Supporting Information: Uncertainty Quantification of Spectral Predictions Using Deep Neural Networks

Sneha Verma [1], Nik Khadijah Nik Aznan [2], Kathryn Garside [2] and Thomas J. Penfold [1]

List of Figures

[1]Chemistry - School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK, tom.penfold@ncl.ac.uk

[2]Research Software Engineering Group - School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

List of Tables

S1 Theory and Computational Details

S1 Neural Networks

In this work, we have used two DNN architecture. The first is based upon the multilayer perceptron (MLP) model and closely follows the network described in ref. [1]. Briefly, this MLP model comprises an input layer, two hidden layers, and an output layer. All layers are dense, *i.e.* fully connected, and each hidden layer performs a nonlinear transformation using the hyperbolic tangent (*ReLU*) activation function. The input layer comprises N neurons to accept a feature vector of length N encoding the local environment around the absorbing atom. This featurisation is performed *via* dimensionality reduction using the wACSF descriptor of Gastegger and Marquetand *et al.* [2]. In this Article, the input layer contains 97 neurons as the feature vector is comprised of a global (G^1) function, 32 radial (G^2) functions, and 64 angular (G^4) functions.

The first and second hidden layers comprises 512 neurons. The output layer comprises 226 neurons from which the discretised K-edge XANES spectrum is retrieved after regression. The internal weights, $\mathbf{W}$, are optimised *via* iterative feed-forward and backpropagation cycles to minimise the empirical loss, $J(\mathbf{W})$, defined here as the mean-squared error (MSE) between the predicted, $\mu_{predict}$, and calculated, $\mu_{calculated}$, K-edge XANES spectra over the reference dataset, *i.e.* an optimal set of internal weights, $\mathbf{W}^*$, is sought that satisfies $\underset{\mathbf{W}}{\mathrm{argmin}}(J(\mathbf{W}))$.

Gradients of the empirical loss with respect to the internal weights, $\delta J(\mathbf{W})/\delta \mathbf{W}$, were estimated over minibatches of 32 samples and updated iteratively according to the Adaptive Moment Estimation (ADAM) [3] algorithm. Throughout an annealed learning rate. The learning rate was initially set to 2×10^{-3} and was reduced by a factor of 2 every 100 epochs. The internal weights were initially set according to ref. [4]. Unless explicitly stated in this Article, optimisation was carried out over 500 iterative cycles through the network commonly termed *epochs* meaning the final learning rate was 1.3×10^{-4}. Regularisation was implemented to minimise the propensity of overfitting.

Our second network is based upon a convolutional neural network. After featurisation, as described above, this network contains two 1D convolutional layers, each of which is convolved with its own set of 16 filters. In both cases, the kernel size is 3 and the stride for the cross-correlation is 2. The output of these convolutional layers is passed into a MLP, containing two hidden layers and an output layer. All layers are dense, *i.e.* fully connected, and each hidden layer performs a nonlinear transformation using the hyperbolic tangent (*ReLU*) activation function. The hidden layers have a dimension of 256. The output layer comprises 226 neurons from which the discretised K-edge XANES spectrum is retrieved after regression. All others hyperparameters are the same as the MLP model.

The DNN is programmed in Python 3 with Pytorch [5]. The Atomic Simulation Environment [6] (*ase*) API is used to handle and manipulate molecular structures. The code is publicly available under the GNU Public License (GPLv3) on GitLab. [7]

S2 Datasets

Our datasets containing X-ray absorption site geometries (*samples*) of first-row transition metal (Ti-Zn) complexes have been harvested from the transition metal Quantum Machine (tmQM) dataset [8,9]. K-edge XANES spectra (*labels*) for these structures were calculated using multiple scattering theory (MST) as implemented in the FDMNES [10]. This led us to develop nine independent reference datasets, one for each first-row transition metal (Ti-Zn) X-ray absorption edge; the number of samples contained in the reference datasets scales from as few as $\sim$1100

(V) to ~8660 (Ni). 250 samples from each reference dataset were isolated at random to form *held-out* testing datasets (evaluated post-optimisation only). These datasets were previously presented in ref. [1] and can be obtained from ref. [9].

S3 Uncertainty Quantification Methods and Analysis

In the subsections below, the methods for uncertainty quantification are outlined. To assess their performance we use a number of metrics namely: coverage, correlation between mean squared error and standard deviation, calibration curves and sharpness.

Coverage refers to the percentage of calculated spectral data points from the *held-out* set which fall within $\pm 3\sigma$ of the average predicted spectra and therefore provides a direct assessment for the uncertainty quantification method to recognise poor predictions. However, while useful, the coverage metric can only be used if the true spectrum is calculated, which would negate the purpose for the use of DNN for future applications. Consequently, we have also used the correlation between the mean-squared error (MSE) between predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ spectra against the uncertainty (σ) calculated for every point in the spectrum and averaged over each spectrum.

Calibration, sometimes referred to as honesty, displays the frequency of expected points in each prediction interval relative to the predicted fraction of points in that interval [11]. In other words, if a model is calibrated the predicted probability should be equal to the expected probability in that quantile. This is achieved by comparing the expected cumulative distribution of points within a given confidence interval to the observed cumulative distributions of the predicted data. A plot of the observed cumulative distributions against the theoretical cumulative distributions is called a calibration curve. A perfectly calibrated model would have normalised residuals that are Gaussian, which would yield a diagonal calibration line. Finally, the sharpness is used to define the size of the uncertainty, with it being defined as the mean of the σ calculated over all data points in the files. To simultaneously take into account the dispersion of the uncertainty estimates, we also calculate the coefficient of variation (C_v):

$$C_v = \frac{\sqrt{\dfrac{\sum_{n=1}^{N}(\sigma_n - \overline{\sigma})^2}{N-1}}}{\overline{\sigma}} \tag{1}$$

where σ_n is the predicted standard deviation of point n, $\overline{\sigma}$ is the average value of σ_n and N is the total number of test points. Low values of C_v indicate a narrow dispersion of uncertainty estimates.

S3.1 Ensembling

The initial weights in the XANES DNN are drawn from the uniform distribution according to ref. [4]. However, there is no guarantee that there exists a single unique solution to the problem of finding a set of internal weights, $\mathbf{W}^*$, which satisfy $\underset{\mathbf{W}}{\operatorname{argmin}}(J(\mathbf{W}))$. An ensemble of statistically-initialised XANES DNNs optimised using the same reference dataset can consequently yield different $\mathbf{W}^*$. Deep ensembling involves optimising multiple instances of a machine learning model with the same reference dataset but with a different statistical initialisation of the internal weights. The ensemble of N machine learning models is then used to produce N independent predictions for each sample in the *held-out* testing dataset from which a mean prediction and standard deviation for each sample can be derived. The latter quantifies the ensemble uncertainty. The concept behind this interpretation is that different $\mathbf{W}^*$ obtained from

6

the N models in the ensemble will tend towards similar spectral predictions when the structural inputs are comparable to those used in the training data. This is because each instance weights, even if different, are optimised for comparable data. In contrast, if the structural inputs are dissimilar to the structures used to train the model, each of the N independent predictions will be more affected by the specificity's of $\mathbf{W}^*$, thus the higher standard deviation will be observed.

S3.2 Bootstrap Resampling

The bootstrap resampling technique, illustrated in Figure 1 in the main text, is used to estimate statistics on a population by sampling a dataset with replacement, and can be used to describe the uncertainty associated with incomplete training data. [12] N machine learning models are optimised using N reference datasets sampled with replacement from the original reference dataset; each one of these is the same size as the original reference dataset and, consequently, may contain repeated instances of the same sample. The bootstrap resampling technique provides increased dataset diversity to each instance of the machine learning model. N independent instances of the machine learning model optimised using N bootstrapped reference datasets are then used to produce N independent predictions for each sample in the "*held-out*" testing dataset from which a mean prediction and standard deviation for each sample can be derived. In contrast to ensembling, bootstrap resampling relies on injecting diversity into the reference dataset, rather than simply sampling diverse machine learning model configurations. Throughout this work N=15, which as shown in Figure S1 below for the Fe K-edge XANES spectra modelled using the MLP model, was selected as it achieved convergence in the median coverage.

As a final note, while both approaches require running multiple models, the predictions from them are close to instantaneous and so the time difference between 1 and 15 predictions is insignificant (<1 secs). The increased computational burden associated with an ensemble of models only occurs during the training of multiple models. However, as demonstrated by us in ref. [1], these models can be optimised in as little as a minute using an off-the-shelf commercial-grade central processing unit (CPU) (AMD Ryzen Threadripper 3970X, 3.7-4.5 GHz) or graphics processing unit (GPU) (NVIDIA RTX 3070, 5888 CUDA cores, 1.5-1.7 GHz) and so the difference between 1 and 15 models is not a limiting factor.
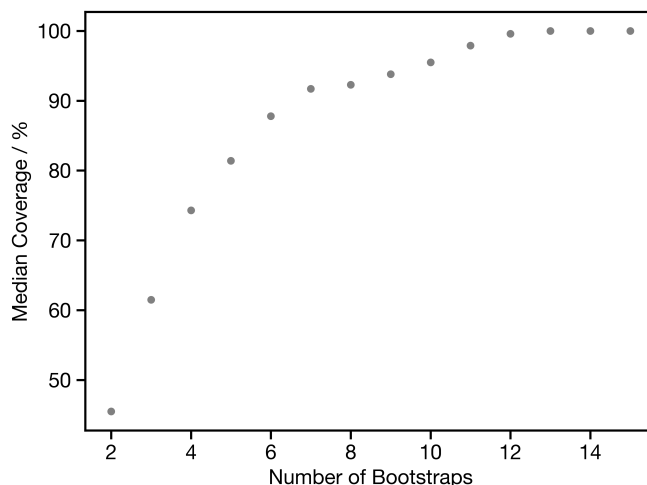
Figure S1: Plot of Median Coverage for the Fe K-edge XANES spectra as a function of the number of models. Calculated using the Bootstrap resampling approach and the MLP network.

