Quantifying similarity for spectra with a large number of overlapping transitions: Examples from soft x-ray spectroscopy

Erik Källman[a,*], Mickaël G. Delcey[a], Meiyuan Guo[a], Roland Lindh[b,c], Marcus Lundberg[a,**]

[a] *Department of Chemistry-Ångström Laboratory, Uppsala University, SE-751 20 Uppsala, Sweden*
[b] *Department of Chemistry-BMC, Uppsala University, P.O. Box 576, SE-751 23 Uppsala*
[c] *Uppsala Center for Computational Chemistry (UC_3), Uppsala University, P.O. Box 596, SE-751 24 Uppsala, Sweden*

Abstract

Theoretical simulations are frequently used to assign electronic and geometric structure from spectral fingerprints. However, such assignments are prone to expectation bias. Bias can be reduced by using numerical measures of the similarity between calculated and experimental spectra. However, the commonly used point-wise comparisons cannot handle larger deviations in peak position. Here a weighted cross-correlation function is used to evaluate similarity scores for soft x-ray spectra of first-row transition metals. These spectra consist of hundreds of overlapping resonances, which makes spectral decomposition difficult. They are also challenging to model, leading to significant errors in both peak position and intensity. It is first shown how the choice of weight-function width can be related to the modeling errors. The method is then applied to evaluate the sensitivity of multiconfigurational wavefunction and charge-transfer multiplet simulations to model choices. The approach makes it possible to assess the reliability of assignments from spectral fingerprinting.

Keywords: X-ray spectroscopy, Spectral similarity, Correlation functions,

*Corresponding author
**Corresponding author
Email addresses: erik.kallman@kemi.uu.se (Marcus Lundberg), marcus.lundberg@kemi.uu.se (Marcus Lundberg)

1. Introduction

A fundamental aspect of electronic structure theory is to describe, explain, and ideally predict the outcome of experiments. In many cases the key target is the system energy, which determines the relative stability of different species. The accuracy and reliability of a method can then be evaluated using benchmarks, which makes it possible to systematically compare methods and improve performance. For complex systems where energy cannot be accurately calculated, or when studying systems far from equilibrium, simulations of spectroscopic data are important to predict electronic and geometric structure. However, unlike calculations of energies, the agreement between simulated and experimental spectra is often evaluated by visual inspection. This introduces significant expectation bias. It also limits the possibilities to compare different systems and methods. To properly establish the reliability of assignments based on spectral simulations requires unbiased numerical estimates of similarity. The applicability of different similarity measures depends on the characteristics of the spectra that are being evaluated, e.g., NMR,[1, 2, 3], infrared,[4], UV/Vis,[5] or x-ray diffraction.[6, 7, 8]

First-row transition metal systems, with their near-degenerate 3d electron levels, are prominent examples of systems with close-lying electronic states. This can lead to difficulties in calculating both electronic ground states and reaction pathways.[9, 10] Metal L-edge ($2p \rightarrow 3d$ excitations) x-ray absorption spectroscopy (XAS) is an element-specific probe of the 3d orbitals and is widely used to extract electronic structure information.[11, 12, 13, 14] However, with both 2p and 3d open shells. these systems can have very high densities of final states.[15] Individual transitions are further broadened through the limited lifetime of the 2p core hole state, which leads to complicated spectral shapes that cannot easily be decomposed into spectral contributions.

Broadly speaking, two different approaches are be used to simulate L-edge

XAS spectra of transition metal coordination complexes, with and without system-dependent parameters. For the former, a widely used method is the semi-empirical charge-transfer multiplet (CTM) model.[11] Here model parameters are fitted to maximize the similarity between simulation and experiment, and the values of these parameters give detailed information about metal-ligand interactions.[16, 12] Methods without system-dependent parameters can be used to identify complexes through spectral fingerprints.[17, 18] For molecular systems commonly used approaches are density-functional theory (DFT),[19, 20, 21, 22] and multiconfigurational wavefunction methods.[23, 24, 25, 26, 27, 28].

Both parameter fitting and fingerprinting requires a quantitative measure of similarity with as little bias as possible. The similarity analysis should ideally:

1. have an upper and lower bound, making it transferable between data sets and methods.
2. be able to account for errors both in energy and intensity.
3. be applicable to cases of partially or completely overlapping peaks.
4. have a smooth response to changes in spectral shape.

Commonly used measures of the difference between data sets like euclidean distance (ED) and root-mean-square-deviation (RMSD) are simple to visualize but dependent on the intensity scale and are not bounded. The cosine similarity (CS) and the coherence (COH) have previously been used in the field of spectroscopy and signal analysis, and have the same bounds for vectors of real, positive numbers: [0,1].[4, 5, 29, 30] The CS has previously been used to compare L-edge XAS spectra of iron-porphyrin systems.[18] However, like other vector distance methods, it fails to account for deviations in peak position larger than the line width. When performing the similarity analysis, a simulated peak in the wrong position results in both deviations from the lack of intensity at the right position and further deviations from the added intensity in the wrong position. In other words, a shifted peak results in twice the error compared to peaks not being present in the spectrum at all. This is clearly an undesirable feature when comparing different simulation approaches.

It has been shown that the distance measures can be expressed in terms of a cross-correlation function.[1, 2] Similarity between neighboring points in a spectrum can then be accounted for by weighting the cross-correlation integrals using a matrix expression.[31] Alternative weighting functions, such as triangle, a Gaussian or Lorentzian function, using only a single parameter makes it easy to adapt to different spectral characteristics.[7, 8] The normalized integral of the weighted cross-correlation function fulfill all four criteria listed above.

An alternative is to bin the spectra into regions, represented by their integrated intensity, and compare region by region.[3] This removes the problem of energy shifts as long as the corresponding peaks stay in the same region. For L-edge XAS spectra a straightforward application is use different bins for the different parts of the edge, essentially giving the branching ratio.[32]. However, within each part of the edge binning is more challenging due to the large number of overlapping transitions. Another alternative would be to represent the x-ray spectrum as a linear combination of component functions and comparing the parameters of those functions, such as done in the Gaussian Mixture Models (GMM).[33, 34] However, for spectra with hundreds of resonances it is difficult to find consistent linear combinations if the compared spectra are significantly different, e.g., if experimental peaks are missing in the simulated spectra.

Here we evaluate the use of the normalized integral of the weighted cross-correlation function for the first time to quantify similarity in metal L-edge XAS spectra. We analyze how the width can be chosen in relation to the expected errors in the modeling and then apply the similarity analysis to evaluate the sensitivity of multiconfigurational wavefunction simulations using the restricted active space (RAS) formalism.[35, 36] Finally, the similarity measure will also be used to fit the ligand-field splitting parameter in a CTM model.

2. Theoretical background

An extensive description of the weighted cross-correlation function can be found in the original reference.[7] For self-consistency, the most relevant equa-

tions required to understand its basic properties are summarized below.

