicting Cr valence

The energy intervals are given in the Cr and V K-edge photon energy units before alignment. The choice of exact points in the spectrum as descriptors is based on their predictive power. Different points and regions of a spectrum have different importance for analysis and prediction of a target property. To estimate that, we depict the dataset of Cr K-edge spectra colored by their formal valence (Figure S2) or coordination number (Figure S3) together with impurity-based feature importance analysis of the ExtraTrees classifier as implemented in scikit-learn[1].

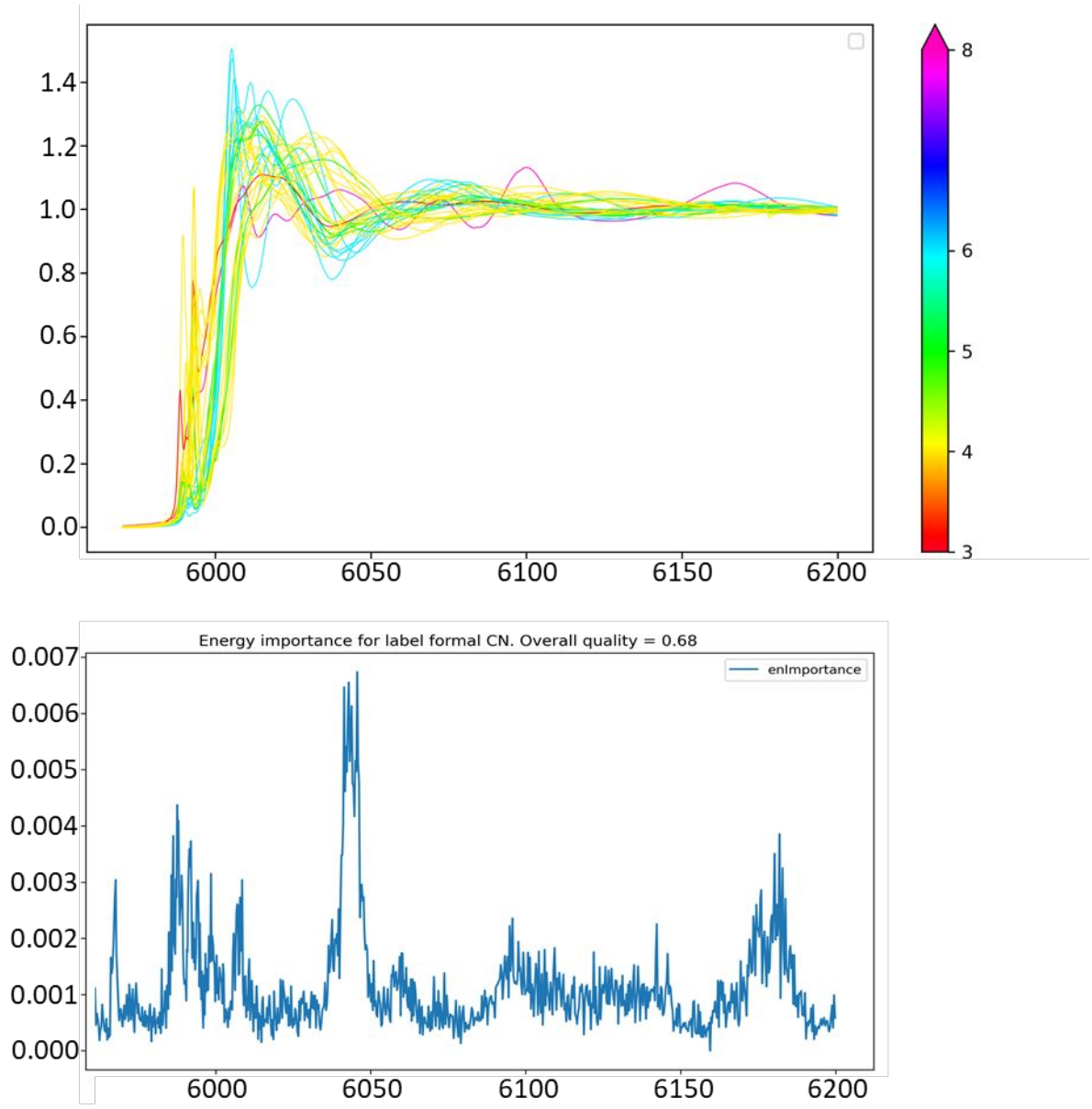
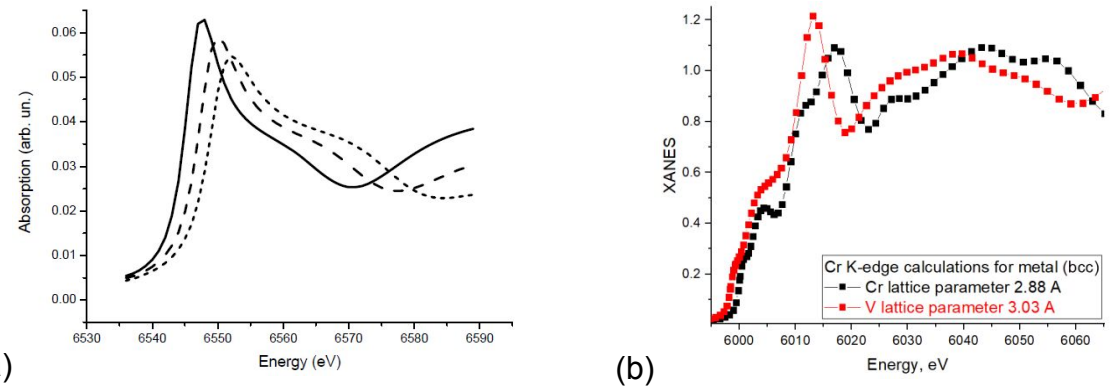


Figure S3. Dataset of Cr K-edge reference spectra and importance of individual points in the spectra for predicting Cr coordination number

2. Difficulties in the energy alignment of absorption edges from different chemical elements.

K-edge X-ray absorption spectra of different chemical elements have similar shape if metal local atomic structures are similar. However formally similar compounds, e.g. V2O3 and Cr2O3 or V and Cr have different lattice parameters that affect position of absorption edge and thus complicate relative alignments of libraries of different chemical elements. This problem was first discussed by P.Glatzel et al.[10] and shown in Figure S4:



(a) (b)

Figure S4. (a) Theoretical Mn K-edge XANES spectra for MnO_6 octahedron with Mn-O distances 2.17 Å (solid line), 2.00 Å (dashed line) and 1.88 Å (short dashed line). Reproduced from[10]. (b) Theoretical Cr K-edge XANES spectra calculated from bcc structure of metallic Cr for two lattice parameters: from 2.88 Å (as in Cr metal) and 3.03 Å (as in V metal).

Figure S4(b) shows that one can't align libraries of Cr and V compounds using positions of their metallic references since lattice parameters in Cr and V metals are different. Therefore, in the main text we used empirical value 522.3 eV to align Cr and V libraries of spectra, while the energy difference between foil absorption edges provided lower accuracy in terms of metal charge classification (Figure S5).

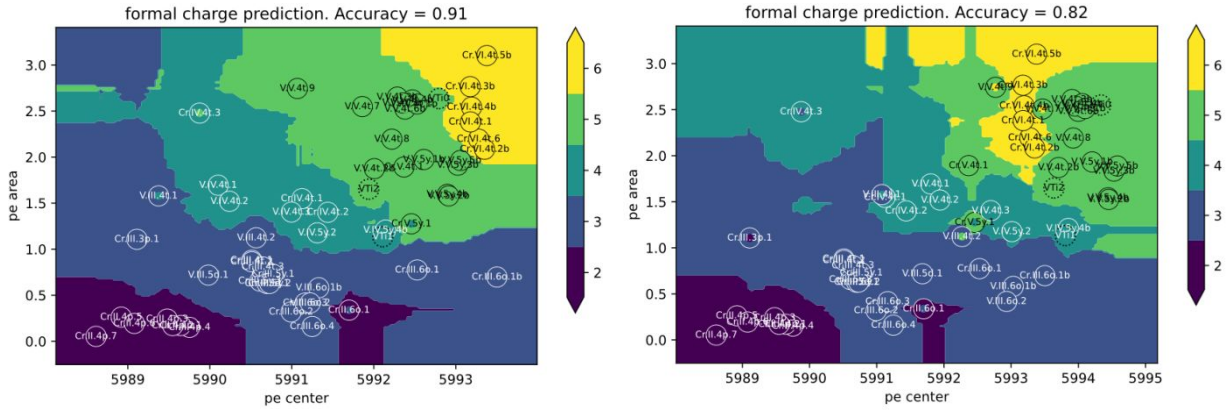


Figure S5. Classification scatter plots based on pre-edge descriptors of spectra. Cr and V libraries were aligned by using energy shift 522.3 eV (left) and 524 eV (right). The value 524 eV was obtained from difference in the Cr and V absorption edges defined by their metal foils and such shif demonstrates worse accuracy.

