White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/214042/

Version: Published Version

Article:
Penfold, Thomas J, Watson, Luke, Middleton, Clelia et al. (2024) Machine Learning Strategies for the Accurate and Efficient Analysis of X-ray Spectroscopy. Machine Learning: Science and Technology. 021001. ISSN: 2632-2153

https://doi.org/10.1088/2632-2153/ad5074

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Machine Learning: Science and Technology

TOPICAL REVIEW • OPEN ACCESS

Machine-learning strategies for the accurate and efficient analysis of x-ray spectroscopy

To cite this article: Thomas Penfold *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 021001

View the article online for updates and enhancements.

You may also like

- Deep learning methods for Hamiltonian parameter estimation and magnetic domain image generation in twisted van der Waals magnets
Woo Seok Lee, Taegeun Song and Kyoung-Min Kim
- Investigating the ability of PINNs to solve Burgers' PDE near finite-time blowup
Dibyakanti Kumar and Anirbit Mukherjee
- The twin peaks of learning neural networks
Elizaveta Demyanenko, Christoph Feinauer, Enrico M Malatesta et al.

This content was downloaded from IP address 144.32.225.234 on 27/06/2024 at 13:22

IOP Publishing Mach. Learn.: Sci. Technol. 5 (2024) 021001 https://doi.org/10.1088/2632-2153/ad5074

MACHINE LEARNING
Science and Technology

CrossMark

OPEN ACCESS

RECEIVED
24 November 2023

REVISED
17 April 2024

ACCEPTED FOR PUBLICATION
24 May 2024

PUBLISHED
7 June 2024

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

TOPICAL REVIEW

Machine-learning strategies for the accurate and efficient analysis of x-ray spectroscopy

Thomas Penfold[1,*], **Luke Watson**[1], **Clelia Middleton**[1], **Tudur David**[1], **Sneha Verma**[1], **Thomas Pope**[1], **Julia Kaczmarek**[1] and **Conor Rankine**[2]

[1] Chemistry, School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom
[2] Department of Chemistry, University of York, York YO10 5DD, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: tom.penfold@newcastle.ac.uk

Keywords: x-ray spectroscopy, deep neural networks, multiple scattering theory, uncertainty, interpretability

Abstract

Computational spectroscopy has emerged as a critical tool for researchers looking to achieve both qualitative and quantitative interpretations of experimental spectra. Over the past decade, increased interactions between experiment and theory have created a positive feedback loop that has stimulated developments in both domains. In particular, the increased accuracy of calculations has led to them becoming an indispensable tool for the analysis of spectroscopies across the electromagnetic spectrum. This progress is especially well demonstrated for short-wavelength techniques, e.g. core-hole (x-ray) spectroscopies, whose prevalence has increased following the advent of modern x-ray facilities including third-generation synchrotrons and x-ray free-electron lasers. While calculations based on well-established wavefunction or density-functional methods continue to dominate the greater part of spectral analyses in the literature, emerging developments in machine-learning algorithms are beginning to open up new opportunities to complement these traditional techniques with fast, accurate, and affordable '*black-box*' approaches. This Topical Review recounts recent progress in data-driven/machine-learning approaches for computational x-ray spectroscopy. We discuss the achievements and limitations of the presently-available approaches and review the potential that these techniques have to expand the scope and reach of computational and experimental x-ray spectroscopic studies.

1. Introduction

Spectroscopy is an indispensable and ubiquitous tool for the investigation of the electronic, magnetic, and geometric structures of molecules and materials. Rapid developments in instrumentation and experimental techniques, (including improvements in spatiotemporal resolution, in particular) [1–4] alongside the development of increasingly sophisticated analysis based upon detailed theory (i.e. computational spectroscopy) [5, 6] have had a marked impact on a broad range of research fields across the natural sciences and beyond.

Propelled by continuous improvements in hardware, software, and infrastructure, computational spectroscopy has become an indispensable tool for the modern spectroscopist that is capable of providing predictions—and, consequently, interpretations—of experimental spectroscopic observables across the electromagnetic spectrum. The predictive power of computational spectroscopy is perhaps best showcased within x-ray spectroscopy [7–10] where the transformative effects of next-generation light sources [11, 12] are pushing the limits of the technique, facilitating new insights into the structure and dynamics of molecules and materials as well as opening up new possibilities across a wide range of research fields [13–22]. The remarkable progress in x-ray spectroscopy continues to stimulate concomitant progress in theoretical techniques to ensure that data can be accurately and affordably analysed, setting up one of the most effective experiment-theory feedback loops [23].

© 2024 The Author(s). Published by IOP Publishing Ltd

The surging popularity of x-ray spectroscopy makes it crucial that a broad range of computational techniques are available to support the analysis of the experimental data recorded. Towards this goal, there has been rapid progress in first-principles computational chemical strategies based upon both wavefunction [24–30] and density-functional (DFT) [31, 32] methods. An increased understanding of the mechanisms responsible for the form of the experimental observables (e.g. the factors governing x-ray spectral lineshape) alongside the increased availability of data (e.g. from our ability to perform more numerous and more sophisticated computational calculations) has opened up new opportunities to develop data-driven/machine-learning approaches to complement the traditional techniques within computational spectroscopy [33]. Machine-learning models have the potential to rapidly and precisely predict properties and observables, often with very sparse external input. Consequently, they have begun to find extensive application across various fields, including materials, catalyst, and drug design [34–37], chemical reaction forecasting [38], and atomistic modelling [39–42].

In this Topical Review, we describe and illustrate recent progress in data-driven approaches for x-ray spectroscopy, outlining the present achievements and limitations as well as the scope for these techniques. We initially begin with a background to x-ray spectroscopy, describing the important aspects of the theory which any machine learning (ML) model will need to capture. This is followed by a review of the recent progress in all aspects of the ML models developed to date for x-ray spectroscopy, and an outline of opportunities and areas for future work. Our review focuses upon the potential of machine-learning for x-ray spectroscopy, but the core principles and challenges described herein are transferable to many other types of spectroscopy. In addition, it is hoped that this Article will provide a guide for researchers new to ML in their development of an understanding of the advantages and limitations of the methods available: to support this, example problems and datasets are made available at [43–45].

2. Background to x-ray spectroscopy

X-ray spectroscopy offers valuable insights into the composition, structural characteristics, and electronic properties of matter. The most widely-used techniques in this domain are x-ray photoelectron spectroscopy (XPS), x-ray absorption spectroscopy (XAS), and x-ray emission spectroscopy (XES); these are illustrated schematically in figure 1. While XAS and XES are bulk-sensitive techniques, XPS interrogates the electronic structure of a material at (or near to) the surface.

X-ray spectroscopy involves the measurement of the interaction of x-ray radiation with matter. The cross-section associated with this interaction generally diminishes with increasing energy but displays clear, discrete steps at specific energies—absorption edges—that correspond to the ionisation thresholds of the core electrons in different (low-lying) orbitals.

XPS measures the kinetic energy of (photo)electrons ejected subsequent to the interaction of a material with x-ray radiation at an energy greater than the ionisation threshold (i.e. with energy sufficient to liberate a (photo)electron; figure 1(a)). The (photo)electron carries the crucial information in the XPS experiment, and its short inelastic mean free path limits the sensitivity of XPS to the surface only; electrons located at greater depth in the material under study are unable to escape the bulk, even if they have been ionised on interaction with the x-ray radiation. The XPS experiment hence provides element-specific information about the chemical state, the electronic structure, and the density of electronic states in the material.

XAS measures the absorption of x-ray radiation—a process by which high-energy, core-hole-excited states are created and through which it is possible to probe the unoccupied electronic states of the material (figure 1(b)). On the lower-energy (pre-edge) front of the absorption edge, the XAS spectrum is shaped by the electronic structure of the unoccupied valence orbitals of the material under study and by the oxidation state of the absorbing atom. Resonances at slightly higher energies ($<$50 eV) in the x-ray absorption near-edge structure (XANES) region of the XAS spectrum contain information about the three-dimensional (geometric) structure around the absorbing atom(s). Resonances at even higher energies ($>$50 eV) above the absorption edge comprise the extended x-ray absorption fine structure (EXAFS) region of the XAS spectrum which, due to the shorter wavelength of the excited (photo)electrons, contains highly-local information about the coordination number(s) of the absorbing atom(s) and the coordination distances between this atom and its immediate (bonded) neighbours.