S2 Supplementary Results

Edge	$\Delta\mu$	IQR	Coverage	$\rho^{all}_{\Delta\mu-\sigma}$	$\rho^{spec}_{\Delta\mu-\sigma}$	Sharp.	C_v
	MLP using Deep Ensembles
Ti	5.0	4.1	99.1	0.84	0.81	0.03	0.93
V	4.5	7.1	95.6	0.79	0.94	0.02	1.01
Cr	3.7	4.1	94.4	0.79	0.86	0.02	1.18
Mn	4.0	3.8	97.8	0.82	0.90	0.02	1.09
Fe	4.5	4.0	99.6	0.82	0.85	0.02	0.96
Co	4.1	3.8	99.3	0.81	0.88	0.02	1.02
Ni	4.0	3.7	99.7	0.82	0.89	0.02	0.92
Cu	3.7	3.5	100.0	0.81	0.88	0.01	0.91
Zn	3.2	3.0	100.0	0.79	0.87	0.01	1.27
	CNN using Bootstrap Resampling
Ti	5.6	3.9	98.8	0.85	0.71	0.03	0.88
V	5.6	6.2	99.6	0.81	0.93	0.03	1.04
Cr	4.0	4.4	96.9	0.80	0.86	0.02	1.20
Mn	5.5	3.7	97.9	0.82	0.82	0.02	1.07
Fe	5.8	3.4	97.7	0.82	0.74	0.02	0.82
Co	5.6	4.1	97.8	0.81	0.75	0.20	0.89
Ni	4.9	3.6	93.4	0.83	0.73	0.02	0.87
Cu	5.1	2.8	96.9	0.79	0.72	0.01	0.83
Zn	3.8	2.9	97.3	0.80	0.84	0.01	1.09
	CNN using Deep Ensembles
Ti	5.7	3.4	95.6	0.85	0.73	0.03	0.89
V	5.2	6.5	96.5	0.81	0.92	0.02	1.02
Cr	3.8	4.1	91.1	0.80	0.84	0.02	1.22
Mn	4.4	3.3	94.2	0.81	0.81	0.02	1.05
Fe	5.2	3.6	93.6	0.83	0.72	0.02	0.81
Co	4.7	3.1	94.2	0.81	0.73	0.01	0.85
Ni	5.1	2.7	92.0	0.81	0.69	0.02	0.89
Cu	4.1	2.4	93.4	0.79	0.70	0.01	0.86
Zn	3.7	2.8	92.0	0.80	0.84	0.01	1.05

Table S1: Summary of the median percentage errors, $\Delta\mu$ (%), interquartile range, coverage and the Pearson correlation between $\Delta\mu$ and σ for each energy point ($\rho^{all}_{\Delta\mu-\sigma}$) and each spectra ($\rho^{spec}_{\Delta\mu-\sigma}$). The sharpness (Sharp.) and coefficient of variation (C_v) for the uncertainty prediction are also shown as discussed above.

S1 Correlation between Mean Square Error and σ

Figure S2: Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ Ti K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.

Figure S3: Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ V K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.

