ChemComm

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2023, **59**, 7100

Received 24th April 2023,
Accepted 16th May 2023

DOI: 10.1039/d3cc01988h

rsc.li/chemcomm

Uncertainty quantification of spectral predictions using deep neural networks†

Sneha Verma,[a] Nik Khadijah Nik Aznan,[b] Kathryn Garside[b] and Thomas J. Penfold *[a]

We investigate the performance of uncertainty quantification methods, namely deep ensembles and bootstrap resampling, for deep neural network (DNN) predictions of transition metal K-edge X-ray absorption near-edge structure (XANES) spectra. Bootstrap resampling combined with our multi-layer perceptron (MLP) model provides an accurate assessment of uncertainty with >90% of all predicted spectral intensities falling within $\pm 3\sigma$ of the true values for *held-out* data across the nine first-row transition metal K-edge XANES spectra.

Supervised/unsupervised machine-learning (ML) models that are able to learn patterns in big data have transformed many aspects of modern life, and their impact is beginning to be felt strongly in computational chemistry.[1] Performant ML models that are able to make accurate predictions of properties and observables instantaneously have already been leveraged to great effect in areas such as materials, catalyst, and drug design,[2] chemical reaction prediction,[3] and atomistic modeling.[4]

Accurate ML models make it possible to accelerate the prediction of a property of interest and the quality of these models is often only limited by the quality and quantity of the data used for training. Consequently, when a model is broadly applied it is likely that out-of-sample data will be encountered. In such cases the ability of these models to recognise and quantify the uncertainty associated with a specific prediction becomes crucial. Uncertainty within ML models can arise from inherent noise or be related to what a model does not yet know. These are typically referred to as *aleatoric* and *epistemic* uncertainty, respectively.[5] The former could arise from the absence of a single solution to the problem of finding a set of internal weights for a model and therefore reoptimising a model multiple times will generate a distribution of weights and consequently predictions. The latter, is associated with incomplete training data, *i.e.* where the training and testing data follow different distribution patterns.

Understanding and accurately assessing the uncertainty arising from both sources is a key piece of information required within the ML workflow if they are to become widely adopted by research communities, especially in supporting non-experts users to rapidly assess the reliability of their data. However, accurate uncertainty quantification also provides a strategy for targeted approaches for growing a given training set, based upon unsatisfactorily high uncertainties,[6] especially important when data is time-consuming or expensive to acquire. The importance and aforementioned benefits of obtaining an accurate quantification of uncertainty has led to a significant research effort in the field and the development of several techniques including; Monte Carlo dropout,[7] deep ensembles,[8] bootstrap resampling[9] and Bayesian neural networks.[10]

Recently progress in X-ray science driven by the emergence of facilities capable of delivering high-brilliance ultrashort pulses of X-rays[11] has been the driving force for significant progress in ML models capable of predicting X-ray spectroscopic observables from an input property or structure.[12,13] Very recently, works extending these models to quantify uncertainty have also been demonstrated for X-ray emission[14] and absorption at the C, N and O K-edges.[15] Consequently, in this article, we apply the deep ensembles and bootstrap resampling methods to our DNN used to predict the lineshape of first-row transition metal K-edge XANES spectra[16,17] and demonstrate its ability to assess pointwise uncertainty using only information about the local coordination geometry of the transition metal complexes. Details of the DNN and training sets are provided in the ESI.†

Fig. 1 shows a schematic of the bootstrap resampling method[9] used to assess uncertainty. Here, N machine learning models are optimised using N reference datasets ($\mathscr{D}_i$) sampled from the original reference dataset ($\mathscr{D}$). Each $\mathscr{D}_i$ has the same number of samples as $\mathscr{D}$ and consequently the random sampling used to generate $\mathscr{D}_i$ means that while each occurrence will

[a] *Chemistry – School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK. E-mail: tom.penfold@ncl.ac.uk*
[b] *Research Software Engineer Group, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3cc01988h