XES, by contrast, probes the occupied states of the material through measurement of x-ray radiation emitted when the core-hole state collapses and the core-hole is filled by electrons from the occupied states (figure 1(c)). XES spectra typically exhibit sensitivity to the charge and spin state(s) of the absorbing atom. In the case of Valence-to-core XES (VtC-XES) [46], through the information that the technique provides on the character of the highest-energy occupied (valence) electron orbitals involved in the Valence-to-Core-filling transition, the nature of the bonding between the absorbing atom and its coordinated neighbours is also unveiled.

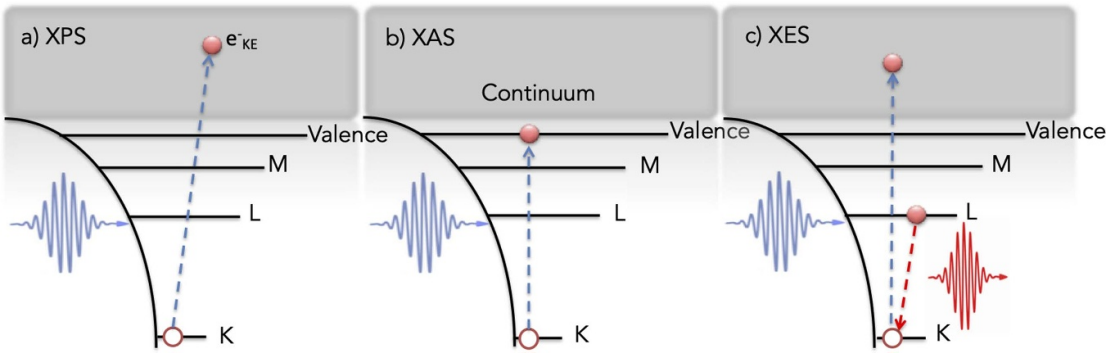

Figure 1. Schematic energy level diagrams of core-hole spectroscopies discussed herein: (a) x-ray photoelectron spectroscopy (XPS), (b) x-ray absorption spectroscopy (XAS), and (c) (non-resonant) x-ray emission spectroscopy (XES).

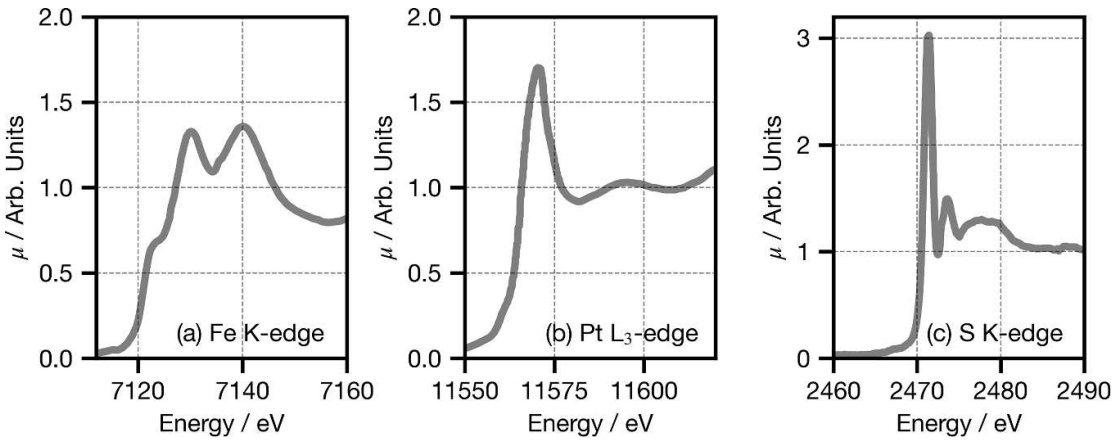

Figure 2. Examples of (a) Fe K-, (b) Pt L_3-, and (c) S K-edge XAS spectra. (a) is the Fe K-edge XAS spectrum of $[Fe(bpy)_3]^{2+}$; (b) is the Pt L_3-edge XAS spectrum of $[Pt_2(P_2O_5H_2)_4]^{4-}$; and (c) is the S K-edge XAS spectrum of C_4H_4S. The XAS spectra were obtained via digitisation from (a) [47], (b) [48] and (c) [49].

The precise information encoded in a given x-ray spectrum is dependent on the element and the absorption edge that the spectrum was measured for. This is illustrated in figure 2, which shows examples of Fe K-, Pt L_3- and S K-edge XAS spectra. The most dominant features in transition metal K-edge XAS spectra represent structural (e.g. geometric) information, with the strongest spectral features appearing at—or slightly above—the absorption edge (e.g. >7125 eV in figure 2(a). XAS spectral features corresponding to transitions from core orbitals into low-lying unoccupied valence states appear in the pre-edge of the XAS spectrum and correspond to dipole-forbidden ($3d \leftarrow 1s$) transitions; these transitions consequently provide limited insight into the electronic configuration of the absorbing atom because they typically manifest spectral features that are both broad and weak [e.g. the feature(s) <7120 eV in figure 2(a)]. In contrast, both the Pt L_3- (figure 2(b)) and S K-edge (figure 2(c)) XAS spectra show strong spectral features at the rising absorption edge; these spectral features correspond to dipole-allowed $5d \leftarrow 2p$ (at the Pt L_3-edge) and $3p \leftarrow 1s$ (at the S K-edge) transitions, and probe effectively the electronic structure of the unoccupied valence states. At energies above these electronic transitions, the yet-higher-energy XAS spectral features correspond to transitions into diffuse, delocalised continuum states above the ionisation threshold which—like the above-ionisation XAS spectral features in the Fe K-edge (figure 2(a)) XAS spectrum—also encapsulate structural information.

Simulating XPS requires the calculation of core electron binding energies. This can be carried out straightforwardly via, e.g. a two-step Δ-self-consistent-field (ΔSCF) [50–53] approach which, practically, requires calculating the energy of the electronic ground state and the energy of the core-hole-excited state so that the difference (the eponymous Δ) can be obtained. The accuracy of a ΔSCF calculation is determined consequently and principally by the description of core orbitals and their response after removal of an electron; this is influenced by factors such as, e.g. the inclusion of (scalar) relativisitic effects, and the choice of basis set. Simulating XES is—in principle, at least—less straightforward as it is a second-order spectroscopy (i.e. it measures the x-ray radiation emitted when an electron fills a core hole created by the

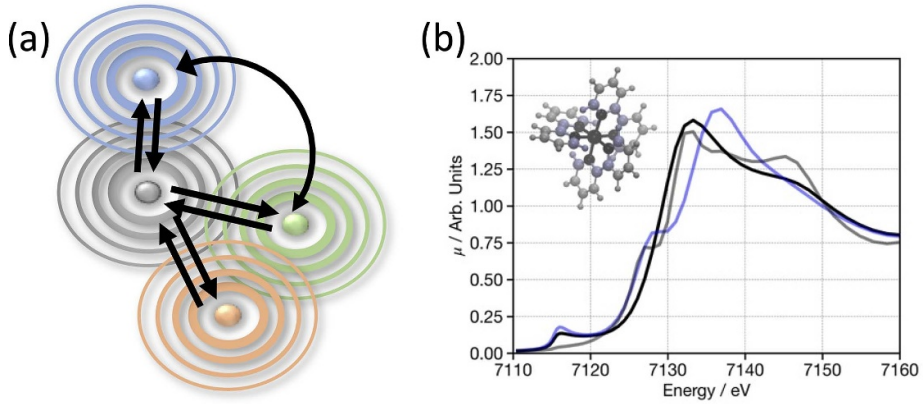

Figure 3. (a) Pictorial representation of single and multiple scattering pathways of an outgoing (photo)electron wave (grey) as it interacts with neighbouring atoms. (b) A '*shell-by-shell*' theoretical simulation of the Fe K-edge XAS spectrum of tris(bipyridine)iron(II), $[Fe(bpy)_3]^{2+}$ (inset), using a radius of 2 (black), 4 (blue), and 6 Å (grey) around the Fe absorption site.

excitation of a core electron into the continuum). An additional consideration over and beyond, e.g. the ΔSCF approach is that the electron orbitals of the intermediate core-hole state from which emission takes place will undergo relaxation relative to the initial (e.g. ground electronic) state from which absorption takes place as the former experience a greater nuclear charge. Although this influences the absolute energies of emission, the effect on the XES spectral lineshapes is not so great at all and can often be neglected to good approximation [54]. With this approximation put into practice, it is possible to simulate XES spectra using a one-electron approach which requires only the energy differences and transition strengths between electron orbitals to be calculated [55]. While effective in some cases, especially for VtC-XES, XES spectra calculated using a one-electron approach cannot model multi-electron phenomena, e.g. multiplet effects, that influence XES spectral lineshape [56–58]. Towards this objective, there has been a significant quantity of work aimed at developing semi-empirical [59–62] and first-principles [63–68] computational spectroscopic strategies for incorporating multiplet effects on XES (and XAS) spectral lineshapes.