Figure S4: Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ Cr K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.
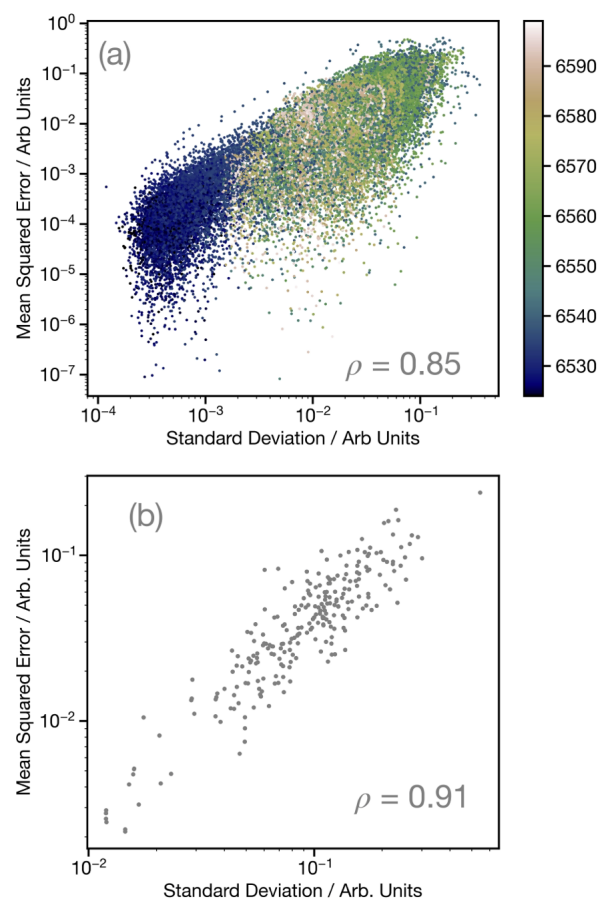
Figure S5: Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ Mn K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.
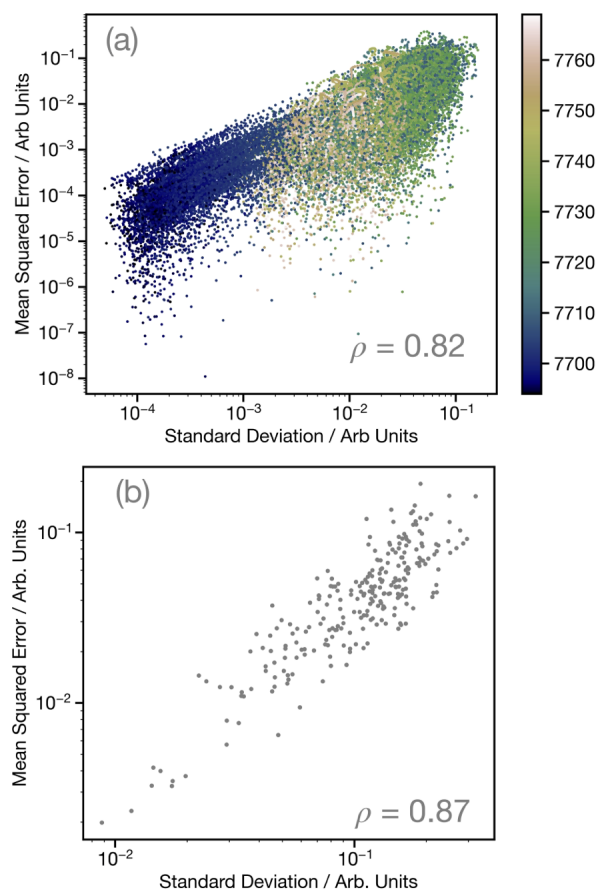
Figure S6: Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ Co K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.
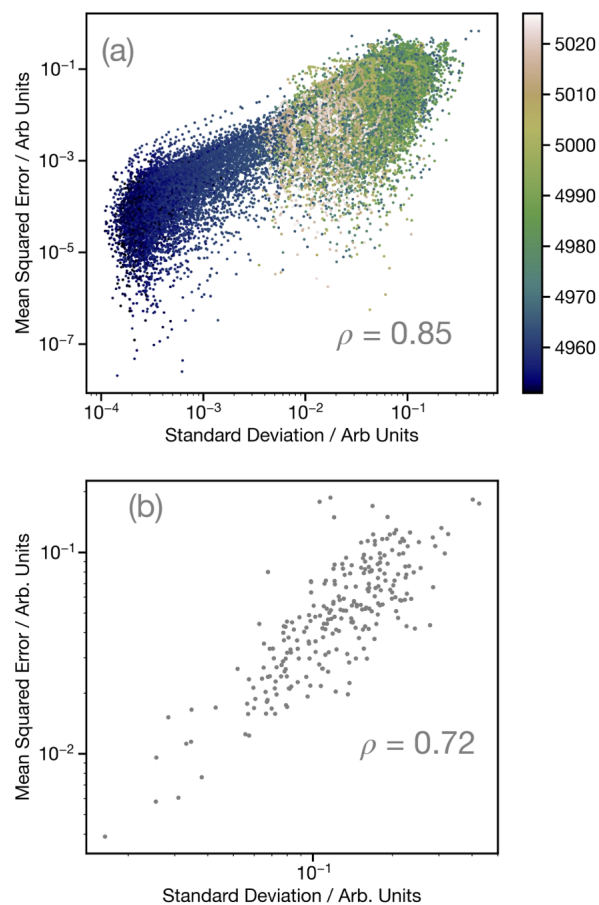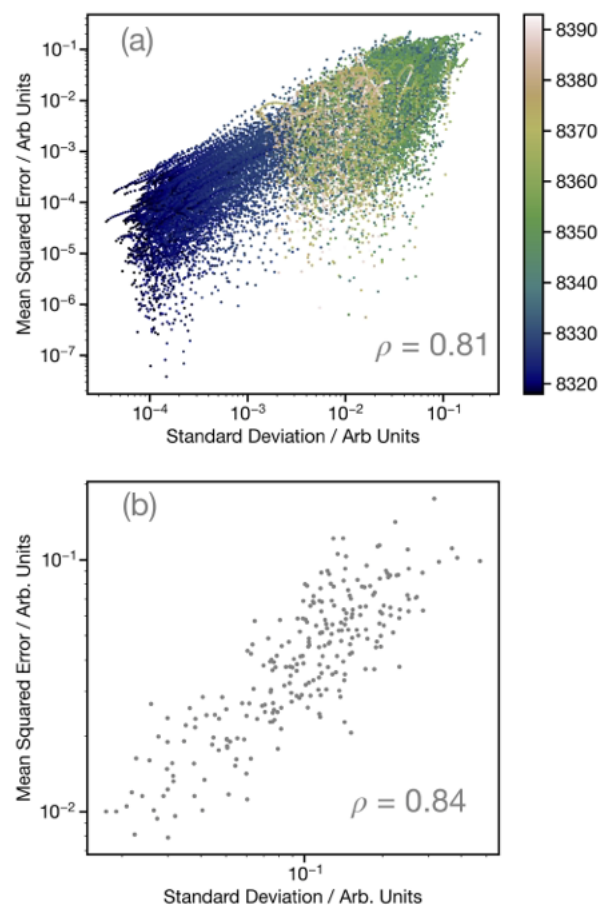
Figure S7: Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ Ni K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.
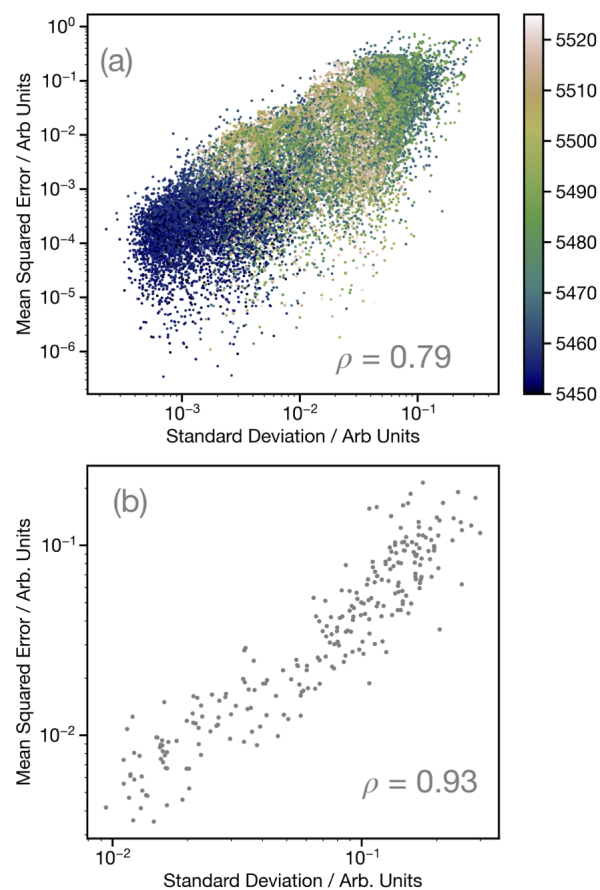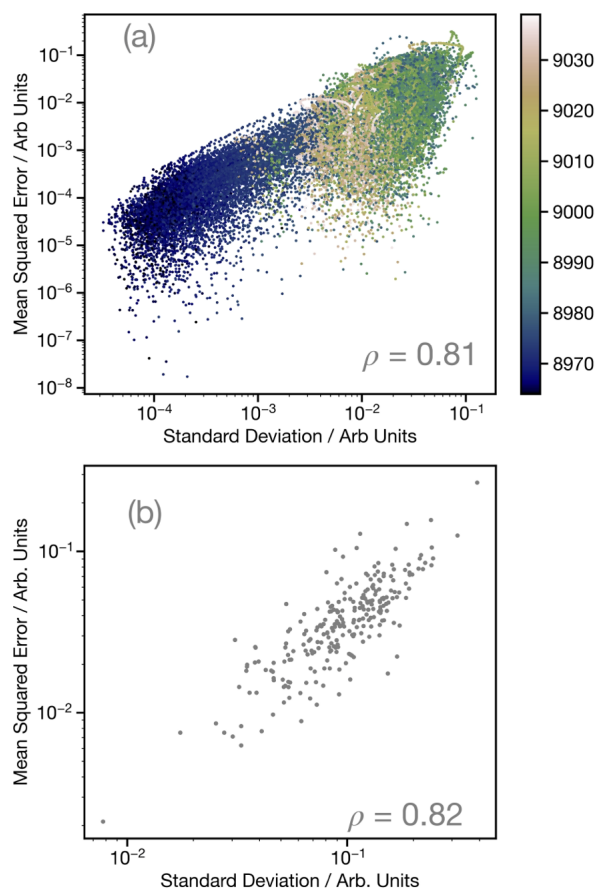
Figure S8: Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ Cu K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.
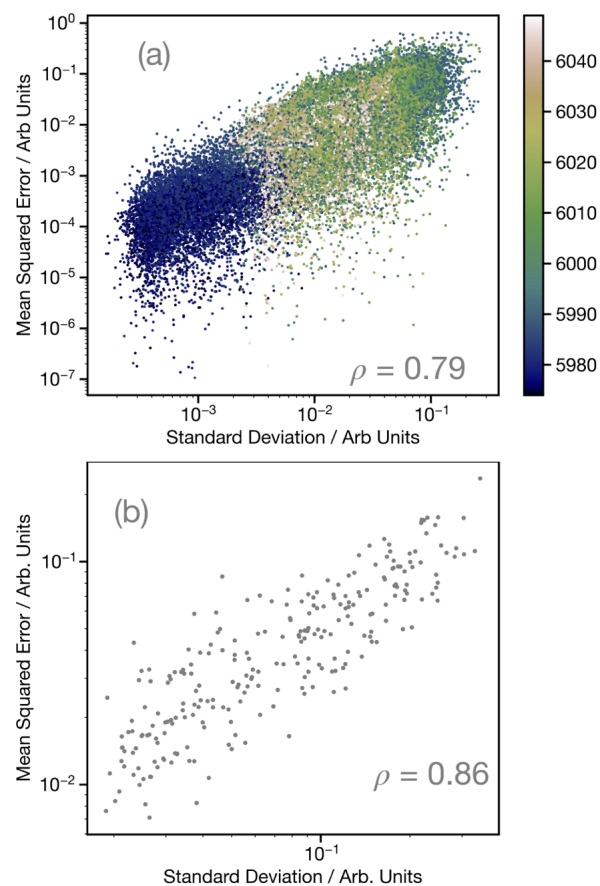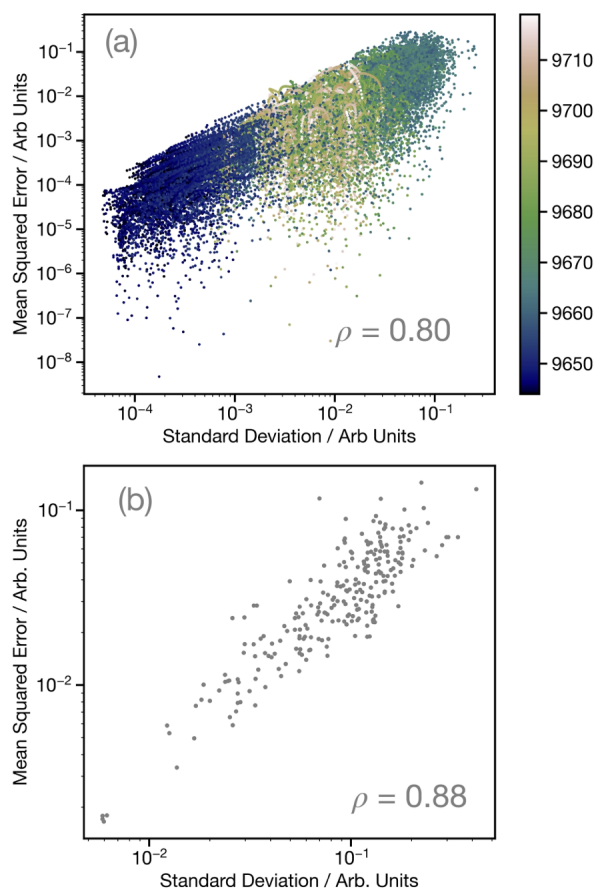
Figure S9: Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{predicted}$, and calculated, $\mu_{calculated}$ Zn K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.