The distance measures between two functions f and g, representing e.g., experimental and theoretical spectra of some type, can be expressed in terms of the cross-correlation function c_{fg}:

$$c_{fg}(r) = \int f(x)g(x+r)dx. \tag{1}$$

For example, a simple point-wise comparison can be found by using the cross-correlation function for $r = 0$. As defined above, c_{fg} depends on the intensity scale used in the spectra, and does not have an upper or lower bound. To address this, and to make any prescaling of the spectra unnecessary, equation 1 can be normalized to a desired limit of [0,1]:

$$\widetilde{c}_{fg}(r) = \frac{c_{fg}(r)}{(c_{ff}(r)c_{gg}(r))^{1/2}}, \tag{2}$$

using the self-correlation functions. It can be shown that both the Pearson product-moment correlation function, as well as the cosine similarity, are versions of the normalized correlation function at $r = 0$, thus being unable to account for peak-shift effects.[7]

As an equivalent statement, we can integrate $c_{fg}(r)$ over the entire range of r, but instead use a weight function $w(r) = \delta(r)$ the Dirac Delta function, to only account for points at $r = 0$:

$$\widetilde{c}_{fg,w} = \frac{\int c_{fg}(r)w(r)dr}{\left(\int c_{ff}(r)w(r)dr \int c_{gg}(r)w(r)dr\right)^{1/2}} \ . \tag{3}$$

From equation 3 it is clear how $\widetilde{c}_{fg,w}$ can be extended to a more general expression for $w(r)$.

In this article we consider the weight function to be an approximation to the deviation between experimental and calculated spectra. Assuming that there are no systematic errors in the modeling, which is reasonable for a well-chosen model, a good approximation would then be a normal distribution:

$$w_{norm}(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right) \ . \tag{4}$$

where is μ the peak position, here set to zero and σ defines the standard deviation. In the following text, we will describe the width function using its full-width-at-half-maximum (FWHM), which relates to σ by a factor of 2.355. An infinitely narrow function corresponds to the point-wise comparison, while infinite width, i.e., a constant function ($w(r) = 1$) as the weight function, would give unity scores for any two spectra. It is thus important to choose the width to properly account for simulation errors, while at the same time retain the capability to separate spectra. Thus, we define the similarity function S_{fg} used in this article:

$$S_{fg} = \frac{\int c_{fg}(r) w_{norm}(r) dr}{\left(\int c_{ff}(r) w_{norm}(r) dr \int c_{gg}(r) w_{norm}(r) dr \right)^{1/2}} \ . \tag{5}$$

To get a visual understanding of equation 1 and the product $c(r)w(r)$, Figure 1 shows the weighted correlation function for two simple model spectra. Both spectra contain two identical Gaussian-shaped peaks with a FWHM of 0.5 eV, but the position of the second peak deviates by 1 eV. The cross-correlation function has one maximum at 0 eV, corresponding to match of the first peak, and another at -1 eV, where the second peaks match. The local maximum in the weighted cross-correlation function $c(r)w(r)$ close to -1 eV illustrates how this measure includes information about the similarity between the shifted peaks.

3. Methods

Comparisons between calculated and experimental spectra have been made for four complexes with well known electronic structure. In this way, any deviation between experiment and theory can be directly attributed to limitations in the spectral simulations. These complexes are $[Fe^{II}(CN)_6]^{4-}$ (ferrocyanide), $[Fe^{III}(CN)_6]^{3-}$ (ferricyanide), $[Fe^{III}(Cl)_6]^{3-}$ (ferric chloride), and $Mn^{III}(acac)_3$ (acac=acetylacetonate). The three iron complexes have octahedral, or close-to octahedral (O_h) symmetry but very different ligands, with CN^- a strong ligand that gives a low-spin state, while Cl^- is weak ligand that gives rise to a high-spin state. The acac ligand also gives rise to a weak ligand field and $Mn^{III}(acac)_3$

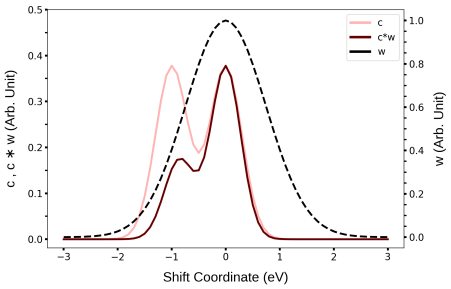

Figure 1: Correlation between a pair of two-peak spectra where the position of second peak differs by 1 eV. The graph shows the correlation function ($c(r)$), the Gaussian weight function ($w(r)$), and their product and their product $c(r)w(r)$, plotted as a function of the integration variable of the correlation function. The full-width-at-half-maximum of $w(r)$ is 1.75 eV. The maxima of $c(r)$ and $c(r)w(r)$ are set equal to facilitate a visual comparison.

has a Jahn-Teller distorted symmetry near D_{4h} coordination symmetry and a high-spin ground state.

L-edge XAS data for the iron complexes have been collected in total-electron yield mode as described in previously.[16, 12]. The spectrum for $Mn^{III}(acac)_3$ has been collected in transmission mode using a liquid flatjet.[37] In all cases, background subtraction has been performed to remove the edge jump. The spectral intensity should therefore be directly comparable to calculated oscillator strengths.

The three iron complexes have been simulated with the RAS approach. The ferrocyanide simulations have been performed using OpenMolcas.[38, 39] A ten-orbital valence active space, with five metal 3d-dominated orbitals (t_{2g} and e_g) two filled ligand-dominated orbitals of e_g symmetry, and three unfilled orbitals of t_{2g} symmetry have been used. The latter can be either metal 4d-type or ligand-dominated orbitals.[39] The structure is taken from a CASPT2 optimization using the ANO-RCC-VTZP basis set,[40, 41] which gives Fe-C distances of 1.918 Å and C-N distances of 1.196 Å. The reference spectrum is calculated using multi-state (MS) RASPT2.[36] with the same valence active space and ANO-RCC-VTZP basis set. The number of final states per irreducible representation

and spin multiplicity is set to 60. The initial state is a singlet and both singlet and triplet final states have been used. For the PT2 calculations, an imaginary shift of 0.3 hartree has been applied together with the the default ionization-potential electron-affinity (IPEA) shift of 0.25 hartree.[42, 43] The RAS state-interaction (RASSI) approach has been used to calculate spin-orbit coupling and electric dipole oscillator strengths.[35]. RAS spectra of ferricyanide and ferric chloride are taken from reference [39]. The manganese complex has been calculated with the CTM model, as described in reference [37].

Spectra of iron complexes are generated from the calculated oscillator strengths using a Lorentzian broadening with a FWHM of 0.4 and 0.8 eV for the L_3 and L_2 edges respectively, convoluted with an experimental Gaussian broadening of 0.4 eV. The corresponding values for the manganese complexes is 0.2 and 0.7 for the L_3 and L_2 edges, with an experimental broadening of 0.3 eV. Spectra are aligned using a global energy shift to maximize the similarity. For plotting purposes, the simulated spectral intensities are scaled uniformly to match the integrated intensity of the experiment, but this does not affect the analysis as the chosen similarity measure is independent of the scaling factor.

