ROYAL SOCIETY
OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

# Emulator-based decomposition for structural sensitivity of core-level spectra

J. Niskanen[1], A. Vladyka[1], J. Niemi[1] and C.J. Sahle[2]

[1]Department of Physics and Astronomy, University of Turku, 20014 Turun yliopisto, Finland
[2]European Synchrotron Radiation Source, 71 Avenue des Martyrs, 38000 Grenoble, France

AV, 0000-0003-1770-8139

**Cite this article:** Niskanen J, Vladyka A, Niemi J, Sahle CJ. 2022 Emulator-based decomposition for structural sensitivity of core-level spectra. *R. Soc. Open Sci.* **9**: 220093.
https://doi.org/10.1098/rsos.220093

Received: 28 January 2022
Accepted: 9 May 2022

**Subject Category:**
Chemistry

**Subject Areas:**
materials science/computational physics

**Keywords:**
X-ray absorption spectra, X-ray emission spectra, X-ray photoelectron spectra, machine learning

**Author for correspondence:**
J. Niskanen
e-mail: johannes.niskanen@utu.fi

We explore the sensitivity of several core-level spectroscopic methods to the underlying atomistic structure by using the water molecule as our test system. We first define a metric that measures the magnitude of spectral change as a function of the structure, which allows for identifying structural regions with high spectral sensitivity. We then apply machine-learning-emulator-based decomposition of the structural parameter space for maximal explained spectral variance, first on overall spectral profile and then on chosen integrated regions of interest therein. The presented method recovers more spectral variance than partial least-squares fitting and the observed behaviour is well in line with the aforementioned metric for spectral sensitivity. The analysis method is able to independently identify spectroscopically dominant degrees of freedom, and to quantify their effect and significance.

## 1. Introduction

Owing to orbital localization, core-level spectroscopies are sensitive to structure in the neighbourhood of the excited atomic site. However, the dependence between the atomistic arrangement and the resulting spectra is not straightforward, which complicates the analysis of these spectra [1–4]. A satisfactory solution to this complexity calls for new methods, such as machine learning (ML), that may relieve the computational burden of repeated function evaluations [5]. Here, the inherent lightness of evaluation may, for example, help with problems involving predictions of statistical averages or prediction of spectra for new structures instead of their explicit simulation. Several ML approaches have recently been applied to spectroscopy [6–11], typically to emulate the relations between known molecular/atomic structures and corresponding spectra [8,9]. The possibility to predict structural variations in the crystals using extended X-ray absorption fine

This article has been edited by the Royal Society of Chemistry, including the commissioning, peer review process and editorial aspects up to the point of acceptance.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6011537.

ROYAL SOCIETY OF CHEMISTRY

THE ROYAL SOCIETY
PUBLISHING

structure has also been demonstrated [7]. Moreover, prediction of X-ray absorption near-edge structure based on descriptors of the molecular structure has been recently shown with a high accuracy [10].

In this work, we turn to the question of how to apply an accurate ML emulator to the interpretation of core-level spectra in terms of the underlying atomistic structure. We develop an ML-based dimensionality reduction of the structural parameter space based on most covered spectral variance, and apply the method to simulations for three types: X-ray photoelectron spectra (XPS), X-ray emission spectra (XES), and X-ray absorption spectra (XAS). To interpret the findings, we present a metric to measure spectral sensitivity to structural change, and as a result we consistently identify regions of higher and lower spectroscopic structural sensitivity with the different methods.

# 2. Methods

## 2.1. Data and emulators

The number of electrons and the nuclear configuration, given by the set of all structural parameters **p**, define the electronic Hamiltonian and its spectra. We obtain transition energies and intensities for numerous structures **p** from electronic structure simulations. The transition intensities are approximated as squared lengths of the transition dipole vectors of the velocity form. To account for physical (lifetime, vibrational substructure) and instrumental lineshapes, the resulting 'stick spectrum' is convoluted. This procedure results in a continuous spectrum $\mathbf{S}(\mathbf{p})$, which on a predefined grid presents a vector. The procedure is repeated for a set of points **p** obtained from structural simulations. This work is based on applying ML to the simulated structure–spectrum pairs to create an emulator that approximates the function $\mathbf{S}(\mathbf{p})$ at any **p**.

As our data we use 10 000 snapshots from *ab initio* molecular dynamics (AIMD) trajectories for the $H_2O$ molecule, with initial kinetic energy equivalent to 10 000 K temperature and spectra simulated for these structures. The structural data and the related XAS spectra have been published previously [11]. For the calculation of XAS and XES spectra, we apply transition-potential density functional theory (TP-DFT). For evaluation of the XPS core-level binding energies, and for correction of the onset of XAS spectra, we carry out respective Δ-DFT calculations for the core-hole state energy with respect to the ground state. Here, we assume a high-enough photon energy to result in a constant O 1s ionization cross-section regardless of the structure. All spectra are convoluted with a 1.0 eV Gaussian and are presented on a 0.1-eV-spaced grid (100 points for all cases). The calculations are carried out using CP2K software [12]. For easier comparison with the experiment, the spectra are shifted by −6.0 eV, 2.25 eV, and 1.5 eV for XES, XAS, and XPS, respectively.