To construct a combined library the V- and Cr K-edge XAS spectra were aligned on the relative energy scale by applying a constant energy shift 522.3 eV. This value was selected empirically based on the criteria of improved classification quality of metals' oxidation state and coordination number (see Figure 4a corresponding to the best choice). The classification quality was lower when using a 524.0 eV shift corresponding to the difference between the K-edge energies of Cr and V foils (5989.0 eV versus 5465.0 eV). This can be explained by the differences in the interatomic distances

among Cr and V references in similar local environments, which are known to affect the position of the absorption edge (see Figure 1 in 62 and Section 2 in ESI).

3. Ambiguities in LOOCV for small datasets

Small libraries are fraught with pitfalls when applied to supervised ML[11]. The problems are due to the ML algorithm overfitting, assessment of its quality, and prediction accuracy for an "unknown" sample (as opposed to reference library item). Some ML algorithms are free from overfitting due to the procedure of their construction and the preferential choice of the spectra for small libraries. This is the case of ensemble methods, e.g. Extra Trees [12]. However, relying solely on a single ML approach does not guarantee accurate prediction. This is especially true for small datasets where LOOCV quality and uncertainty of a given prediction have large confidence intervals. To overcome these problems additional information about the spectrum should be taken into account or at least several independent predictions should be averaged.

Figure S6 demonstrates why cross-validation analysis being efficient and mostly used in data science may provide misleading results for small unbalanced datasets. The first example concerns the case when class contains only one item (Figure S6a). Removing this item during LOOCV results in a wrong prediction because the algorithm is re-trained on a library containing no representative of this class. In this case, the cross-validation quality will be poor, but the ML prediction can still be good if unknown data is located near the references. Figure S6b illustrates different but also common case. Each class in Figure S6b contains many library items, but all of them are similar. Removing one of them will result in high cross-validation quality since algorithm is trained on very similar references. However, evaluated LOOCV quality does not guarantee good transferability and proper prediction if the unknown data is far from the references. An example of the library with homogeneously distributed library items inside each class is shown in Figure S6c. In this case, the LOOCV can precisely estimate the approximation quality and MAE. However, if an unknown data point is far from the library items, the uncertainty of such prediction cannot be estimated with the LOOCV approach since all ML algorithms work well for interpolation but often fail in extrapolation.

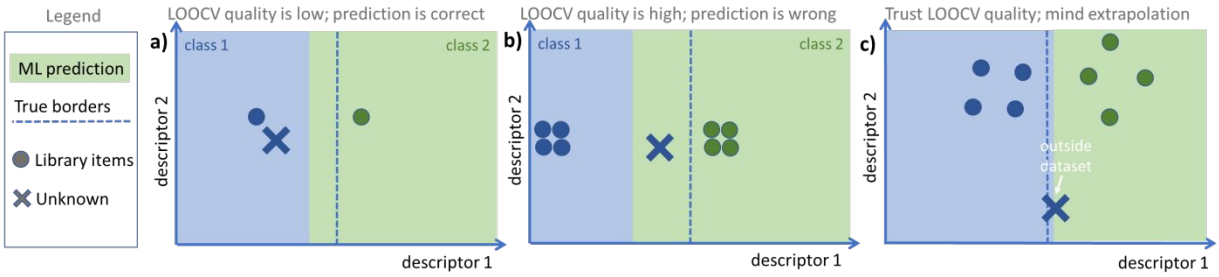
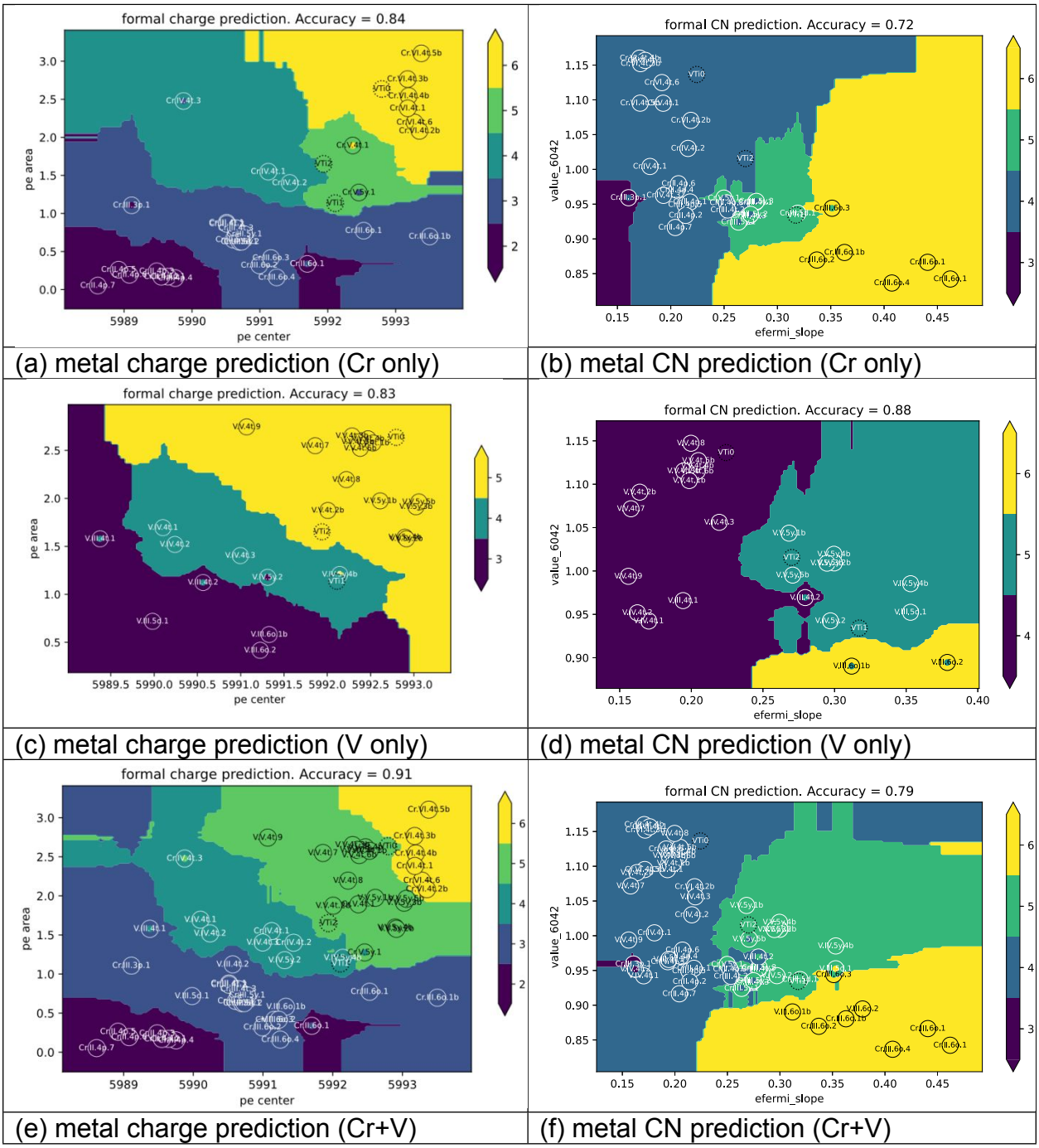


Figure S6. Toy examples of the libraries when the LOOCV approach can be misleading. In panel (a), LOOCV is low, but the prediction (background color behind "unknown") is correct. In panel (b), the LOOCV is high, but the prediction is wrong. In panel (c), the LOOCV correctly describes the uncertainty but should not be used to evaluate performance of the algorithm outside dataset. The background color in all panels is the predicted class by the algorithm trained on all the library items while the dashed line shows the true border between the classes.

Since small size of library is the source of ambiguity in the analysis of uncertainties then augmentation of the library with the chemically diverse references is a way to improve accuracy and generalization ability of the trained algorithm. Figure S7 shows the augmentation of independent Cr-based and V-based libraries by their merging. The visualization is performed with two pairs of descriptors selected for metal charge and CN classification.