In order to estimate typical errors in position and intensity of the peaks, which is required to choose a suitable width of the weight function, calculated and experimental spectra of ferricyanide and ferric chloride are compared in Figure 2. Relative intensities show deviations for some peaks of up to 40%, while the largest errors in position are around 1.0 eV. The spectra are divided up into two edges, L_3 and L_2, split by the 2p spin-orbit coupling in the final state. For ferricyanide the splitting is underestimated by 1.5 eV, probably due to treatment of the strong 2p spin-orbit coupling as a perturbation. This prevents a comparison over the full spectral range. Therefore, similarity analysis has been performed only over the L_3 edge unless otherwise stated.

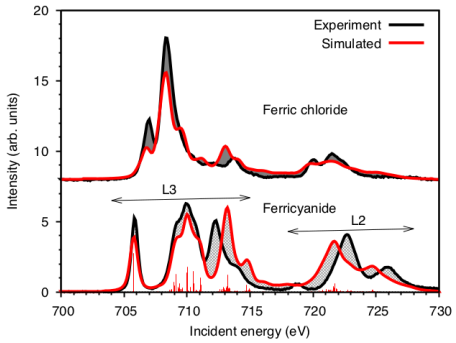

Figure 2: Iron L-edge x-ray absorption spectra of $[Fe^{III}(Cl)_6]^{3-}$ (ferric chloride) and $[Fe^{III}(CN)_6]^{3-}$ (ferricyanide) from experiment and restricted active space simulations.[16, 12, 39] The shaded areas represent the differences between the experiment and theory in a point-wise comparison.

4. Results and Discussion

The results section is divided into five parts. The first part is a study of the effects of the width of the weight function on the qualitative similarity analysis, and how the width relates to the expected errors in the simulations. The second part applies the similarity measure to RAS L-edge XAS spectra of ferrocyanide. This includes an analysis of the the sensitivity to model choices and how that affects comparisons to experiment. The third part contains a similar analysis for previously published simulations of ferric complexes. This is followed by an analysis of CTM modeling and how the parameters in this model can be selected by maximizing similarity. Finally, a discussion of the additional potential applications is included.

4.1. Analysis for two-peak model spectra

Before discussing how the similarity S_{fg} depends on errors in peak position in the case of spectra with a large number of overlapping transitions, it is illustrative to revisit the model two-peak spectra analyzed in Figure 1, where only one peak is different between reference and trial spectra. The position of the second peak in the trial spectra is shifted, starting from the position where

both peaks in the trial spectrum perfectly overlaps with the peaks in the reference spectrum, until the point where the second peak of the trial spectrum has no overlap with the reference spectrum. This process is repeated for a series of widths of the weight function, producing a graph of $S_{fg}(x)$, see Figure 3). These results are then compared to a case where the second peak is completely missing from the trial spectrum. Note that the two Gaussians were chosen sufficiently far apart (8 eV) that the similarity would not contain any contribution originating from overlap between the two. Because of this, the similarity between the reference spectrum and the one with a missing peak appears nearly constant for the Gaussian width used, while it still converges to 1 for an infinitely wide weight function.

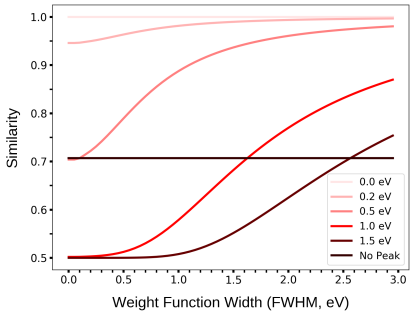

Figure 3: Similarity scores for two-peak reference and trial spectra depending on the width of the weight function. Results are shown for different deviations in the position of the second peak of the trial spectrum. Results are also given for a trial spectrum that completely lacks the second peak.

At zero width of the weight function, the trial spectrum that lacks a second peak gives a higher similarity (0.707) than trial spectra that contain both peaks if the error in energy is larger than 0.5 eV, see Figure 3). As shown in Figure 2 above, peak errors of 1 eV are not uncommon even with advanced models. For an error of this magnitude, the similarity is only 0.5, significantly smaller than that of a trial spectrum that lacks this peak. Ideally, simulations that predict the presence of a spectral feature should be more similar than those that lack those

features, even if the energy is not correctly predicted. Alternatively phrased, the similarity measure should be more forgiving with respect to quantitative methodological errors, i.e., that the peak is shifted, than qualitative modeling errors resulting in absent peaks. This gives a basis for selecting the width in relation to the typical quantitative errors of the model.

Figure 3 shows that with a 0.5 eV error in peak position, a width above 0.3 eV is required to reach higher similarity than the "no peak". When the error increases to 1.0 eV, as observed for the RAS calculations above, a significantly higher width (1.7 eV) is required to reach the same threshold. At the same time, a very large width decreases the information that can be extracted from the similarity analysis, with the extreme case of an infinite width giving the same similarity irrespective of the error in energy. This trend can be seen in Figure 3 where the similarity of different spectra becomes more similar with higher widths. Thus, for RAS calculations in the present implementation, a width of 1.75 eV, just above the 1.7-eV threshold, is a reasonable choice.

A more advanced analysis can be made by simultaneously accounting for errors in both energy and intensity in the two-peak example discussed above. This gives two-dimensional surfaces of $S_{fg}(\Delta(E), \Delta(I))$, see Figure 4. By analyzing the contours along those surfaces, it is possible to select a width taking into account the specific error profile of any given calculation, which in the case of RAS calculations is around 40% errors in intensity and 1-eV errors in position.

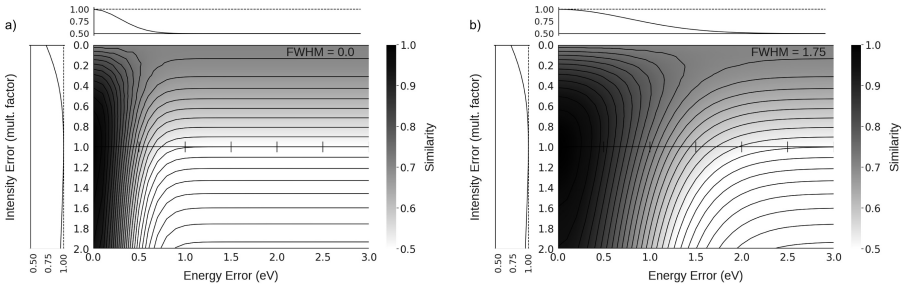

Figure 4: Similarity surfaces obtained from two-peak spectra where the both intensity and peak position of the second peak in the trial spectra are varied. a) Spectrum with 0.0 eV width of the weight function. b) Spectrum with 1.75 eV width of the weight function.

The function obtained from the cut along the intensity axis in Figure 4 represents the similarity when only intensity is changed and can be understood by forming a simplified expression for S_{fg}. Assuming the area under a single peak is A, the area of the two-peak reference spectrum is $2A$, while area of the trial spectra are $A + s * A$ where s is a scaling factor of the intensity of the second peak. Substituting this into equation 5 one obtains:

$$S_{fg} = \frac{(1+s)}{(2+2s^2)^{1/2}}, \tag{6}$$

which in the limit of s $\rightarrow$ 0 converges to $1/2^{1/2}$, as seen in Figure 3.

