A Deep Neural Network for the Rapid Prediction of X-ray Absorption Spectra

C. D. Rankine,[†] M. M. M. Madkhali,[†,‡] and T. J. Penfold[*,†]

[†]*Chemistry - School of Natural and Environmental Sciences, Newcastle University, Newcastle-upon-Tyne, NE1 7RU, UK.*

[‡]*Department of Chemistry - College of Science, Jazan University, Jazan, Saudi Arabia.*

E-mail: tom.penfold@ncl.ac.uk,conor.rankine@ncl.ac.uk

Abstract

X-ray spectroscopy delivers strong impact across the physical and biological sciences by providing end-users with highly-detailed information about the electronic and geometric structure of matter. To decode this information in challenging cases, *e.g. in operando* catalysts, batteries, and temporally-evolving systems, advanced theoretical calculations are necessary. The complexity and resource requirements often render these out of reach for end-users, and therefore data are often not interpreted exhaustively, leaving a wealth of valuable information unexploited. In this paper, we introduce supervised machine learning of X-ray absorption spectra, by developing a deep neural network (DNN) that is able to estimate Fe K-edge X-ray absorption near-edge structure spectra in less than a second with no input beyond geometric information about the local environment of the absorption site. We predict peak positions with sub-eV accuracy and peak intensities with errors over an order of magnitude smaller than the spectral variations that the model is engineered to capture. The performance of the DNN is promising, as illustrated by its application to the structural refinement

of iron(II)tris(bipyridine) and nitrosylmyoglobin, but also highlights areas for which future developments should focus.

Introduction

The emergence of high-brilliance light sources, such as 3$^{\text{rd}}$-generation synchrotrons and 4$^{\text{th}}$-generation X-ray free-electron lasers (XFELs), has transformed X-ray absorption spectroscopy (XAS).[1] It is now possible to acquire high-resolution XAS spectra under the most challenging operating conditions; examples include *operando* catalysts[2] and batteries.[3] In addition, one can also follow ultrafast excited-state dynamics in real time by exploiting ultrashort X-ray pulses generated by XFELs.[4,5] The unprecedented level of detail in modern XAS spectra, coupled with ever-increasing data acquisition rates, brings into focus the acute challenge of accurately and efficiently analysing these data to access the quantitative electronic and geometric structural information encoded into each XAS spectrum.

To access the information encoded into the X-ray absorption near-edge structure (XANES) region of an XAS spectrum, high-level theoretical calculations are necessary to capture the complexity of the underlying physics.[6] In amorphous materials, and those under non-equilibrium conditions, *e.g.* *operando* measurements and time-resolved experiments, performing the large number of theoretical calculations required to analyse the data is often precluded by their individual complexity. Indeed, accounting for contributions from all absorption sites in an amorphous material is time-consuming and resource-intensive; the challenge is compounded when the contributions change as a function of time. *In lieu* of an alternative, the *status quo* is to interpret the XANES spectrum superficially using semi-empirical heuristics for estimating changes in geometry,[7] symmetry,[8] and oxidation state[9] at the absorption site. At best, a valuable scientific resource is not exploited to its fullest potential if the XANES spectra cannot be interpreted exhaustively; at worst, the *status quo* leaves the data open to the danger of misinterpretation.

Although the underlying physics connecting the XANES observables to the geometric structure of a material are complex and challenging to simulate efficiently, they are well-understood. Modulations in the absorption cross-section just above the absorption edge are a consequence of the multiple scattering of low-kinetic-energy photoelectrons by atoms neighbouring the absorbing atom.[10] It is therefore possible to develop models for the refinement of the local geometric structure around the absorption site and fit the XANES data, as successfully demonstrated by Smolentsev *et al.* in the development of the FitIT[11,12] code. FitIT is able to interpolate XANES spectra for a given system in a user-defined geometric parameter space, reducing the number of theoretical calculations required to refine the geometric structure. While powerful, it requires a bespoke model to be initiated for each new system.

The contemporary literature indicates a growing interest in generalisable, '*data-driven*' approaches for these kind of analyses.[13] Supervised machine learning/deep learning algorithms[14] have enjoyed considerable success in mapping complex, non-linear relationships without any hand-coded heuristics and it transpires that these algorithms are able to make this particular link: that between the XANES spectrum and the underlying electronic and geometric structure of a material. Timoshenko *et al.*[13,15–20] have worked extensively on automated, machine-led analysis of XAS spectra and have deployed neural networks to extract physical insight from the experimental data, including, but not limited to, information on geometric structure and morphology. In other work, Zheng and Mathew *et al.*[21,22] have leveraged ensemble learning to extract information on coordination geometry and oxidation state from XAS spectra. Overwhelmingly, these are '*forward*' (spectrum-to-property/structure) rather than '*reverse*' (property/structure-to-spectrum) mappings and, where the latter are sparingly demonstrated, they are always system-specific or restricted to a narrow class of systems, *i.e.* a new set of theoretical calculations would be necessitated, and the model would have to be reoptimised, were it to be applied to different systems, assuming that this were even possible under the constraints of the model architecture.

The only authors to have deployed a generally-applicable machine learning model capable of predicting XANES spectra for an arbitrary absorption site (and consequently overcoming system-specificity) are Carbone *et al.*[23] In their recent contribution, the authors introduce a supervised machine learning model for the prediction of XANES at the N and O K edges for small molecules from the QM9 molecular database.[24] Their model successfully learns a structure-to-spectrum mapping using XANES spectra derived from the Materials Project library. However, these XANES spectra have already been convoluted to account for core-hole lifetime broadening, instrument response, and many-body effects, *e.g.* inelastic losses, which leaves little flexibility for experiments performed at different resolution, such as high-energy-resolution fluorescence detected (HERFD) measurements.[25]

Beyond XAS, reviews note the success of machine learning models in bypassing expensive theoretical calculations of scalar properties in computational chemistry and physics,[26,27] but there are very few examples of this being attempted for higher-dimensional data, *e.g.* spectra, with genuine generality. In our present paper we aim to contribute to the wider literature in this respect. We introduce a deep neural network (DNN) for instantaneous estimations of the Fe K-edge XANES spectra of arbitrary systems which only requires the geometry of the local environment around the absorbing atom as an input. We demonstrate that this initial implementation is able to predict peak positions and intensities with good accuracy and illustrate its performance by applying it to the structural refinement of iron(II)tris(bipyridine) and nitrosylmyoglobin.