(a) metal charge prediction (Cr only) | (b) metal CN prediction (Cr only)

(c) metal charge prediction (V only) | (d) metal CN prediction (V only)

(e) metal charge prediction (Cr+V) | (f) metal CN prediction (Cr+V)

Figure S7. Scatter plots demonstrating augmentation of independent Cr and V-based libraries by their merging.

5. Synthesis and characterization of reference molecular compounds

Cr- and V-K-edge XAS Spectra were taken from the previous report, unless stated below. Commercially available materials were purchased from the corresponding supplier and used without further purification. $V(Mes)_3(thf)$[13], $V(CH_2SiMe_3)_4$[14], $V(NMe_2)_4$[15], $VO(CH_2SiMe_3)_3$[16], $Cr(N(SiMe_3)_2)_2(thf)_2$[17] and $Cr(TBOS)_2(tmeda)$[18] were synthesized according to the reported procedure.

Synthesis of $Cr(N(SiMe_3)_2)_2(OPPh_3)_2$ (II.4p.5) (Synth-1): $Cr(N(SiMe_3)_2)_2(thf)_2$ (150 mg, 0.29 mmol) was dissolved in toluene (5 mL), giving a blue solution. Triphenylphosphine oxide ($OPPh_3$, 161 mg, 2 equiv.) was dissolved in toluene 5 mL, and was added to the solution of Cr complex at room temperature, immediately giving a yellow solution. The combined solution was stirred for 30 min and concentrated under vacuo. Recrystallization at -30 °C yielded blue crystal in 49% yield. Obtained blue crystal was used for sc-XRD analysis. ^{1}H NMR (300 MHz, C_6D_6) δ/ppm = 8.11 (br., $v_{1/2}$≈ 200 Hz) 7.49 (br., $v_{1/2}$≈ 20 Hz).

Synthesis of $Cr(TBOS)_3(TPPO)$ (III.4t.1) (Synth-2): $Cr(OSi(OtBu)_3)_3(thf)_2$ (150 mg, 0.15 mmol) was dissolved in C_6H_6 (5 mL) giving pale-blue solution. Triphenylphosphine oxide (TPPO, 42 mg, 1 equiv.) was dissoved in C_6H_6 (5 mL) and was added to the solution of Cr complex at room temperature. The combined blue solution was stirred for 2 h and dried under vacuum, yielding blue-purple powder (89% yield). Recrystallization in n-pentane at -30 °C yielded XRD-quality blue-purple crystals. Obtained crystal was used for sc-XRD analysis. ^{1}H NMR (200 MHz, C_6D_6) δ/ppm = 8.52 (br., $v_{1/2}$≈ 30 Hz), 6.55 (br., $v_{1/2}$≈ 15 Hz), 1.79 (br., $v_{1/2}$≈ 40 Hz)

Synthesis of $CrO_3(py)_2$ (Synth-3, Precursor of $CrO_3(TPPO)$): $CrO_3(py)_2$ was synthesized as a precursor of CrO_3(TPPO) (VI.4t.6), according to reported procedure[19]. ^{1}H NMR (200 MHz, C_6D_6) δ/ppm = 8.15 (br, 4H, *o*-py), 6.62 (t, 2H, $^3J_{HH}$ = 6.6 Hz, *p*-py), 6.28 (t, 4H, $^3J_{HH}$ = 5.7 Hz, *m*-py).

Synthesis of $CrO_3(OPPh_3)$ (VI.4t.6) (Synth-4): $CrO_3(py)_2$ (50 mg, 0.19 mmol) was dissolved in C_6H_6 (10 mL) giving red solution. Tiphenylphosphine oxide (54 mg, 0.19 mmol, 1.0 equiv.) was dissolved in C_6H_6 (5 mL) and was added to the solution of Cr complex, immediately giving a yellowish-orange solution. The mixture was stirred for 3 min followed by filtration. The liquid filtrate was concentrated in vacuo, and excess amount of *n*-pentane was added giving orange precipitation. The powder was washed with *n*-pentane (2 mL x 3 times), dried under vacuo yielding orange powdery product (61 mg, 83% yield). Crystallization from benzene solution layered with *n*-pentane at -30 °C yielded orange crystal which was used for sc-XRD analysis. ^{1}H NMR (200 MHz, C_6D_6) δ/ppm = 7.70 (dd, $^3J_{HH}$ = 7.4 Hz, 6H, *o*-Ph) 7.10-6.90 (m, 9H, *m*/*p*-Ph). ^{13}C(^{1}H) NMR (50 MHz, C_6D_6): δ/ppm = 132 (*o*-Ph), 131 (*p*-Ph), 128 (*m*-Ph). We could not observe ^{13}C NMR signal of C_{ipso} due to the low solubility of the complex. ^{31}P(^{1}H) NMR (81 MHz, C_6D_6): δ/ppm = 30.

Single-crystal X-ray Diffraction Analysis

X-ray diffraction experiments were performed on a Rigaku XtaLAB Dualflex Synergy-S diffractometer equipped with a Rigaku HyPix-6000HE detector using copper (1.54184 Å) radiation. Suitable crystals were selected, protected by polybutene oil, mounted under a cold nitrogen stream, and datasets were collected at 100 K. The data collection and reduction were performed using the CrysAlisPro software, respectively. Structure solution and refinement were performed with SHELXT[20] and SHELXL[21], respectively, embedded in Olex2[22]. All non-hydrogen atoms were refined anisotropically.

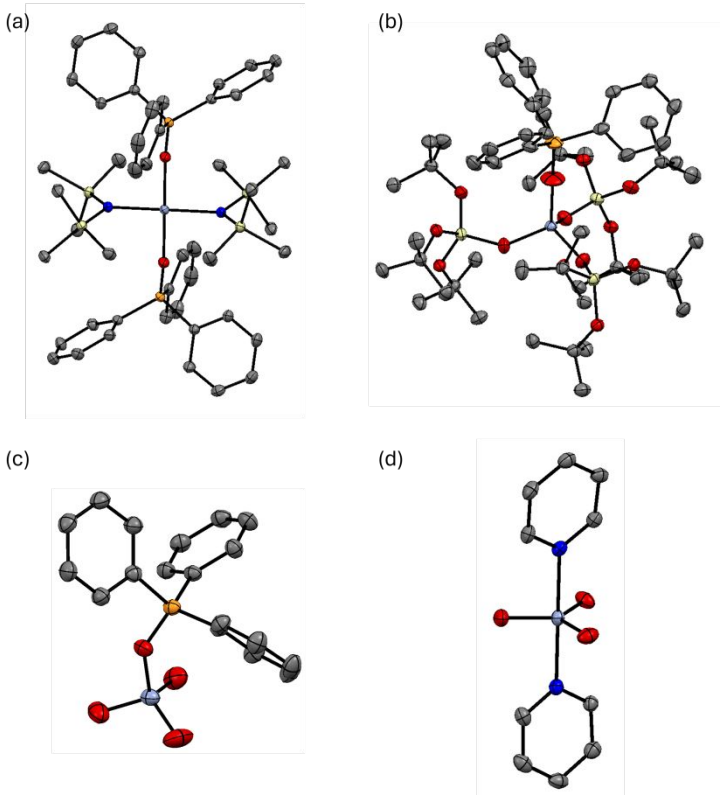


Figure S8. Molecular structures of (a) $Cr(HMDS)_2(TPPO)_2$ (II.4p.5), (b) $Cr(TBOS)_3(TPPO)$ (III.4t.1), (c) $CrO_3(TPPO)$ (VI.4t.6) and $CrO_3(py)_2$ with 50% thermal ellipsoids. All hydrogen atoms were omitted for clarity. For (d), one of two molecules in a unit cell is depicted.