Now looking at the full map, it is clear that with zero width of the weight function, the similarity gradient drops fast in the incident energy direction, see Figure 4a. Here a 40% error in intensity gives the same similarity as a 0.1 eV error in energy. As the width increases to 1,75 eV, the contours extend further along the energy direction, see Figure 4b. Here an 40% intensity error corresponds to an energy error of 0.3 eV, which indicates that the similarity measure will still favor peaks with relatively small energy errors, although to a much smaller degree than the point-wise measurement.

4.2. Applications to a real two-peak example: Ferrocyanide L-edge XAS

For a method to be predictive, the results should ideally show a small sensitivity to the choice of any parameters introduced in the calculation. CASPT2/RASPT2 is an *ab-initio* method but the results depend strongly on the choice of the active space. The importance of this choice is already well known, and will not be elaborated upon further here. However, there are also other model choices that also affect the results, such as the basis set, the number of final states and the geometry. In addition, RASPT2 contains one parameter, the IPEA shift, introduced to improve the relative energies of states with different spin multiplicities.[43] However, the optimum value varies between systems.[44, 45] It is therefore important to understand the sensitivity of the results to the IPEA shift and other model choices.

The L-edge XAS spectrum of ferrocyanide provides an excellent application of the similarity method as each edge consists of two major peaks, see Figure 5. These peaks have been assigned as coming from e_g, and π^* respectively, with π^* representing an antibonding molecular orbital between iron t_{2g} and the π^* orbitals of the CN^- ligands.[12] Ferrocyanide is a commonly used model system and spectra have been simulated previously with RAS, both for L-edge XAS as well as for L and K-edge resonant inelastic x-ray scattering (RIXS).[46, 48, 47] However, none of these studies have performed a systematic test of the sensitivity of the simulations.

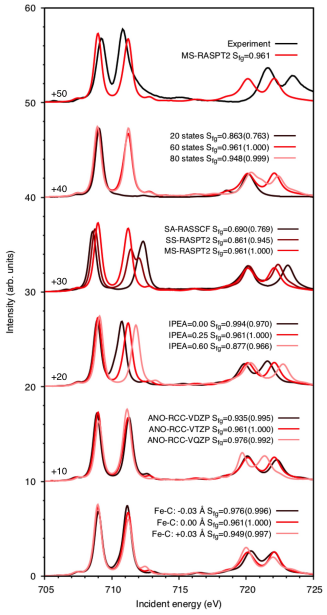

Figure 5: Comparisons of method choices for RAS simulations of L-edge XAS spectra of $[Fe^{II}(CN)_6]^{4-}$ (ferrocyanide). The similarity score (S_{fg}) is given compared to the experiment with comparisons to the reference spectrum in parenthesis. All values calculated with a FWHM of the weight function of 1.75 eV. The reference value for the Fe-C distance is 1.918 Å. Additional tests are shown in SI Figures S1-S4.

In the L_3 edge, the experimental spectrum shows two peaks at 709.2 and 710.8 eV, with the second π^* peak slightly higher in intensity. This is very similar to the spectra collected for ferrocyanide in aqueous solution collected

using partial-fluorescence yield.[46] The reference RAS calculation is performed with MS-RASPT2 as described in detail in the Methods section. It shows the same distinct two-peak structure in both edges, but the 1.6-eV peak distance is overstimated by 0.65 eV and the π^* peak is slightly lower in intensity. Without applying a weight function the similarity is 0.863, much lower than previous simulation results for heme complexes (0.975-0.990).[18] This is due to the error in the relative position of intense π^* peak, as discussed for the model two-peak example in Figure 3. Adding a 1.75-eV weight function gives a similarity score of 0.961, see Figure 5. This can be used as a reference point to estimate the sensitivity of the simulation results. Effects are considered small if they change S_{fg} by less than ± 0.02, moderate up to ± 0.05, large up to ± 0.1, and otherwise very large.

The first choice when performing the calculations is the number of final states. The number of states must be large enough so that all intense transitions are included.. However, as the calculations scale almost linearly with the number of states, including a large number leads to significant computational cost compared to ground state calculations even with recent improvements in the CI algorithm.[15] In addition, the quality of the state-average orbital optimization decreases when including too many states. Starting with 20 states in each of the four irreducible representations, for a total of 80 singlet and 80 triplet states, does not reproduce the π^* peak, see Figure 5. This gives a situation similar to the model two-peak example above, where the second peak is missing. For the 20-state spectrum the similarity is 0.863, much higher than in the model example (0.707). This can be explained by the relatively small energy difference between the two peaks in the experiment, which allows for overlap with both e_g and π^* peaks at the same time. Increasing to 40 states leads to problems in the spin-orbit coupling calculation but with 60 states the second peak appears. Further increasing to 80 states leads to a minor increase in intensity of the second peak but also increases the peak splitting, thus leading to a slightly lower similarity (0.948). The 60-state calculation is therefore considered converged with respect to the number of states and is used in all other compar-

isons. Including also quintet final states leads to almost no spectral changes in the L_3 edge, see Figure SI 1.

In many calculations, the cost is dominated by the second-order perturbation step. The current calculations show that the effect can also be very important. Simulations at the RASSCF level gives a peak splitting of 4 eV, in line with previously reported results at this level,[46] and a similarity score of only 0.690, see Figure 5. Significant improvement is achieved at the state-specific (SS) RASPT2 level, although the π^* peak is too broad with reduced peak intensity, leading to a similarity of 0.861. Finally, going from SS to MS-RASPT2 gives a narrower and more intense $\pi*$ peak, which leads to another significant increase in similarity to the 0.961 value reported above. The very large effects of the PT2 treatment can be explained by looking in more detail at the active space. It consists of ten valence orbitals, two ligand-dominated σ, the five metal 3d, and three of the $\pi*$ orbitals.[39] However, the latter are part of a 24-orbital π-$\pi*$ system of the six CN^- ligands, which gives an unbalanced active space and thus large effects of dynamical correlation.

The CASPT2/RASPT2 algorithm includes an IPEA shift introduced to correct spin-state energetics and an imaginary shift to reduce problems with intruder states. Decreasing the value of the IPEA shift from the default 0.25 to 0.00 leads to smaller $e_g - \pi^*$ gap and better agreement with experiment (0.994), while increasing the IPEA shift leads to lower similarity, see Figure 5. The large sensitivity to the IPEA shift is a consequence of the corresponding effects of dynamical correlation with an incomplete active space. The imaginary shift has a similar, but smaller, effect on the simulated spectra, see Figure SI 2.

The cost of the important PT2 step increases significantly with the size of the basis set. For ferrocyanide increasing the basis set from double to quadruple-ζ basis consistently increases the similarity to experiment, with a similarity of 0.976 for the largest basis set, see Figure 5. The simulations are moderately sensitive to basis set but there is no clear indication of basis-set convergence.