Our analysis relies on ML and the ability to predict spectra at new points in the configurational space, here defined by three degrees of freedom: H–O–H bond angle $\alpha$, and the shorter and longer O–H bond lengths $b_s$ and $b_l$, respectively. We select the ML spectroscopic emulators in a fashion similar to that of Niskanen *et al.* [11]. In brief, we examine polynomial models with the orders from 2 to 9, and multilayered perceptrons (MLP) with 2–5 hidden layers and 5–500 neurons in each layer, and use mean-squared error as a metric of the training quality for a set of 8000 data points. The `scikit-learn` [13] Python package is used. Based on cross-validation performance scores, we use an MLP emulator for XES, and polynomial emulators for XAS and XPS in the later stages of the analysis, carried out with a completely separate test set of 2000 samples. However, due to the wiggly behaviour of the MLP isosurfaces for XES spectra, we use the smoother-behaving polynomial emulators to produce all the plots in figure 1.

## 2.2. Spectral sensitivity metric

We measure structural sensitivity as the rate of change of spectrum $\mathbf{S}(\mathbf{p})$ at structural parameter point **p**. For vector-valued function **S**, we define the metric

$$M_{\mathrm{grad}}(\mathbf{p}) := \frac{\|\mathbf{J}_{\mathbf{S}}(\mathbf{p})\|_2}{\|\mathbf{S}(\mathbf{p}_{\mathrm{cen}})\|_2}, \tag{2.1}$$

where

$$[\mathbf{J}_{\mathbf{S}}(\mathbf{p}')]_{ij} = \left.\frac{\partial S_i}{\partial p_j}\right|_{\mathbf{p}=\mathbf{p}'}. \tag{2.2}$$

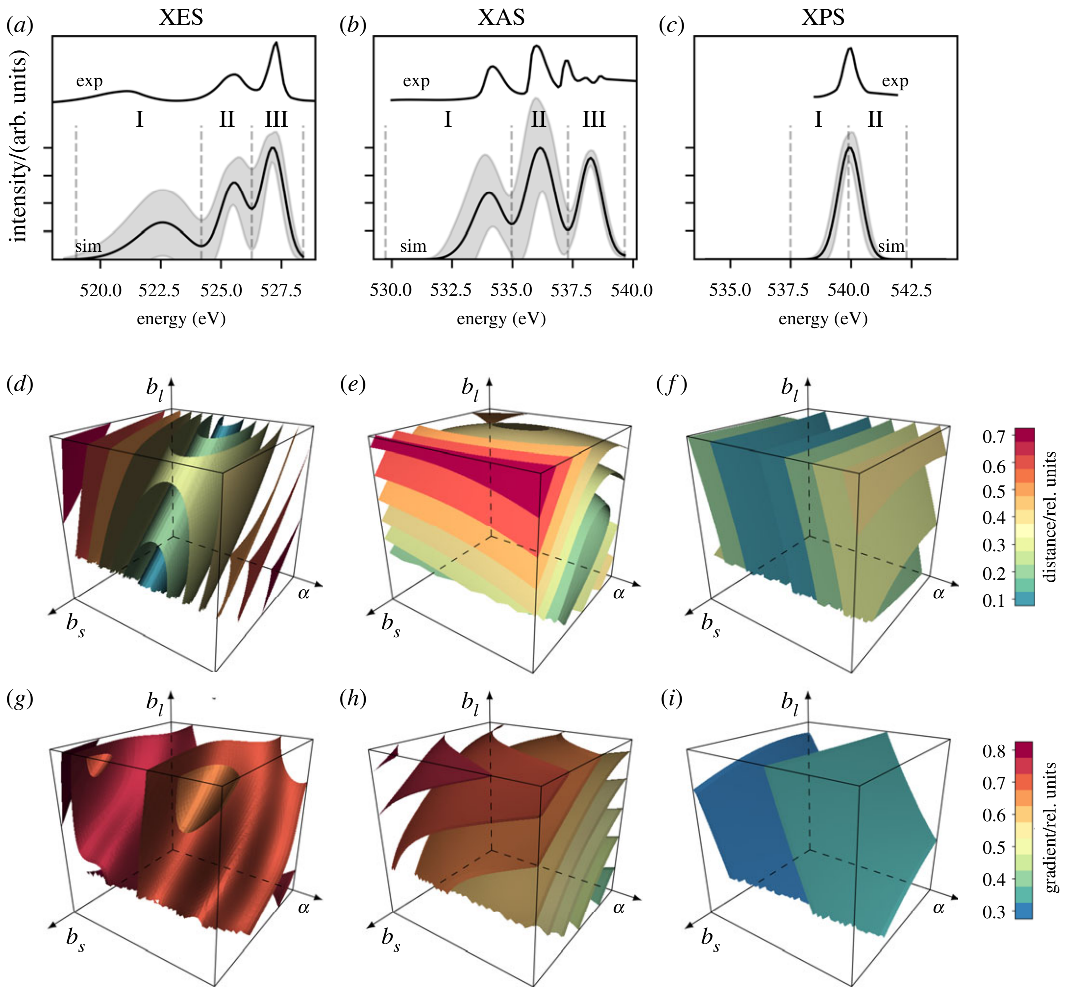

**Figure 1.** Spectra of the $H_2O$ molecule in the training dataset: (*a–c*) the mean spectrum is shown in black and the shaded area depicts ±1 s.d. from the mean; dashed lines indicate the regions of interest (ROIs I, II and III) for the coarsened spectra; digitized experimental spectra from [14–17] are shown for comparison; and simulated spectra have been shifted for the best match with the experiments. Structural sensitivity of these spectra: (*d–f*) Cartesian distance difference $M_{\text{diff}}$ and (*g–i*) Jacobian norm $M_{\text{grad}}$. Since polynomial approaches behave smoother, they were used also for the plots of XES. The ranges of the parameters shown are $\pm\sigma$ from the mean of the training set. For details, see text.