S2 Calibration curves and Distribution of Predicted Standard Deviations

Figure S10: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the Ti K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.
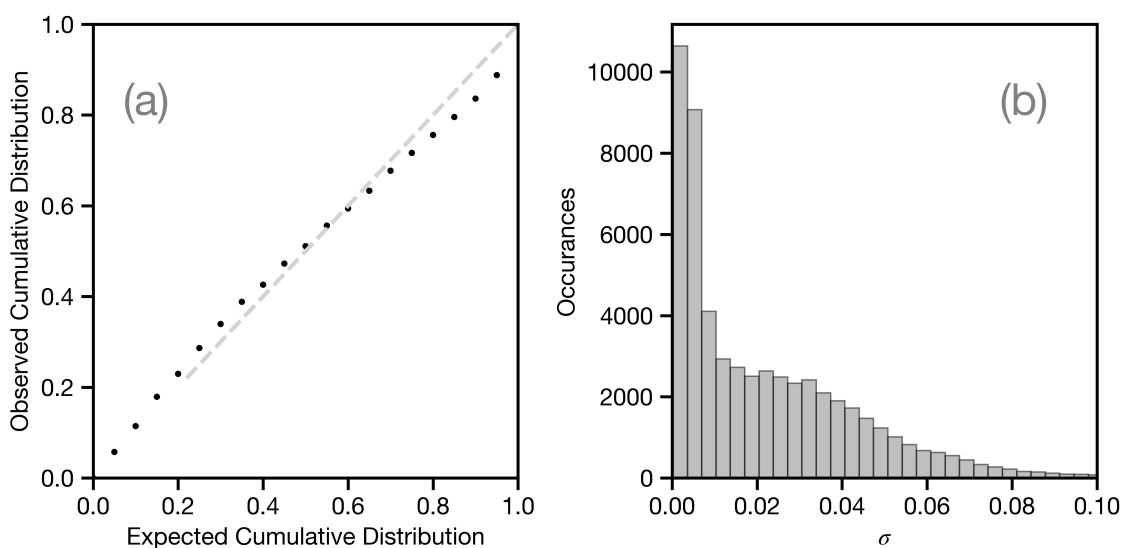
Figure S11: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the V K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.
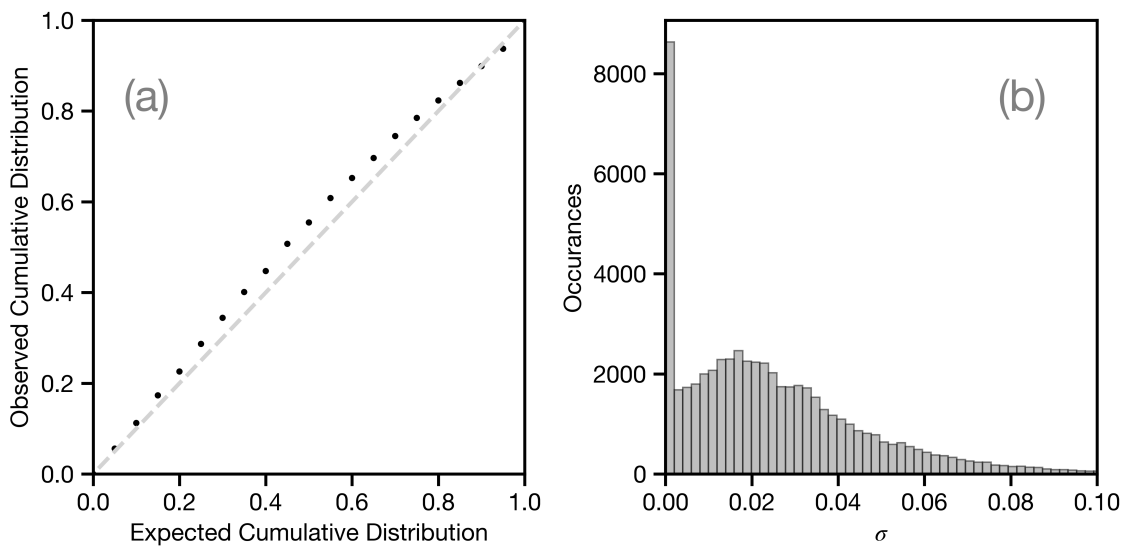
Figure S12: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the Cr K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.
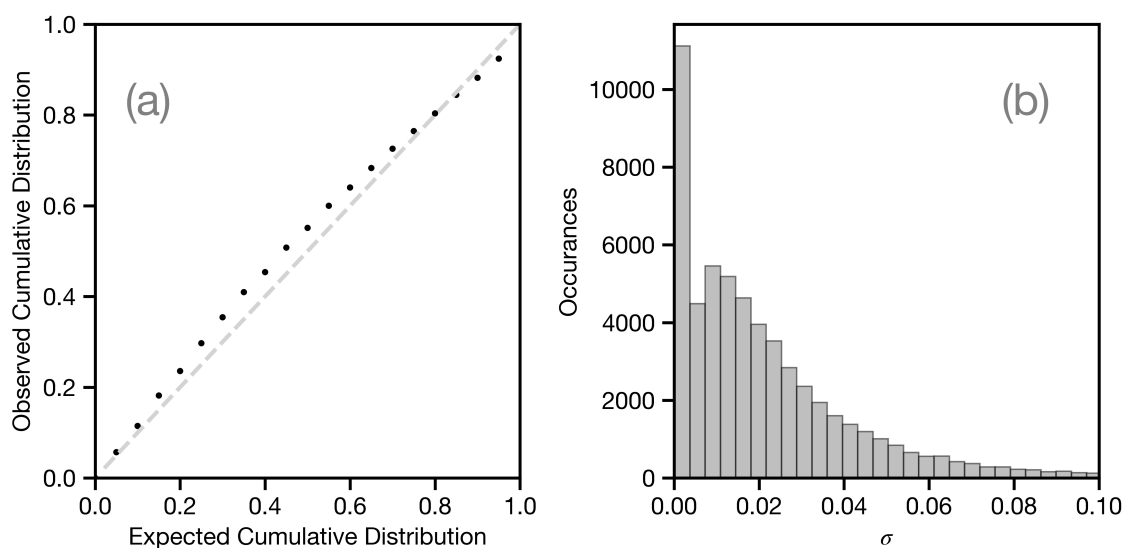
Figure S13: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the Mn K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.
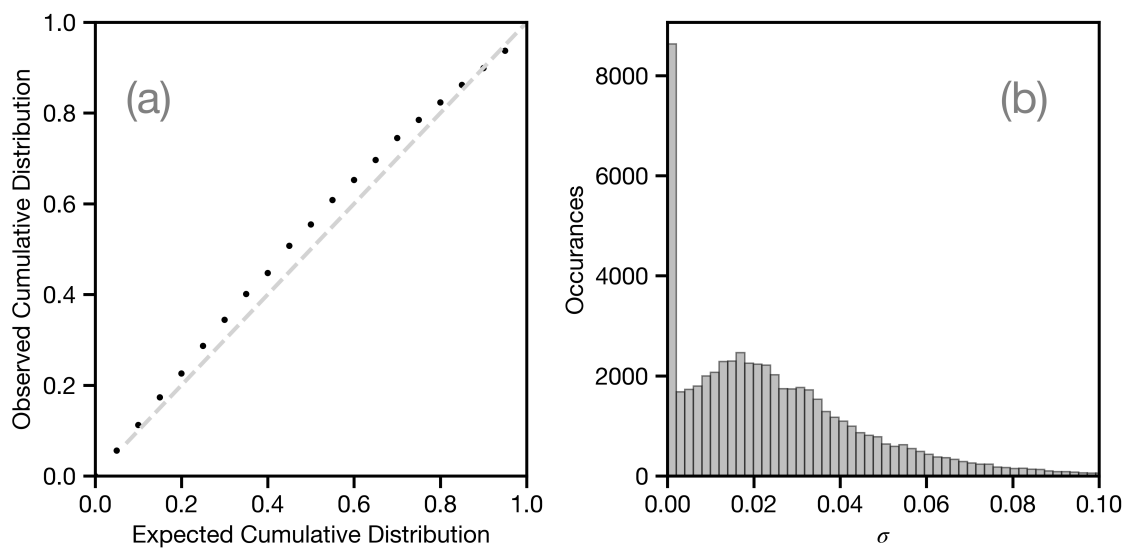
Figure S14: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the Fe K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.
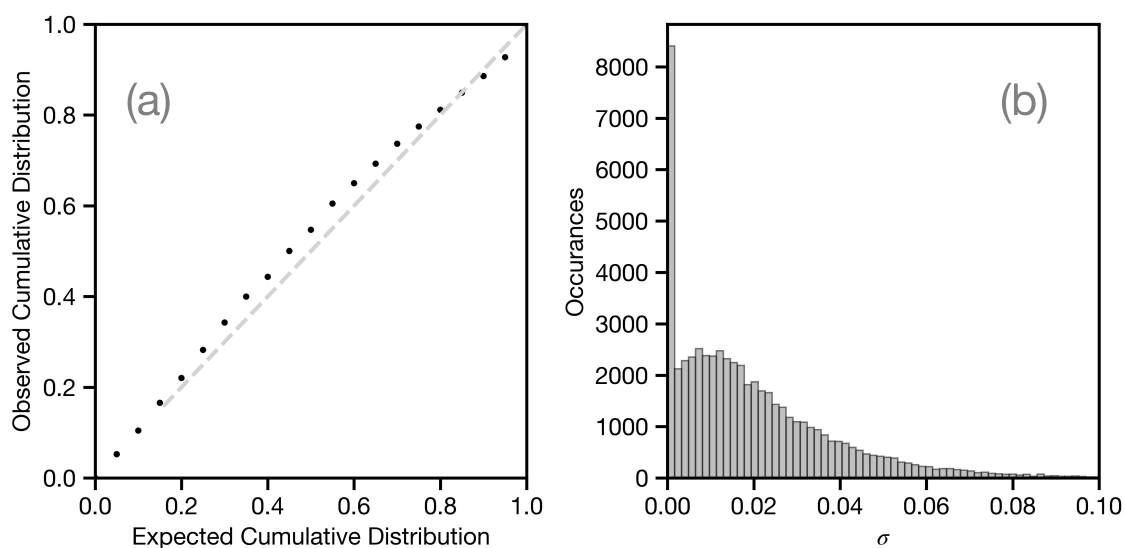
Figure S15: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the Co K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.
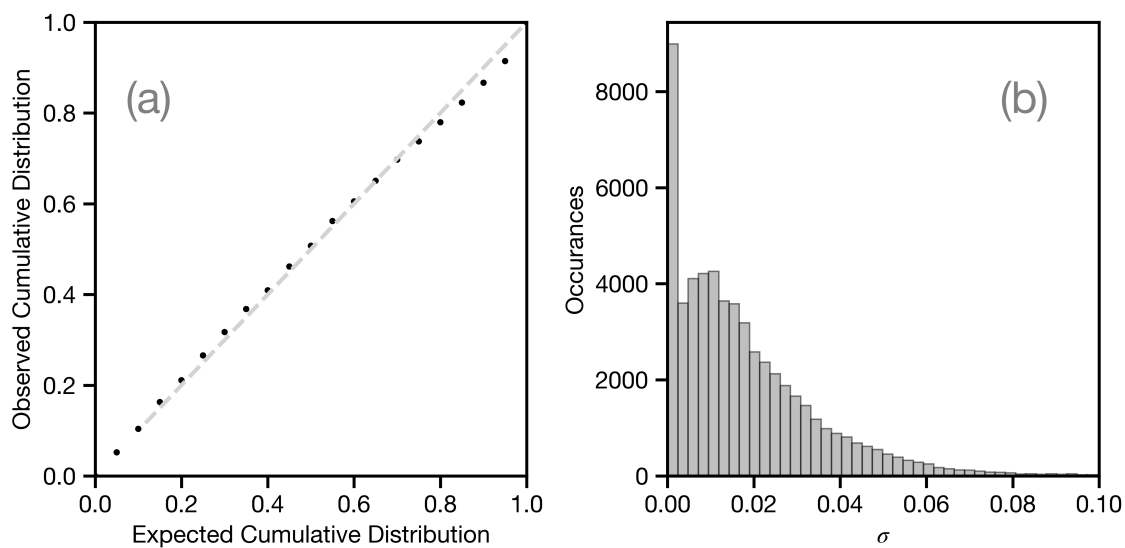
Figure S16: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the Ni K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.
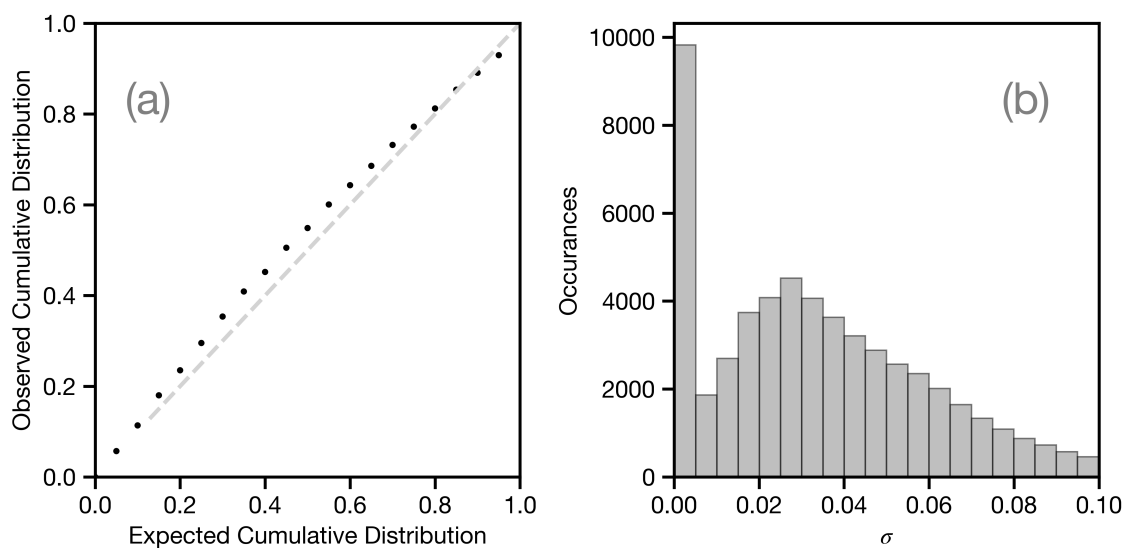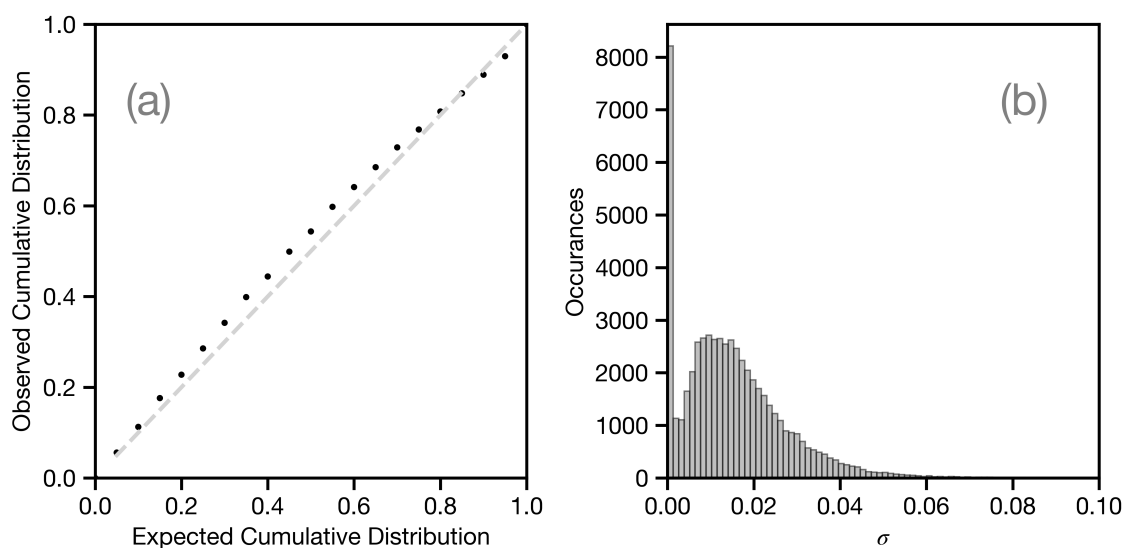
Figure S17: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the Cu K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.
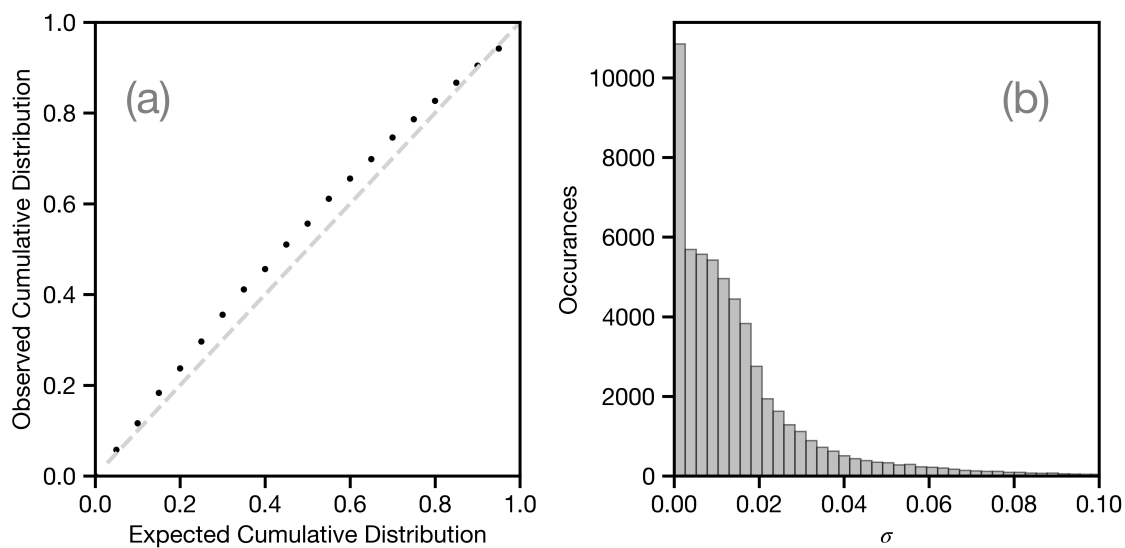
Figure S18: Calibration curve (a) and histogram of all predicted standard deviations used in the sharpness analysis (b) at the Zn K-edge XANES spectra. Calculated using the Bootstrap resampling approach and the MLP network.