Table S3. Crystal data and data collection parameters for $Cr(HMDS)_2(TPPO)_2$ (II.4p.5), $Cr(TBOS)_3(TPPO)$ (III.4t.1), $CrO_3(TPPO)$ (VI.4t.6) and $CrO_3(py)_2$

	$Cr(HMDS)_2(OPPh_3)_2$	$Cr(OTBOS)_3(OPPh_3)$	$CrO_3(OPPh_3)$	$CrO_3(py)_2$
CCDC No.	2386689	2386667	2386690	2386691
empirical formula	$C_{48}H_{66}CrN_2O_2P_2Si_4$	$C_{54}H_{96}CrO_{13}PSi_3$	$C_{18}H_{15}CrO_4P$	$C_{10}H_{10}CrN_2O_3$
formula weight	929.32	1120.59	378.27	258.20
crystal system	Monoclinic	Trigonal	Monoclinic	Triclinic
space group	*P2₁/n*	*R-3*	*P2₁/n*	*P-1*
a, Å	12.5012(9)	21.2744(8)	9.46640(10)	7.94290(10)
b, Å	14.4251(8)	21.2744(8)	18.1713(2)	8.8767(2)
c, Å	14.5268(10)	24.0927(10)	10.2540(2)	15.5630(4)

$\alpha\theta$, deg.	90	90	90	77.086(2)
β, deg.	110.802(3)	90	104.023(2)	81.771(2)
γ, deg.	90	120	90	86.323(2)
V, Å^3	2448.9(3)	9443.5(6)	1711.29(4)	1057.94(4)
Z	2	6	4	4
$D_{calcd,}$ g/cm^3	1.260	1.1822	1.468	1.621
μ, mm^{-1}	0.436 [Mo-Kα]	0.319 [Mo-Kα]	6.550 [Cu-Kα]	8,871
T, K	100.0	100.0	100.0	100.0
crystal size, mm	0.1 × 0.06 × 0.03	0.2 x 0.1 x 0.04	0.2 x 0.07 x 0.02	0.3 x 0.1 x 0.04
2θ range for data collection (deg.)	4.658 to 68.154	4.74 to 83.32	9.734 to 160.116	5.878 to 159.68
no. of reflections measured	63911	318438	24826	15226
unique data (R_{int})	9997 (0.0677)	14379 (0.0879)	3688 (0.0365)	4498 (0.0337)
data / restraint / parameters	9997/0/274	14379 / 0 / 226	3688/0/217	4498/ 0 / 289
$R1$ ($I > 2.0\ \sigma(I)$)	0.0491	0.0540	0.0482	0.0341
$wR2$ ($I > 2.0\sigma(I)$)	0.1386	0.1290	0.1384	0.1102
$R1$ (all data)	0.0958	0.1249	0.0496	0.0365
$wR2$ (all data)	0.1799	0.2104	0.1395	0.1126
GOF on F^2	1.129	1.211	1.110	0.959
$\Delta\rho$, e Å^{-3}	0.67 / -0.97	1.69 / -3.71	1.16 / -0.84	0.46 / -0.57

a) $R1 = (\Sigma||Fo| - |Fc||)/(\Sigma\,|Fo|)$ b) $wR2 = [(\Sigma\, w(Fo^2-Fc^2)^2)/(\Sigma\, w(Fo^4))]^{1/2}$

Table S4. The list of entries in the vanadium and chromium databases. For each sample we report the chemical formula, short name used later in the figures, reference to the synthesis protocol (commercial or synth-1, synth-2, etc. methods). Among several options available for commercial standards, we tried to choose the most reliable.

#	Chemical formula	Short Name	Type[a]	Formal charge	CN	Synthesis[b]
1	V	V.0.8c.1b	bulk	0	8	comm
2	$V(OSi(OtBu)_3)_3(OPPh_3)$	V.III.4t.2	mol	3	4	[23]
3	$V(Mes)_3(thf)$	V.III.4t.1	mol	3	4	[13]
4	$V(OSi(OtBu)_3)_3(thf)_2$	V.III.5d.1	mol	3	5	[23]
5	$V(acac)_3$	V.III.6o.2	mol	3	6	comm
6	V_2O_3	V.III.6o.1b	bulk	3	6	comm
7	$V(CH_2SiMe_3)_4$	V.IV.4t.1	mol	4	4	[14]
8	$V(NMe_2)_4$	V.IV.4t.2	mol	4	4	[15]
9	$V(OSi(OtBu)_3)_4$	V.IV.4t.3	mol	4	4	[23]
10	$VOSO_4$	V.IV.5y.4b	bulk	4	5	comm
11	$VO(acac)_2$	V.IV.5y.2	mol	4	5	comm
12	V_2O_4	V.IV.6o.1b	bulk	4	6	comm
13	$VO(O^iPr)_3$	V.V.4t.7	mol	5	4	comm
14	$VO(OSi(OtBu)_3)_3$	V.V.4t.8	mol	5	4	[23]
15	$BiVO_4$	V.V.4t.2b	bulk	5	4	comm
16	$aNaVO_3$	V.V.4t.1b	bulk	5	4	comm
17	NH_4VO_3	V.V.4t.6b	bulk	5	4	comm

18	KVO_3	V.V.4t.4b	bulk	5	4	comm
19	K_3VO_4	V.V.4t.3b	bulk	5	4	comm
20	$VO(CH_2SiMe_3)_3$	V.V.4t.9	mol	5	4	[16]
21	Na_3VO_4	V.V.4t.5b	bulk	5	4	comm
22	V_2O_5	V.V.5y.5b	bulk	5	5	comm
23	$bNaVO_3$	V.V.5y.1b	bulk	5	5	comm
24	NaV_6O_{15}	V.V.5y.3b	bulk	5	5	comm
25	$Na_{1.16}V_2O_5$	V.IV.5y.1b	bulk	4	5	comm
26	$(NH_4)_6V_{10}O_{28}\times6H_2O$	V.V.5y.4b	bulk	5	5.2	comm
27	$Na_6V_{10}O_{28}\times18H_2O$	V.V.5y.2b	bulk	5	5.4	comm
1	Cr	Cr.0.8c.1b	bulk	0	8	comm
2	$Cr(CO)_6$	Cr.0.6o.1	mol	0	6	comm
3	$Cr(O(tBu)_2C_6H_3)_2(thf)_2$	Cr.II.4p.1	mol	2	4	[24]
4	$Cr(acac)_2$	Cr.II.4p.2	mol	2	4	[25, 26]
5	$[Cr(OSi(OtBu)_3)_2]_2$	Cr.II.4p.3	mol	2	4	[17]
6	$Cr(N(SiMe_3)_2)_2(thf)_2$	Cr.II.4p.4	mol	2	4	[17]
7	$Cr(N(SiMe_3)_2)_2(OPPh_3O)_2$	Cr.II.4p.5	mol	2	4	Synth-1
8	$Cr(OSi(OtBu)_3)_2(tmeda)_2$	Cr.II.4p.6	mol	2	4	[18]
9	$Cr(OSi(OtBu)_3)_2(XyNC)_4$	Cr.II.6o.1	mol	2	6	[27]
10	$Cr(O(tBu)_2C_6H_3)_2(XyNC)_2$	Cr.II.4p.7	mol	2	4	[27]
11	$Cr(OSi(OtBu)_3)_3(OPPh_3)$	Cr.III.4t.1	mol	3	4	Synth-2
12	$Cr(OSi(OtBu)_3)_3(XyNC)_2$	Cr.III.6o.1	mol	3	6	[27]
13	μ-O-$(Cr(OSi(OtBu)_3)_2)_2$	Cr.III.5y.1	mol	3	5	[17]
14	$Cr(OSi(OtBu)_3)_3$	Cr.III.5y.2	mol	3	5	[24]
15	$Cr(OSi(OtBu)_3)_3(thf)_2$	Cr.III.5d.1	mol	3	5	[28]
16	$Cr(OSi(OtBu)_3)_3(dme)$	Cr.III.5y.3	mol	3	5	[24]
17	$Cr(OCMe_2CH_2OMe)_3$	Cr.III.6o.2	mol	3	6	[29]
18	$Cr(POSS)(thf)_3$	Cr.III.6o.3	mol	3	6	[24]
19	$Na[Cr(OSi(OtBu)_3)_4]$	Cr.III.4t.2	mol	3	4	[24]
20	$Cr(acac)_3$	Cr.III.6o.4	mol	3	6	comm
21	$Cr(CH(SiMe_3)_2)_3$	Cr.III.3p.1	mol	3	3	[30]
22	Cr_2O_3	Cr.III.6o.1b	bulk	3	6	comm
23	K-Kryptofix $[Cr(OSi(OtBu)_3)_4]$	Cr.III.4t.3	mol	3	4	[24]
24	$Cr(OtBu)_4$	Cr.IV.4t.1	mol	4	4	[24]
25	$Cr(OSi(OtBu)_3)_4$	Cr.IV.4t.2	mol	4	4	[24]
26	$Cr(CH_2tBu)_4$	Cr.IV.4t.3	mol	4	4	[31]
27	$CrO(OTBOS)_3$	Cr.V.4t.1	mol	5	4	[24]
28	$Na[CrO(O_2CC(CH_3)_2O]$	Cr.V.5y.1	mol	5	5	[24]
29	$CrO_2(OSi(OtBu)_3)_2$	Cr.VI.4t.1	mol	6	4	[24]
30	CrO_3	Cr.VI.4t.2b	bulk	6	4	comm
31	CrO_3[c]	Cr.VI.4t.5b	bulk	6	4	comm
32	Na_2CrO_4	Cr.VI.4t.3b	bulk	6	4	comm
33	K_2CrO_4	Cr.VI.4t.4b	bulk	6	4	comm
34	$CrO_3(OPPh_3)$[d]	Cr.VI.4t.6	mol	6	4[d]	Synth-4

[a] bulk: bulk material, mol: molecular material.