Theoretical and Computations Details

Dataset

Our dataset comprises 9040 Fe absorption site geometry and theoretical XANES spectrum pairs. Fe absorption site geometries were harvested from the Materials Project Database[28] *via* the Materials Project API;[29] one Fe absorption site geometry was selected arbitrarily

per Fe-containing material.

The Fe absorption site geometries, defined by an arbitrary cutoff radius, were featurised for input as a two-body pair/radial distribution curve (RDC):

$$\sum_{i}^{N}\sum_{j>i}^{N} Z_i Z_j e^{-\alpha(r_{ij}-\boldsymbol{R})^2} \tag{1}$$

The summation runs over all internuclear pairs ij separated by some distance, r_{ij}. Z_i and Z_j are the atomic numbers of i and j, respectively. $\boldsymbol{R}$ is a vector obtained by discretising a linear interpolation between zero and twice the cutoff radius around the absorption site (defining the maximum pairwise distance that can be accommodated by the RDC), and α is a smoothing parameter. Throughout this work, the following parameters were used for featurisation: $\alpha = 10.0$, and $\boldsymbol{R} = 0.0 \xrightarrow{1.2} 800.0$ pm. RDCs were standarised for input into our DNN; Fe K-edge XANES spectra were subject to post-edge normalisation.

Deep Neural Network

Our DNN is based on the multilayer perceptron model and is programmed in Python 3 with Tensorflow/Keras.[30,31] The model comprises an input layer of 680 neurons (into which an RDC is input) and an output layer of 240 neurons (from which the estimated XANES spectra is retrieved). There are four dense hidden layers between the input layer and output layer; the initial hidden later contains 1150 neurons, and each subsequent hidden layer contains 10% fewer neurons than the preceding hidden layer. Each hidden layer performs non-linear transformations using a hyperbolic tangent (tanh) activation function. Regularisation is implemented to minimise overfitting; batch standardization and dropout are applied at each hidden layer. The probability, p, of dropout is set to 0.15.

Our DNN optimises *c.a.* three million internal weights *via* sequential feed-forward and backpropagation cycles. Gradients of a MSE loss function with respect to the internal weights are updated iteratively according to the Adaptive Moment Estimation (ADAM)

algorithm. Gradients are estimated over minibatches of 48 samples. The learning rate for the ADAM algorithm, η, is set to 3×10^{-4}. Our DNN learns from, and casts estimations of, unconvoluted Fe K-edge XANES spectra. The estimated XANES spectra are convoluted subsequently with an arctangent function in a post-processing step.

By repeated K-fold cross-validation with an 80:20 in-sample/out-of-sample split, we assess the performance of our DNN on the entire dataset while keeping the data used for evaluation always out-of-sample, *i.e.* unseen by our DNN.

The optimal hyperparameters for our DNN were determined *via* 500 cycles of Bayesian optimisation as implemented in the GPyOpt[32,33] module.

Simulation of XANES Spectra

Fe K-edge XANES spectra for all Fe absorption site geometries were calculated using multiple scattering theory as implemented in the FDMNES package.[34] A self-consistent muffin-tin potential with a cutoff radius of 6 Å around the Fe atom was used; the interaction with the X-ray field was described by the electric quadrupole approximation, and scalar relativistic effects were included.

To compare to experimental XANES spectra, computed cross-sections are routinely convoluted to account for core-hole lifetime broadening, instrument response, and many-body effects, *e.g.* inelastic losses. The convolution is performed here as a post-processing step and employs an energy-dependent arctangent function ($\boldsymbol{\Gamma}$):

$$\boldsymbol{\Gamma} = \Gamma_i + \Gamma_f \left(\frac{1}{2} + \frac{1}{\pi} \arctan \left(\frac{\pi}{3} \frac{\Gamma_f}{E_w} \left(\frac{\boldsymbol{E} - E_f}{E_c} - \frac{E_c^2}{(\boldsymbol{E} - E_f)^2} \right) \right) \right) \tag{2}$$

$\boldsymbol{\Gamma}$ is defined over the energy scale, $\boldsymbol{E}$, relative to the Fe K-edge of the XANES spectrum as per specification of the core-level and final-state widths (Γ_i and Γ_f, respectively), the centre and width of the arctangent function (E_c and E_w, respectively), and the Fermi energy (E_f). This is an empirical model, closely related to the Seah-Dench formalism.[35] Our

implementation is the same as that implemented in the FDMNES package.[34] The following parameters were used throughout for arctangent convolution: $\Gamma_i = 1.25$ eV, $\Gamma_f = 15.0$ eV, $E_c = 30.0$ eV, $E_w = 30.0$ eV, $E_f = -5.0$ eV, and $\boldsymbol{E} = -35.0 \overset{0.25}{\rightarrow} 75.0$ eV.

Results

Figure 1 shows six out-of-sample DNN estimations and theoretical XANES spectra. The top three XANES spectra (Figs. 1a, 1b, and 1c) are selected from the first centile, Mean Squared Error (MSE) $< 2.0 \times 10^{-4}$, and the bottom three XANES spectra (Figs. 1d, 1e, and 1f) are selected from the ninety-ninth centile (MSE $> 1.2 \times 10^{-1}$) when performance is ranked over all out-of-sample DNN estimations by MSE. The XANES spectra have been selected to represent broadly across spectral line-shapes, elemental composition, and unit cell packing density.