Finally, the sensitivity to the Fe-C and C-N distances have been evaluated. For most systems, geometries are determined from density-functional theory

optimizations, which in general gives errors in metal-ligand bond distances of 0.03 Å or less.[49] Decreasing the Fe-C bond distance by 0.03 Å correctly gives the π^* peak as more intense than the e_g one and increases similarity to 0.976. The effect is opposite for the C-N distance, but that effect is small, see Figure SI 3. When going beyond the ligands, the simplest approach is to include the environment using a polarized continuum model (PCM) model. Considering the large negative charge of the model (-4) it could be expected that adding a solvent environment would be important but instead it has no measurable effect, see Figure SI 4. The reason is the strong attraction between hole-electron pair. This leads to almost complete screening and virtually no effects on the electron density at the surface of the molecule.[14, 37]

4.3. Applications to ferric complexes

To test whether the previous observations for RAS simulations of ferrocyanide are general, the same analysis is performed for the previously published simulations of ferricyanide and ferric chloride shown in Figure 2.[39] Ferricyanide has three peaks, which can be assigned to t_{2g}, e_g, and π^* respectively. As was the case for ferrocyanide, the relative energy of the final π^* peak is overestimated. Ferric chloride has one major and two minor peaks for which positions are relatively well described with RAS. Figure 6 evaluated the extent to which the simulated RAS spectra change when modifying the model, both compared to internal references and the experimental spectra.

In general, ferricyanide shows larger sensitivity to model choices than ferric chloride. The main reason is the challenge in modeling the cyanide $\pi*$ peak. The difference is clear from when looking at the effects of the IPEA shift. For ferric chloride, the similarity compared to the RAS reference changes very little, see Figure 6a. When comparing to experiment the similarity increases marginally with increasing IPEA shift, but stays between 0.982 and 0.987. Ferricyanide shows much larger effects of changing the IPEA shift and effects on the similarity scores and ten times larger than ferric chloride. The effect of changes in IPEA shift are easy to analyze and are similar to what was previously seen for

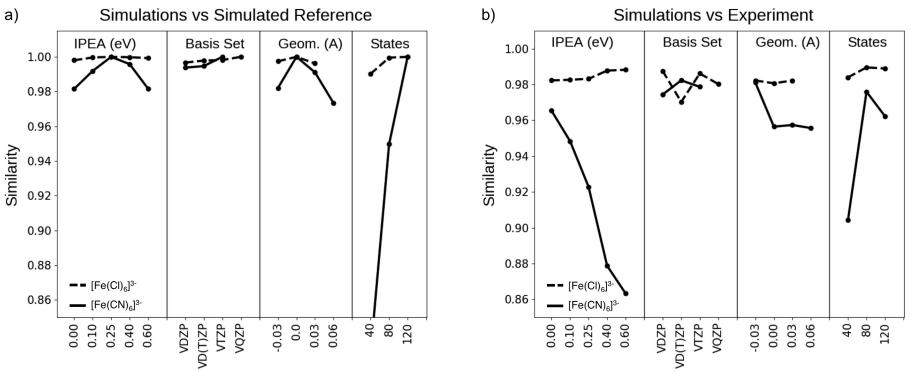

Figure 6: Similarity scores for ferric chloride (dashed lines) and ferricyanide (solid lines) spectra. a) Analysis of the sensitivity of modeled spectra in comparison with a reference calculation. b) Analysis of the accuracy of the modeled spectra in comparison to experimental data. Similarity values are calculated using a 1.75-eV width of the weight function. Selected results for other widths are shown in the Supporting Information, see Figures S5-S8. Due to timing and convergence issues, the series use different combinations of active spaces, number of final states, and basis sets, and results are therefore not transferable across series.[39]

ferrocyanide. The energy of the π^* peak increases with the shift, see Figure 7 and already the default value of 0.25 overestimates its energy, as seen in Figure 2. Again, the large effects of the IPEA shift is a diagnostic of an incomplete active space for the CN^- ligands. In contrast, the valence levels of Cl^- are all filled and there is no strong internal correlation, so the 10-orbital active space of $[FeCl_6]^{3-}$ is a much better approximation. This is also seen in the small effect of changing the IPEA in Figure 6.

For both systems there is a strong dependence on the number of final states, especially for ferricyanide where spectra change significantly even above 80 states per irreducible representation and spin multiplicity. The spectra calculated with 40 states lack the high-energy π^* peak in both edges, see Figure 8. Increasing the number of states to 80 recovers this peak, albeit with relatively low intensity. Finally, with 120 states the intensity of the π^* peak increases and a high-energy shoulder appears. The need for more states in ferricyanide compared to ferrocyanide is due to additional coupling with the open t_{2g} shell. As

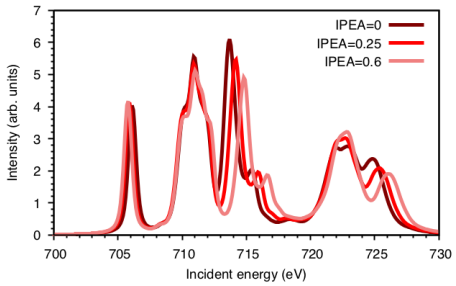

Figure 7: RAS simulated L-edge XAS spectra of ferricyanide with different values of the IPEA shift. Data from reference [39].

the energy difference between the e_g and π^* is larger than in the experiment, a point-wise comparison shows that the 120-states calculation with the converged π^* peak is worse than the 80-state one that underestimates its intensity, see Figure 9. The 120-states result is also only marginally better than the 40-state one that completely lacks the peak. Increasing the width decreases the difference between the 80 and 120-states calculations, and it becomes clear that these two simulations are significantly better than the 40-states simulation. The use of a width function thus works to alleviate the effect of peak shifts also in real spectral data.

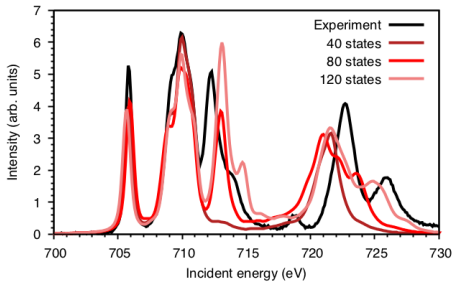

Figure 8: RAS simulated L-edge XAS spectra of ferricyanide with different number of final states per irreducible representation and spin multiplicity compared to experiment. Data from reference [39].

For both systems, the effects of increasing the size of basis set are small. There is no clear convergence with respect to basis set and a larger basis set

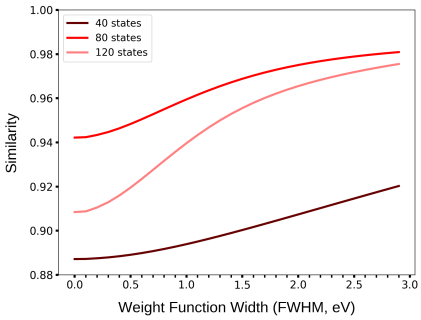

Figure 9: Similarity scores between experiment and RAS simulations with different number of final states for ferricyanide L-edge XAS as functions of the width of the weight function.

does not always improve the comparison to experiment, see Figure 6. When it comes to starting geometry, deviations of iron-ligand distances of 0.03 Å lead to moderate spectral changes for ferricyanide but small effects for ferric chloride.