[b] reference numbers for corresponding synthesis are shown. Comm: commercially available.

[c] removed from the library due to possible contamination with H_2O.

[d] While initially attributed to the 5-coordinated species this compound was finally characterized as 4-coordinated with only one $OPPh_3$ ligand coordinating Cr.

6. Acquisition and pre-processing of spectra

Data collection

Experimental Cr- and V K-edge X-ray absorption spectra were acquired at the SuperXAS beamline at the Swiss Light Source (PSI, Villigen, Switzerland), operating at 400 mA and 2.4 GeV. The beamline is equipped with a 2.9 T superbend magnet, Si collimating mirror at 2.5 mrad, channel-cut Si(111) quickXAS monochromator, and Rh-coated toroidal mirror. XAS spectra of V and Cr reference samples for the library were acquired either in transmission mode with 15 cm long ionization chambers or in fluorescence mode using solid-state detectors. Air-sensitive samples were sealed in a glovebox prior to measurements.

The series of V K-edge XAS spectra of VOx species in bilayered 5% V2O5/15%TiO2/SiO2 were measured in fluorescence mode inside an operando reactor [35]. A temperature-programmed reduction (TPR) by ethanol (1.6 vol % EtOH in He 6.0, total flow 50 mL/min) was performed in the temperature interval of 100–400 °C with a heating rate of 5 °C/min. Prior to the ethanol TPR experiment, a standard pretreatment in an oxygen-containing atmosphere (400 °C in an oxygen-containing flow (20 vol % O2 in He, 50 mL/min) at a rate of 12 °C/min and dwelling for 1 h) was conducted. Fluorescence XAS spectra in operando cell were recorded using a PIPS diode (Mirion Technologies) as a detector. The Si(111) channel-cut monochromator was oscillating with a frequency of 1 Hz, which corresponds to a repetition rate of 2 scans/s. Prior to each data acquisition, the X-ray energy was calibrated by measuring vanadium (for V K-edge at 5465 eV) in transmission mode by moving the sample temporarily out of the beam. The intensities of the incident and transmitted beam were measured using 15 cm long ionization chambers filled with 500 mbar N2 and 500 mbar He.

Radiation sensitive samples were cooled to 100 K and fluorescence spectra were measured using PIPS diode (Mirion Technologies) as a fluorescence silicon drift Ketek detector. To avoid contact with air, sensitive samples were sealed in a glovebox. Pressed pellets with optimized thickness for transmission detection diluted with boron nitrile were sealed in two aluminized plastic bags (Polyaniline (15 μm), polyethylene (15 μm), Al (12 μm), polyethylene (75 μm) from Gruber-Folien GmbH & Co. KG (Straubing, Germany), using an impulse sealer inside a glovebox. The outer aluminized plastic bag was removed right before the measurement. Powder samples for fluorescence detection were filled in quartz capillaries (0.01 mm wall thickness, 0.9 mm outer diameter; Hilgenberg GmbH) under inert atmosphere. With the addition of quartz wool, the powder inside the capillaries was slightly compressed to exclude particle migration while cooling. The capillaries were sealed with Apiezon vacuum grease and wax (M&I Materials Ltd), stored in glass tubes under an argon atmosphere, and opened just before measurement. To prevent beam damage, three individual spots of 5 minutes each were averaged both for XAS transmission and fluorescence modes.

Preprocessing and alignment

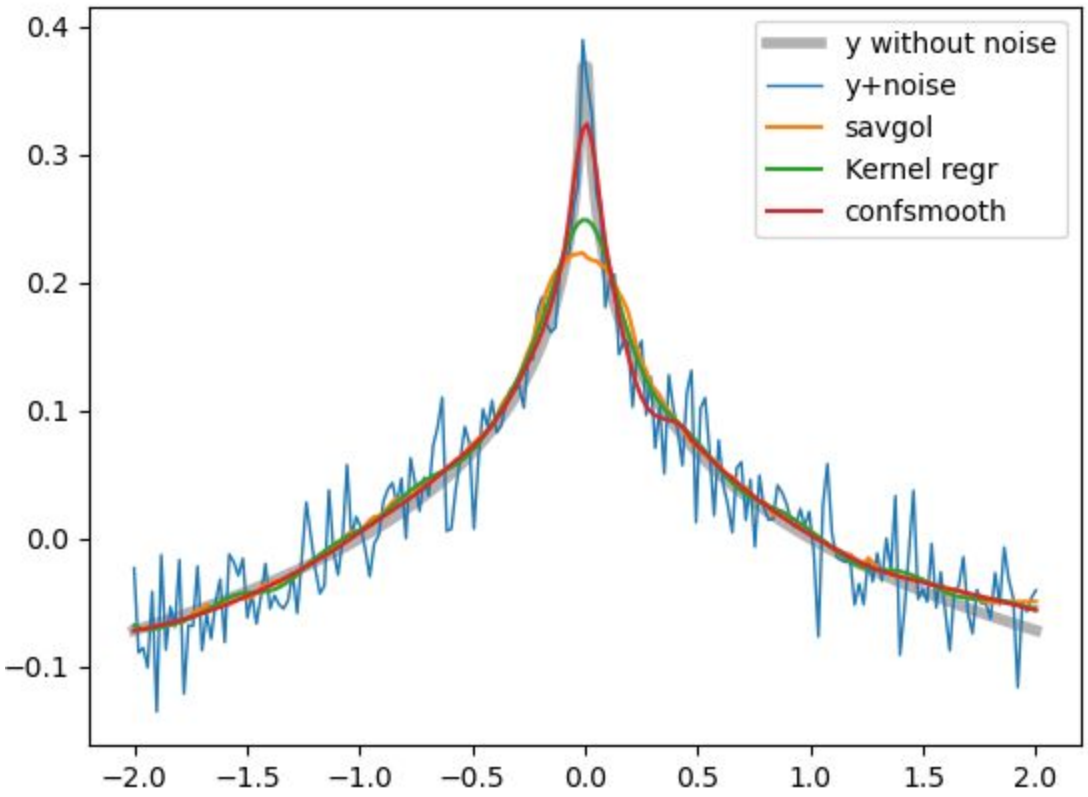
The acquired data were aligned using standard procedures of the *IFEFFIT* program package [32, 33] and further processed by routines of *PyFitIt* software [34]. At the first step, the

energy of monochromator was calibrated. The energy alignment of the spectra was performed using metal foil of the same element. Background subtraction and normalization was performed by means of modified *MBACK* algorithm [35, 36] implemented originally in the Larch package [37]. The parameters of polynomial degree were common for most of reference spectra and studied catalyst, however some entries in the library have higher intensity of the background and required adjustment of energy intervals in MBACK for good polynomial fit. The original procedure of MBACK was modified to fit the spectrum to the Heaviside function instead of tabulated exponentially decaying absorption coefficient. For the background function, a series of Legendre polynomials centered at K-edge energy were used for approximation of spectra in the post-edge region (50-250 eV, m=3 for polynomial order). The complementary error function was replaced with a parabola in the pre-edge region (between -150 and -20 eV), for better stability when only a short energy interval was available. To account for the highly non-linear pre-edge which often results from Compton scattering in the measurement window of an energy-discriminating detector, MBACK adds a complementary error function to the polynomial. The parameters of the error function are hard to fit when a short energy interval is available. For this reason, the quadratic polynomial vanishing at the absorption edge is preferable. The smoothness of the piecewise normalized polynomial was provided by convolution with gaussian kernel.