Each channel in the spectrum $\mathbf{S}$ is defined by the structural parameters $\mathbf{p}$. Thus, each row in the Jacobian gives the gradient of the particular energy channel with respect to structure. Spectral sensitivity with respect to a given structural parameter is given by the length of the according column vector. To classify points in the configuration space, we focus on the square norm of the whole Jacobian matrix. Since we compare different spectroscopies, normalization with the spectrum at the centre of the data $\mathbf{p}_{\text{cen}}$ set is applied.

An alternative metric is spectral deviation from that at the centre of the training set

$$M_{\text{diff}}(\mathbf{p}) := \frac{\|\mathbf{S}(\mathbf{p}) - \mathbf{S}(\mathbf{p}_{\text{cen}})\|_2}{\|\mathbf{S}(\mathbf{p}_{\text{cen}})\|_2} \tag{2.3}$$

Numerical calculations on a grid rely on evaluation of the ML predictor.

## 2.3. Emulator-based component analysis

The algorithm carries out a step-wise component vector (CV) search for dimensionality reduction in the structural parameter space, with the criterion to maximize the explained spectral variance together with the components of the previous steps. For a set of $N$ parameter points $\{\mathbf{p}_i\}_{i=1}^{N}$, this is achieved by projection on CVs optimized for the purpose. For each step $k$ ($k = 1, 2, \ldots$), a unit vector $\hat{\mathbf{v}}_k$ is searched so that generalized covered variance

$$\rho = 1 - \frac{\mathrm{tr}(\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}})}{\mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A})} \tag{2.4}$$

is maximized. Here, matrix $\mathbf{A}$ contains the true spectra of the original data points as its row vectors $\mathrm{A}_i$. The corresponding predicted spectra for projected data points are given as row vectors of matrix

$$\mathrm{A}_i^{(\mathrm{pred})} = \mathbf{S}^{(\mathrm{pred})}\left(\sum_{j=1}^{k}(\hat{\mathbf{v}}_j \cdot \mathbf{p}_i)\,\hat{\mathbf{v}}_j\right) \tag{2.5}$$

where function $\mathbf{S}^{(\mathrm{pred})}$ is an ML-based emulator capable of predicting spectra for previously unseen structures and

$$\tilde{\mathbf{A}} = \mathbf{A} - \mathbf{A}^{(\mathrm{pred})}. \tag{2.6}$$

We apply the orthonormality constraint $\hat{\mathbf{v}}_k \cdot \hat{\mathbf{v}}_j = \delta_{kj}$ to the CVs, and as the result of the procedure, a set of orthonormal projection vectors is obtained so that they always maximize the generalized covered spectral variance $\rho$ up to the given order $k$. We apply an overall factor ±1 for the CVs to point towards increasing intensity.

The generalized covered variance $\rho$ accounts for the goodness score in the spectrum space and is necessitated by the nonlinearity of spectrum prediction operation $\mathbf{S}^{(\mathrm{pred})}$. When applied to a data matrix from a projection in the same linear space, the definition reduces to that of covered variance used, for example, in principal component analysis. Due to its definition, $\rho$ may obtain negative values for notably bad predictions as the value zero corresponds to errors with the magnitude of the variance of the known data. We see no problem in alternatively using the remaining unexplained spectral variance $1 - \rho$ as an error metric in a minimization problem for vectors $\hat{\mathbf{v}}_k$.

## 2.4. Partial least-squares fits using SVD

We adapt an approach based on singular value decomposition (PLSSVD) [18] owing to its straightforward simplicity and to orthogonality of the CVs. Here, the partial least-squares fit is applied to data in matrices $\mathbf{X}$ and $\mathbf{Y}$ that contain standardized structural parameters and the corresponding standardized spectra in their row vectors. A linear fit is applied between the component scores of left and right eigenvectors for each order of the decomposition. As a result, an approximation of data

$$\mathbf{Y} \approx \mathbf{X}\sum_{j=1}^{k} U^{(j)} c_j V^{(j)\mathrm{T}} \tag{2.7}$$

is obtained. In the equation, $U^{(j)}$ and $V^{(j)}$ denote the left and right eigenvectors (column vectors) corresponding to the eigenvalue $\lambda_j$ ordered in descending fashion. As the data are standardized in each of their dimensions, the covariance matrix reads directly

$$\mathrm{cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{U}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_k)\,\mathbf{V}^{\mathrm{T}} \tag{2.8}$$

from which the matrices $\mathbf{U}$,$\mathbf{V}$ and diag($\lambda_1, \ldots, \lambda_k$) are obtained by singular value decomposition. The procedure thus gives basis vectors on which to project the data $\mathbf{X}$ and $\mathbf{Y}$.