S3 Sample Spectra from Held-out Set: Bootstrap Resampling using MLP model

Figure S19: Example K-edge XANES spectra for Ti-containing samples. The upper three panels show K-edge XANES spectra from the 0th-10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th-55th percentiles, *i.e.* around the median. The lower three panels show K-edge XANES spectra from the 90th-100th percentiles, *i.e.* the lowest performance. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples.

Figure S20: Example K-edge XANES spectra for V-containing samples. The upper three panels show K-edge XANES spectra from the 0th-10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th-55th percentiles, *i.e.* around the median. The lower three panels show K-edge XANES spectra from the 90th-100th percentiles, *i.e.* the lowest performance. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples.

Figure S21: Example K-edge XANES spectra for Cr-containing samples. The upper three panels show K-edge XANES spectra from the 0th-10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th-55th percentiles, *i.e.* around the median. The lower three panels show K-edge XANES spectra from the 90th-100th percentiles, *i.e.* the lowest performance. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples.
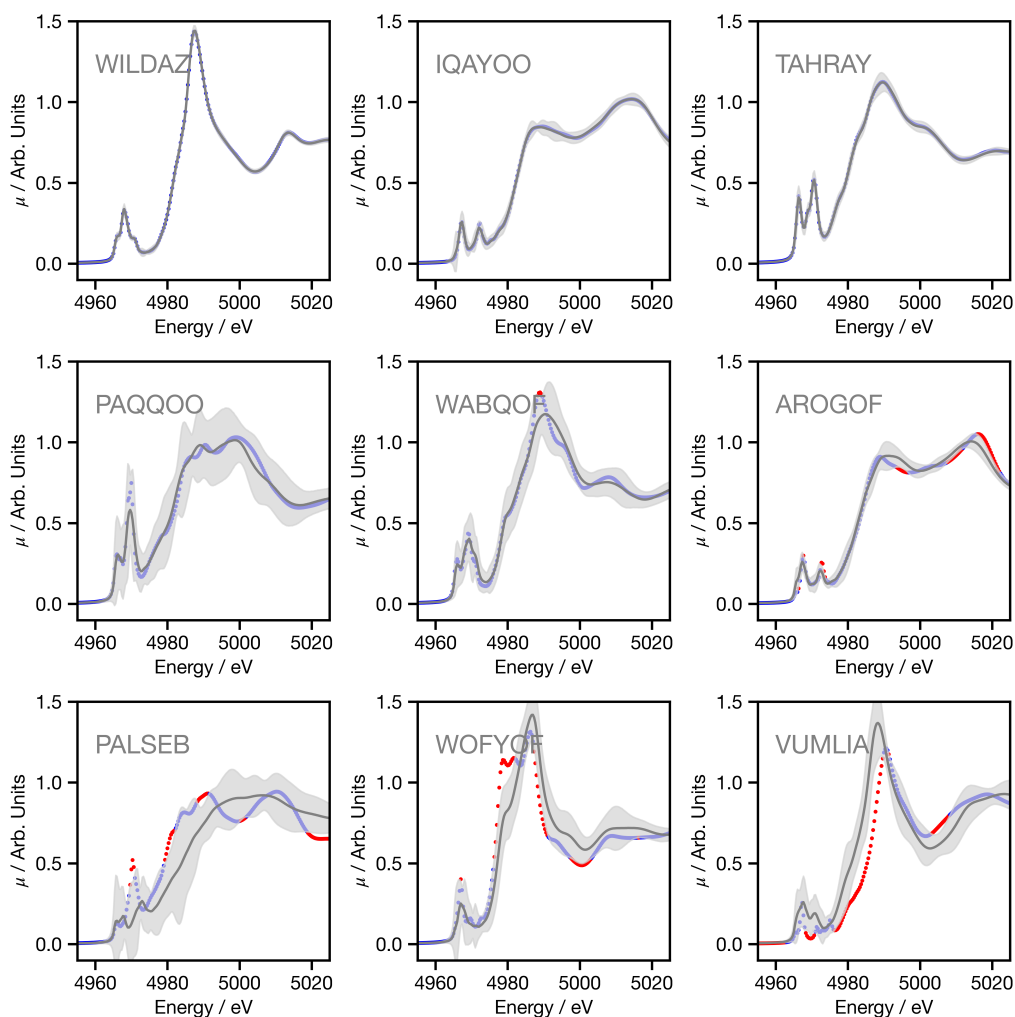
