Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Electronic Supplementary Information (ESI)

Machine Learning Powered by Principal Component Descriptors as the Key for Sorted Structural Fit of XANES

A. Martini[1,2]*, A. A. Guda[1]*, S. A. Guda[1,3]*, A. Bugaev[1], O. V. Safonova[4], A. V. Soldatov[1]

[1]The Smart Materials Research Institute, Southern Federal University, 344090 Sladkova 178/24 Rostov-on-Don, Russia
[2]Department of Chemistry, INSTM Reference Center and NIS and CrisDi Interdepartmental Centers, University of Torino, Via P. Giuria 7, I-10125 Torino, Italy
[3]Institute of mathematics, mechanics and computer science, Southern Federal University, 344090 Milchakova 8a, Rostov-on-Don, Russia
[4]Paul Scherrer Institute, 5232 Villigen PSI, Switzerland

Corresponding Authors: *A. M.: andrea.martini@unito.it, *A. A. G.: guda@sfedu.ru, *S. A. G.: gudasergey@gmail.com

1. FDMNES parameters employed for the convolution procedure

In the simulations we used finite the difference approach (FDM). The variable energy step used in the calculations was 0.02 eV near Fermi level and 2 eV after 30 eV above the edge. The screening was represented by one electron on the 4s orbital of the Cu absorbing atom. The calculations were performed using the real Hedin, Lundqvist and Von Barth potential.[1, 2] In the small spheres around the atoms and in the outer sphere, the potential and wave function were expanded to spherical harmonics choosing the maximum value of angular momentum *l* as $kr = \sqrt{l_{max}(l_{max}+1)}$, in which *k* is the photo electron wave vector and *r* is a radius of the sphere.

The following parameters were selected for the convolution of the calculated spectrum employing energy dependant arctangent shape of the Lorentzian profile (details can be found in the FDMNES program manual):

Structure	Gamma hole (eV)	Ecent, (eV)	Elarge, (eV)	Gamma max (eV)	E Fermi (eV)
$[Cu(I)(NH_3)_2]^+$ $[Cu(II)(NH_3)_4]^{2+}$	1.5	30	30	15	8981.3
$CuCeO_2$	1.6	50	50	15	8981.0

Table S1: Lorentian convolution parameters employed to obtain the spectra reported in **Figures 2**, **6** and **13** (a) of the main text.