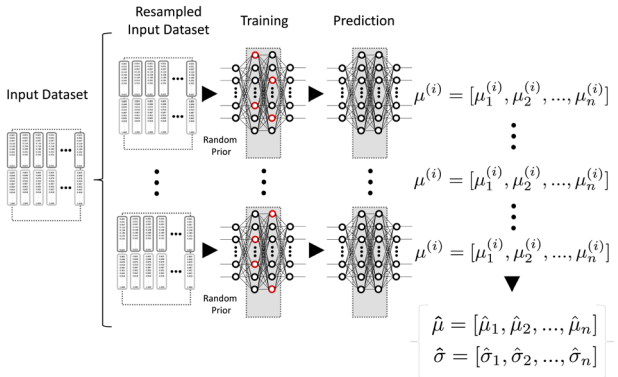

Fig. 1 Schematic of bootstrap resampling: *N* models are optimised using *N* reference datasets sampled with replacement from the original reference dataset; each one is the same size as the original reference dataset and, consequently, may contain repeated instances of the same sample. These *N* models are used to produce *N* independent predictions from which a mean and standard deviation can be derived. The red circles indicate neurons models in which dropout has been activated. The deep ensembling approach takes the same form, without resampling the training data.

only include a fraction of the instances from $\mathscr{D}$, they will contain repeats. The underlying assumption of this approach is that if $\mathscr{D}$ is a good approximation of the underlying distribution, each $\mathscr{D}_i$ will also be.[18] The models trained using the independent $\mathscr{D}_i$ training sets and a different random initialisation of weights for each occurrence are then used to produce *N* independent predictions of the "*held-out*" dataset from which a mean and standard deviation for each sample prediction can be computed. Throughout, the number of independent instance was set to 15 to achieve convergence (See Fig. S1, ESI†). The deep ensembles method uses the same approach, but in each case the models are trained using the same reference dataset, $\mathscr{D}$, variation in the models therefore arise only from the random initialisation of the weights as the models are trained. Consequently, deep ensembles will only capture *aleatoric* uncertainty associated with the natural variations in the model, while bootstrap resampling will also capture uncertainty associated with the structure of the dataset.

Fig. 2 shows histograms of the coverage, defined the percentage of calculated spectral data points which fall within $\pm 3\sigma$ of the average predicted spectra. This has been computed using the spectra predicted using the MLP model and bootstrap resampling approach for the *held-out* sets of each of the 9 first-row transition metal K-edge XANES spectra. The median coverage obtained from these histograms is tabulated, alongside the median percentage error, $\Delta\mu$ and corresponding interquartile range are shown in the Table 1. Across the nine first-row transition metal reference datasets, the median coverage is 100%, with the exception of V and Cr which are 98.7% and 97.3%, respectively. The average P_{90}, *i.e.* the probability that 90% of the data points for a *held-out* sample falls within $\pm 3\sigma$ of the prediction, is $\sim$93.6%, while the average P_{80} is $\sim$97.7%. The median $\Delta\mu$ is typically $<5\%$, with an average of 4.4%, in excellent agreement with ref. 16, indicating that the use to bootstrap resampling approach to assess uncertainty does not impact the overall performance of the network. As shown in Table S1 (ESI†), the MLP combined with deep ensembles method demonstrates a slightly improved overall performance, with an average median $\Delta\mu$ $\sim$4.1%. However, this small improvement is offset by a clear reduction in the median coverage to $\sim$98.4%. As described in the ESI†, we have also implemented bootstrap resampling or deep ensembles methods with a convolutional neural network (CNN). While comparable, the performance of the CNN is slightly worse than the MLP, although it is noted that the latters performance is achieved with a reduced number of internal weights, 114 000 *vs.* 414 208.