Figure S22: Example K-edge XANES spectra for Mn-containing samples. The upper three panels show K-edge XANES spectra from the 0th-10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th-55th percentiles, *i.e.* around the median. The lower three panels show K-edge XANES spectra from the 90th-100th percentiles, *i.e.* the lowest performance. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples.
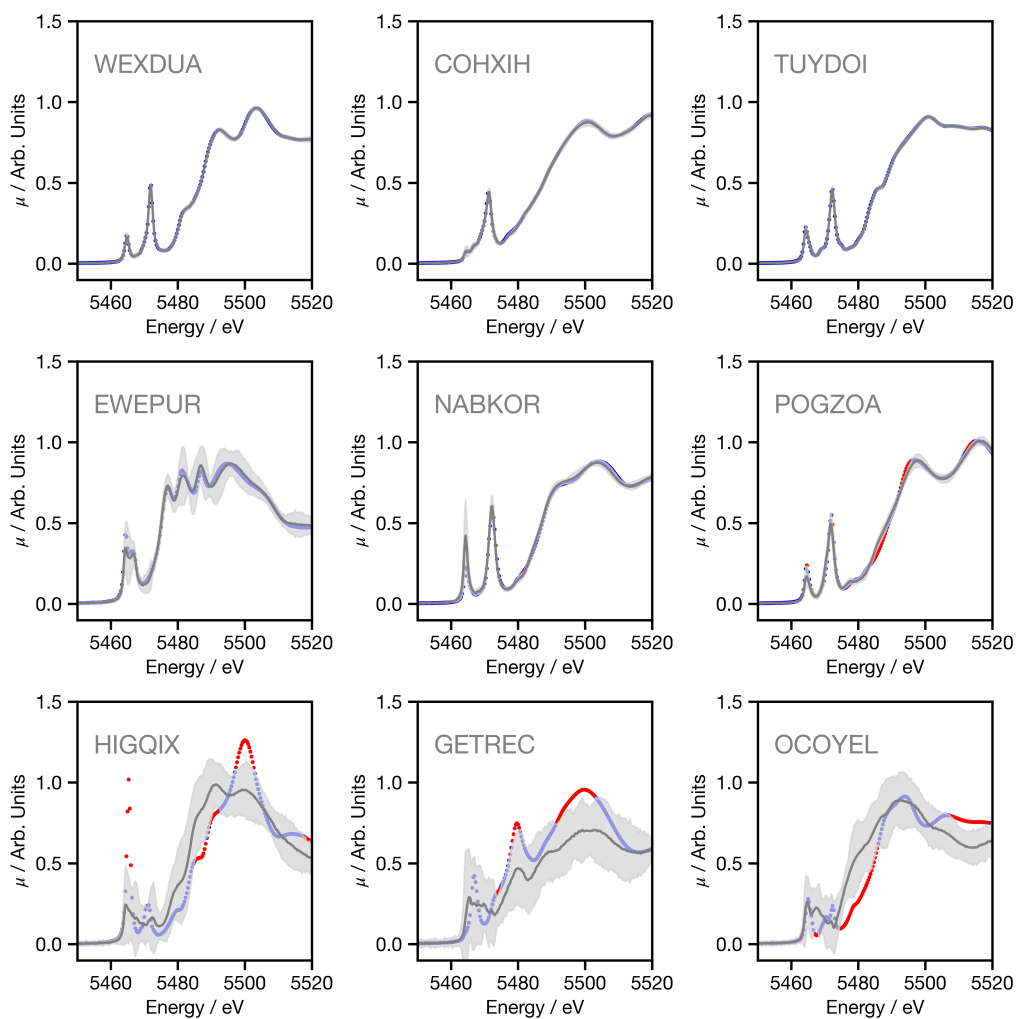
Figure S23: Example K-edge XANES spectra for Co-containing samples. The upper three panels show K-edge XANES spectra from the 0th-10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th-55th percentiles, *i.e.* around the median. The lower three panels show K-edge XANES spectra from the 90th-100th percentiles, *i.e.* the lowest performance. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples.
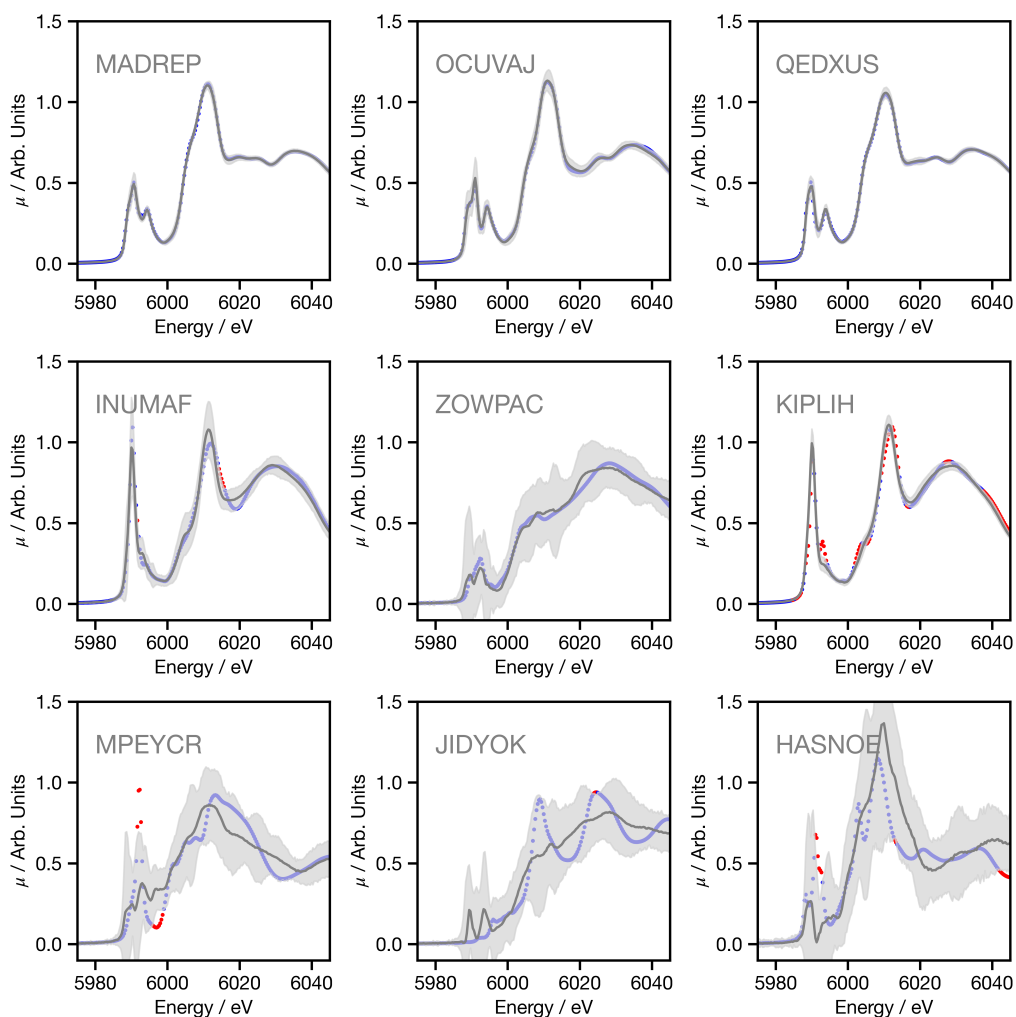
Figure S24: Example K-edge XANES spectra for Ni-containing samples. The upper three panels show K-edge XANES spectra from the 0th-10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th-55th percentiles, *i.e.* around the median. The lower three panels show K-edge XANES spectra from the 90th-100th percentiles, *i.e.* the lowest performance. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples.
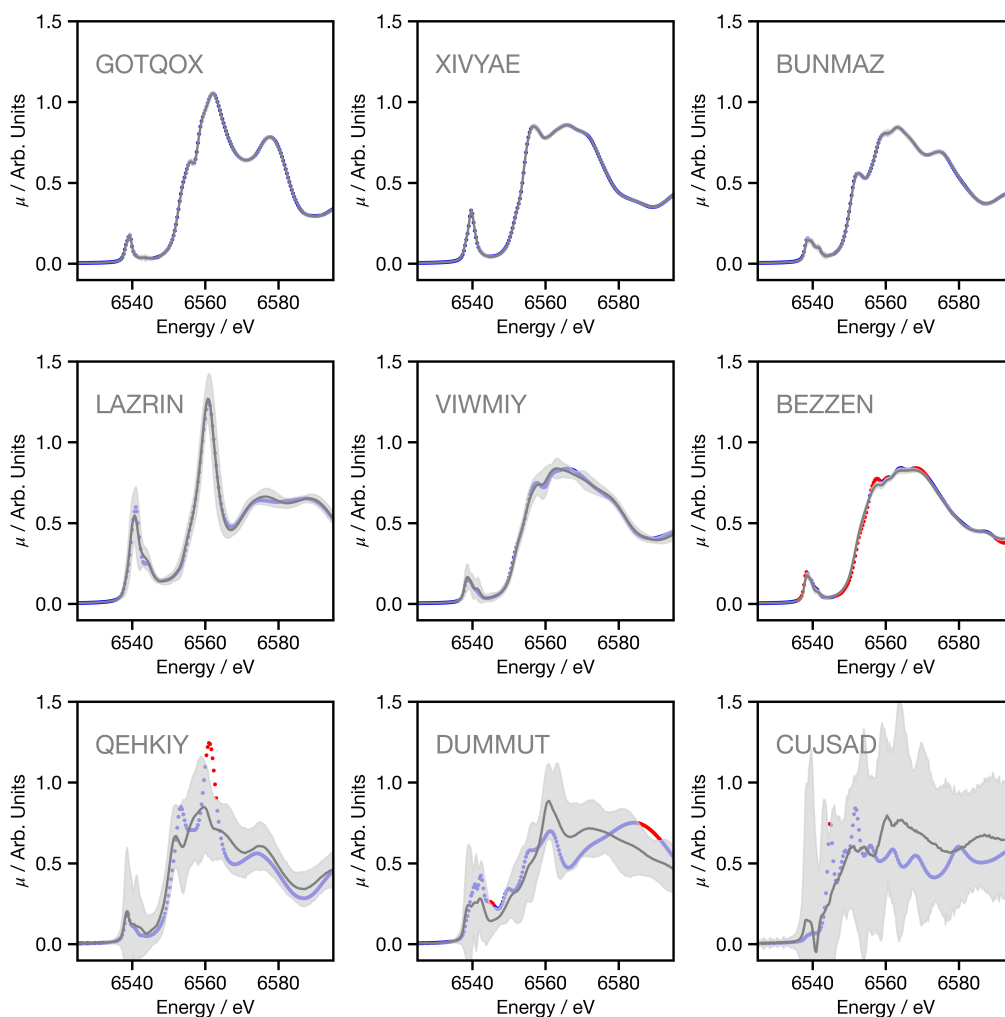