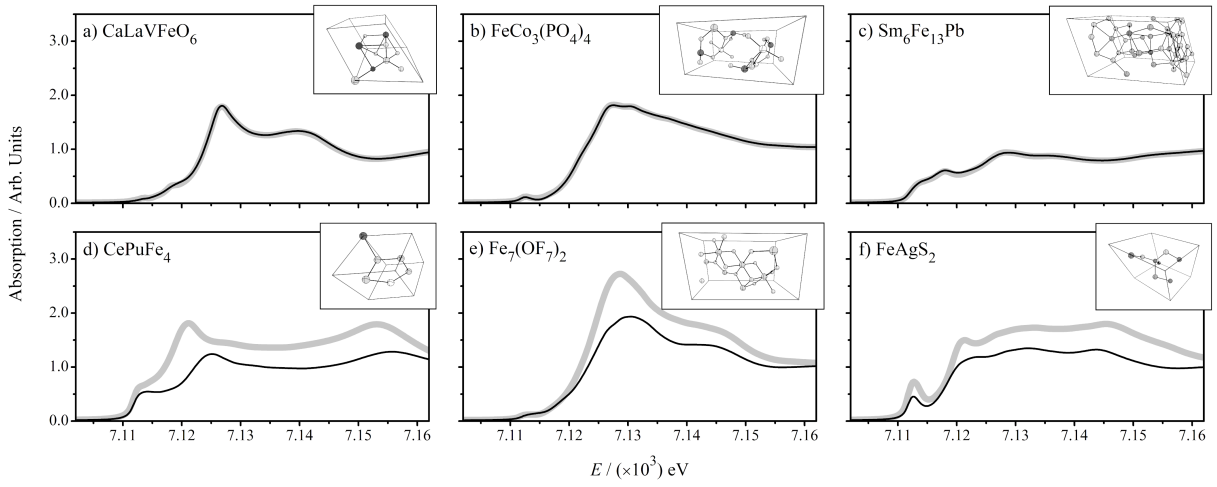

Figure 1: Examples of theoretical target (grey) and out-of-sample DNN-estimated (black) Fe K-edge XANES spectra for absorption sites in a) $CaLaVFeO_6$, b) $FeCo_3(PO_4)_4$, c) $Sm_6Fe_{13}Pb$, d) $CePuFe_4$, e) $Fe_7(OF_7)_2$, and f) $FeAgS_2$. XANES spectra a), b), and c) are drawn from the first centile (MSE $< 2.0 \times 10^{-4}$) and XANES spectra d), e), and f) are drawn from the ninety-ninth centile (MSE $> 1.2 \times 10^{-1}$) when estimations are ranked over all convoluted out-of-sample DNN estimations by MSE.

Estimations drawn from the first centile cannot be distinguished from the target XANES spectra; the average Pearson correlation coefficient between the estimated and target XANES

spectra in this subgroup is >0.99, and peak positions on the energy scale are predicted consistently to sub-eV accuracy. Impressively, even the worst estimations, *i.e.* those drawn from the ninety-nineth centile, reproduce faithfully the spectral shapes of their target; the average Pearson correlation coefficient for these estimations is still >0.95. Estimated peak positions on the energy scale (arguably the metric of principle importance for the spectroscopist) are broadly accurate even here. Indeed, these estimations are only placed in the ninety-ninth percentile by MS because of a) an underestimation of the spectral intensity that compounds across the energy scale, and b) because small peak shifts on the energy scale result in relatively larger errors on the intensity scale.

Figure 2 shows a histogram of the MSE achieved on 9040 estimations of out-of-sample XANES spectra. The median MSE is 5.2×10^{-3} and the lower and upper quartiles are found at 1.8×10^{-3} (-3.5×10^{-3}) and 1.2×10^{-2} ($+9.5 \times 10^{-3}$), respectively. The narrow interquartile range (IQR) of 1.0×10^{-2} and the high positive skewness coefficient of 5.18 attest to the strong and balanced performance of our DNN on out-of-sample estimations across our dataset. For perspective, the median MSE is an order of magnitude smaller than the spectral variation 4.0×10^{-2} (measured relative to the average spectral intensity over the whole spectrum) of our dataset, and corresponds to a median absolute error of <5% relative to the target spectra, allowing for high confidence in qualitative analyses.

The inset in Figure 2 shows a histogram of the MSE achieved on 9040 unconvoluted out-of-sample DNN estimations. These estimations have not been post-processed, *i.e.* an arctangent convolution (see Equation 2) has not been performed. Post-processing the estimated XANES spectra *via* convolution with the arctangent function improves the MSE by an order of magnitude. The median MSE achieved on out-of-sample unconvoluted XANES spectra is 5.4×10^{-2} and the lower and upper quartiles are found at 2.6×10^{-2} (-2.8×10^{-2}) and 1.1×10^{-1} ($+5.2 \times 10^{-2}$), respectively. Beyond the absolute values, the greater spread of the data and the lower coefficient of skewness (2.53) for the MSEs on the out-of-sample unconvoluted XANES spectra demonstrates that the improvement brought about by arctan-