2. Two dimensional cross sections associated to the multipliers functions of the $[Cu(II)(NH_3)_4]^{2+}$ complex

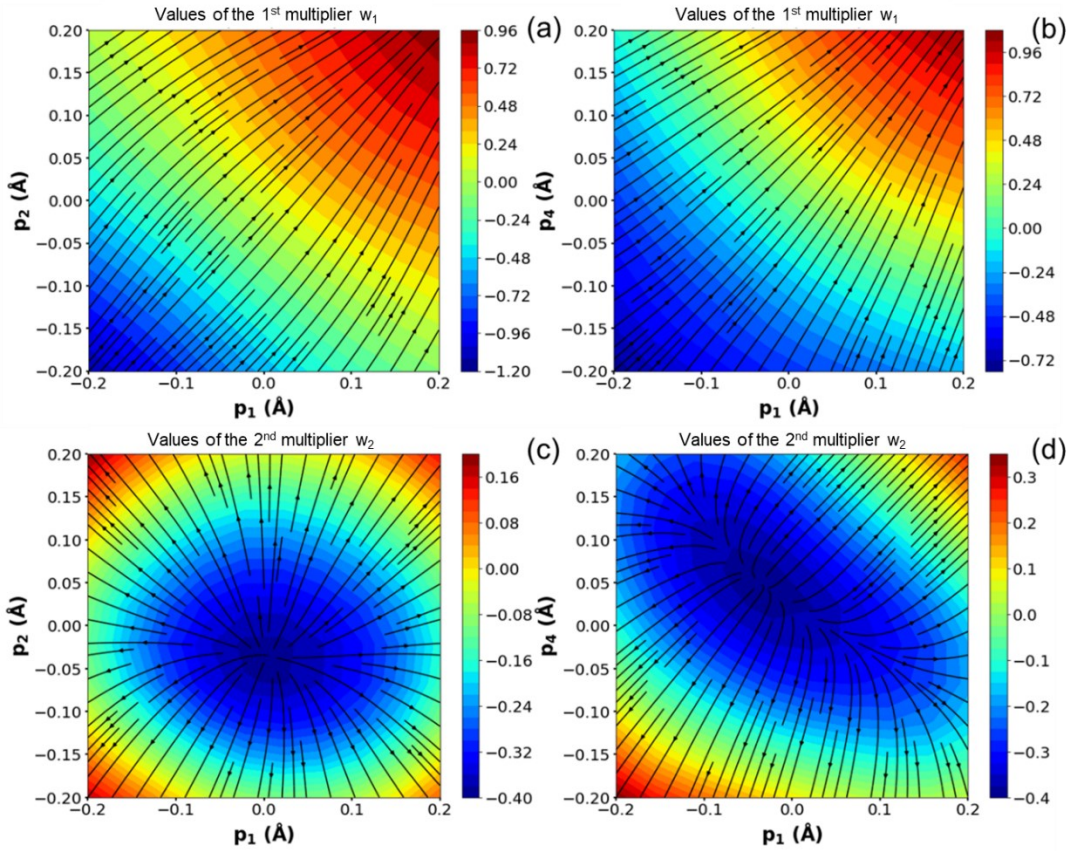

Figure S1: Four representative cross-sections corresponding to the first (a),(b) and second (c),(d) PCA multipliers functions: $w_1(p_1, ..p_4)$ and $w_2(p_1, ..p_4)$ of the $[Cu(II)(NH_3)_4]^{2+}$ complex. The arrows represent the gradient field emerging from the multipliers surfaces. Each panel has been obtained by keeping fixed to the null the variation of remaining two parameters

3. Analytical transfer to the optimal coordinates

Starting from the direction, characterizing the linear expression of w_1 obtained through the linear ridge regression, we normalized it and considered furthermore three new vectors of coefficient enabling to constitute an orthonormal basis in R^4. The set of coefficients defines a 4×4 transformation matrix which has been employed to convert the old variables/structural parameters (p_1,p_2,p_3,p_4) in the new-ones (d_0,d_1,d_2,d_3). Afterwards the latter have been used to rewrite the second multiplier (quadric) function w_2 as w_2'. The requirement that $w_1(p_1,p_2,p_3,p_4) = w_1^{exp}$ lead to the following relation: $d_0 = w_1/\|k\|$, where $k = (k_1,k_2,k_3,k_4)$ are the coefficient multiplying the linear terms (p_1,p_2,p_3,p_4) in the w_1. As introduced before, this constraint reduces by one the dimension of the ellipsoidal equation of the second multiplier function w_2'. The expression of w_2' was then written in canonical form with the related three axis expressed as a function of the initial variables (p_1,p_2,p_3,p_4) and indicated in the main text as d_1, d_2, d_3 (for the $[Cu(II)(NH_3)_4]^{2+}$) and with a further direction d_4 for the $CuCeO_2$ case of study.

4. Quality of estimation of the ML-based approximations

Regressor	$[Cu(I)(NH_3)_2]^+$	$[Cu(II)(NH_3)_4]^{2+}$	$CuCeO_2$
RBF	w_1:99.9	w_1: 99.7%	w_1: 99.4%
	w_2:97.8	w_2: 98.1 %	w_2: 93.4%
Ridge		w_1: 96.7%	w_1: 96.6%
Ridge Quadric		w_2: 97.9%	w_2: 84%

Table S2: R^2 score accuracy associated to the different regressors employed in the main text to approximate the XANES multipliers w_1 and w_2 as a function of the variation of the selected set of structural parameters (p_1, …, p_n). The quantities reported in the table have been obtained through a ten-fold cross-validation approach.

References

1. A. A. Guda, S. A. Guda, A. Martini, A. L. Bugaev, M. A. Soldatov, A. V. Soldatov and C. Lamberti, *Radiation Physics and Chemistry*, 2019, 108430.
2. Y. Joly, *Phys. Rev. B*, 2001, **63**, 125120.