Looking at the similarity results for the RAS simulation of all three iron complexes, it is clear that avoiding large errors in peak energies has the largest effect, changing the similarity by 0.3 units for ferrocyanide. It is thus recommended to calculate spectra with dynamical correlation using the MS-RASPT2 algorithm. The sensitivity to PT2 method parameters is large for all systems (0.1) with an incomplete active space, and small (0.02) otherwise. Another important consideration is to make sure that enough states are included to describe the full spectrum with some very large effects of up to 0.2 units. As the basis set size has a moderate effect (max 0.04), it is recommended to reduce the size of the basis set to afford enough final states to reach convergence. Finally, considering the moderate effects of metal-ligand distance (up to 0.03) it can be efficient to optimize selected distances using a high-level method as a ground-state optimization of a few coordinates is less expensive than a full spectral calculation. The other tested factors have small effects on simulated spectra.

4.4. Fitting of the semi-empirical CTM model

The CTM model typically involves varying several parameters to obtain the best possible fit between model and experiment. In a three-configuration model, already highly-symmetric systems belonging to the O_h point group have seven different parameters: the ligand-field splitting, the energy differences between the configurations, and the off-diagonal matrix elements in different irreducible representations.[11, 12] The number of parameters increase further with descent in symmetry. The final choice of parameters is therefore often guided by previous knowledge about the electronic structure. This is critical to restrict the parameter space, but also introduces some expectation bias.

Similarity analysis is thus clearly useful in selection of parameters, here exemplified for the fitting of a single variable, the ligand-field splitting of $Mn^{III}(acac)_3$ as shown in Figure 10. For zero width, the highest similarity is obtained for a 10Dq value of 1.5 eV, identical to the original assessment in reference [37]. Compared to estimates from optical spectroscopy of 2.1-2.4 eV, this is a significant reduction. The effect is argued to arise from localization of the 3d orbitals in the presence of the core hole. Plotting the similarity clearly motivates the choice of a smaller 10 Dq value for the core excited state and this choice is not very sensitive to the width function, at least for this complex.

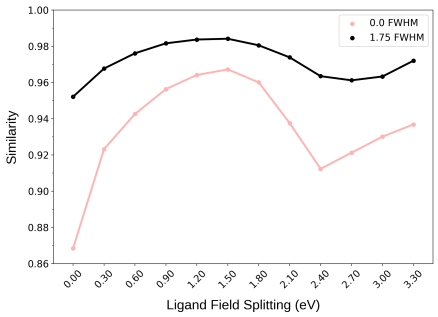

Figure 10: Similarity scores with different widths of the weight function for comparisons between experiment and CTM simulations of $Mn^{III}(acac)_3$ L-edge XAS as a function of the ligand-field splitting.

4.5. Potential applications

A promising application is the possibility to estimate confidence intervals for assignments of electronic and geometric from spectral fingerprints. Further, the accuracy of different methods for simulating the spectra can be compared, giving a better picture of strengths and weaknesses of different approaches. Although the RAS method has been applied to a sizeable number of systems, it has not yet been used systematically enough to get a statistical analysis of the error distributions. Further, the dependence on the choice of active space makes it challenging to get directly transferable data points. However, for methods that can quickly generate spectra with comparable quality, there is considerable potential for using appropriate similarity measures.

To take a few recent examples, benchmarking studies for K-edge XAS data of hundred thousand different materials have been simulated multiple scattering calculations and evaluated with the Pearson correlation coefficient.[50] The possibility to generate large data sets also opens up for machine learning approaches to spectroscopy. A recent application used a deep tensor neural network to predict optical spectroscopic using data sets of more than 132 thousand organic molecules. [51] They also used a point-wise approach, the root-mean-square error, to assess the quality of the data set. This forces the neural networks to prioritize excitation energies, which could be problematic for spectra where individual resonances overlap more strongly, as is the case for x-ray spectroscopy. Here the weighted cross-correlation function could be highly useful for evaluating the performance of predicting x-ray spectra.

For RAS simulations, a future use can be to get numerical relations between computational cost and changes in spectral shape, which is required to optimize modeling protocols. Further extensions of the analysis would be to apply it to two-dimensional data sets, such as resonant inelastic x-ray scattering (RIXS) in metal L- and K-edges.[24, 47]

5. Conclusions

Numerical analysis of the similarity between spectra avoids the expectation bias of visual comparisons. For spectra that are difficult to separate into components and where simulations give significant errors in predicted peak positions, the weighted cross-correlation function provides principal benefits compared to point-wise measures. The analysis has one important parameter, the width of the weight function, which should be chosen in relation to the errors in the simulations. For RAS calculations, with errors in peak positions of up to 1 eV, a width function of 1.75 eV ensures that a calculated spectrum with a peak shifted in energy gives a higher similarity than a spectrum that completely lacks that spectral feature.

This weighted cross-correlation is first applied to RAS simulations of three iron model systems. This numerical analysis shows effects from varying the number of final states and from the adding dynamical correlation are very large. The sensitivity to PT2 method parameters is large for all systems with an incomplete active space, and small otherwise. Moderate effects are found from varying basis set size and metal-ligand distance (up to 0.03 Å). Other factors, like the use of a PCM solvent environment, have small effects on simulated spectra. Computational time is best spent calculating dynamical correlation using MS-RASPT2 with a smaller basis, including a high number of final states and ensuring geometries are accurate. The similarity varies slowly with the width, but a value of 1.75-eV gives significant differences compared to point-wise comparisons. An example is the analysis of the state-dependence in ferricyanide simulations, where calculations that include the signature $\pi*$ peak gets significantly higher similarity than the one that lacks the peak, which was not the case in the point-wise comparison. The weighted cross-correlation function has the potential to replace point-wise comparisons in a number of applications where both peak position and intensity are key variables in spectral analysis.

6. Acknowledgments

We acknowledge financial support from the Knut and Alice Wallenberg Foundation (Grant No. KAW-2013.0020). M.D. and M.L acknowledge support from the Foundation Olle Engkvist Byggmastare. R.L. acknowledge financial support from the Swedish Research Council (VR, grant no. 2016-03398). The computations were performed on resources provided by SNIC through the National Supercomputer Centre at Linköping University (Tetralith) under project snic-2018-3-575 and snic2019-3-586.

References

[1] D. S. Stephenson, G. Binsch, Automated analysis of high-resolution nmr spectra. i. principles and computational strategy, J. Magn. Reson 37 (1980) 395–407.

[2] D. S. Stephenson, G. Binsch, Automated analysis of high-resolution nmr spectra. ii. illustrative applications of the computer program davins, J. Magn. Reson 37 (1980) 409–430.