The coefficients $c_j$ are obtained from a linear least-squares fit between projected data points $\mathbf{X}U^{(j)}$ and $\mathbf{Y}V^{(j)}$ for each order $j = 1, 2, \ldots$. The constant term in the fits is negligible and the first-order coefficient is assigned $c_j$. As an example, the results of the fits for the overall spectrum case are depicted in the electronic supplementary material. For comparison of the PLSSVD fit results, generalized explained variance metrics are evaluated for decompositions cumulatively incremented up to order $k$, as given by equation (2.7). An overall factor ±1 is applied for the PLSSVD structural space basis vectors to point towards increasing intensity.

# 3. Results and discussion

Although a static classical nuclei model is used, the appearance of the studied spectra of the $H_2O$ molecule in figure 1*a–c* are in agreement with the respective experiments [14–17]. The emulators trained on the sampled AIMD structures and corresponding spectra allow for easy and computationally light evaluation of the data on a mesh grid. We applied this capability to calculate the square norms of spectral deviation from that of the mean structure, as depicted in figure 1*d–f*. In addition, numerical differentiation of an emulator for the spectrum $\mathbf{S}(\mathbf{r})$ is a computationally light task on a mesh grid. Here, each partial derivative gives the rate of change for each channel in a spectrum $\mathbf{S}(\mathbf{r})$ at point $\mathbf{r}$ with respect to each structural parameter. The square norms of the Jacobian matrices $[\mathbf{J}_{\mathbf{S}}(\mathbf{r}')]_{ij} = \partial S_i/\partial r_j|_{\mathbf{r}=\mathbf{r}'}$ presented in figure 1*g–i* indicate strongest spectral changes in specific directions for each method. Normalization by the spectrum at the mean structure $\mathbf{r}_{\text{cen}}$ is applied in both cases to allow for a direct comparison.

The spectra show differing structural behaviour, with more variation in XES and XAS than XPS, also indicated by the channel-wise one standard deviation drawn together with the spectra. Figure 1*e*,*h* reveals that XAS is most sensitive to the symmetric stretch. This is seen as the largest isovalue surface being located at large $b_l$ and $b_s$ values, with little variation along the bond angle $\alpha$. On the other hand, the XPS spectrum changes most at high bond angles, as seen in figure 1*f*,*i*: isosurfaces are oriented parallel to the $b_l$–$b_s$ plane. From this view, XES is expected to be most sensitive to all structural parameters in the system, being least affected by the asymmetric stretch as seen in figure 1*d*,*g*. Here, the cartesian distance difference has a low-value isosurface region intersecting the plot of figure 1*d*, but the overall rate of change still has high isosurface values throughout the plot of figure 1*g*.

Spectroscopic data can be seen as two correlated datasets: one for structures and one for the corresponding spectra. One way to analyse the interdependencies in such data is provided by partial least-squares (PLS) fitting [19,20], and a variant of this family of methods has already been applied to binding energies in XPS in aqueous solution [1]. In PLS algorithms, latent variables connecting the two datasets are searched for using only existing data points. However, we show that the relation of structure and spectra may be investigated more deeply with the help of an ML-based emulator that is capable of making accurate and computationally light predictions of new data. Indeed, for a set of parameters defining the Hamiltonian, the spectra are defined as a function. We use the aforementioned capabilities of a good emulator and make a step-wise parameter-space decomposition, where the search for structural space CVs is guided by covering of maximal variance in the spectrum space. Because the search for each CV consists of an iterative solution of an optimization problem, the lightness of evaluation of the emulator is essential. Moreover, this emulator-based component analysis (ECA) routine relies on prediction of spectra on new data, i.e. projected data points in the standardized structural parameter space.

When compared with the results of PLS implemented on eigenvectors from singular value decomposition of the covariance matrix (PLSSVD) [18], the ECA algorithm is able to explain more spectral variance with a decomposition to a given order (table 1). Consequently, explained structural variance for ECA may be less than for the PLSSVD. We understand this by the design principle of ECA to search for directions that matter the most for spectra, with no emphasis on covered structural variance. Moreover, the nonlinearity of ECA allows for a tighter match with the data than linear methods. The first CVs of the methods agree in interplay of all structural parameters, in opposing directions for angle and bond lengths for XES. Likewise, the overall shape of XAS is agreed to be dominantly affected by the bond lengths, and the XPS is virtually completely explained by the H–O–H angle. The results are also depicted in figure 2 and these findings are consistent with the spectral sensitivity metrics presented in figure 1.

Interpretation of experimental core-level spectra is complicated by unavoidable inaccuracy of the spectrum simulations. As a solution to the problem, we have previously proposed an analysis of spectral regions of interest (ROI) that are identifiable in both experimentation and theory [2,3,11,21,22]. In such a line of thought, it is argued that the risk of overanalysis is reduced, as the procedure would naturally focus on confirmedly reproduced spectral features. An alternative approach to assess uncertainties in simulated X-ray spectra has been presented by Bergmann *et al.* [23]. By studying the spectral response to slight structural distortions, their method results in error bars for calculated spectra for more reliable interpretation of the experiment.

We analysed the behaviour of ROIs marked in figure 1*a–c* with two approaches: simultaneous and independent for each ROI. A joint treatment of ROIs revealed that some regions dominated the component analysis at the cost of the others. This occurred due to different overall variances in the ROI intensities seen in figure 1*a–c*. For example, the optimization of the first CV became dictated by XES ROI I, which resulted in highly sub-optimal description of ROI III intensity. Therefore, we conclude that interpretation of ROIs is best done by individual fitting, i.e. analysing each ROI separately.