The simulation of XAS presents a particular problem: that of treating accurately the final (electronically-excited core-hole) state. The diffuse, delocalised continuum states above the ionisation threshold are challenging to incorporate properly within computational simulations of XAS spectroscopy [69]. Simulation of higher-energy windows in the XAS spectrum, e.g. in the EXAFS domain, is typically carried out using the EXAFS equation [69] in which the scattering $\chi(k)$ is expressed as:

$$\chi(k) = \sum_{\gamma} \frac{N_\gamma S_0^2 F_\gamma(k,R)}{kR_\gamma^2} e^{-2R_\gamma/\lambda(k)} e^{-2\sigma^2 k^2} \sin(2kR_\gamma + \phi_\gamma). \quad (1)$$

where γ is the scattering path index with degeneracy N_γ; $F_\gamma(k,R)$ is the backscattering amplitude; R_γ is the '*half-path*' distance [i.e. half the length of the round-trip of the electron from the absorber to the neighbouring atom(s) and back]; σ^2 is the squared Debye-Waller factor; $\lambda(k)$ is the energy-dependent mean free path; and S_0^2 is an amplitude reduction factor which accounts for many-body effects. Usually, the first step towards obtaining a quantitative description of the structure is achieved using either a Fourier [70, 71] or wavelet [72–75] transform of the (experimentally-acquired) EXAFS signal, yielding a pseudo-radial distribution. The low computational cost of calculations using equation (1) (which can typically be completed in a matter of seconds) is such that accurate first-principles computational spectroscopic simulations and analyses of EXAFS data are common in the literature. Consequently, there has been little to no obvious motivation for developing ML models to simulate EXAFS spectra. However, there are plenty of examples of ML models having been applied to analyse automatically EXAFS spectra, [76, 77], assist EXAFS fitting, [78] and invert EXAFS spectra to obtain directly structural parameters of interest [79–82].

Simulation of the lower-energy windows in the XAS spectrum, e.g. close to the absorption edge in the XANES region, is commonly carried out under Multiple Scattering (MS) theory (represented pictorially in figure 3(a)). Under MS theory, Fermi's Golden Rule is re-expressed using Green's functions [83–85]:

$$\mu(E) \sim -\frac{1}{\pi} \Im \sum_i |\langle i | \epsilon \cdot \mathbf{r} G(\mathbf{r}, \mathbf{r}'; E) \epsilon \cdot \mathbf{r}' | i \rangle| \quad (2)$$

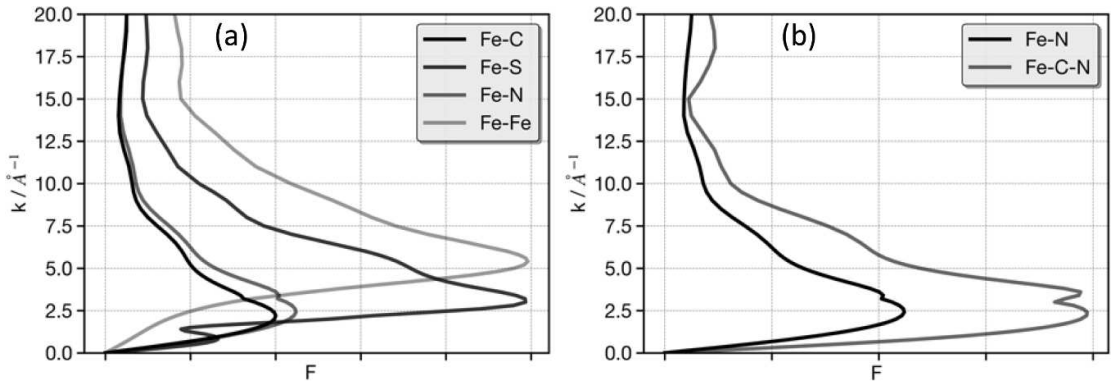

Figure 4. Backscattering amplitude, F, as a function of the momentum of the (photo)electron, k, for single (a) and multiple (b) scattering pathways.

where $G(\mathbf{r}, \mathbf{r}', E)$ is the energy-dependent Green's function propagator with amplitude moving from $\mathbf{r}$ to $\mathbf{r}'$. This approach is computationally efficient as it condenses the sum over the final states into a Green's function propagator which is expressed effectively as a MS path expansion [86]:

$$G = G^0 + G^0 t G^0 + G^0 t G^0 t G^0 + \cdots \tag{3}$$

G^0 describes the propagation of the (photo)electron wave between two atomic sites, and t describes how the wave scatters from a neighbouring atom. Consequently, the first term in equation (3) (G^0) accounts for the atomic-like background (i.e. the XAS spectrum of the isolated atom) while the subsequent terms ($G^0 t G^0$, $G^0 t G^0 t G^0$, *etc*) account for the fine structure of the oscillations in the XAS spectra that arise from the interaction of the (photo)electron wave with neighbouring atoms. Each term is expanded to an increasing order, i.e. the term ($G^0 t G^0$) describes all single-scattering events (i.e. scattering events involving a single neighbouring atom), the term ($G^0 t G^0 t G^0$) includes MS processes involving two neighbouring atoms, and so on.

MS theory is applicable both above and below the ionisation threshold, although the theory is formulated in terms of electron scattering, since the scattering order of the expansion simply reflects how much the final state deviates from that of an isolated atom [69]. In addition, the representation of continuum states in terms of MS pathways facilitates the intuitive interpretation of spectral features using a '*shell-by-shell*' analysis [87, 88]. In such an analysis, a series of theoretical simulations are carried out in which the cutoff radius around the absorption site (and, by extension, the number of neighbouring atoms taken into account) is successively expanded (an example is given in figure 3(b). Beyond providing an insight into the origin of spectral features in an XAS spectrum, the '*shell-by-shell*' analysis also provides a potential approach to assess the performance of, and feature importance in, an ML model; this is discussed in greater detail later in this Topical Review.

Finally, as shown in equation (3), it is important to understand how the photoelectron scatters from a particular atom, which can be characterised by the backscattering amplitude. The outgoing photoelectron wave is scattered principally by the bound electrons of the neighbouring atoms and the scattering is consequently enhanced under resonant conditions (i.e. where the electron orbital energy is equal to the energy of the photoelectron). This makes backscattering amplitude an element-specific quantity that is proportional to both the momentum of the photoelectron, k, and the distance between the two atoms, r_{IJ}. The former relationship is illustrated in figure 4(a) which shows the backscattering amplitude, F, as a function of the momentum of the photoelectron, k, for four pairs of atoms: Fe–Fe, Fe–S, Fe–N, and Fe–C. Figure 4(a) shows three distinct maxima in F as a function of k for Fe–N and Fe–C that are attributable to scattering from electrons in the 2p (*ca.* 3 Å^{-1}), 2 s (*ca.* 4 Å^{-1}), and 1 s (*ca.* 11 Å^{-1}) electron orbitals. For Fe–S, a larger backscattered amplitude is observed with additional features which are attributable to scattering from electrons in the 3 s (*ca.* 5.5 Å^{-1}) and 3p (*ca.* 3.5 Å^{-1}) electron orbitals observed. Fe–Fe presents an additional feature corresponding to scattering from electrons in the 3d (*ca.* 5 Å^{-1}) electron orbital. Typically, the importance of a scattering pathway decreases with its increasing order (where order corresponds to the number of scattering events). However, as shown in figure 4(b): an enhanced backscattering amplitude is observed at second-order for a linear bond, e.g. Fe–C–N geometry [89]. In such a scenario, the lineshape of F as a function of k remains consistent, yet the magnitude of F is increased due to the focusing effect [90].