The noise level in diluted samples with low metal concentration was reduced by means of confidence-based smoothing. We developed this approach to smooth the experimental spectra and preserve the features of the spectra. To distinguish peaks from noise the algorithm receives noise standard deviation noise level as input. It can be scalar - common for all points, or vector. The user also sets the confidence level. Confidence Based Smoothing use piecewise polynomial approximation with adaptive piece size, which is chosen to meet condition:

$$\mathrm{erf}\left(\frac{Error}{\sqrt{2}NoiseLevel}\right) < 1 - confidence$$

The algorithm also checks pairs, threes, fours, etc of consecutive error values. If, for example, confidence = 0.99 and four consecutive errors are positive and more than 0.04, than their joint probability is less than 0.00766. So, these four should be treated as the signal, and we have to decrease approximation piece size.



confidence-based smoothing algorithm: https://github.com/gudasergey/confsmooth
supplementary: https://stackoverflow.com/questions/43700404/curve-smoothing-preserving-peaks-and-valleys/75950175#75950175

Extracting spectrally pure components from XAS series versus 'blind' analysis

The XAS spectra of typical supported metal species are challenging to analyze due to a structural heterogeneity (presence of multiple species at once) and overlapping XANES features of different species. The analysis can sometimes be facilitated if the structure of supported metal species depends on metal loading and applied synthetic methods or can evolve under reaction conditions. Such changes can be followed by XAS producing a series of XANES spectra, which can then be decomposed into signals of spectrally pure components (e.g., via principal component [34, 38] or multivariate curve resolution (MCR) analysis) [39]. Identifying spectral components partially removes the structural uncertainty, ideally leading to pure phases. However, it enormously increases the amount of experimental work. In this study, we test two approaches for the analysis of the same experimental data. The studied dataset consists of previously reported series of V K-edge XANES spectra of bilayered 5% V_2O_5/15%TiO_2/SiO_2 containing VO_x species with sub-monolayer loading, which change their structure (oxidation state and local coordination) upon TPR in 1.6 vol % EtOH in He. Within the first approach, we analyzed the structure of spectrally pure components resulting from the MCR analysis of the XANES spectral series. Based on our previous work, we consider that the spectral components are likely pure species. In the second 'blind' approach, we analyze each XANES spectrum in the series separately, ignoring all preliminary knowledge about the

catalyst structure and MCR analysis. This is done to assess the performance of ML algorithms and the uncertainty of the structural predictions (See section 3.5 in “Results in discussion”).

Pre-edge subtraction algorithm

The calculation of pre-edge peaks by subtracting the contribution of the main edge from XANES is done in accordance with the classical technique used in the programs Larch[37], XANES dactyloscope[40]. The peaks interval is cut out from the baseline fit interval, and a baseline is fitted to the remaining points. The relative strengths of the pre-edge peaks are then obtained by the baseline subtraction from the initial XANES spectrum. For baseline approximation, the XANES dactyloscope developer uses splines; in Larch, it is possible to choose from the Lorentzian, Gaussian, or Voigt shapes, with the optional addition of a constant, linear, or quadratic function.

When processing large spectral databases, we encountered the problem of the time-consuming determination of the optimal fit parameters. A user must manually adjust parameters for each spectrum. To automate this process, we worked out a new algorithm. For the baseline fit, we use a simple piecewise linear model consisting of two linear functions. The first is fitted along the left part of the baseline fit interval with the cut-out peak interval, the second - along the right. Linear models make the fitting procedure stable with respect to the changes in parameters and the spectrum.

The pre-edge peaks significantly change the neighboring points of the spectrum that are used for the baseline fit. This is especially evident for intense peaks and interferes with the fit when, due to the shoulder, the right boundary of the fit interval has to be shifted very close to the pre-edge peaks. To eliminate the contribution of the pre-edge peaks to the spectrum, we fit the relative pre-edge peaks spectrum with a Cauchy function located in the center of the peak interval, the height and width of which are selected during the fit. Thus, to determine the baseline, we are forced to use some approximation to the relative pre-edge peaks spectrum, which led us to the following iterative algorithm:

1) Iterating through all tangents to the spectrum S, we find the best initial approximation to the pre-edge peak interval and the relative pre-edge peaks spectrum relPE. As the best, we take the tangent with the maximum relative peak area between two consecutive tangent points.
2) In the loop:
a) fit the current approximation of the relative pre-edge peaks spectrum with the Cauchy function CF in the center of the peak interval
b) subtract the found Cauchy function from the spectrum: S* := S-CF
c) find a piecewise linear approximation LA for the resulting curve S* on the baseline fit interval with the peak interval cut out
d) calculate a new approximation to the relative pre-edge peaks by subtracting the piecewise linear baseline from the original spectrum: relPE := S - LA
e) return to the step a)

Evaluating the algorithm for the Cr-V spectrum database shows that two iterations are already enough to obtain the sufficient baseline approximation.

Implementation of the LCF algorithm

For the common fingerprint analysis procedure – linear combination fitting (LCF) – we use original implementation, found to be more efficient in cases of multiple components fitting with iteration on the reference library. Herein, for the given spectra of the pure components $\mu_1(e),\ldots,\mu_n(e)$ LCF analysis finds concentration $c_1,\ldots,c_n$, so that spectra mixture $c_1\mu_1(e) + \ldots + c_n\mu_n(e)$ fits best the unknown spectrum $\mu^*(e)$.

$$c_1\mu_1(e) + \ldots + c_n\mu_n(e) \approx \mu^*(e)$$

Choice of L_2 norm to compare spectra results in the optimization problem:

$$\int_a^b (c_1\mu_1(e) + \ldots + c_n\mu_n(e) - \mu^*(e))^2 de \rightarrow \min_{c_i:c_i \geq 0,\ \sum_{i=1}^{n} c_i = 1}$$

This problem is closely related to the non-negativity constrained least-squares [41]:

$$\sum_{k=1}^{K} w_k(c_1\mu_1(e_k) + \ldots + c_n\mu_n(e_k) - \mu^*(e_k))^2 \rightarrow \min_{c_i:c_i \geq 0}$$

with some grid $e_0,e_1,\ldots e_K$ and weights $w_k = e_k - e_{k-1}$, k=1,…,K. To make the sum of the resulting concentrations equal to one, we add an extra term to the optimized function:

$$\sum_{k=1}^{K} w_k(c_1\mu_1(e_k) + \ldots + c_n\mu_n(e_k) - \mu^*(e_k))^2 + w_{K+1}(c_1 + \ldots + c_n) \rightarrow \min_{c_i:c_i \geq 0}$$

with high weight $w_{K+1} = 100*\max(w_1,\ldots,w_K)$. This NNLS problem we then solve using scipy.optimize.nnls function with default parameters.

7. DFT and spectra simulations

DFT and FDMNES calculations settings

We have simulated the possible local environments of V single site on the most energetically stable (202) surface of TiO2 (Figure 7), calculating theoretical spectrum for each one (See the section 7 in the SI for the details). First, the V atom was embedded in the dangling oxygen bonds of (202)TiO_2, and resulting geometry was optimized, producing the V(V) structural model. Then the V(III) geometry was obtained from it by addition of hydrogen atom in the second coordination shell of V and subsequent geometry optimization.

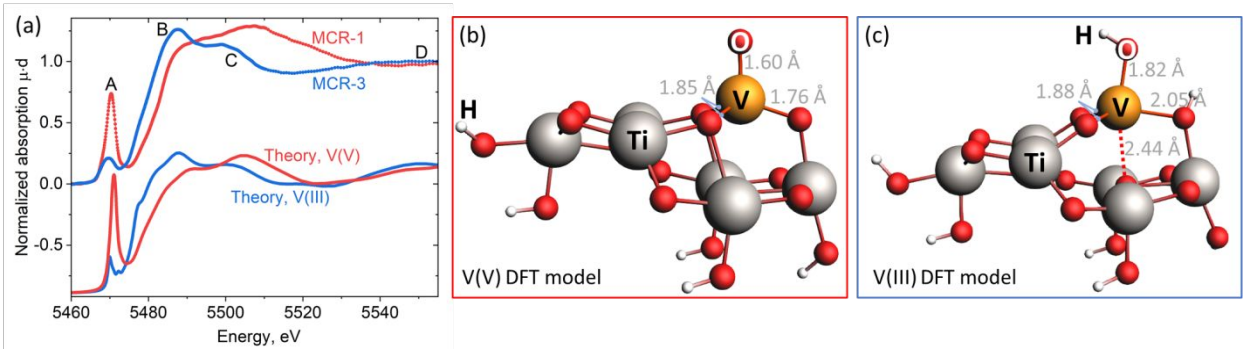


Figure 7. (a) Comparison of the simulated spectra V K-edge XANES spectra of V(V) and V(III) species with the MCR-1 and MCR-3 and corresponding DFT-optimized structural models of V(V) (b) and V(III) (c) species on TiO_2 (202) surface.