The results of individual analyses for each ROI are presented in figure 3 and in table 2. When performed this way, the first CVs explain on average (87 ± 14)% of ROI intensity variance with the mean structural covered variance of (38 ± 7)%, as indicated by figure 3*b–c*. The first PLSSVD CVs

**Table 1.** Analysis of the overall shape of spectra in increasing order of decomposition: cumulative fractional explained variance in spectral ($\sigma^2_{\text{spec}}$) and structural ($\sigma^2_{\text{stru}}$) space and the corresponding CVs in the standardized parameter space.

| | $k$ | $\sigma^2_{\text{spec}}$ | $\sigma^2_{\text{stru}}$ | $\alpha$ | $b_l$ | $b_s$ | $\sigma^2_{\text{spec}}$ | $\sigma^2_{\text{stru}}$ | $\alpha$ | $b_l$ | $b_s$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ECA | | | | | PLSSVD | | | | |
| XES | 1 | 0.74 | 0.41 | [0.88 | −0.34 | −0.32] | 0.38 | 0.47 | [0.77 | −0.44 | −0.47] |
| | 2 | 1.00 | 0.84 | [−0.47 | −0.65 | −0.59] | 0.51 | 0.85 | [−0.64 | −0.54 | −0.55] |
| | 3 | 1.00 | 1.00 | [0.00 | −0.67 | 0.74] | 0.51 | 1.00 | [0.01 | −0.72 | 0.69] |
| XAS | 1 | 0.75 | 0.50 | [0.16 | 0.66 | 0.74] | 0.50 | 0.50 | [0.07 | 0.74 | 0.67] |
| | 2 | 0.91 | 0.67 | [−0.20 | 0.75 | −0.63] | 0.53 | 0.84 | [−0.98 | 0.17 | −0.09] |
| | 3 | 1.00 | 1.00 | [−0.97 | −0.05 | 0.25] | 0.58 | 1.00 | [0.18 | 0.65 | −0.74] |
| XPS | 1 | 0.99 | 0.29 | [0.96 | 0.26 | 0.03] | 0.89 | 0.32 | [−0.99 | −0.17 | −0.05] |
| | 2 | 1.00 | 0.80 | [0.14 | −0.42 | −0.90] | 0.88 | 0.78 | [0.17 | −0.93 | −0.33] |
| | 3 | 1.00 | 1.00 | [−0.23 | 0.87 | −0.44] | 0.88 | 1.00 | [0.01 | −0.34 | 0.94] |

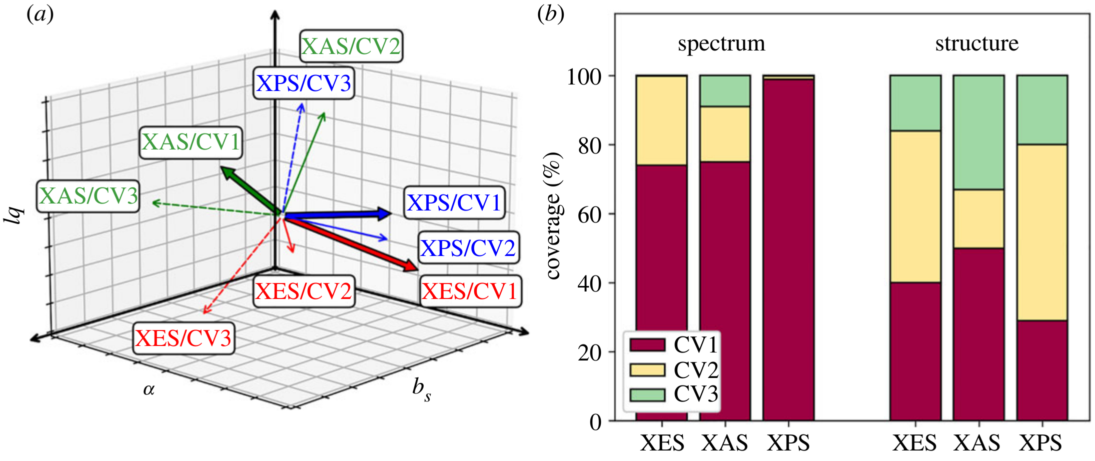

**Figure 2.** ECA of the full spectra. (*a*) Orientation of the component vectors; different colours indicate the type of spectroscopy and line type depicts the components. (*b*) Ratios of explained variances for spectrum and for structure.