[3] E. P. Lorant Bodis, Alfred Ross, A novel spectra similarity measure, Chemom. Intell. Lab. Syst 85 (1) (2007) 1 – 8.

[4] E. Debie, E. De Gussem, R. K. Dukor, W. Herrebout, L. A. Nafie, P. Bultinck, A confidence level algorithm for the determination of absolute configuration using vibrational circular dichroism or raman optical activity, Chem. Phys. Chem 12 (8) (2011) 1542–1549.

[5] T. Bruhn, A. Schaumloeffel, Y. Hemberger, G. Bringmann, Specdis: Quantifying the comparison of calculated and experimental electronic circular dichroism spectra, Chirality 25 (4) (2013) 243–249.

[6] S. L. Lawton, L. S. Bartell, Application of the overlap integral in x-ray diffraction powder pattern recognition, Powder Diffr 9 (2) (1994) 124–135.

[7] R. de Gelder, R. Wehrens, J. A. Hageman, A generalized expression for the similarity of spectra: application to powder diffraction pattern classification, J. Comput. Chem 22 (3) (2001) 273–289.

[8] J. A. Hageman, R. Wehrens, R. De Gelder, L. M. C. Buydens, Powder pattern indexing using the weighted crosscorrelation and genetic algorithms, J. Comput. Chem 24 (9) (2003) 1043–1051.

[9] K. Pierloot, Transition metals compounds: Outstanding challenges for multiconfigurational methods, Int. J. Quantum Chem 111 (13) (2011) 3291–3301.

[10] K. D. Vogiatzis, M. V. Polynski, J. K. Kirkland, J. Townsend, A. Hashemi, C. Liu, E. A. Pidko, Computational approach to molecular catalysis by 3d transition metals: challenges and opportunities, Chem. Rev 119 (4) (2018) 2453–2523.

[11] F. de Groot, Multiplet effects in X-ray spectroscopy, Coordin. Chem. Rev 249 (1) (2005) 31–63.

[12] R. K. Hocking, E. C. Wasinger, F. M. de Groot, K. O. Hodgson, B. Hedman, E. I. Solomon, Fe L-edge XAS studies of $K_4[Fe(CN)_6]$ and K3[$Fe(CN)_6$]: A direct probe of back-bonding, J. Am. Chem. Soc 128 (32) (2006) 10442–10451.

[13] J. K. Kowalska, B. Nayyar, J. A. Rees, C. E. Schiewer, S. C. Lee, J. A. Kovacs, F. Meyer, T. Weyhermler, E. Otero, S. DeBeer, Iron L2, 3-Edge X-ray Absorption and X-ray Magnetic Circular Dichroism Studies of Molecular Iron Complexes with Relevance to the FeMoco and FeVco Active Sites of Nitrogenase, Inorg. Chem 56 (14) (2017) 8147–8158.

[14] M. Kubin, M. Guo, T. Kroll, H. Löchel, E. Källman, M. L. Baker, R. Mitzner, S. Gul, J. Kern, A. Föhlisch, A. Erko, U. Bergmann, V. K. Yachandra, J. Yano, M. Lundberg, P. Wernet, Probing the Oxidation State of Transition Metal Complexes: A Case Study on How Charge and Spin

Densities Determine Mn L-Edge X-ray Absorption Energies, Chem. Sci 9 (33) (2018) 6813–6829.

[15] M. G. Delcey, L. K. Sørensen, M. Vacher, R. C. Couto, M. Lundberg, Efficient calculations of a large number of highly excited states for multi-
545 configurational wavefunctions, J. Comput. Chem 40 (2019) 1789–1799.

[16] E. C. Wasinger, F. M. De Groot, B. Hedman, K. O. Hodgson, E. I. Solomon, L-edge X-ray absorption spectroscopy of non-heme iron sites: experimental determination of differential orbital covalency, J. Am. Chem. Soc 125 (42) (2003) 12894–12906.

550 [17] P. Wernet, K. Kunnus, I. Josefsson, I. Rajkovic, W. Quevedo, M. Beye, S. Schreck, S. Gruebel, M. Scholz, D. Nordlund, W. Zhang, R. W. Hartsock, W. F. Schlotter, J. J. Turner, B. Kennedy, F. Hennies, F. M. F. de Groot, K. J. Gaffney, S. Techert, M. Odelius, A. Foehlisch, Orbital-specific mapping of the ligand exchange dynamics of Fe(CO)5 in solution,
555 Nature 520 (7545) (2015) 78–81.

[18] M. Guo, E. Källman, R. V. Pinjari, R. C. Couto, L. Kragh Sørensen, R. Lindh, K. Pierloot, M. Lundberg, Fingerprinting electronic structure of heme iron by ab initio modeling of metal l-edge x-ray absorption spectra, J. Chem. Theory Comput 15 (1) (2019) 477–489.

560 [19] M. Stener, G. Fronzoni, M. de Simone, Time dependent density functional theory of core electrons excitations, Chem. Phys. Lett 373 (1) (2003) 115 – 123.

[20] W. Hua, G. Tian, G. Fronzoni, X. Li, M. Stener, Y. Luo, Fe l-edge x-ray absorption spectra of fe (ii) polypyridyl spin crossover complexes from
565 time-dependent density functional theory, J. Phys. Chem. A 117 (51) (2013) 14075–14085.

[21] T. F. Stetina, J. M. Kasper, X. Li, Modeling l2, 3-edge x-ray absorption

spectroscopy with linear response exact two-component relativistic time-dependent density functional theory, J. Chem. Phys 150 (23) (2019) 234103.

[22] M. Roemelt, D. Maganas, S. DeBeer, F. Neese, A combined DFT and restricted open-shell configuration interaction method including spin-orbit coupling: Application to transition metal L-edge X-ray absorption spectroscopy, J. Chem. Phys 138 (20) (2013) 204101.

[23] R. Klooster, R. Broer, M. Filatov, Calculation of x-ray photoelectron spectra with the use of the normalized elimination of the small component method, Chem. Phys 395 (2012) 122–127.

[24] I. Josefsson, K. Kunnus, S. Schreck, A. Föhlisch, F. de Groot, P. Wernet, M. Odelius, Ab initio calculations of x-ray spectra: Atomic multiplet and molecular orbital effects in a multiconfigurational scf approach to the l-edge spectra of transition metal complexes, J. Phys. Chem. Lett 3 (23) (2012) 3565–3570.

[25] S. I. Bokarev, M. Dantz, E. Suljoti, O. Kühn, E. F. Aziz, State-dependent electron delocalization dynamics at the solute-solvent interface: Soft-x-ray absorption spectroscopy and ab initio calculations, Phys. Rev. Lett 111 (8) (2013) 083002–083007.

[26] R. V. Pinjari, M. G. Delcey, M. Guo, M. Odelius, M. Lundberg, Restricted active space calculations of l-edge x-ray absorption spectra: From molecular orbitals to multiplet states, J. Chem. Phys 141 (12) (2014) 124116.