Figure S25: Example K-edge XANES spectra for Cu-containing samples. The upper three panels show K-edge XANES spectra from the 0th-10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th-55th percentiles, *i.e.* around the median. The lower three panels show K-edge XANES spectra from the 90th-100th percentiles, *i.e.* the lowest performance. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples.
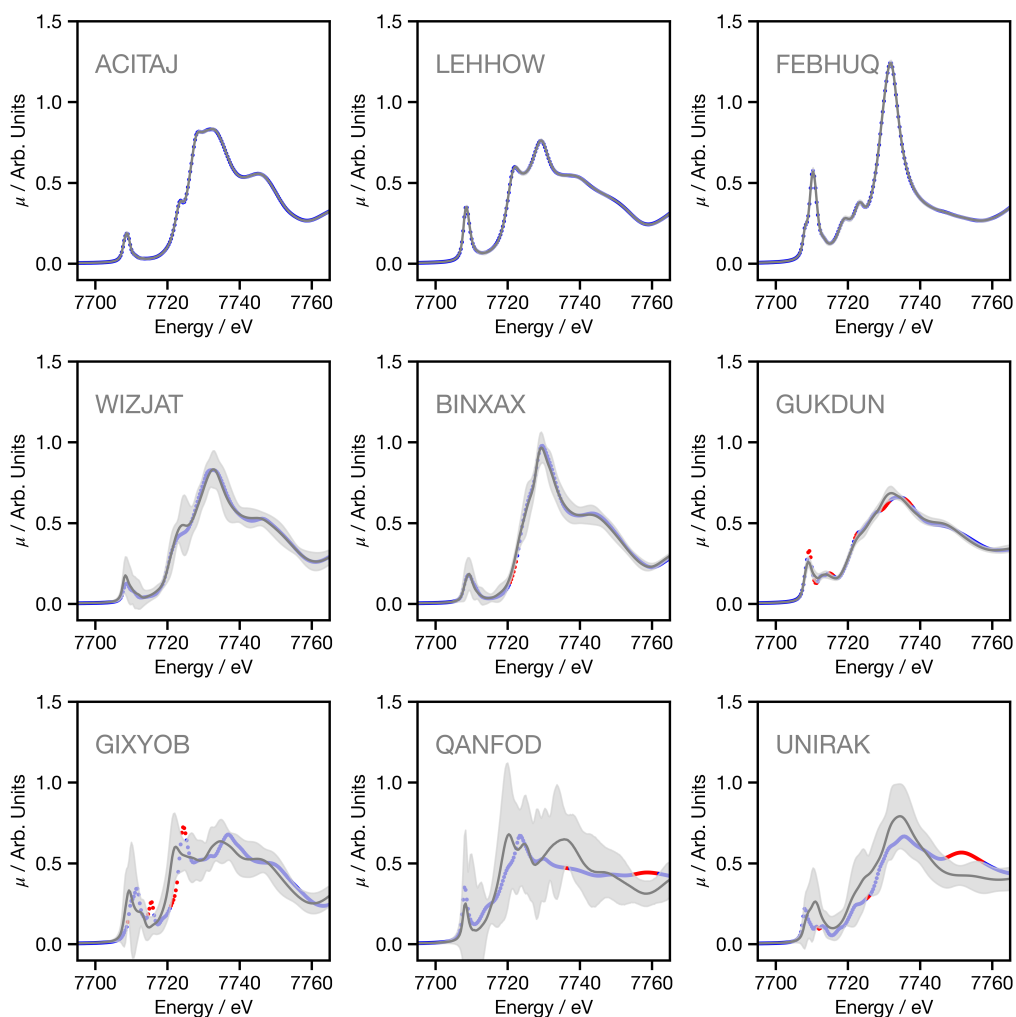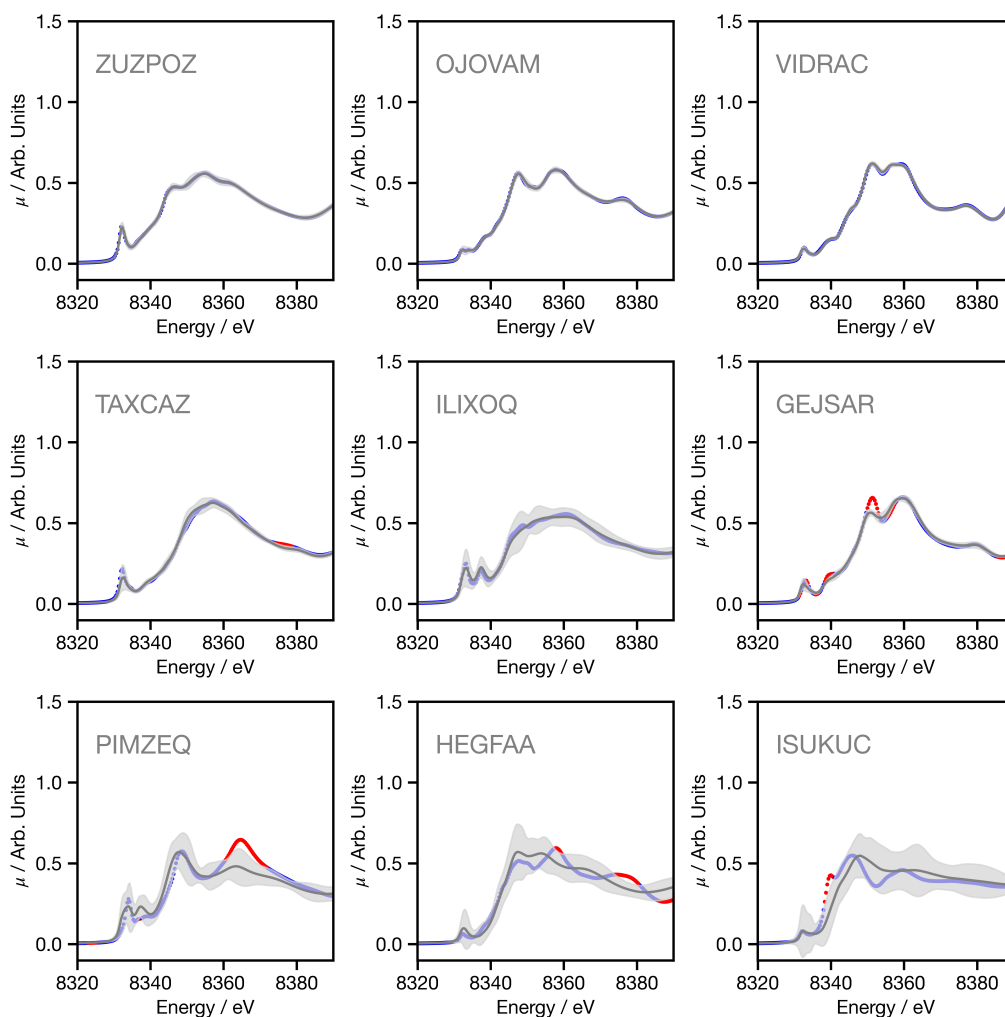
Figure S26: Example K-edge XANES spectra for Zn-containing samples. The upper three panels show K-edge XANES spectra from the 0th-10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th-55th percentiles, *i.e.* around the median. The lower three panels show K-edge XANES spectra from the 90th-100th percentiles, *i.e.* the lowest performance. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples.
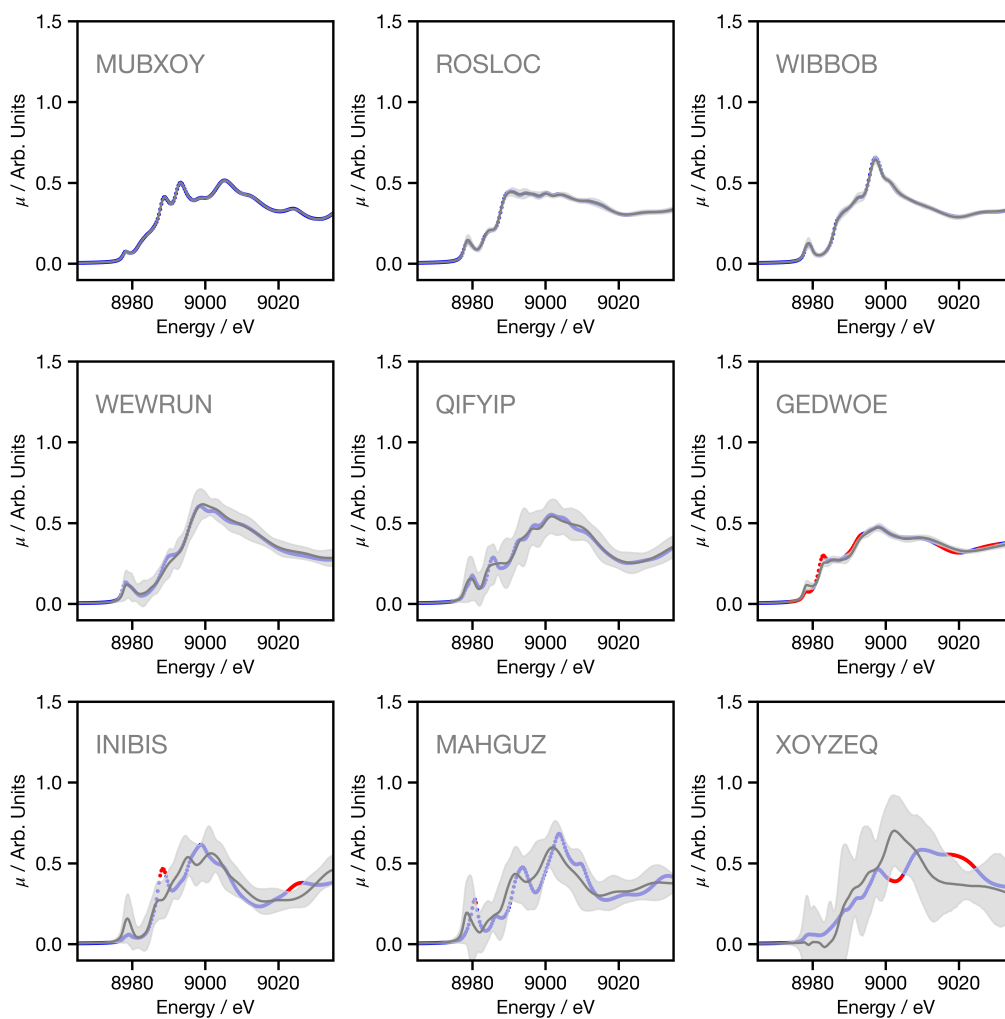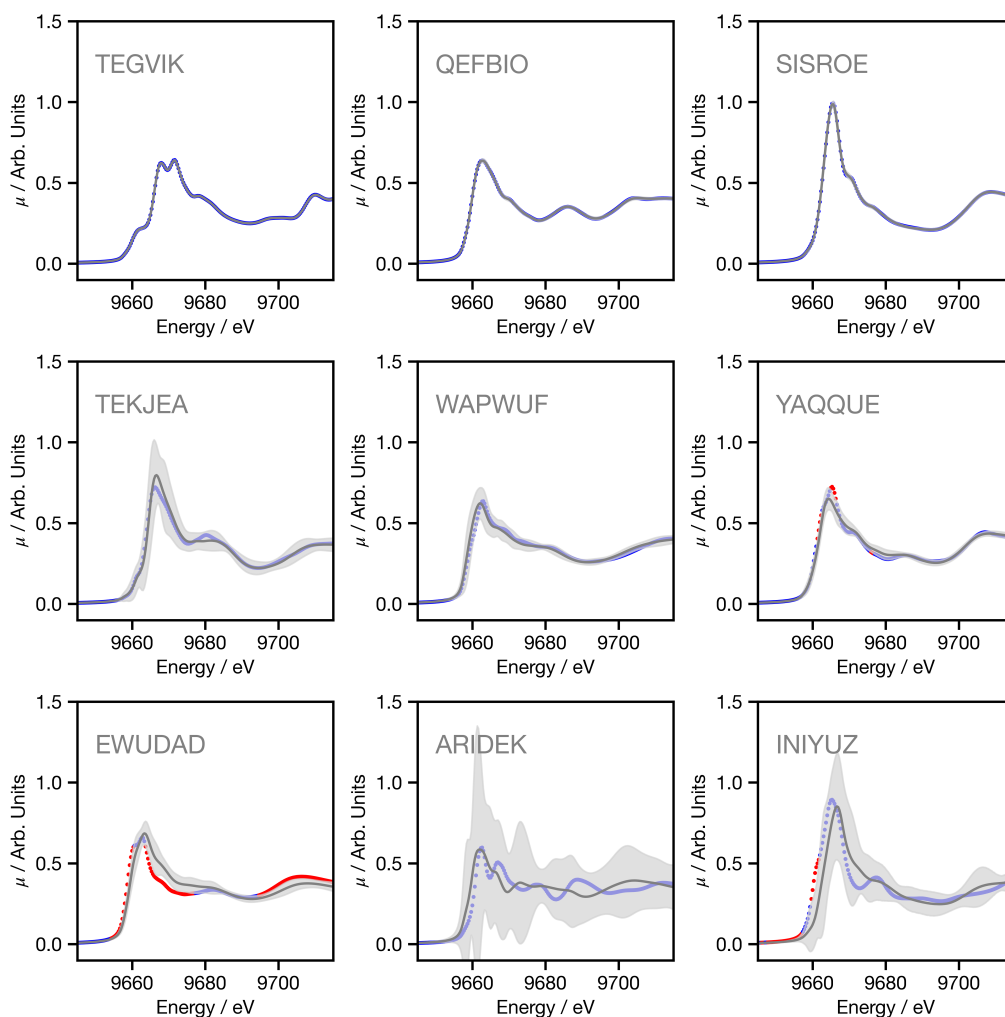
Figure S27: Example K-edge XANES spectra for Fe-containing samples which are out of sample and therefore significantly different from the training sets. These includes complexes with multiple metal sites and heavy elements, such as Ru and Tc not in the original training set. The six-character labels in the lower right of each panel are the Cam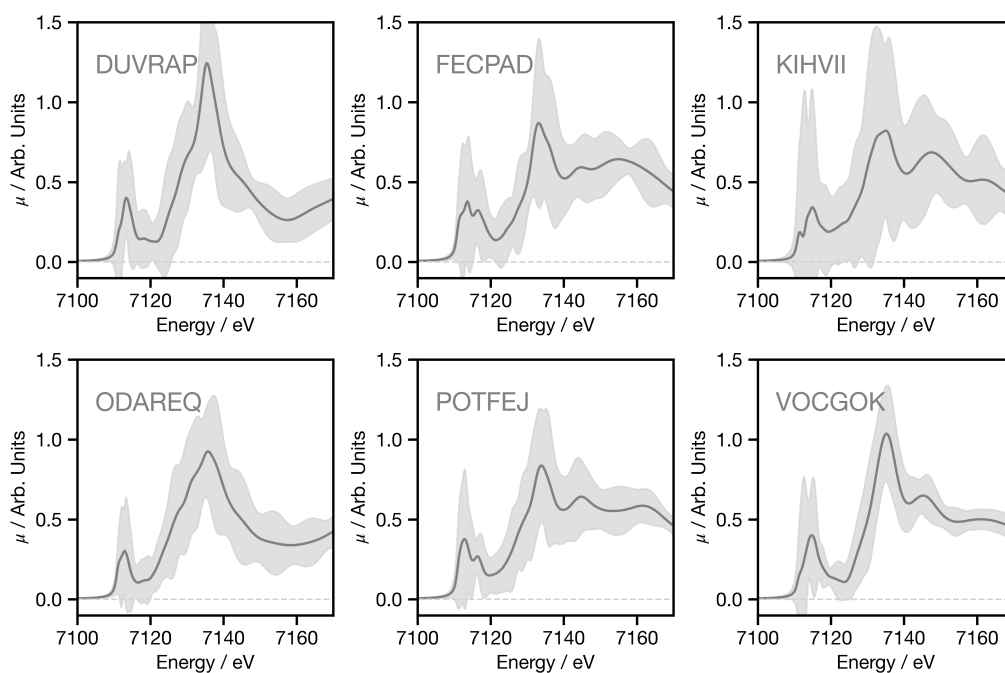bridge Structural Database (CSD) codes for the samples, the xyz files are uploaded with the supporting information. All spectra exhibit, as expected, a larger σ, indicating that the uncertainty quantification works for out of sample species.