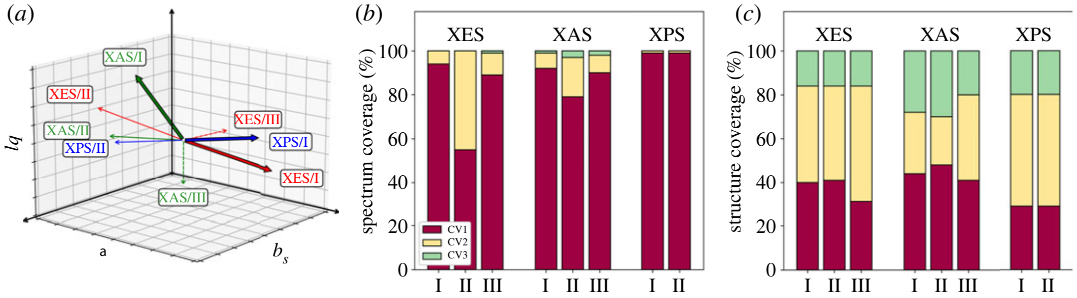

**Figure 3.** ROI-wise ECA of the spectra. (*a*) Orientation of the first component vectors; different colours indicate the type of spectroscopy and line type depicts the ROI. (*b*) Ratios of explained spectral variances. (*c*) Ratios of explained structural-parameter variances.

show a weaker (68 ± 27)% performance for covered spectral variance but cover (42 ± 9)% of the structural variance. Standard deviations are given as the uncertainties above.

The CVs were oriented along the increase of corresponding ROI intensity. Whereas this is a trivial task for linear models, defining the positive direction is more complicated for ECA, because of nonlinear and possibly oscillatory behaviour of intensity along the component (see electronic supplementary material). Our analysis reports dominant dependence on the H–O–H angle of all ROIs in XES spectra: based on the first CVs intensity transfer to ROI II is expected with inward bending. The ROIs in XAS are mostly affected by the bond lengths, and, for example, ROI I intensity is found to be increased with further elongation of the longer bond. Last, the sensitivity of XPS to the H–O–H bond angle only is recovered, as intensity is shifted to lower binding energies with increasing bend angles.

In the $H_2O$ molecule that we use as the pilot system, there are only three nuclear degrees of freedom. It is therefore relevant to ask what would change if a problem with more degrees of freedom, such as a liquid, was to be studied. We turn to this question next.

All other things being equal, a more complicated system can be expected to require a more complicated emulator architecture. This naturally will require larger training (and test) datasets that should cover the whole region of prediction [11], i.e. accessible structural space. The field of ML provides measures how to evaluate the model and the number of required training points, by, for example, studying the learning curves. For the water molecule alone, a simple three-dimensional grid evaluation would have been feasible. However, for more complicated systems, the number of dimensions would prohibit such a raw approach. We see (AI)MD and Monte Carlo simulations as feasible ways to generate structures, as the achieved sampling cuts out a large portion of the inaccessible structural space by design. These considerations are complicated by the note that the complexity of an emulator architecture depends also on how well behaving a function the spectral response is. Last, it remains a case-dependent question of how much precision loss is tolerated in the process.

**Table 2.** Component-wise ECA analysis of the ROI intensities: cumulative fractional explained variance in spectral ($\sigma^2_{\text{spec}}$) and structural ($\sigma^2_{\text{stru}}$) space and the corresponding CVs in the standardized parameter space. The CVs are oriented along increasing ROI intensity based on a linear fit on the predicted data for projection along the CV in question only.

| | $k$ | $\sigma^2_{\text{spec}}$ | $\sigma^2_{\text{stru}}$ | $\alpha$ | $b_l$ | $b_s$ | $\sigma^2_{\text{spec}}$ | $\sigma^2_{\text{stru}}$ | $\alpha$ | $b_l$ | $b_s$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECA | | | | | | PLSSVD | | | | |
| XES | ROI I | | | | | | | | | | |
| | 1 | 0.94 | 0.40 | [0.90 | −0.31 | −0.32] | 0.32 | 0.53 | [0.39 | −0.65 | −0.65] |
| | 2 | 1.00 | 0.84 | [−0.44 | −0.67 | −0.59] | | | | | |
| | 3 | 1.00 | 1.00 | [−0.04 | 0.67 | −0.74] | | | | | |
| | ROI II | | | | | | | | | | |
| | 1 | 0.55 | 0.41 | [−0.89 | 0.33 | 0.31] | 0.24 | 0.32 | [−0.90 | −0.29 | −0.32] |
| | 2 | 1.00 | 0.84 | [−0.46 | −0.62 | −0.64] | | | | | |
| | 3 | 1.00 | 1.00 | [−0.02 | −0.71 | 0.70] | | | | | |
| | ROI III | | | | | | | | | | |
| | 1 | 0.88 | 0.31 | [0.84 | 0.43 | 0.32] | 0.69 | 0.36 | [0.70 | 0.49 | 0.53] |
| | 2 | 0.99 | 0.84 | [−0.53 | 0.62 | 0.57] | | | | | |
| | 3 | 1.00 | 1.00 | [0.03 | −0.65 | 0.76] | | | | | |
| XAS | ROI I | | | | | | | | | | |
| | 1 | 0.92 | 0.45 | [−0.42 | 0.88 | 0.25] | 0.88 | 0.52 | [−0.38 | 0.76 | 0.53] |
| | 2 | 0.99 | 0.72 | [−0.15 | 0.20 | −0.97] | | | | | |
| | 3 | 1.00 | 1.00 | [0.90 | 0.44 | −0.05] | | | | | |
| | ROI II | | | | | | | | | | |
| | 1 | 0.79 | 0.48 | [−0.15 | 0.28 | 0.95] | 0.58 | 0.49 | [−0.24 | 0.38 | 0.89] |
| | 2 | 0.97 | 0.70 | [−0.14 | −0.95 | 0.26] | | | | | |
| | 3 | 1.00 | 1.00 | [0.98 | −0.09 | 0.18] | | | | | |