While the coverage demonstrates the performance of the uncertainty quantification methods, it can only be used as a metric if the calculated spectrum exists, which would negate the purpose for the use of DNN for future applications. Consequently, Fig. 3a shows parity plots of the mean-squared error (MSE) between the predicted, $\mu_{\text{predicted}}$, and calculated, $\mu_{\text{calculated}}$ spectra against σ for every energy point of the 250 *held-out* spectra at the Fe K-edge. Fig. 3b shows the corresponding plots with the MSE and σ calculated for each spectrum in the *held-out* set, rather than each energy point. The corresponding plots for the other elements are shown in Fig. S2–S9 (ESI†). The Pearson correlation (ρ) for all elements is shown in Table 1 and indicates a strong correlation with an average ρ = 0.82 and 0.84 observed across the 9 transition metal datasets for energy point and spectrum, respectively.

Fig. 3a indicates that this correlation exhibits two distinct regions of uncertainty as a function of energy. Firstly, the low energy region (blue) is associated with the pre-edge region of the spectrum and therefore exhibits a smaller MSE and smaller σ as the spectral intensity is weak. At higher energies ($>$7115 eV) the variation in the spectral intensity is larger and therefore so is σ. In agreement with the observations of Ghose *et al.*[15] Fig. 3a also shows, for a given σ, a dispersion of data points towards lower MSE. This is consistent with a model underconfidence, *i.e.* the predictions can be accurate (*i.e.* small MSE), but still produce a significant σ. This underconfidence is confirmed using the calibration curves shown in Fig. S10–S18 (ESI†) which predominantly show the data slightly above the diagonal calibration line.[19] However, it is stressed that during assessment of models, underconfidence is much preferred to overconfidence and overall this analysis shows σ can be used as an assessment of the error of a predicted spectrum. Fig. S10–S18 (ESI†) also show histograms of the σ calculated and used to assess the sharpness and coefficient of variation in Table 1 and Table S1 (ESI†). In all cases, the σ exhibits a narrow distribution, with an average σ (sharpness) substantially smaller than the spectral variations the model is attempting to predict.

Fig. 4 shows illustrative examples of Fe K-edge XANES spectra taken from around the median (45th–55th percentile), lower (0th–10th percentile) and upper (90th–100th percentile) when performance is ranked over all *held-out* DNN predictions by MSE. The corresponding plots for the other elements are shown in Fig. S9–S16 (ESI†). The dark grey line show the average predicted spectrum, the light grey region represents $\pm 3\sigma$ of the predicted spectrum and the dotted line is the