3. Representing x-ray absorption sites and spectra in ML models

A key element in the development of any high-performing ML model is the implementation of an optimal representation of the data—that is, one which is at once compact, pertinent and comprehensive. Indeed, the choice of representation is often critical in enabling the model to develop effectual and cogent interpretations of the relationship between input and output data. In this section we discuss, and supply examples to illustrate the importance of, structural and spectral representations used in ML for x-ray spectroscopy.

3.1. Representing x-ray absorption sites (featurisation)

An ML model that operates on atomic structures must map each atomistic system, i.e. the atomic identities and their Cartesian (x, y, z) coordinates, onto some sort of suitable (typically lower-dimensional) representation, or '*feature vector*', through featurisation [91]. A (supervised) ML model might then learn the mapping between the feature vector(s) and the target property (e.g. a structure → spectrum mapping). X-ray spectroscopy is a local spectroscopic technique in that it is sensitive to the local atomic environment around the absorption site, and so an ML model should carry out featurisation subject to the constraints that the feature vector is:

- local, such that it does not encode the entire molecular structure—rather, it encodes the immediate molecular structure at an arbitrary point up to a maximum (radial) cutoff distance, usually up to *ca.* 6 AA;
- invariant to transformations that do not alter the target property, e.g. translations and rotations of the three-dimensional structure within Cartesian coordinate space, or permutations of the atomic indexing scheme;
- unique, such that it should vary when the target property varies;
- general, such that it can be applied to any atomistic system;
- efficient, such that it should not take a long time to construct or parse programmatically.

There exist several representations for which these criteria are (at least largely) fulfilled [91]. The criteria of locality, invariance, and efficiency are the least challenging to fulfil; generality (across the periodic table) is less frequently fulfilled and often trades off against efficiency.

3.1.1. RDC

The radial distribution curve [RDC; also known as the pair distribution function (PDF)] is a simple local descriptor that encodes the space around an x-ray absorption site via dimensionality reduction of the three-dimensional space to a histogram of atomic densities, f_{RDC}, as a function of the radial distance, r. f_{RDC} is defined as:

$$f_{\mathrm{RDC}} = \sum_{i}^{n}\sum_{j>i}^{n} Z_i Z_j \exp^{-\alpha\left(r_{ij}-r\right)^2} \tag{4}$$

where Z is nuclear charge and r_{ij} is the Euclidean distance between atoms i and j. f_{RDC} is defined over an auxiliary real-space grid, r, and smoothed using Gaussian-type functions with full-width half-maxima (FWHM) moderated by the parameter α. The RDC fulfils the criteria of locality; invariance with respect to atomic indexing, translation, and rotation; generality; and is computationally efficient to construct and parse programmatically. It is also straightforward to extend the canonical RDC so as to construct a property-weighted RDC by changing Z for an alternative atomic property, e.g. the electron affinity, [93, 94] among other possibilities. A limitation of the RDC is that it only contains two-body terms (i.e. between the x-ray absorption site and all atoms) which are insufficient alone to characterise completely the three-dimensional molecular geometry, and—consequently—it does not fulfil the criterion of uniqueness. The RDC is compared against alternative featurisation approaches in table 1, and it displays comparatively poor performance.

3.1.2. wACSF

Higher-order terms that are not included in the RDC but that are nonetheless necessary to characterise completely the three-dimensional molecular geometry (e.g. those that describe three- and four-body relationships) can be incorporated into a feature vector of weighted atom-centered symmetry functions (wACSF). wACSF are an extension of the ACSF of Behler [95, 96] and are designed to lend the ACSF descriptor to molecular systems that contain any arbitrary number of different types of atoms. Indeed, the limitation of the canonical ACSF descriptor (as with the SOAP and LMBTR descriptors) is that the feature vector is stratified by atom type (i.e. chemical element). While this guarentees that the ACSF feature vector fulfils the criterion of invariance with respect to permutation of the atomic indices, it also means that the size

Table 1. Performance at the Fe K-edge using the structure → spectrum XANESNET MLP network and different approaches to featurisation. Performance was assessed according to the median percentage error between predicted and target XAS spectra for 250 held-out structure/spectrum pairs. The interquartile range associated with the percentage error is given in brackets. The held-out structure-spectrum pairs are the same as those used in [92] and were selected via random partitioning of the full dataset. XAS spectra were represented on a discretised energy grid. All input files, XANESNET MLP network details, and associated datasets are publicly available at [45].

Structural rep.	Input length	Network weights	Performance/%
RDC	121	441 058	14.8 (8.4)
wACSF	97	428 770	4.4 (3.6)
^{2}MSR	38	398 562	4.7 (3.9)
^{2,3}MSR	76	418 018	3.9 (3.0)
^{2,3,4}MSR	114	437 474	3.9 (3.0)
^{2,3,4,5}MSR	152	463 918	3.8 (3.1)
^{2,3}AR-MSR	722	748 770	3.7 (3.0)

of the canonical ACSF descriptor grows commensurately with the number of different atom types that the descriptor encodes. This can make it challenging to apply the canonical ACSF descriptor to datasets containing many different atom types, ultimately limiting the generality. A wACSF feature vector for an atom, i, can be constructed by concatenating a global (G^1), N radial (G^2; two-body), and M angular (G^4; three-body) terms, which have the functional forms:

$$G_i^1 = \sum_{j \neq i} f_c\left(r_{ij}\right) \tag{5}$$

$$G_i^2 = \sum_{j \neq i} Z_j \cdot f_c\left(r_{ij}\right) \cdot \exp^{-\eta\left(r_{ij}-\mu\right)^2} \tag{6}$$

$$G_i^4 = 2^{1-\zeta} \sum_{j \neq i} \sum_{k \neq i,j} Z_j Z_k \cdot \left(1 + \lambda \cos\left(\theta_{jik}\right)\right)^{\zeta} \cdot f_c\left(r_{ij}\right) \cdot f_c\left(r_{ik}\right) \cdot f_c\left(r_{jk}\right) \cdot \exp^{-\eta\left(r_{ij}-\mu\right)^2} \cdot \exp^{-\eta\left(r_{ik}-\mu\right)^2} \cdot \exp^{-\eta\left(r_{jk}-\mu\right)^2} \tag{7}$$

where i, j, and k are atomic sites, Z_i is the nuclear charge of atom i, r_{ij} is the distance between atoms i and j, and θ_{jik} is the angle between atoms j, i, and k. f_c is a radial cutoff function ensuring the functions go to zero where $r_{ij} \geqslant r_c$. r_c is usually chosen to be approximately 6.0 Å. In practice the global G^1 wACSF is often omitted from the feature vector, and the feature vector is not limited to wACSFs encoding up to three-body terms; wACSFs encoding higher-order relationships can be constructed (see, for example, the definitions given by Behler in [96]).

In comparison to the simple RDC descriptor (for which the only parameter is σ; equation (4)), wACSFs have a number of parameters (η, μ, λ, and ζ) which have to be determined empirically [92]. A number of automated parameter tuning strategies are effective at determining η, μ, λ, and ζ, however [e.g. intelligent-sampling/Bayesian approaches, decomposition (principal component analysis (PCA)), [97] and genetic algorithm [98] optimisation], and—in practice—tuning does not present an obstacle to the application of wACSF. A feature vector of wACSFs displays significant improvement in performance over an RDC representation (14.8% *vs.* 4.4% see table 1) without any substantial increase in the computational overhead associated with translating the molecular structure into the descriptor. The use of wACSF also tends to confer the advantage of improved compactness in the feature vector, reducing the propensity of the ML model for overfitting.