(*Continued.*)

**Table 2.** (*Continued.*)

| | $k$ | $\sigma^2_{\text{spec}}$ | $\sigma^2_{\text{stru}}$ | $\alpha$ | $b_l$ | $b_s$ | $\sigma^2_{\text{spec}}$ | $\sigma^2_{\text{stru}}$ | $\alpha$ | $b_l$ | $b_s$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECA | | | | | | PLSSVD | | | | |
| | ROI III | | | | | | | | | | |
| | 1 | 0.90 | 0.42 | [−0.33 | −0.86 | −0.39] | 0.80 | 0.51 | [−0.04 | −0.76 | −0.65] |
| | 2 | 0.98 | 0.80 | [0.92 | −0.20 | −0.33] | | | | | |
| | 3 | 1.00 | 1.00 | [0.20 | −0.47 | 0.86] | | | | | |
| XPS | ROI I | | | | | | | | | | |
| | 1 | 0.99 | 0.29 | [0.97 | 0.26 | 0.02] | 0.98 | 0.32 | [0.99 | 0.16 | 0.03] |
| | 2 | 1.00 | 0.80 | [0.13 | −0.38 | −0.92] | | | | | |
| | 3 | 1.00 | 1.00 | [0.23 | −0.89 | 0.40] | | | | | |
| | ROI II | | | | | | | | | | |
| | 1 | 0.99 | 0.29 | [−0.97 | −0.26 | −0.02] | 0.98 | 0.32 | [−0.99 | −0.16 | −0.03] |
| | 2 | 1.00 | 0.80 | [−0.13 | 0.38 | 0.92] | | | | | |
| | 3 | 1.00 | 1.00 | [−0.23 | 0.89 | −0.40] | | | | | |

The idea of using decomposition is to provide interpretation of spectroscopic data learned by an emulator. The aim is to identify dominant trends in a complicated structure–spectrum relation, with inherent loss of information. In this work, we used a linear transformation around a well-identifiable centre to identify relevant directions of spectral sensitivity. For more complicated data such as liquids, these centres may be numerous or a continuous valley of regions may appear—possibly with varying local spectral behaviour. As one potential way to solve the problem, a manifold approach might be used. In such an approach, locally linear variations would be studied together with additional parameters defining the local neighbourhood, e.g. particular molecular isomer. Such parametrizations could be made by energy criteria, by abundance of points in an MD trajectory, or by principal component or clustering analysis of the structural data. However, for spectral data that is severely wiggly or heavily scattered over the accessible structural space, it is hard to see any interpretation method to be able to draw correct universal trends from, as inverting the structure–spectrum function becomes impossible. It seems that a structural-information bottleneck can be reached in at least two ways: first, due to insensitivity of the probe to certain structural variation and, second, due to the back-and-forth wiggle of the spectra in the structural parameter space.

## 4. Conclusion

Spectroscopically relevant structural variability can be captured by decomposition techniques. Using ML-based emulators allows for decomposition of structural space based on explained spectral variance; this is an approach that outperforms partial least-squares fitting both in spectral coverage and structural selectivity. The presented ECA method relies on accurate and computationally light prediction of spectra for new structures enabled by ML emulators, the development of which is currently an active field of research. Application of this analysis on ROIs in the spectrum may provide a direct interpretation of experimentally observed and theoretically reproduced spectral change. Our results manifest X-ray spectra forming a bottleneck for structural information, some of which is not recoverable from them. Whereas high sensitivity might be beneficial for a detailed analysis of structure, sensitivity to only a few structural parameters may be used for identification of the related structural classes by their spectroscopic fingerprints. On the other hand, spectroscopic methods that are heavily sensitive to many parameters may require a statistical approach.

Data accessibility. Raw data and related codes are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.dncjsxm1m [24]. The data are provided in the electronic supplementary material [25].

Authors' contributions. J.N.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, validation, writing—original draft, writing—review and editing; A.V.: data curation, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; J.Niemi: validation; C.S.: conceptualization, data curation, methodology.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. The authors declare no competing interests.

Funding. J.N. is funded by University of Turku and Academy of Finland via project 331234. A.V. acknowledges funding by Academy of Finland via project 331234. J.Niemi declares no funding. C.S. is funded by ESRF.

Acknowledgements. Prof. E. Kukk is thanked for discussions.