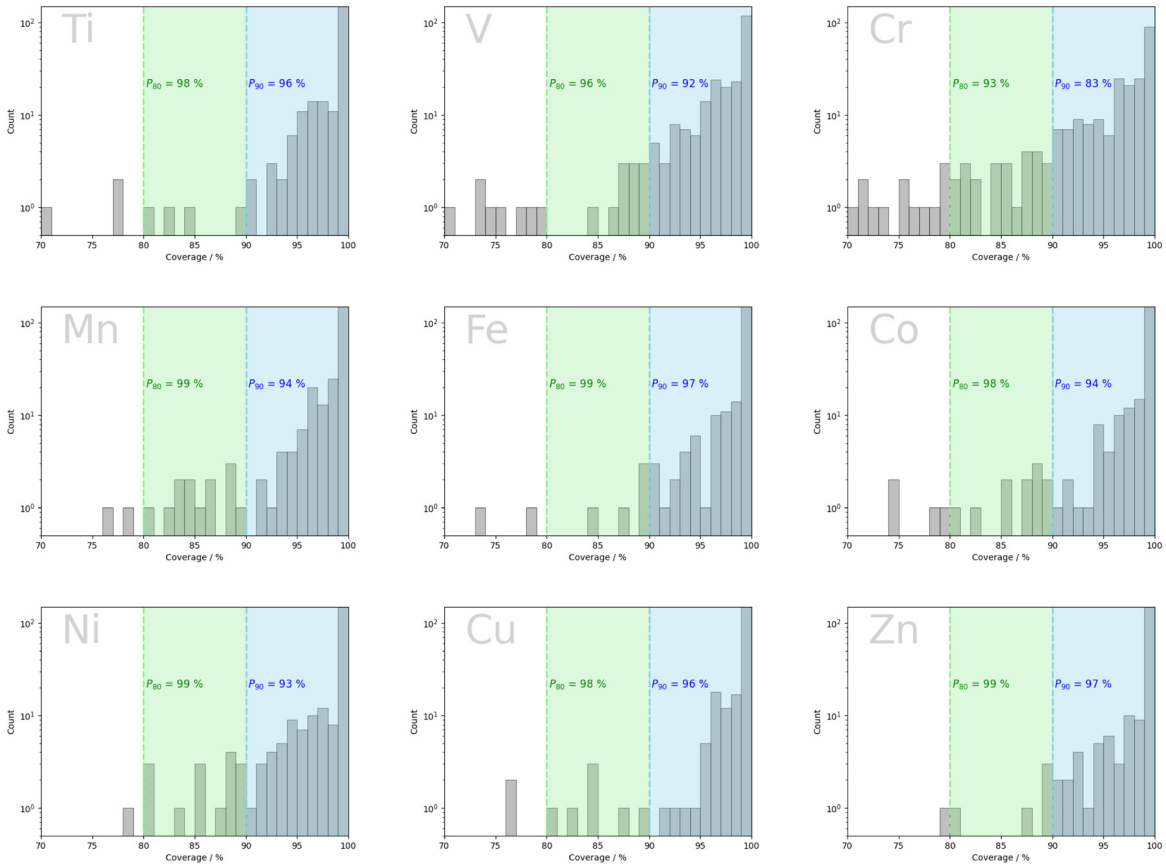

Fig. 2 Histograms of the coverage, defined the percentage of calculated spectral data points which fall within $\pm 3\sigma$ of average spectra, predicted using the multi-layer perceptron model and bootstrap resampling approach. Evaluated on nine *held-out* transition metal test datasets (Ti–Zn) containing 250 randomly selected samples.

Table 1 Summary of the median percentage errors, $\Delta\mu$ (%), interquartile range and coverage (Cover.). Pearson correlation between $\Delta\mu$ and σ for each energy point ($\rho^{\text{all}}_{\Delta\mu-\sigma}$) and each spectra ($\rho^{\text{spec}}_{\Delta\mu-\sigma}$). The sharpness (Sharp.) and coefficient of variation (C_v) for the uncertainty prediction are also shown and discussed in the ESI. Results obtained using the MLP and the bootstrap resampling method

Edge	$\Delta\mu$	IQR	Cover.	$\rho^{\text{all}}_{\Delta\mu-\sigma}$	$\rho^{\text{spec}}_{\Delta\mu-\sigma}$	Sharp.	C_v
Ti	5.3	5.0	100.0	0.85	0.72	0.04	0.85
V	4.6	7.5	98.7	0.79	0.93	0.02	1.03
Cr	3.9	5.3	97.3	0.79	0.86	0.02	1.12
Mn	4.3	4.7	100.0	0.85	0.91	0.02	1.06
Fe	5.0	5.0	100.0	0.83	0.77	0.03	0.88
Co	4.5	4.0	100.0	0.82	0.87	0.02	1.06
Ni	4.5	4.0	100.0	0.81	0.84	0.02	0.96
Cu	4.0	3.5	100.0	0.81	0.82	0.01	0.82
Zn	3.7	3.3	100.0	0.80	0.88	0.01	1.15

calculated spectrum. For the latter, the dots are blue if they fall within $\pm 3\sigma$ of the average predicted spectrum, and red if not. The first row, *i.e.* the best performers within the top 0th–10th percentile, exhibit accurate predictions with a small 3σ, consistent with the model confidence. In contrast, the third row, *i.e.* the worse predictions within the lowest (90th–100th) percentile, exhibit a significant increase in 3σ, expected given the larger MSE of the predictions. Finally, the second row which