The initial configuration (Figure 7b, V(V) DFT model) is described by a short vanadyl bond with the length of 1.60 Å and three bonds to lattice oxygens with lengths of 1.76-1.85 Å. To simulate vanadium reduction hydrogen atoms terminated two oxygens in the metal first coordination sphere. Upon reduction to the V(III) state, VO_x-species lose the vanadyl group and move closer to the surface, forming additional V-O bonds (Figure 7c), thus explaining the increase in coordination number predicted in Table 2. The simulated V K-edge XANES spectra of V(V) and V(III) structures produced by DFT show good agreement with the experimental data.

To represent possible local environments of V single site on the TiO_2 layer, we pick the structure of the most energetically stable (202) surface of TiO_2 (Figure 7 of the main text). First, the V atom was embedded in the dangling oxygen bonds of (202)TiO_2, and the resulting geometry was optimized, producing the V(V) structural model. In the next step the V(III) geometry was obtained from relaxed V(V) by terminating oxygen atoms in the first coordination shell of V with hydrogens and subsequent geometry optimization (see the table in section 7.2). All calculations were performed with density functional theory using TPSS exchange-correlation functional[42] and Slater-type orbitals triple-ζ TZP basis set as implemented in ADF2022 software[43] with default settings for the SCF and geometry convergence.

For these structures theoretical V K-edge XANES spectra were calculated by FDMNES[44, 45] code within the full potential finite difference method. 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.

Coordinates of optimized structures

Table S5 represents coordinates of DFT-optimized V- and III-valent structures from Figure 7 of the main text. The data format is inherited from ".xyz" data format: atom type as a string, followed by x, y, and z coordinates of atom in angstroms.

Table S5. The coordinates of the V- and III-valent vanadium species on the (202)TiO_2 surface.

5-valent structural model				3-valent structural model			
V	1.26895	4.712674	5.049403	V	1.144947	4.71778	5.149928
O	1.729942	3.240399	6.058608	O	1.337974	3.247317	6.318072
O	3.849011	2.789861	3.816066	O	3.847499	2.839383	3.776874
H	3.921975	4.78268	11.68657	H	3.944923	4.70949	11.68965
O	3.470585	1.984732	7.411456	O	3.426725	1.923513	7.206416
O	-0.3276	4.718986	4.963151	O	-0.55309	4.463222	4.533374
Ti	2.017018	2.828292	7.972948	Ti	2.017018	2.828292	7.972948
Ti	3.902518	2.828292	5.615448	Ti	3.902518	2.828292	5.615448
O	2.496761	4.716626	8.213335	O	2.613769	4.725301	8.251858
O	3.884035	4.713276	5.302124	O	3.578187	4.713832	5.400097
O	1.427389	2.986005	9.692438	O	1.360886	3.034563	9.683629
Ti	3.902418	4.713692	3.257798	Ti	3.902418	4.713692	3.257798
Ti	2.017048	6.599322	7.972648	Ti	2.017048	6.599322	7.972648
Ti	2.017048	4.713822	10.33015	Ti	2.017048	4.713822	10.33015
Ti	3.902548	6.599322	5.615148	Ti	3.902548	6.599322	5.615148
O	5.787918	2.828192	5.998048	O	5.787918	2.828192	5.998048
H	6.409756	4.303544	3.51376	H	6.222945	3.839931	2.858967
O	3.816501	6.64042	3.820037	O	3.808885	6.551306	3.766449
O	5.788048	4.713822	2.874898	O	5.788048	4.713822	2.874898
O	1.733844	6.189045	6.064734	O	1.323185	6.182565	6.304413
H	5.966275	6.756522	6.94686	H	5.93516	6.866581	6.926878
O	3.467204	7.443726	7.427567	O	3.423259	7.500969	7.198345
O	1.40781	6.421892	9.706575	O	1.365033	6.405206	9.680892
H	5.973726	2.682464	6.946162	H	5.937517	2.588799	6.933349
O	3.902448	4.713722	10.71332	O	3.902448	4.713722	10.71332
O	2.016948	4.713722	12.30532	O	2.016948	4.713722	12.30532
O	5.787948	6.599222	5.998318	O	5.787948	6.599222	5.998318
H	1.423805	5.345228	12.75473	H	1.156245	4.713853	12.76536
O	1.940096	4.693537	3.424163	O	1.864799	4.698179	3.228048
				H	-1.14455	3.945417	5.109757
				H	1.269637	4.133632	2.699866

References

(1) Torrisi, S. B.; Carbone, M. R.; Rohr, B. A.; Montoya, J. H.; Ha, Y.; Yano, J.; Suram, S. K.; Hung, L., Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. npj Comput. Mater. **2020,** *6*, 109, DOI: 10.1038/s41524-020-00376-6

(2) Martini, A.; Guda, A. A.; Guda, S. A.; Bugaev, A. L.; Safonova, O. V.; Soldatov, A. V., Machine Learning Powered by Principal Component Descriptors as the Key for Sorted Structural Fit of XANES. Phys. Chem. Chem. Phys. **2021**, DOI:

(3) Guda, A. A.; Guda, S. A.; Martini, A.; Kravtsova, A. N.; Algasov, A.; Bugaev, A.; Kubrin, S. P.; Guda, L. V.; Sot, P.; Van Bokhoven, J. A., et al., Understanding X-ray Absorption Spectra by Means of Descriptors and Machine Learning Algorithms. npj Comput. Mater. **2021,** *7*, 203, DOI: 10.1038/s41524-021-00664-9

(4) Kozyr, E. G.; Bugaev, A. L.; Guda, S. A.; Guda, A. A.; Lomachenko, K. A.; Janssens, K.; Smolders, S.; De Vos, D.; Soldatov, A. V., Speciation of Ru Molecular Complexes in a Homogeneous Catalytic System: Fingerprint XANES Analysis Guided by Machine Learning. **2021,** *125*, 27844-27852, DOI: 10.1021/acs.jpcc.1c09082

(5) de Groot, F.; Vanko, G.; Glatzel, P., The 1s x-ray absorption pre-edge structures in transition metal oxides. J. Phys. Condens. Matter **2009,** *21*, DOI: 10.1088/0953-8984/21/10/104207

(6) Maaten, L. v. d.; Hinton, G. E., Visualizing Data using t-SNE. JMLR **2008,** *9*, 2579-2605, DOI: api.semanticscholar.org/CorpusID:5855042

(7) https://opentsne.readthedocs.io/en/stable/ (accessed 21.09.2024).

(8) Kraskov, A.; Stögbauer, H.; Grassberger, P., Estimating mutual information. Physical review. E, Statistical, nonlinear, and soft matter physics **2004,** *69*, 066138, DOI: 10.1103/PhysRevE.69.066138

(9) Kozachenko L. F.; N., L. N., Sample Estimate of the Entropy of a Random Vector. Problems Inform. Transmission **1987,** *23*, 95-101, DOI:

(10) Pieter, G.; Grigory, S.; Grant, B., The electronic structure in 3d transition metal complexes: Can we measure oxidation states? J. Phys. Conf. Ser. **2009,** *190*, 012046, DOI: 10.1088/1742-6596/190/1/012046

(11) Hawkins, D. M., The Problem of Overfitting. J. Chem. Inf. Comput. Sci. **2004,** *44*, 1-12, DOI: 10.1021/ci0342472

(12) Geurts, P.; Ernst, D.; Wehenkel, L., Extremely randomized trees. Mach. Learn. **2006,** *63*, 3-42, DOI: 10.1007/s10994-006-6226-1

(13) Vivanco, M.; Ruiz, J.; Floriani, C.; Chiesi-Villa, A.; Rizzoli, C., Chemistry of the vanadium-carbon .sigma. bond. 1. Insertion of carbon monoxide, isocyanides, carbon dioxide, and heterocumulenes into the V-C bond of tris(mesityl)vanadium(III). Organometallics **1993,** *12*, 1794-1801, DOI: 10.1021/om00029a041

(14) Razuvaev, G. A.; Latyaeva, V. N.; Vyshinskaya, L. I.; Drobotenko, V. V., Synthesis and properties of covalent tri- and tetravalent vanadium. J. Organomet. Chem **1981,** *208*, 169-182, DOI: 10.1016/S0022-328X(00)82672-8

(15) Zhang, F.; Song, H.; Zi, G., Synthesis and catalytic activity of group 5 metal amides with chiral biaryldiamine-based ligands. Dalton Trans. **2011,** *40*, 1547-1566, DOI: 10.1039/C0DT01229G

(16) Belov, D. S.; Fenoll, D. A.; Chakraborty, I.; Solans-Monfort, X.; Bukhryakov, K. V., Synthesis of Vanadium Oxo Alkylidene Complex and Its Reactivity in Ring-Closing Olefin Metathesis Reactions. Organometallics **2021,** *40*, 2939-2944, DOI: 10.1021/acs.organomet.1c00425