References

[1] CD Rankine and TJ Penfold. Accurate, affordable, and generalizable machine learning simulations of transition metal x-ray absorption spectra using the xanesnet deep neural network. *The Journal of Chemical Physics*, 156(16):164102, 2022.

[2] Michael Gastegger, Ludwig Schwiedrzik, Marius Bittermann, Florian Berzsenyi, and Philipp Marquetand. wACSF - Weighted Atom-Centered Symmetry Functions as Descriptors in Machine Learning Potentials. *J. Chem. Phys.*, 148:241709, 2018.

[3] Diederik P. Kingma and Jimmy L. Ba. ADAM: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[5] Nikhil Ketkar, Jojo Moolayil, Nikhil Ketkar, and Jojo Moolayil. Introduction to pytorch. *Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch*, pages 27–91, 2021.

[6] Ask Hjorth Larsen, Jens Jorgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, BjØrk Hammer, Cory Hargus, et al. The Atomic Simulation Environment - A Python Library for Working With Atoms. *J. Phys.: Condens. Mat.*, 29:273002, 2017.

[7] XANESNET, 2023, gitlab.com/team-xnet/xanesnet.

[8] David Balcells and Bastian Bjerkem Skjelstad. tmqm dataset-quantum geometries and properties of 86k transition metal complexes. *Journal of chemical information and modeling*, 60(12):6135–6146, 2020.

[9] XANESNET Training Data, 2023, gitlab.com/team-xnet/training-sets.

[10] O Bunău and Yves Joly. Self-consistent aspects of x-ray absorption calculations. *Journal of Physics: Condensed Matter*, 21(34):345501, 2009.

[11] Kevin Tran, Willie Neiswanger, Junwoong Yoon, Qingyang Zhang, Eric Xing, and Zachary W Ulissi. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 1(2):025006, 2020.

[12] Andrew A Peterson, Rune Christensen, and Alireza Khorshidi. Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.*, 19:10978–10985, 2017.