3.1.3. Multiple scattering representation

The theoretical treatment of XS under the MS framework is often based on a path expansion to increasing order (section 2; equation (3)). Two additional approaches to featurisation, inspired by MS theory, are the multiple-scattering representation (MSR) and the angle-resolved MSR (AR-MSR); both representations are available in the XANESNET code [43, 45]. The MSR and AR-MSR representations retain the two-body terms from the wACSF representation (with which they share their functional form; equation (6)), although the MSR representation implements an alternative three-body term:

$$S_i^3 = \sum_{j \neq i} \sum_{k \neq i,j} Z_j Z_k \cdot \left|\cos\left(\theta_{jkl}\right)\right| \cdot f_c\left(r_{ijk}\right) \cdot \exp^{-\eta\left(r_{ijk}-\mu\right)^2} \tag{8}$$

and adds a four-body term:

$$S_i^4 = \sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} Z_j Z_k Z_l \cdot |\cos\left(\theta_{ijk}\right)| \cdot |\cos\left(\theta_{jkl}\right)| \cdot f_c\left(r_{ijkl}\right) \cdot \exp^{-\eta\left(r_{ijkl}-\mu\right)^2}. \tag{9}$$

The MSR representation can be extended to an arbitrarily higher order (*cf* the wACSF representation) albeit with an increase in the computational cost associated with constructing the descriptor. A limitation of the MSR representation is that, for these higher orders, a large number of scattering pathways of similar length will be present, as expected from first principles calculations. When these pathways are represented on a single auxiliary radial grid, a significant overlap of terms may arise, resulting in a loss of information and breaking the representational uniqueness. To overcome this limitation, the AR-MSR representation uses both a radial grid and an auxiliary angular grid. The three- and four-body terms in the AR-MSR representation are given as:

$$S_{i,\alpha}^3 = \sum_{j\neq i}\sum_{k\neq i,j} w_j Z_j w_k Z_k \cdot f_c\left(r_{ijkl}\right) \cdot \exp^{-\eta\left(r_{ijk}-\mu\right)^2} \cdot \exp^{-\sigma\left(\cos\left(\theta_{ijk}\right)-\phi\right)^2} \tag{10}$$

$$S_{i,\alpha,\beta}^4 = \sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} w_j Z_j w_k Z_k w_l Z_l \cdot f_c\left(r_{ijkl}\right) \cdot \exp^{-\eta\left(r_{ijkl}-\mu\right)^2} \cdot \exp^{-\sigma\left(\cos\left(\theta_{ijk}\right)-\phi\right)^2} \cdot \exp^{-\sigma\left(\cos\left(\theta_{jkl}\right)-\phi\right)^2}. \tag{11}$$

The AR-MSR representation encodes more information, although at the cost of a longer, less compact feature vector. Assuming a radial grid of 38 points and an angular grid of 18 points, the $S_{\alpha,\beta}^3$ component of the AR-MSR feature vector would have a dimension of 722 and the $S_{\alpha,\beta,\gamma}^4$ component of the AR-MSR feature vector would have a dimension of 10 368; practically, this would necessitate truncation of the order of expansion to be tractable.

Table 1 shows the performance of these structural featurisations at the Fe K-edge. For the MS representation, the inclusion of four- and five-body terms does not improve the performance of the network which, as discussed above, is due to the many overlapping pathways exhibiting the same lengths leading to a loss of information on a single auxiliary radial grid. Upon including angular resolution there is a very small improvement, although this is likely to arise simply from the larger input vector length leading to more free parameters within the network due to the increased resolution of the feature vector. These results suggest that improvement in the performance of the network using these local atom-centred symmetric descriptors is likely to be achieved only by using a larger training set or adopting alternative and/or extended representations.

3.1.4. SOAP

The RDC, wACSF, (AS-)MSR descriptors all leverage an N-body interaction weighting term based on the atomic number, Z, of the atom types involved in the interaction(s) encoded in the feature vector. An alternative option is for the stratification of the feature vector according to atom type (*cf* the ACSF descriptor)—this, in principle, enables the retention of more information. The smooth overlap of atomic positions (SOAP) [99, 100] descriptor encodes the local environment around an x-ray absorption site using an expansion of the Gaussian-smeared atomic density based on spherical harmonics and radial basis functions. The local environment around an x-ray absorption site, i, is characterised by atomic neighbourhood density:

$$\rho_i(r) = \sum_j \exp^{-|\mathbf{r}-\mathbf{r}_{ij}|^2/2\sigma_{\text{atom}}^2} f_{\text{cut}}\left(\mathbf{r}_{ij}\right) \tag{12}$$

$$= \sum_{n=0}^{n_{\max}}\sum_{l=0}^{l_{\max}}\sum_{m=-l}^{m=l} c_{nlm}^i g_n(\mathbf{r})\, Y_{lm}(\mathbf{r}) \tag{13}$$

where $\mathbf{r}_{ij}$ are the vectors pointing to the neighbouring atoms; σ_{atom} is a parameter corresponding to the size of the atoms, and f_{cut} is the cutoff function. The expansion on the second line uses spherical harmonics and a set of orthonormal radial basis functions, g_n, limited by the number of radial and angular basis functions defined using $n_{\max}$ and $l_{\max}$. Accumulating the expansion coefficients, a power spectrum can be defined as:

$$p_i^{ss'nn'l} = \pi \sqrt{\frac{8}{2l+1}} \sum_m \left(c_{snlm}\right)^* c_{s'n'lm}. \tag{14}$$

The full feature vector is constructed by concatenating the elements, $p_i^{ss;nn'l}$, for all unique pairs of atoms, all unique pairs of radial basis functions, n, n' up to $n_{\max}$, and the angular degree values l up to $l_{\max}$.

SOAP has seen much success, however those who wish to apply it must be aware of a potential drawback of this descriptor in that descriptor length scales drastically with the number of species due to be described. The descriptor length of SOAP is expressed [101]:

$$\mathcal{L} = \frac{1}{2} n_{\max} S_n \left(n_{\max} S_n + 1\right) \left(l_{\max} + 1\right) \tag{15}$$

where S_n is the number of atomic species, $n_{\max}$ is the number of radial basis functions and $l_{\max}$ is the number of angular basis functions. We can see that the size of the SOAP feature vector scales quadratically with the number of elements. This issue can be somewhat mitigated by exploiting sparsity, *i.e.* the SOAP feature vector is sparse with respect to elements, so even for S_{total} elements across a given dataset, only those present in a given input need to be considered when computing an individual descriptor. This will not only reduces the space required to store representations, but also reduces the number of model parameters. Other effective compression strategies do exist [102] , although they have yet to be investigated in the context of in x-ray spectroscopy.

3.1.5. MBTR

The many-body tensor representation (MBTR) [103] combines the '*bag-of-bonds*' [104] and Coulomb matrix [105] representations to overcome their shortcomings (including their non-uniqueness, discontinuity, and limited generality). The MBTR descriptor is usually expressed in terms containing atomic numbers (k_1), (inverse) distances between atoms (k_2), and the cosine of angles between atoms (k_3). For x-ray spectroscopy, the feature vector only needs to be calculated for atom combinations including the central atom, this corresponds to the Local MBTR (LMBTR) representation. In this case, k_1 is not used, so the two and three body terms are expressed:

$$\text{MBTR}_i^2 = \sum_{j \neq i} w_1^{i,j} \frac{1}{\sigma_1 \sqrt{2\pi}} \exp^{-\frac{\left(x - k_2\left(R_i, R_j\right)\right)^2}{2\sigma_2^2}} \tag{16}$$

$$\text{MBTR}_i^3 = \sum_{j \neq i} \sum_{k \neq i,j} w_1^{i,j,k} \frac{1}{\sigma_1 \sqrt{2\pi}} \exp^{-\frac{\left(x - k_3\left(R_i, R_j, R_k\right)\right)^2}{2\sigma_3^2}}. \tag{17}$$

Here σ is the standard deviation of the Gaussian kernel and x runs over a predefined range of values covering the possible values for k_n. R_i is the position vector of atom i and w is a weighting function that is used to control the significance of different terms.

The output of the LMBTR descriptor, in common with SOAP and the ACSF descriptor, is stratified according to the involved chemical elements, making the vector length dependent on the number of elements considered. For training containing a lot of different elements, the input vector, if uncompressed, will become very large. Resultantly, applications of the descriptor to date have focused upon specific systems containing one or two elements. Kwon *et al* [106] used the LMBTR descriptor, alongside ACSF and SOAP to directly predict XANES spectra of amorphous carbon from structural descriptors. They found that for this case LMBTR outperforms ACSF and SOAP. The authors ascribe this improvement to the explicit inclusion of bond lengths and angles which influence XANES spectra. Hirai *et al* [107] used linear regression of the input descriptors (LMBRT, SOAP and ACSF) to predict and interpret the XANES spectra of amorphous Si and SiO_2 with SOAP displaying the lowest mean squared error.