[27] A. Chantzis, J. K. Kowlska, D. Maganas, S. DeBeer, F. Neese, Ab initio wave function-based determination of element specific shifts for the efficient calculation of x-ray absorption spectra of main group elements and first row transition metals, J. Chem. Theory Comput 14 (7) (2018) 3686–3702.

[28] M. Lundberg, M. G. Delcey, Multiconfigurational approach to x-ray spectroscopy of transition metal complexes, in: Transition Metals in Coordination Environments, Vol. 29, Springer, 2019, pp. 185–217.

[29] J. C. Shaw, An introduction to the coherence function and its use in eeg signal analysis, J. Med. Eng. Technol 5 (1981) 273–288.

[30] M. C.-C. Miguel Angel Guevara, Eeg coherence or eeg correlation?, Int. J. Psychophysiol 23 (1996) 145–153.

[31] H. Karfunkel, B. Rohde, F. Leusen, R. Gdanitz, G. Rihs, Continuous similarity measure between nonoverlapping xray powder diagrams of different crystal modifications, J. Comput. Chem 14 (10) (1993) 1125–1135.

[32] B. Thole, G. Van der Laan, Branching ratio in x-ray absorption spectroscopy, Phys. Rev. B 38 (5) (1988) 3158.

[33] K. Pearson, Contributions to the mathematical theory of evolution, Philos. Trans. R. Soc. London 185 (1984) 71–110.

[34] G. V. Adam Tauman Kalai, Ankur Moitra, Disentangling gaussians, Commun. ACM 55 (2012) 113–120.

[35] P.-Å. Malmqvist, B. O. Roos, B. Schimmelpfennig, The restricted active space (RAS) state interaction approach with spin-orbit coupling, Chem. Phys. Lett 357 (3) (2002) 230–240.

[36] P.-Å. Malmqvist, K. Pierloot, A. R. M. Shahi, C. J. Cramer, L. Gagliardi, The restricted active space followed by second-order perturbation theory method: Theory and application to the study of cuo2 and cu2o2 systems, J. Chem. Phys 128 (20) (2008) 204109.

[37] M. Kubin, M. Guo, M. Ekimova, M. L. Baker, T. Kroll, E. Källman, J. Kern, V. K. Yachandra, J. Yano, E. T. Nibbering, M. Lundberg, P. Wernet, Direct Determination of Absolute Absorption Cross Sections at the L-Edge of Dilute Mn Complexes in Solution Using a Transmission Flatjet, Inorg. Chem 57 (9) (2018) 5449–5462.

[38] I. Fdez. Galvn, M. Vacher, A. Alavi, C. Angeli, F. Aquilante, J. Autschbach, J. J. Bao, S. I. Bokarev, N. A. Bogdanov, R. K. Carlson, L. F. Chibotaru, J. Creutzberg, N. Dattani, M. G. Delcey, S. S. Dong,

A. Dreuw, L. Freitag, L. M. Frutos, L. Gagliardi, F. Gendron, A. Giussani, L. Gonzlez, G. Grell, M. Guo, C. E. Hoyer, M. Johansson, S. Keller, S. Knecht, G. Kovaevi, E. Kllman, G. Li Manni, M. Lundberg, Y. Ma, S. Mai, J. P. Malhado, P. Å. Malmqvist, P. Marquetand, S. A. Mewes, J. Norell, M. Olivucci, M. Oppel, Q. M. Phung, K. Pierloot, F. Plasser, M. Reiher, A. M. Sand, I. Schapiro, P. Sharma, C. J. Stein, L. K. Srensen, D. G. Truhlar, M. Ugandi, L. Ungur, A. Valentini, S. Vancoillie, V. Veryazov, O. Weser, T. A. Wesoowski, P.-O. Widmark, S. Wouters, A. Zech, J. P. Zobel, R. Lindh, Openmolcas: From source code to insight, J.Chem. Theory Comput. 15 (11) (2019) 5925–5964. doi:10.1021/acs.jctc.9b00532.

[39] R. V. Pinjari, M. G. Delcey, M. Guo, M. Odelius, M. Lundberg, Cost and sensitivity of restricted active-space calculations of metal l-edge x-ray absorption spectra, J. Comput. Chem 37 (5) (2016) 477–486.

[40] B. O. Roos, R. Lindh, P.-Å. Malmqvist, V. Veryazov, P.-O. Widmark, Main Group Atoms and Dimers Studied with a New Relativistic ANO Basis Set, J. Phys. Chem. A 108 (15) (2004) 2851–2858.

[41] B. O. Roos, R. Lindh, P.-Å. Malmqvist, V. Veryazov, P.-O. Widmark, New relativistic ANO basis sets for transition metal atoms., J. Phys. Chem. A 109 (29) (2005) 6575–6579.

[42] N. Forsberg, P.-Å. Malmqvist, Multiconfiguration perturbation theory with imaginary level shift, Chem. Phys. Lett 274 (1) (1997) 196–204.

[43] G. Ghigo, B. O. Roos, P.-Å. Malmqvist, A modified definition of the zeroth-order Hamiltonian in multiconfigurational perturbation theory (CASPT2), Chem. Phys. Lett 396 (1) (2004) 142–149.

[44] S. Vancoillie, H. Zhao, M. Radon, K. Pierloot, Performance of caspt2 and dft for relative spin-state energetics of heme models, J. Chem. Theory. Comput 6 (2) (2010) 576–582.

[45] J. P. Zobel, J. J. Nogueira, L. González, The ipea dilemma in caspt2, Chem. Sci 8 (2) (2017) 1482–1499.

[46] N. Engel, S. I. Bokarev, E. Suljoti, R. Garcia-Diez, K. M. Lange, K. Atak, R. Golnak, A. Kothe, M. Dantz, O. Kühn, E. F. Aziz, Chemical bonding in aqueous ferrocyanide: Experimental and theoretical xray spectroscopic study, J. Phys. Chem. B 118 (6) (2014) 1555–1563.

[47] M. Guo, E. Kallman, L. K. Sørensen, M. G. Delcey, R. V. Pinjari, M. Lundberg, Molecular orbital simulations of metal 1s2p resonant inelastic x-ray scattering, The Journal of Physical Chemistry A 120 (29) (2016) 5848–5855.

[48] K. Kunnus, W. Zhang, M. G. Delcey, R. V. Pinjari, P. S. Miedema, S. Schreck, W. Quevedo, H. Schroeder, A. Föhlisch, K. J. Gaffney, M. Lundberg, M. Odelius, P. Wernet, Viewing the valence electronic structure of ferric and ferrous hexacyanide in solution from the fe and cyanide perspectives, J. Phys. Chem. B 120 (29) (2016) 7182–7194.

[49] M. Bühl, H. Kabrede, Geometries of transition-metal complexes from density-functional theory, J. Chem. Theory Comput 2 (5) (2006) 1282–1290.

[50] K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J. J. Rehr, S. P. Ong, K. A. Persson, High-throughput computational x-ray absorption spectroscopy, Sci. Data 5 (2018) 180151.

[51] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, P. Rinke, Deep learning spectroscopy: Neural networks for molecular excitation spectra, Adv. Sci 6 (9) (2019) 1801367.