## References

1. Ottosson N *et al.* 2011 On the origins of core-electron chemical shifts of small biomolecules in aqueous solution: insights from photoemission and ab initio calculations of glycine (aq). *J. Am. Chem. Soc.* **133**, 3120–3130. (doi:10.1021/ja110321q)
2. Niskanen J *et al.* 2016 Sulphur K$\beta$ emission spectra reveal protonation states of aqueous sulfuric acid. *Sci. Rep.* **6**, 21012. (doi:10.1038/srep21012)
3. Niskanen J, Sahle CJ, Gilmore K, Uhlig F, Smiatek J, Föhlisch A. 2017 Disentangling structural information from core-level excitation spectra. *Phys. Rev. E* **96**, 013319. (doi:10.1103/PhysRevE.96.013319)
4. Vaz da Cruz V *et al.* 2019 Probing hydrogen bond strength in liquid water by resonant inelastic X-ray scattering. *Nat. Commun.* **10**, 1013. (doi:10.1038/s41467-019-08979-4)
5. Hutson M. 2020 AI shortcuts speed up simulations by billions of times. *Science* **367**, 728. (doi:10.1126/science.367.6479.728)
6. Timoshenko J, Lu D, Lin Y, Frenkel AI. 2017 Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles. *J. Phys. Chem. Lett.* **8**, 5091–5098. (doi:10.1021/acs.jpclett.7b02364)
7. Timoshenko J, Anspoks A, Cintins A, Kuzmin A, Purans J, Frenkel AI. 2018 Neural network approach for characterizing structural transformations by X-ray absorption fine structure spectroscopy. *Phys. Rev. Lett.* **120**, 225502. (doi:10.1103/PhysRevLett.120.225502)
8. Timoshenko J, Frenkel AI. 2019 'Inverting' X-ray absorption spectra of catalysts by machine learning in search for activity descriptors. *ACS Catalysis* **9**, 10 192–10 211. (doi:10.1021/acscatal.9b03599)

9. Ghosh K, Stuke A, Todorović M, Jørgensen PB, Schmidt MN, Vehtari A, Rinke P. 2019 Deep learning spectroscopy: neural networks for molecular excitation spectra. *Adv. Sci.* **6**, 1801367. (doi:10.1002/advs.201801367)
10. Carbone MR, Topsakal M, Lu D, Yoo S. 2020 Machine-learning X-ray absorption spectra to quantitative accuracy. *Phys. Rev. Lett.* **124**, 156401. (doi:10.1103/PhysRevLett.124.156401)
11. Niskanen J, Vladyka A, Kettunen JA, Sahle CJ. 2021 Machine learning in interpretation of electronic core-level spectra. (http://arxiv.org/abs/2104.02374).
12. Hutter J, Iannuzzi M, Schiffmann F, VandeVondele J. 2014 CP2K: atomistic simulations of condensed matter systems. *WIREs Comput. Mol. Sci.* **4**, 15–25. (doi:10.1002/wcms.1159)
13. Pedregosa F *et al.* 2011 Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.
14. Fransson T, Harada Y, Kosugi N, Besley NA, Winter B, Rehr JJ, Pettersson LGM, Nilsson A. 2016 X-ray and electron spectroscopy of water. *Chem. Rev.* **116**, 7551–7569. (doi:10.1021/acs.chemrev.5b00672)
15. Weinhardt L, Benkert A, Meyer F, Blum M, Wilks RG, Yang W, Bär M, Reinert F, Heske C. 2012 Nuclear dynamics and spectator effects in resonant inelastic soft x-ray scattering of gas-phase water molecules. *J. Chem. Phys.* **136**, 144311. (doi:10.1063/1.3702644)
16. Nordlund D, Ogasawara H, Andersson KJ, Tatarkhanov M, Salmerón M, Pettersson LG, Nilsson A. 2009 Sensitivity of X-ray absorption spectroscopy to hydrogen bond topology. *Phys. Rev. B - Conden. Matter Mater. Phys.* **80**, 2–5. (doi:10.1103/PhysRevB.80.233404)
17. Sankari R *et al.* 2003 Vibrationally resolved O 1s photoelectron spectrum of water. *Chem. Phys. Lett.* **380**, 647–653. (doi:10.1016/j.cplett.2003.08.108)
18. Bookstein FL, Sampson PD, Streissguth AP, Barr HM. 1996 Exploiting redundant measurement of dose and developmental outcome: new methods from the behavioral teratology of alcohol. *Dev. Psychol.* **32**, 404–415. (doi:10.1037/0012-1649.32.3.404)
19. Geladi P. 1988 Notes on the history and nature of partial least squares (PLS) modelling. *J. Chemom.* **2**, 231–246. (doi:10.1002/cem.1180020403)
20. Wegelin JA. 2000 A survey of partial least squares (PLS) Methods, with Emphasis on the Two-Block Case. Technical Report 371 Department of Statistics, University of Washington, Seattle.
21. Sahle CJ, Niskanen J, Gilmore K, Jahn S. 2018 Exchange-correlation functional dependence of the O 1*s* excitation spectrum of water. *J. Electron. Spectrosc. Relat. Phenom.* **222**, 57–62. (doi:10.1016/j.elspec.2017.09.003)
22. Niskanen J *et al.* 2019 Compatibility of quantitative X-ray spectroscopy with continuous distribution models of water at ambient conditions. *Proc. Natl Acad. Sci. USA* **116**, 4058–4063. (doi:10.1073/pnas.1815701116)
23. Bergmann TG, Welzel MO, Jacob CR. 2020 Towards theoretical spectroscopy with error bars: systematic quantification of the structural sensitivity of calculated spectra. *Chem. Sci.* **11**, 1862–1877. (doi:10.1039/C9SC05103A)
24. Niskanen J, Vladyka A, Niemi J, Sahle CJ. 2022 Data from: Emulator-based decomposition for structural sensitivity of core-level spectra. Dryad Digital Repository. (doi:10.5061/dryad.dncjsxm1m)
25. Niskanen J, Vladyka A, Niemi J, Sahle CJ. 2022 Emulator-based decomposition for structural sensitivity of core-level spectra. Figshare. (doi:10.6084/m9.figshare.c.6011537)