3.1.6. On the explicit inclusion of electronic information

The feature vectors explored in this section all explicitly encode nuclear geometric information and rely on implicit encoding of electronic information, i.e. allowing the ML model to infer/establish the connection between the nuclear and electronic structure through relationships in the dataset. This is, in some sense, the natural consequence of the fact that the construction of a purely geometric feature vector is computationally inexpensive while the computation of the electronic structure is expensive. In the context of x-ray spectroscopy, XAS spectra at the transition metal K edges principally contain structural information, encoded via the scattering of the x-ray (photo)electrons (section 2), and so purely geometric feature vectors

are easy to justify. In contrast, XES spectra [108, 109] and XAS spectra recorded at other (transition metal) absorption edges [e.g. the Pt $L_{2/3}$ edges, or the S K-edge (figure 2)], encode a wealth of electronic information by virtue of the selection rules and orbital-to-orbital transitions that are measured. Consequently, the question of whether electronic structural information (e.g. orbital information) should be included explicitly in the feature vector alongside the nuclear structural information is a natural one and, at present, one that requires further investigation. Watson *et al* [110] demonstrated that there remains a sufficiently strong implicit link between geometric and electronic structural information to develop a sufficiently accurate ML model at the Pt $L_{2/3}$ edges using a purely geometric feature vector viathe wACSF representation. The authors noted, however, that the error in the ML XAS spectral predictions was largest close to the $L_{2/3}$ absorption edges, i.e. in the spectral window which contains the greatest wealth of electronic information.

The literature contains examples of effective quantum-inspired representations which include electronic structural information: these include the representation used in molecular-orbital-basis ML (MOB-ML); [111] the F (Fock), J (Coulomb), and K (exchange) matrices (FJK) representation; [112] the spectrum-of-approximated-Hamiltonian-matrices (SPA^{H}M) representation; [113] and the matrix-of-orthogonalised-atomic-orbital-coefficients representation [114]. However, these representations are not specifically directed towards ML for x-ray spectroscopy and, while potentially suitable, have not been applied to problems in this domain to date.

An alternative approach, trialled by Lüder [115], addresses the challenge associated with the inclusion of electronic structural information in (transition metal) L-edge XAS spectra using a ML model motivated by the multiplet theory framework. In this work the model is trained on theoretically-simulated L-edge XAS spectra with the objective of enabling the relative energies of the 3*d* orbitals and the Coulomb and exchange interactions to be extracted from experimental L-edge XAS spectra of transition metal complexes [60]. Middleton *et al* [116] have also addressed the challenge associated with the inclusion of electronic structural information through the partial density-of-states (p-DOS) descriptor for ML x-ray spectroscopy. The approach presented by Middleton *et al* is based on an expression of Fermi's Golden Rule within the one-electron approximation and the dipole approximation:

$$\sigma = \frac{4\pi^2}{\omega} \sum_f |\langle \phi_{un} | \hat{\mathcal{D}} | \phi_{1s} \rangle|^2 \delta \left(E_i - E_f + \omega \right). \tag{18}$$

Under this approximation, a transition dipole moment will only be non-zero if the selection rule, $\Delta L = \pm 1$, is satisfied and if there is sufficient spatial overlap between the initial and final states. Consequently, by taking advantage of the localised nature of the initial core-hole state, an approximate spectrum can be obtained using the partial density of states corresponding to dipole-allowed transitions from the core orbital. For example, at the sulphur K-edge (as in [116]), this corresponds to ($p \leftarrow s$) electronic transitions. The p-DOS descriptor encodes information about the density of states on the absorbing sulphur atom from a minimal basis set in conjunction with a guess (i.e. unconverged) electronic configuration. To this end, this descriptor introduces a quantum-inspired representation for ML specifically tailored towards the simulation of x-ray spectra. The form of the p-DOS descriptor is directly inspired by the spectral shapes within the single-particle and dipole approximations and enables, for the first time, the inclusion of explicit electronic information of the absorbing atom into structural featurisation.

3.1.7. Molecular graph representations

This section has so far only explored manually constructed, or '*hand-crafted*', feature vectors of fixed dimensions ('*molecular descriptors*'). These representations have been widely applied across the space of chemical ML, motivated by the fact that they are computationally inexpensive to construct, intuitive, interpretable (e.g. through feature importance assessment), and easy to visualise. An alternative (and equally intuitive) approach, adopted increasingly commonly across the space of chemical ML, is based on using molecular graphs as input [117–120].

In the field of ML interatomic potentials, Batatia *et al* [121] exploited the graph representation within the MACE method, a message passing neural network to achieve high accuracy machine-learned potentials, where the use of higher-order terms (*messages*) led to an improved learning rate. The MACE approach extends the atomic cluster expansion (ACE) method [122] and achieves encoding of high-order many-body information of the nuclear structure in a computationally efficient manner. This approach has been applied to inter-atomic potentials [123], and recently to the modelling of infrared, Raman, and sum-frequency generation spectra [124]. It has not yet been used to simulate x-ray spectroscopic observables.

For x-ray spectroscopy, graph-based representations have not yet been widely applied. Carbone *et al* [125] implemented an approach based upon graph neural networks operating at the O and N K-edge. Their featurisation included an adjacency matrix describing atomic connectivity, a list of atom features (absorber,

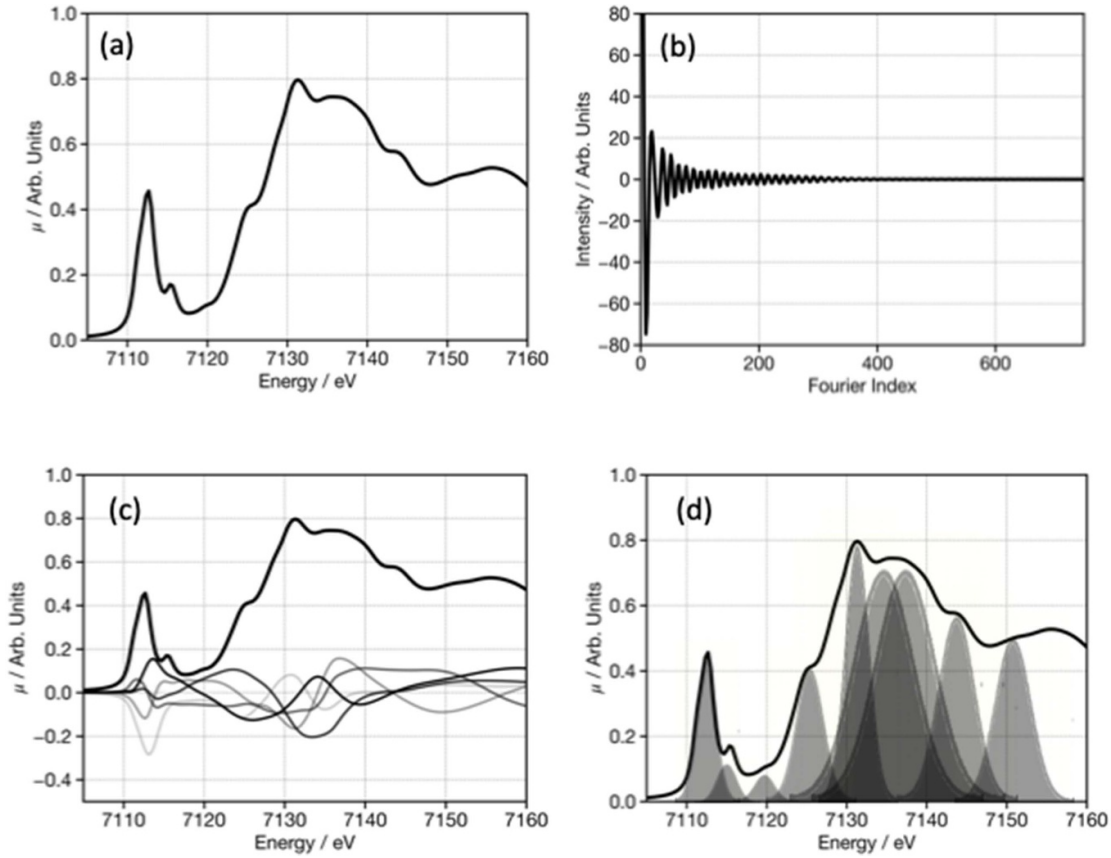

Figure 5. Schematic showing different XAS spectral representations: (a) energy space, (b) Fourier space, (c) principle component space, and (d) as a sum of Gaussian (or Lorentzian) functions.

atom type, hybridization, donor or acceptor status), and a list of bond features (bond type and length). Using this, the authors demonstrated that the resulting network could predict spectra with 90% accuracy of the predicted spectral peak locations being within 1 eV of the expected energy, very comparable to the performance achieved by Rankine and Penfold [92], although this did not specifically take advantage of the message passing framework to encode higher-order information. A similar approach was recently adopted by Kotobi *et al* [126] in which the authors focused on developing an explainable network. Indeed, using feature attribution the authors were able to quantify the contribution of each atom to peaks in the spectrum, which subsequently could be compared to orbitals involved in the transitions.