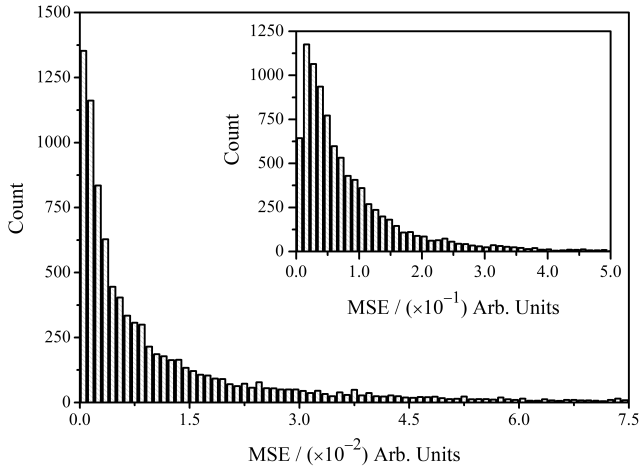

Figure 2: Histogram of the MSE achieved on 9040 convoluted out-of-sample DNN estimations. All estimations are made on unconvoluted XANES spectra; an arctangent convolution is applied as a post-processing step. The inset is a histogram of the MSE prior to arctangent convolution.

gent convolution is not uniform for all estimations in the dataset. Outlier estimations with high MSE are improved to a greater extent than other estimations; in all cases, the outlier estimations have intense pre-edge peaks before arctangent convolution, and we find a strong correlation between the intensity of these pre-edge peaks and the improvement in MSE after arctangent convolution.

Parity plots of the difference between the estimated and target peak positions on the energy (E_{Target} and $E_{\text{Est.}}$, respectively) and intensity (μ_{Target} and $\mu_{\text{Est.}}$, respectively) scales are presented in Figs. 3a and 3b, respectively. Strong linear relationships between E_{Target} and $E_{\text{Est.}}$, and μ_{Target} and $\mu_{\text{Est.}}$, are evidenced by their coefficients of determination, R^2, which are 0.986 and 0.973, respectively. Histograms of the mean absolute error (MAE) between E_{Target} and $E_{\text{Est.}}$ (ΔE), and μ_{Target} and $\mu_{\text{Est.}}$ ($\Delta\mu$), are shown in Figs. 3c and 3d, respectively. The median ΔE and $\Delta\mu$ are 0.45 eV and 3.7×10^{-2}, respectively; low, in each case, and reflective of the predictive power of our DNN $> 90\%$ of all prominent peaks in the target XANES spectra are reproduced in the estimations.

Figure 4 shows the convergence of our DNN as a function of real time and the number of forward passes through the dataset. It is possible to optimise our DNN to effective

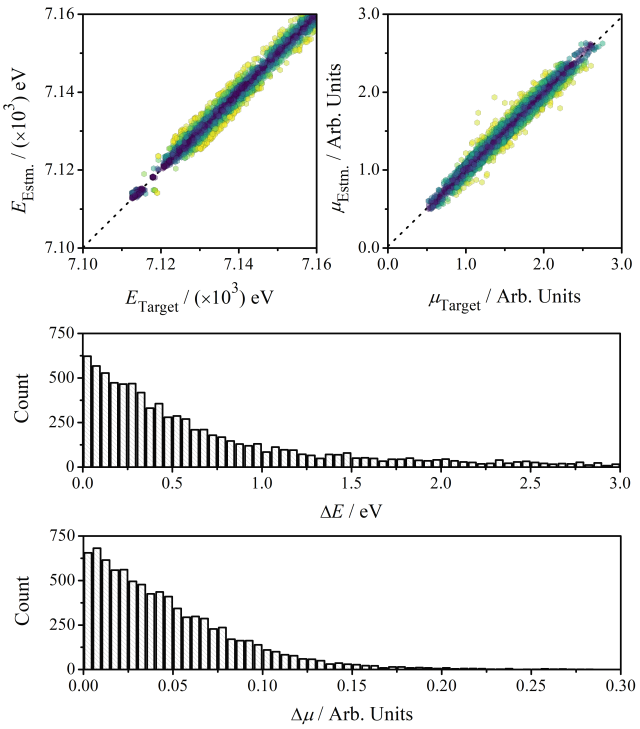

Figure 3: Parity plots of estimated and target peak positions on the a) energy (E_{Target} and $E_{\text{Estm.}}$, respectively) and b) intensity (μ_{Target} and $\mu_{\text{Estm.}}$, respectively) scales. Histograms of the MAEs, c) ΔE and d) $\Delta\mu$, between estimated and target peak positions on the energy and intensity scales, respectively.

convergence in < 500 forward passes through the dataset and this can be achieved in as little as five or ten minutes (real time) using off-the-shelf, consumer-grade hardware (two nVidia RTX 2080 Ti GPUs connected *via* an nVidia NVLink). This illustrates that once the data have been curated, our DNN could be quickly reoptimised to estimate XANES spectra at other absorption edges and/or for other elements.

Figure 5 shows the convergence of our DNN as a function of the number of in-sample XANES spectra used in the learning process. The MSE on the out-of-sample XANES spectra improves monotonically as the in-sample allocation of XANES spectra is increased; likewise, the standard errors for the MSE evaluations decrease as the DNN is given access to a larger in-sample reference database to learn from. Convergence is not entirely achieved in the limit of the current dataset (*ca.* 8000 in-sample XANES spectra); Figure 5 indicates that there is still scope to improve further on the results communicated here by growing our dataset.

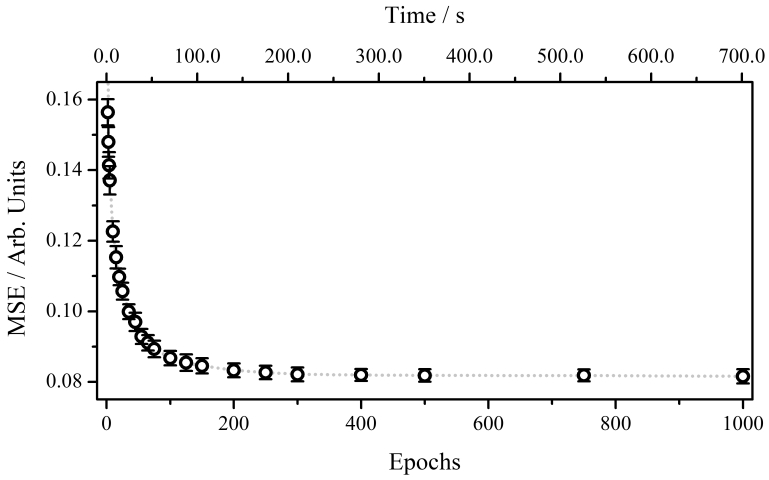

Figure 4: Evolution of the MSE as a function of real time and the number of forward passes through our dataset ('epochs') Data points are averaged over 100 K-fold cross-validated evaluations; error bars indicate one standard deviation. No post-processing has taken place; XANES spectra have not been convoluted with the arctangent function.