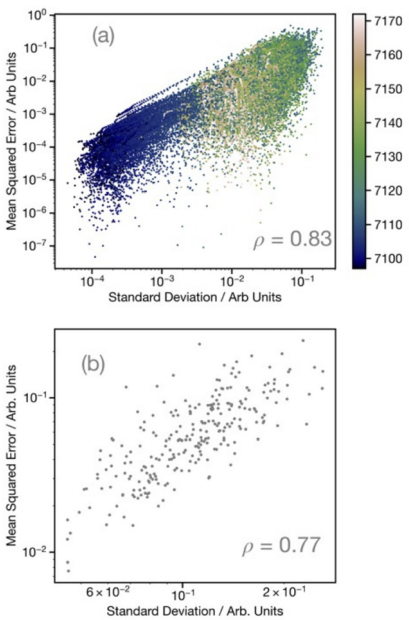

Fig. 3 Parity plots for every data point (a) and spectrum (b) of the mean-squared error (MSE) between the predicted, $\mu_{\text{predicted}}$, and target, $\mu_{\text{calculated}}$ K-edge XANES spectra against the standard deviation, σ, calculated using the Bootstrap resampling approach and the MLP network. The colours in (a) represent the energy over the full spectral range. Inset each plot are the Pearson correlation.

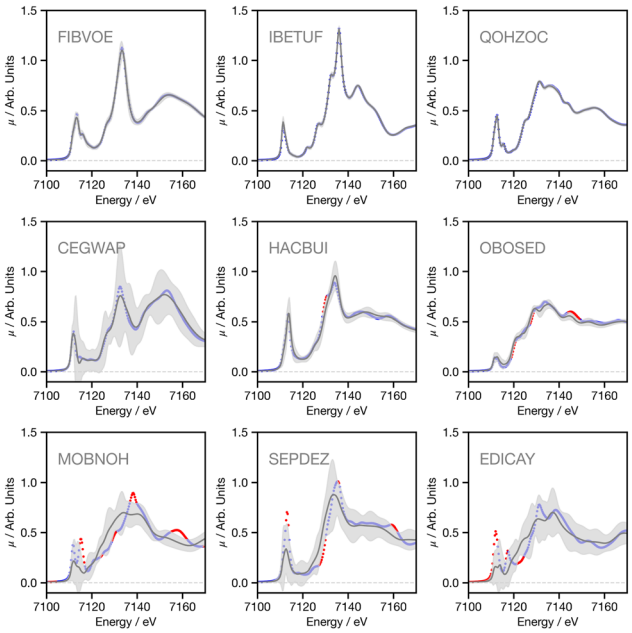

Fig. 4 Example K-edge XANES spectra for Fe-containing samples. The upper three panels show K-edge XANES spectra from the 0th–10th percentiles, *i.e.* the best performers when *held-out* set is ranked by MSE. The centre three panels show K-edge XANES spectra from the 45th–55th percentiles, *i.e.* around 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.

shows examples around the median illustrates a small MSE and for HACBUI and OBOSED a corresponding small 3σ. In contrast, despite being a good prediction, CEGWAP exhibits a large 3σ, consistent with the underconfidence discussed above. The source of this underconfidence is likely to be the presence of linear bonds, such as CO or CN. The structures can be obtained from ref. 20. In XANES spectra the scattering pathways along these linear bonds play a much more important role than similar structures containing non-linear bonds due to the focusing effect.[21] Indeed, such bonds feature in 60% of the samples exhibiting the largest σ from the held-out data, and consequently are clearly challenging for the model to predict. Finally, Fig. S27 (ESI†) shows the predicted spectra for 6 example complexes including multiple metal sites and heavy elements, such as Ru and Tc not in the original training set. As expected, all spectra exhibit, a large uncertainty, consistent with the model being unable to accurately predict samples which differ so much from the training data, highlighting that our uncertainty quantification approach remain valid for out of distribution samples.