3.2. Representing x-ray spectra

Besides featurisation of the (local) atomic structure around the x-ray absorption site, the spectrum, $\mu(E)$, can also be represented in several ways. As with structural representation, the selection of representation for the x-ray spectrum influences both the size of a neural network (i.e. the number of free parameters) and its performance. The most common approach is discretisation, $\mu_i = \mu(E_i)$, where E_i represents an individual spectral energy point in the discretisation. While conceptually simple, and used in most models to date, this approach does have two potential limitations: (i) a large number of points may be required to resolve sharp peaks in spectra. (ii) small spectral shifts of narrow bands to slightly different positions in the spectrum can transfer intensity from one output neuron to a neighbour. While this may correspond to a relatively small change in spectroscopic lineshape, a machine-learning algorithm will be unable to differentiate these small shifts in position from more pertinent changes in intensity which result from a truly spectroscopically distinct peak. As such spectral shifts can arise from very small changes in the input structure, applying the grid-discretisation technique reduce the correlation between inputs (i.e. structures) and outputs (i.e. spectra) from the perspective of the ML algorithm, and so risks the development of a model which has not robustly encoded a valid relationship between variant inputs and meaningful changes in spectroscopic features. Other options to represent the spectra are illustrated in figure 5 and include polynomial regression, cosine transform, Gaussian fitting and PCA. While polynomials can also be used to represent x-ray spectra in a lower-dimensional form, a polynomial representation typically lacks generalisability in practice, as a high-order polynomial is required to fit all of the x-ray spectra in the dataset satisfactorily and encountering

Table 2. Performance at the Fe K-edge using the XANESNET MLP network as a function of the spectral representation, assessed using 250 *held-out* structure-spectrum pairs. The structure-spectrum pairs used in the *held-out* set are the same as those used in [92] and were selected at random from the full training set and never seen by the network. While the nature of the *held-out* set will influence the performance reported, this data which has never been seen by the network provides indicative performance. Structure represented using the wACSF descriptor. Input files and associated data are available at [45].

Spectral rep.	Output values	Network weights	Rel. performance
Energy discretisation	226	428 770	4.4 (3.6)
$TDCT_{50}$	50	338 482	5.5 (3.8)
$TDCT_{30}$	30	328 222	6.2 (3.9)
$TDCT_{15}$	15	320 527	6.7 (4.0)
$TDCT_{10}$	10	317 962	8.4 (4.4)
PCA_{50}	50	338 482	4.0 (3.0)
PCA_{30}	30	328 222	4.0 (3.2)
PCA_{15}	15	320 527	4.2 (3.2)
PCA_{10}	10	317 962	4.5 (3.3)

numerical instability is commonplace. Consequently, polynomial expansion is usually performed for the small energy intervals of the spectrum instead of the whole spectrum as recently demonstrated by Torrisi *et al* [127] in combination with a random forest ensembling ML model.

Table 2 shows the performance for the three spectral representations discussed in this section. We have excluded the Gaussian basis representation as we have found that the construction of the representation from an x-ray spectrum is time-intensive compared to the alternatives and, unless using a dense grid of Gaussians, comparatively worse in performance. However, Chen *et al* [128] have recently demonstrated the advantages of the Gaussian basis representation for the reverse ('spectrum-to-structure') problem, although in this case, the authors found that a cumulative distribution function representation of the x-ray spectrum achieved the highest degree of accuracy and transferability.

Table 2 demonstrates that the energy grid discretisation and PCA representations provide the best performance when assessed using the held-out datasets presented in [44]. A PCA representation, even when reducing the dimensionality of the x-ray spectrum to as few as 10 components, achieves performance comparable to energy grid discretisation while also reducing the size of the (MLP) network by >100 000 free parameters. However, we note that the PCA space is dependent on the set of spectra from which it is calculated. In addition, as this reduces the spectrum to coefficients of basis vectors over the whole spectrum, the poor prediction of one coefficient influences the whole spectrum. For this representation, we observe that some of the poor performers were significantly worse than those using energy discretisation, owing to this global effect of the coefficients predicted by the model. For the cosine transform, while the spectra are formally reproducible, the coefficients for the higher-frequency components approach zero. Consequently, we adopt the Truncated Discrete Cosine Transform (TDCT), which includes only the first N coefficients and assumes the remaining coefficients are zero. For TDCT($N = 50$), the performance is only slightly worse than the energy discretisation and PCA approaches, but in contrast to the latter, it shows a much faster decline in performance as N is decreased.

The performance of the PCA representation in table 2 highlights the potential advantage of dimensionality reduction and establishing descriptors not only for the input geometry but also the spectrum. This is especially important in the context of spectral inversion, i.e. for models seeking to extract structure from an input spectrum. Tetef *et al* have gone beyond the linear PCA method and employed non-linear approaches including t-Distributed Stochastic Neighbor Embedding (t-SNE) [129] and uniform manifold approximation (UMAP) [130] which could be used to perform clustering and classification analysis of both XES and XAS spectra. Routh *et al* [131] and Liang *et al* have employed constructed spectral descriptors based upon the latent space of an autoencoder. Importantly in [131], the authors not only generated spectral descriptors based upon the autoencoder but were also able to interpret the latent space representations highlighting the physical insight they can provide. Beyond mathematical deconstructions, Guda *et al* [132] used chemical intuition to develop XANES descriptors based upon edge position, intensities, positions, and curvatures of minima and maxima which they could demonstrate correlation to structural parameters such as coordination number and first shell bond lengths.

Finally, for models seeking to transform structures into spectra, when representing any calculated spectrum, it is also important to consider spectral broadening. Figure 2 illustrates that x-ray spectra are typically broad in comparison to, for example, optical and vibrational spectroscopies. Consequently, the calculated spectra must be transformed by incorporating factors including core-hole-lifetime broadening and instrument response [110, 133] to enable them to be compared to the experiment. An example of the

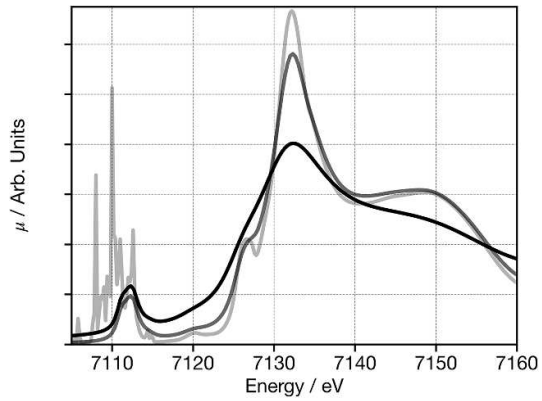

Figure 6. Example absorption cross-sections for the Fe K-edge of $C_{20}H_{18}FeN_6S_2$ (CCSD code: ABITEM) calculated using multiple scattering theory implemented within the FDMNES package [134] (a) without any post-processing (light grey), (b) broadened with a fixed-width Lorentzian function (FWHM = 1.25 eV, grey), and (c) broadened with an arctangent convolution model (black). See [110] for a more in-depth discussion on this.

influence this has is shown in figure 6 and can be added as a pre-processing or post-processing step in the ML models.