To evidence the generality and sensitivity of the DNN, we now apply it to experimental XANES data for structures far outside our dataset by performing a structural refinement on iron(II)tris(bipyridine), $[Fe(bpy)_3]^{2+}$,[36] and nitrosylmyoglobin (MbNO),[37] respectively. Figures 6a and 6c show comparisons between the experimental, theoretical, and DNN-estimated XANES spectra of $[Fe(bpy)_3]^{2+}$ and MbNO, respectively. The DNN-estimated XANES spectra are in qualitative agreement with the experimental spectra; the MSEs relative to experiment are 5.1×10^{-2} and 1.3×10^{-2}, respectively, in line with the median of the MSE reported in Figure 2. The prominent peaks in the XANES spectra of $[Fe(bpy)_3]^{2+}$ and MbNO are estimated with eV accuracy with respect to their experimental positions, and other features in proximity to the absorption edge, *e.g.* the shoulders to the left of the prominent peaks, are reproduced qualitatively. The DNN-estimated XANES spectra for$[Fe(bpy)_3]^{2+}$ and MbNO are consequently on par with the out-of-sample estimations from our dataset.

We further demonstrate that our DNN has sufficient sensitivity to small changes in internuclear distance to allow for rudimentary refinement of geometric structures. Figures 6b and 6d show the MSE between the DNN-estimated and experimental XANES spectrum as a function of Fe-to-N coordination distances, rFe-N, for $[Fe(bpy)_3]^{2+}$ and MbNO, respectively.

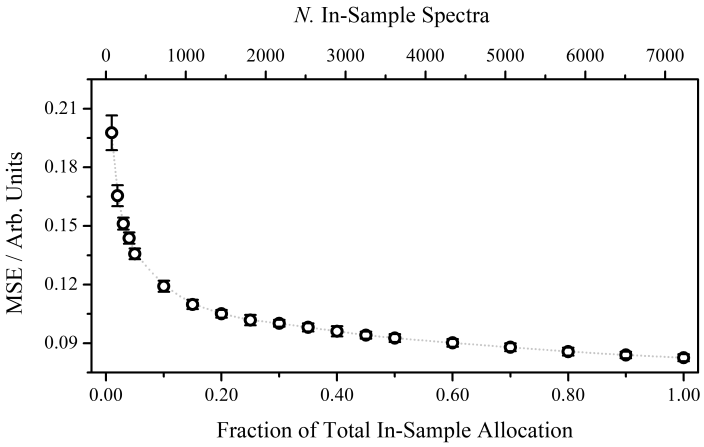

Figure 5: Evolution of the MSE as a function of the number of in-sample spectra accessible to our DNN during the learning process. Data points are averaged over 100 K-fold cross-validated evaluations; error bars indicate one standard deviation. No post-processing has taken place; XANES spectra have not been convoluted with the arctangent function.

For $[Fe(bpy)_3]^{2+}$ (Figure 6b), evaluation of the MSE along a symmetric stretching coordinate, rFe-N_{bpy}, produces a minimum in MSE at 1.91 Å - a result in agreement with the value of 1.93 Å obtained from high-level calculations[40] and 1.96 Å obtained from X-ray diffraction.[38] Similarly, evaluation of the MSE along the rFe-N_{NO}, rFe-$N_{porphyr.}$, and rFe-$N_{hist.}$ stretching coordinates in MbNO (Figure 6d) produces minima at 1.67, 1.98, and 2.31 Å, respectively. These values that are in reasonable agreement with previous XANES analysis[37] values of 1.83, 2.01, and 2.04 Å, and could serve as a starting point for a more rigorous refinement. Figure 6d shows that our DNN is more sensitive to distortions along rFe-$N_{porphyr.}$ than either rFe-N_{NO} or rFe-$N_{hist.}$. This is a consequence of the choice of featurisation, since the intensity of the dominant peak in the RDC is affected to a greater extent by changing four internuclear distances (rFe-$N_{porphyr.}$) along a symmetric stretching coordinate than by changing either of the other internuclear distances (rFe-N_{NO} or rFe-$N_{hist.}$) individually. Being able to refine independently these three internuclear distances attests to the sensitivity of our DNN because rFe-N_{NO}, rFe-$N_{porphyr.}$, and rFe-$N_{hist.}$ all appear under the same peak in the RDC; our DNN is sensitive to the subtle differences in peak shape and intensity communicated through the featurisation. Furthermore, there are no Fe absorption sites in our dataset that resemble