(17) Conley, M. P.; Delley, M. F.; Siddiqi, G.; Lapadula, G.; Norsic, S.; Monteil, V.; Safonova, O. V.; Copéret, C., Polymerization of Ethylene by Silica-Supported Dinuclear CrIII Sites through an Initiation Step Involving C□H Bond Activation. Angew. Chem., Int. Ed. **2014,** *53*, 1872-1876, DOI: 10.1002/anie.201308983

(18) Werner, D.; Anwander, R., Unveiling the Takai Olefination Reagent via Tris(tert-butoxy)siloxy Variants. **2018,** *140*, 14334-14341, DOI: 10.1021/jacs.8b08739
(19) Cameron, T. S.; Clyburne, J. A.; Dubey, P. K.; Grossert, J. S.; Ramaiah, K.; Ramanatham, J.; Sereda, S. V., Compounds of chromium(VI): The pyridine □ chromic anhydride complex, benzimidazolinium dichromate, and three 2-alkyl-1H-benzimida -zolinium dichromates. Can. J. Chem. **2003,** *81*, 612-619, DOI: 10.1139/v03-042
(20) Sheldrick, G., SHELXT - Integrated space-group and crystal-structure determination. Acta Crystallogr. A **2015,** *71*, 3-8, DOI: 10.1107/S2053273314026370
(21) Sheldrick, G., Crystal structure refinement with SHELXL. Acta Crystallogr. C **2015,** *71*, 3-8, DOI: 10.1107/S2053229614024218
(22) Dolomanov, O. V.; Bourhis, L. J.; Gildea, R. J.; Howard, J. A. K.; Puschmann, H., OLEX2: a complete structure solution, refinement and analysis program. J. Appl. Crystallogr. **2009,** *42*, 339-341, DOI: 10.1107/S0021889808042726
(23) Zabilska, A.; Clark, A. H.; Moskowitz, B. M.; Wachs, I. E.; Kakiuchi, Y.; Copéret, C.; Nachtegaal, M.; Kröcher, O.; Safonova, O. V., Redox Dynamics of Active VOx Sites Promoted by TiOx during Oxidative Dehydrogenation of Ethanol Detected by Operando Quick XAS. JACS Au **2022,** *2*, 762-776, DOI: 10.1021/jacsau.2c00027
(24) Trummer, D.; Searles, K.; Algasov, A.; Guda, S. A.; Soldatov, A. V.; Ramanantoanina, H.; Safonova, O. V.; Guda, A. A.; Copéret, C., Deciphering the Phillips Catalyst by Orbital Analysis and Supervised Machine Learning from Cr Pre-edge XANES of Molecular Libraries. JACS **2021,** *143*, 7326-7341, DOI: 10.1021/jacs.0c10791
(25) Cotton, F. A.; Rice, C. E.; Rice, G. W., The crystal and molecular structures of bis(2,4-pentanedionato)chromium. **1977,** *24*, 231-234, DOI: https://doi.org/10.1016/S0020-1693(00)93880-5
(26) Ocone, L. R.; Block, B. P.; Collman, J. P.; Buckingham, D. A., Anyhdrous Chromium(II) Acetate, Chromium(II) Acetate 1-Hydrate, and Bis(2,4-Pentanedionato)Chromium (II). In *Inorg. Synth.*, 1966; pp 125-132.
(27) Kakiuchi, Y.; Shapovalova, S.; Protsenko, B.; Guda, S.; Safonova, O. V.; Guda, A.; Copéret, C., Influence of strong π-acceptor ligands on Cr-K-edge X-ray absorption spectral signatures and consequences for the interpretation of surface sites in the Phillips catalyst. Catal. Sci. Technol. **2024,** *14*, 3682-3690, DOI: 10.1039/D3CY01692G
(28) Ciborska, A.; Chojnacki, J.; Wojnowski, W., Bis(tetra-hydro-furan-κO)tris-(tri-tert-butoxy-siloxy)chromium(III). **2007,** *63*, m1103-m1104, DOI: https://doi.org/10.1107/S1600536807011038
(29) Herrmann, W. A.; Huber, N. W.; Anwander, R.; Priermeier, T., Monomere flüchtige Alkoxide von Chrom und Bismut. **1993,** *126*, 1127-1130, DOI: https://doi.org/10.1002/cber.19931260510
(30) Barker, G. K.; Lappert, M. F.; Howard, J. A. K., Silylmethyl and related complexes. Part 6. Preparation, properties, and crystal and molecular structure of tris[bis(trimethylsilyl)methyl]-chromium(III);the chemistry of related compounds of titanium(III), vanadium(III), zirconium(IV), and hafnium(IV). **1978**, 734-740, DOI: 10.1039/DT9780000734
(31) Schulzke, C.; Enright, D.; Sugiyama, H.; LeBlanc, G.; Gambarotta, S.; Yap, G. P. A.; Thompson, L. K.; Wilson, D. R.; Duchateau, R., The Unusual Stability of Homoleptic Di- and Tetravalent Chromium Alkyls. **2002,** *21*, 3810-3816, DOI: 10.1021/om020237z
(32) Ravel, B.; Newville, M., ATHENA and ARTEMIS Interactive Graphical Data Analysisusing IFEFFIT. **2005**, 1007, DOI: 10.1238/physica.topical.115a01007
(33) Newville, M., IFEFFIT : interactive XAFS analysis and FEFF fitting. **2001,** *8*, 322-324, DOI: doi:10.1107/S0909049500016964
(34) Martini, A.; Guda, S. A.; Guda, A. A.; Smolentsev, G.; Algasov, A.; Usoltsev, O.; Soldatov, M. A.; Bugaev, A.; Rusalev, Y.; Lamberti, C., et al., PyFitit: The software for quantitative analysis of XANES

spectra using machine-learning algorithms. Comput. Phys. Commun. **2020,** *250*, DOI: 10.1016/j.cpc.2019.107064
(35) Weng, T.-C.; Waldo, G. S.; Penner-Hahn, J. E., A method for normalization of X-ray absorption spectra. J. Synchrotron Radiat. **2005,** *12*, 506-510, DOI: 10.1107/S0909049504034193
(36) Lee, J. C.; Xiang, J.; Ravel, B.; Kortright, J.; Flanagan, K., Condensed matter astrophysics: a prescription for determining the species-specific composition and quantity of interstellar dust using x-rays. ApJ **2009,** *702*, 970, DOI: 10.1088/0004-637X/702/2/970
(37) Newville, M., Larch: An Analysis Package for XAFS and Related Spectroscopies. J. Phys. Conf. Ser. **2013,** *430*, 012007, DOI: 10.1088/1742-6596/430/1/012007
(38) Martini, A.; Guda, A. A.; Guda, S. A.; Dulina, A.; Tavani, F.; D'Angelo, P.; Borfecchia, E.; Soldatov, A. V., Estimating a Set of Pure XANES Spectra from Multicomponent Chemical Mixtures Using a Transformation Matrix-Based Approach. In *Synchrotron Radiation Science and Applications. Springer Proceedings in Physics*, Di Cicco, A.; Giuli, G.; Trapananti, A., Eds. Springer: 2021; Vol. 220, pp 65-84.
(39) de Juan, A.; Jaumot, J.; Tauler, R., Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. Anal. Methods **2014,** *6*, 4964-4976, DOI: 10.1039/C4AY00571F
(40) Klementiev, K. V. XANES dactyloscope for Windows, freeware: www.desy.de/~klmn/xanda.html.
(41) Lawson, C. L.; Hanson, R. J., *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics: 1995.
(42) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E., Climbing the Density Functional Ladder: Nonempirical Meta--Generalized Gradient Approximation Designed for Molecules and Solids. **2003,** *91*, 146401, DOI: 10.1103/PhysRevLett.91.146401
(43) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Guerra, C. F.; Van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T., Chemistry with ADF. J. Comput. Chem. **2001,** *22*, 931-967, DOI: 10.1002/jcc.1056
(44) Guda, S. A.; Guda, A. A.; Soldatov, M. A.; Lomachenko, K. A.; Bugaev, A. L.; Lamberti, C.; Gawelda, W.; Bressler, C.; Smolentsev, G.; Soldatov, A. V., 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. **2015,** *11*, 4512-4521, DOI: 10.1021/acs.jctc.5b00327
(45) Joly, Y., X-ray absorption near-edge structure calculations beyond the muffin-tin approximation. Phys. Rev. B **2001,** *63*, 125120, DOI: 10.1103/PhysRevB.63.125120