In summary, we have implemented and assessed methods for uncertainty quantification of XANES spectra. We have demonstrated the ability of our DNN, which provide accurate predictions of the spectral lineshapes, to assess the uncertainty from each prediction, with $>90\%$ of all predicted data points falling within $\pm 3\sigma$ of the target calculated spectrum for *held-out* sets. Our results also demonstrate a strong correlation between MSE and σ, meaning that it can be used as a metric to accurately assess the uncertainty for a particular prediction. These results will not only allow non-experts to assess the reliability of the data, but it also provides a strategy for specifically targeted approaches for growing a training set, based upon unsatisfactorily high uncertainties.

This research made use of the Rocket High Performance Computing service at Newcastle University. T. J. P would like to thank the EPSRC for an Open Fellowship (EP/W008009/1) and research grant number EP/X035514/1. The authors acknowledge the Leverhulme Trust (Project RPG-2020-268).

Conflicts of interest

The authors have no conflicts of interest to declare.

References

1 M. Ceriotti, C. Clementi and O. Anatole von Lilienfeld, *Introduction: machine learning at the atomic scale*, 2021.
2 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha and T. Wu, *et al.*, *Nat. Mater.*, 2016, **15**, 1120–1127.
3 A. F. de Almeida, R. Moreira and T. Rodrigues, *Nat. Rev. Chem.*, 2019, **3**, 589–604.
4 P. O. Dral and M. Barbatti, *Nat. Rev. Chem.*, 2021, **5**, 388–405.
5 G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li and W. H. Green, *J. Chem. Inf. Model.*, 2020, **60**, 2697–2717.
6 L. I. Vazquez-Salazar, E. D. Boittier and M. Meuwly, *Chem. Sci.*, 2022, **13**, 13068–13084.
7 Y. Gal and Z. Ghahramani, International conference on machine learning, 2016, pp. 1050–1059.
8 R. Hu, Q. Huang, S. Chang, H. Wang and J. He, *Appl. Intell.*, 2019, **49**, 2942–2955.
9 A. A. Peterson, R. Christensen and A. Khorshidi, *Phys. Chem. Chem. Phys.*, 2017, **19**, 10978–10985.
10 Y. Kwon, J.-H. Won, B. J. Kim and M. C. Paik, *Comput. Stat. Data Anal.*, 2020, **142**, 106816.
11 K. Asakura, K. J. Gaffney, C. Milne and M. Yabashi, *Phys. Chem. Chem. Phys.*, 2020, **22**, 2612–2614.
12 C. D. Rankine and T. J. Penfold, *J. Phys. Chem. A*, 2021, **125**, 4276–4293.
13 C. Middleton, C. Rankine and T. J. Penfold, *Phys. Chem. Chem. Phys.*, 2023, **25**, 13325–13334.
14 T. Penfold and C. Rankine, *Mol. Phys.*, 2022, e2123406.
15 A. Ghose, M. Segal, F. Meng, Z. Liang, M. S. Hybertsen, X. Qu, E. Stavitski, S. Yoo, D. Lu and M. R. Carbone, *Phys. Rev. Res.*, 2023, **5**, 013180.
16 C. Rankine and T. Penfold, *J. Chem. Phys.*, 2022, **156**, 164102.
17 XANESNET, *2023*, ***gitlab.com/team-xnet/xanesnet***.
18 Y.-P. Li, K. Han, C. A. Grambow and W. H. Green, *J. Phys. Chem. A*, 2019, **123**, 2142–2152.
19 K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing and Z. W. Ulissi, *Mach. Learn. Sci. Technol.*, 2020, **1**, 025006.
20 *XANESNET Training Data*, *2023*, ***gitlab.com/team-xnet/training-sets***.
21 S. Zabinsky, J. Rehr, A. Ankudinov, R. Albers and M. Eller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1995, **52**, 2995.