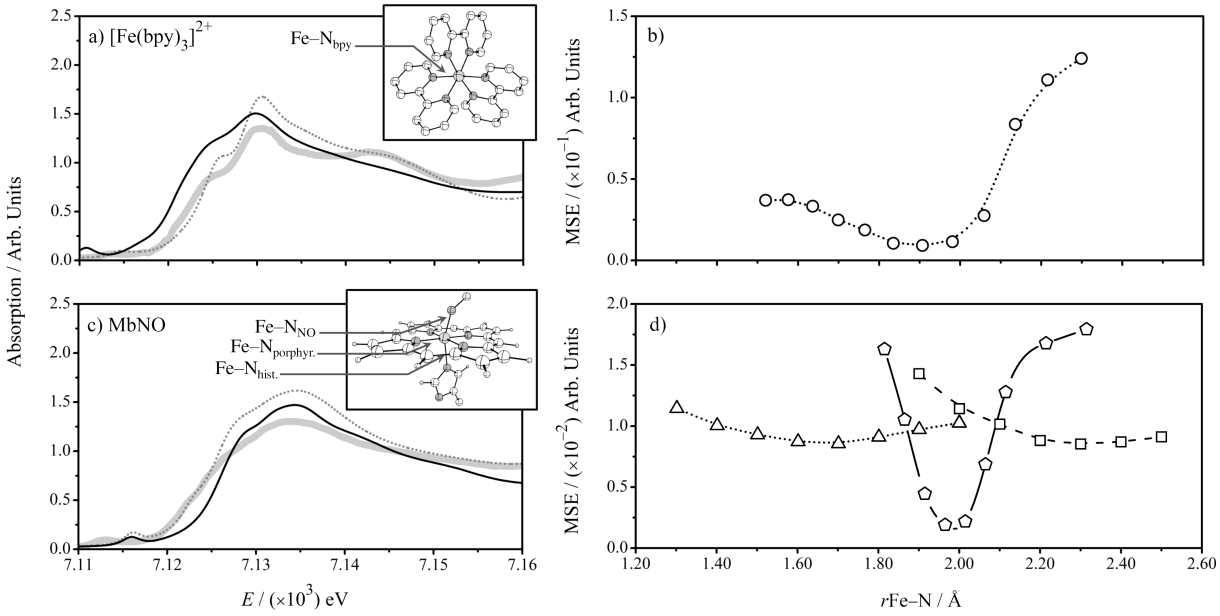

Figure 6: Experimental target (grey; continuous), theoretical target (grey; dotted), and DNN-estimated XANES spectra for a) $[Fe(bpy)_3]^{2+}$ and c) MbNO. MSEs as a function of Fe-to-N coordination distance, rFe-N, for b) $[Fe(bpy)_3]^{2+}$ and c) MbNO. The MSE as a function of rFe-N_{bpy} (dotted; circular markers) is plot in b). The MSE as a function of rFe-N_{NO} (dotted; triangular markers), rFe-$N_{porphyr.}$ (continuous; pentagonal markers), and rFe-$N_{hist.}$ (dashed; square markers) is plot in d).

the Fe absorption sites of $[Fe(bpy)_3]^{2+}$ or MbNO, and there are less than fifty examples of Fe-to-N coordination in our dataset. The performance of our DNN for $[Fe(bpy)_3]^{2+}$ and MbNO supports the conclusion that our machine learning model has learnt how to generalise beyond our dataset, as opposed to memorising the contents, and has begun to learn a deep relationship: how to map the local geometric environment of an arbitrary Fe absorption site to the corresponding XANES features in order to estimate the XANES spectrum.

Discussion and Conclusions

In the previous section we presented the promising performance of our DNN and successful application of this approach to the structural refinement of $[Fe(bpy)_3]^{2+}$ and MbNO. We now discuss the limitations of our DNN, which represent the areas of focus for the future development of the method. It is always the case that a machine learning model is only

as good as the dataset that it is exposed to during the learning process. This paper is intended as a proof-of-principle demonstration, and consequently uses a dataset that was computationally cost-effective to compute under the muffin-tin approximation. Indeed, this limitation is likely to be a contributing factor in the structural refinement of MbNO. However, now that the optimal hyperparameters, learning behaviour, and requisite dataset size have been determined, it is our objective to have our DNN reproduce XANES spectra from higher-level theoretical calculations. Even at higher levels of theory, it is important to acknowledge that disagreements between experiment and theory may still exist and it is important that these situations can be recognised so that the DNN is not treated as a black box by end-users. The objective of the DNN is to supplement and support high-level calculations, not replace them.

Furthermore, although our DNN is suitable for estimating the Fe K-edge XANES of molecular systems far outside our training dataset of perfectly-ordered, homogeneous crystalline systems, the sensitivity of our DNN to irregularities in the bulk such as vacancies, defects, undercoordinated sites, and the effects of lattice stress is not clear. Feedback on systems for which the underlying theory is inappropriate, and on epistemic uncertainty (*i.e.* uncertainty arising from the incompleteness and homogeneity of the training dataset), is essential for end-users and will be the focus of future developments for this method.

Although we have demonstrated that our DNN is sensitive to subtle structural differences (*e.g.* changes in internuclear distance of less than an Ångström near equilibrium), it is necessary to point out that our DNN cannot be expected at present to provide valid estimations of Fe K-edge XANES for structures that are far from equilibrium, limiting its present applicability to very-high-temperature samples and time-resolved XAS experiments. However, we expect that a combination of data augmentation and feedback on epistemic uncertainty will enable our DNN to address these problems in the near future.

In summary, this paper has introduced a DNN capable of estimating Fe K-edge XANES spectra in less than a second from no input beyond geometric information about the local

environment of an arbitrary Fe absorption site. We have demonstrated that not only is our DNN able to predict reliably XANES spectra with qualitative accuracy, it is also able to achieve quantitative (sub-eV) accuracy on peak positions with reference to the target XANES spectra. Our DNN can be trained to convergence on a moderately-sized dataset of theoretical XANES spectra in under ten minutes, and can cast individual estimations in less than a second. We expect that our DNN can be reoptimised to cast estimations of XANES spectra for other elements/absorption edges, and that it will transform consequently the analysis workflow in XAS spectroscopy. Beyond this, we expect that our approach is equally transferable to other spectroscopies. We stress that the objective of the work communicated in this paper is not to replace high-level theoretical calculations. Rather, we anticipate that our DNN will find application in situations where many geometric configurations need to be sampled to simulate accurately the XANES spectrum but the requisite computational resources for these high-level theoretical calculations are not available, *i.e.* in molecular dynamics/nuclear ensemble approaches,[41,42] and in qualitative analysis and screening for high-throughput XAS measurements.

Supporting Data

Data supporting this publication is openly available under an 'Open Data Commons Open Database License'. Additional metadata are available at: http://dx.doi.org/10.25405/data.ncl.12018129

Acknowledgements

The research described in this paper was funded by the Leverhulme Trust (Project RPG-2016-103) and EPSRC (EP/S022058/1, EP/R021503/1, and EP/R51309X/1). CDR is supported by a Doctoral Prize Fellowship (EP/R51309X/1). MMMM thanks Jazan University (KSA) for supporting her study and funding. This research made use of the Rocket High Performance Computing (HPC) service at Newcastle University. CDR additionally thanks the

Alan Turing Institute, *via* which access to the EPSRC-supported (EP/T022205/1) Joint Academic Data Science Endeavour (JADE) HPC cluster was provided under Project JAD029.

References

(1) Van Bokhoven, J. A.; Lamberti, C. *X-ray Absorption and X-ray Emission Spectroscopy: Theory and Applications*; John Wiley & Sons, 2016; Vol. 1.

(2) Lomachenko, K. A.; Borfecchia, E.; Negri, C.; Berlier, G.; Lamberti, C.; Beato, P.; Falsig, H.; Bordiga, S. The Cu-CHA deNO$_x$ Catalyst in Action: Temperature-Dependent NH$_3$-Assisted Selective Catalytic Reduction Monitored by Operando XAS and XES. *J. Am. Chem. Soc.* **2016**, *138*, 12025–12028.

(3) Luo, K.; Roberts, M. R.; Hao, R.; Guerrini, N.; Pickup, D. M.; Liu, Y.-S.; Edström, K.; Guo, J.; Chadwick, A. V.; Duda, L. C. et al. Charge Compensation in 3d Transition Metal Oxide Intercalation Cathodes Through the Generation of Localized Electron Holes on Oxygen. *Nat. Chem.* **2016**, *8*, 684.

(4) Penfold, T. J.; Milne, C. J.; Chergui, M. Recent Advances in Ultrafast X-ray Absorption Spectroscopy of Solutions. *Adv. Chem. Phys.* **2013**, *153*, 1–41.

(5) Kraus, P. M.; Zürch, M.; Cushing, S. K.; Neumark, D. M.; Leone, S. R. The Ultrafast X-ray Spectroscopic Revolution in Chemical Dynamics. *Nat. Rev. Chem.* **2018**, *2*, 82–94.

(6) Joly, Y.; Grenier, S.; Van Bokhoven, J.; Lamberti, C. Theory of X-ray Absorption Near-Edge Structure. *X-ray absorption and X-ray emission spectroscopy: Theory and applications* **2016**, 73–97.

(7) Natoli, C. In EXAFS and Near Edge Structure (A. Bianconi, L. Incoccia and S. Stipcich, eds.). *Springer Series in Chemical Physics* **1983**, *27*, 43.

(8) Westre, T. E.; Kennepohl, P.; DeWitt, J. G.; Hedman, B.; Hodgson, K. O.; Solomon, E. I. A Multiplet Analysis of Fe K-edge 1s-3d Pre-Edge Features of Iron Complexes. *J. Am. Chem. Soc.* **1997**, *119*, 6297–6314.

(9) Mino, L.; Agostini, G.; Borfecchia, E.; Gianolio, D.; Piovano, A.; Gallo, E.; Lamberti, C. Low-Dimensional Systems Investigated by X-ray Absorption Spectroscopy: a Selection of 2D, 1D and 0D Cases. *J. Phys. D* **2013**, *46*, 423001.

(10) Rehr, J.; Ankudinov, A. Progress in the Theory and Interpretation of XANES. *Coord. Chem. Rev.* **2005**, *249*, 131–140.

(11) Smolentsev, G.; Soldatov, A. V. FitIt: New Software to Extract Structural Information on the Basis of XANES Fitting. *Comput. Mater. Sci.* **2007**, *39*, 569–574.

(12) Martini, A.; Guda, S.; Guda, A.; Smolentsev, G.; Algasov, A.; Usoltsev, O.; Soldatov, M.; Bugaev, A.; Rusalev, Y.; Lamberti, C. et al. PyFitit: Software for Quantitative Analysis of XANES Spectra Using Machine Learning Algorithms. *Comput. Phys. Commun.* **2019**, 107064.

(13) Timoshenko, J.; Frenkel, A. I. "Inverting" X-ray Absorption Spectra of Catalysts by Machine Learning in Search for Activity Descriptors. *ACS Catal.* **2019**, *9*, 10192–10211.

(14) Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.

(15) Timoshenko, J.; Lu, D.; Lin, Y.; Frenkel, A. I. Supervised Machine-Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles. *J. Phys. Chem. Lett.* **2017**, *8*, 5091–5098.

(16) Timoshenko, J.; Halder, A.; Yang, B.; Seifert, S.; Pellin, M. J.; Vajda, S.; Frenkel, A. I. Subnanometer Substructures in Nanoassemblies Formed from Clusters under a Reactive Atmosphere Revealed Using Machine Learning. *J. Phys. Chem. C* **2018**, *122*, 21686–21693.

(17) Timoshenko, J.; Ahmadi, M.; Cuenya, B. R. Is There a Negative Thermal Expansion in Supported Metal Nanoparticles? An in Situ X-ray Absorption Study Coupled with Neural Network Analysis. *J. Phys. Chem. C* **2019**, *123*, 20594–20604.

(18) Ahmadi, M.; Timoshenko, J.; Behafarid, F.; Cuenya, B. R. Tuning the Structure of Pt Nanoparticles through Support Interactions: An in Situ Polarized X-ray Absorption Study Coupled with Atomistic Simulations. *J. Phys. Chem. C* **2019**, *123*, 10666–10676.

(19) Timoshenko, J.; Wrasman, C. J.; Luneau, M.; Shirman, T.; Cargnello, M.; Bare, S. R.; Aizenberg, J.; Friend, C. M.; Frenkel, A. I. Probing Atomic Distributions in Mono- and Bimetallic Nanoparticles by Supervised Machine Learning. *Nano Lett.* **2019**, *19*, 520–529.

(20) Liu, Y.; Marcella, N.; Timoshenko, J.; Halder, A.; Yang, B.; Kolipaka, L.; Pellin, M. J.; Seifert, S.; Vajda, S.; Liu, P. et al. Mapping XANES Spectra on Structural Descriptors of Copper Oxide Clusters Using Supervised Machine Learning. *J. Chem. Phys.* **2019**, *151*, 164201.

(21) Zheng, C.; Mathew, K.; Chen, C.; Chen, Y.; Tang, H.; Dozier, A.; Kas, J. J.; Vila, F. D.; Rehr, J. J.; Piper, L. F. J. et al. Automated Generation and Ensemble-Learned Matching of X-ray Absorption Spectra. *NPJ Comput. Mater.* **2018**, *4*, 12.

(22) Mathew, K.; Zheng, C.; Winston, D.; Chen, C.; Dozier, A.; Rehr, J. J.; Ong, S. P.; Persson, K. A. Data Descriptor: High-Throughput Computational X-ray Absorption Spectroscopy. *Sci. Data* **2018**, *5*, 108151.

(23) Carbone, M. R.; Topsakal, M.; Lu, D.; Yoo, S. Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy. *Phys. Rev. Lett.* **2020**, *124*, 156401.

(24) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O.A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data.* **2014**, *1*, 140022.

(25) Glatzel, P.; Bergmann, U. High Resolution 1s Core Hole X-ray Spectroscopy in 3d Transition Metal Complexes- Electronic and Structural Information. *Coord. Chem. Rev.* **2005**, *249*, 65–95.

(26) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comp. Chem.* **2017**, *38*, 1291–1307.

(27) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336–2347.

(28) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G. et al. The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.

(29) Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; Persson, K. A. The Materials Application Programming Interface (API): A Simple, Flexible and Efficient API for Materials Data Based on Representational State Transfer (REST) Principles. *Comput. Mater. Sci.* **2015**, *97*, 209–215.

(30) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. 2015; `tensorflow.org/`.

(31) Keras. 2015; `github.com/keras-team/keras`.

(32) GPy: A Gaussian Process Framework in Python. 2012; `github.com/SheffieldML/GPy`.

(33) GPyOpt: A Bayesian Optimization Framework in Python. 2016; `github.com/SheffieldML/GPyOpt`.

(34) Bunău, O.; Joly, Y. Self-Consistent Aspects of X-ray Absorption Calculations. *J. Phys. Condens. Matter* **2009**, *21*, 345501.

(35) Seah, M.; Dench, W. NPL Report Chem. **1978**, *82*.

(36) Cannizzo, A. ; Milne, C. J. ; Consani, C.; Gawelda, W.; Bressler, Ch.; Van Mourik, F.; Chergui, M. Light Induced Spin Crossover in Fe (II) Based Complexes: The Full Photocycle Unraveled by Ultrafast Optical and X-ray Spectroscopies. *Coord. Chem. Rev.* **2010**, *254*, 2677–2686.

(37) Lima, F. A.; Penfold, T. J.; van der Veen, R. M.; Reinhard, M. and Abela, R.; Tavernelli, I.; Rothlisberger, U. ; Benfatto, M.; Milne, C. J.; Chergui, M. Probing the Electronic and Geometric Structure of Ferric and Ferrous Myoglobins in Physiological Solutions by Fe K-edge Absorption Spectroscopy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 1617–1631.

(38) Dick, S.; Crystal Structure of tris (2, 2'-bipyridine) iron (II) bis (hexafluorophosphate),(C10H8N2) 3Fe (PF6) 2. *Z. Kristallogr. New Cryst. Struct.* **1998**, *213*, 370–370.

(40) Sousa, C.; de Graaf, C.; Rudavskyi, A.; Broer, R.; Tatchen, J.; Etinski, M.; Marian, C. M. Ultrafast Deactivation Mechanism of the Excited Singlet in the Light-Induced Spin Crossover of [Fe (2, 2?-bipyridine) 3] 2+. *Chem. Eur. J.* **2013**, *19*, 17541–17551.

(40) Sousa, C.; de Graaf, C.; Rudavskyi, A.; Broer, R.; Tatchen, J.; Etinski, M.; Marian, C. M. Ultrafast Deactivation Mechanism of the Excited Singlet in the Light-Induced Spin Crossover of [Fe (2, 2?-bipyridine) 3] 2+. *Chem. Eur. J.* **2013**, *19*, 17541–17551.

(41) Katayama, T.; Northey, T.; Gawelda, W.; Milne, C. J.; Vankó, G.; Lima, F. A.; Németh, Z.; Nozawa, S.; Sato, T.; Khakhulin, D. et al. Tracking Multiple Components of a Nuclear Wavepacket in Potoexcited Cu(I)-Phenanthroline Complex Using Ultrafast X-ray Spectroscopy. *Nat. Comm.* **2019**, *10*, 1–8.

(42) Capano, G.; Milne, C.; Chergui, M.; Rothlisberger, U.; Tavernelli, I.; Penfold, T. Probing Wavepacket Dynamics Using Ultrafast X-ray Spectroscopy. *J. Phys. B* **2015**, *48*, 214001.

Graphical TOC Entry