While the spectra without the aforementioned broadening (figure 6, light grey line) retain the most spectral information, the sharp nature of the resonances, especially at low energy can make learning challenging. In contrast, while the fully broadened (figure 6, black line) is the closest representation of experimental spectra, it presumes a specific resolution and therefore lacks the flexibility to model different experimental techniques (e.g. high-energy-resolution fluorescence detection (HERFD) [135] spectroscopy) which offer higher resolution. Consequently, during our previous work [92], our models used spectra containing a minimal core-hole lifetime broadening which represents a midpoint between the two extremes.

4. Types of networks

In the realm of x-ray spectroscopy analysis, ML models aim principally to tackle two challenges: the *forward* (from property/structure to spectrum) and *reverse* (from spectrum to property/structure) mapping problems. Beyond the treatment of these two categories of problems, there exist a broad range of other applications, encompassing such diverse uses as automated diagnostics, data management and cleaning, and even experimental control [136–140]. Whilst the innovative developments leveraging ML for utility in these fields are undoubtedly exciting, a comprehensive delineation and assessment of the works within them is beyond the scope of the present review, and so in this section we discuss treatments of the two principal categories of *forward* and *reverse* mapping problems.

4.1. Forward mapping: structure → spectrum

The focus of ML techniques applied to x-ray spectroscopy has, to date, largely been on the *forward* mapping problem. Here, in a manner akin to quantum chemistry calculations, an input structure is used to predict binding energies for photoemission [141–143], which is converted into the lineshape for XAS [92, 110, 125, 144–148] or XES [108, 129, 149]. These methods have addressed light and heavy elements (e.g. C, N, O, Fe, Mn, Ni, Pt) as well as different absorption edges (e.g. K and $L_{2,3}$). Overall, while the methods differ in the formulation of the network and training sets, they are conceptually similar. All exhibit promising results, and clearly demonstrate an ability to transform easy-to-generate structural properties, such as nuclear geometry, into spectroscopic observables.

ML models seeking to simulate XPS must establish a link between atomic structure and core electron binding energies. To date, most of the work in this area has focused on the analysis of XPS spectra for amorphous structures, which can be imprecise since the disorder can create overlapping bands and broadening peaks. The computational prediction of XPS spectra of such materials requires extensive sampling, which is time-consuming, and therefore ML methods can potentially bridge this gap. Sun *et al* [150] used the LMBTR descriptor with a random forest model to predict XPS at the carbon K-edge specifically for solid-electrolyte interfaces reporting an MSE in peak positions as low as 0.05 eV. Golze *et al* [151] used the SOAP descriptor to develop a kernel regression model that can predict the XPS spectra for CHO-containing molecules and materials. This is achieved using a comprehensive database of calculated core-binding energies at DFT and GW levels of theory. Their work is implemented within an openly available

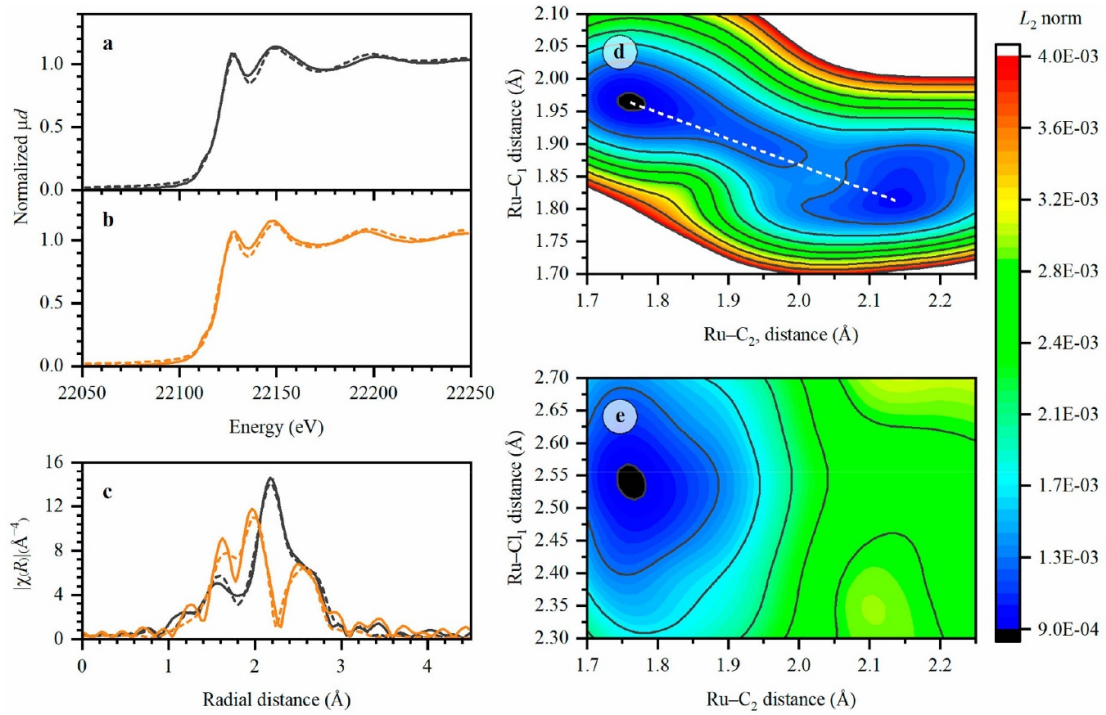

Figure 7. (a), (b) XANES and FT-EXAFS (c) spectra of $[RuBr_2(CO)_3]_2$ (black) and $[RuCl_2(CO)_3]_2$ (orange). The solid lines are experimental data and the dashed lines are best fits obtained using PyFitIt. (d), (e) show contour plots of the error (L_2-norm) as a function of key structural parameters. Reprinted with permission from [153]. Copyright (2021) American Chemical Society.

XPS prediction server, nancarbon.fi/xps, which highlights the accessibility accurate ML models can provide. In this work, the authors found ~10 000 training samples were required to achieve an MSE below 0.02 eV which suggests this approach could be more broadly applied to different elements and edges. ML models for XPS have to date focused upon lighter elements as extensive theoretical work means that computational simulations used to generate the training sets are most accurate in this energy range [51]. For heavier elements, there is an increased significance of relativistic effects and the self-interaction error associated with the approximate treatment of exchange in density functional-based methods making developing training sets as accurate as the errors achievable using the ML models above a challenge [53, 152].

For XAS, as the underlying relationship between the input structure and spectroscopic observables is well-understood (see section 2) there has been a large number of works aimed at developing models connecting the two using a variety of levels of sophistication. Amongst the most widely used is the FitIT [154] code developed by Smolentsev *et al* This approach uses a multi-dimensional interpolation of spectra calculated within a user-defined structural parameter space to develop a model which can subsequently be used to optimise structures by fitting XAS spectra within the defined structural parameter. This limits the number of calculations needed to achieve a detailed spectral interpretation. However, while powerful, it requires a bespoke model to be initiated for each new system. Recently Martini *et al* [148] have extended this method to produce PyFitIt software which incorporates multiple ML algorithms including ridge regressions, decision trees and neural networks, which have been used for both XANES [155, 156] and EXAFS [79] spectra. As an illustrative example, figure 7 shows the application of PyFitIt to refine the structures of dimeric $[RuX_2(CO)_3]_2$ (X = Cl, Br) complexes. The structure was refined using 5 structural degrees of freedom focused upon first coordination shell bond length and the model within this space developed using a training set of ~9000 spectra, i.e. a fairly comprehensive coverage of nuclear configuration space. The authors also demonstrated that the model developed can determine the uncertainty of the predicted structures and associated confidence. These methods provide a powerful approach that is highly adaptable to a wide variety of models. However, a limitation is its lack of generality meaning that it requires a bespoke model to be initiated for each new system studied.

To increase generality, several works have implemented DNN to predict spectral lineshapes. figure 8 shows an illustration of the general workflow for the *forward* mapping approach using DNN. While this directly refers to the XANESNET method [43], describing the approach adopted in [92], the general principles remain broadly relevant across all approaches in this field. Firstly, structures ('*samples*') from datasets such as tmQM [157], QM9 [105, 158, 159] and materials project [160] are used to calculate the

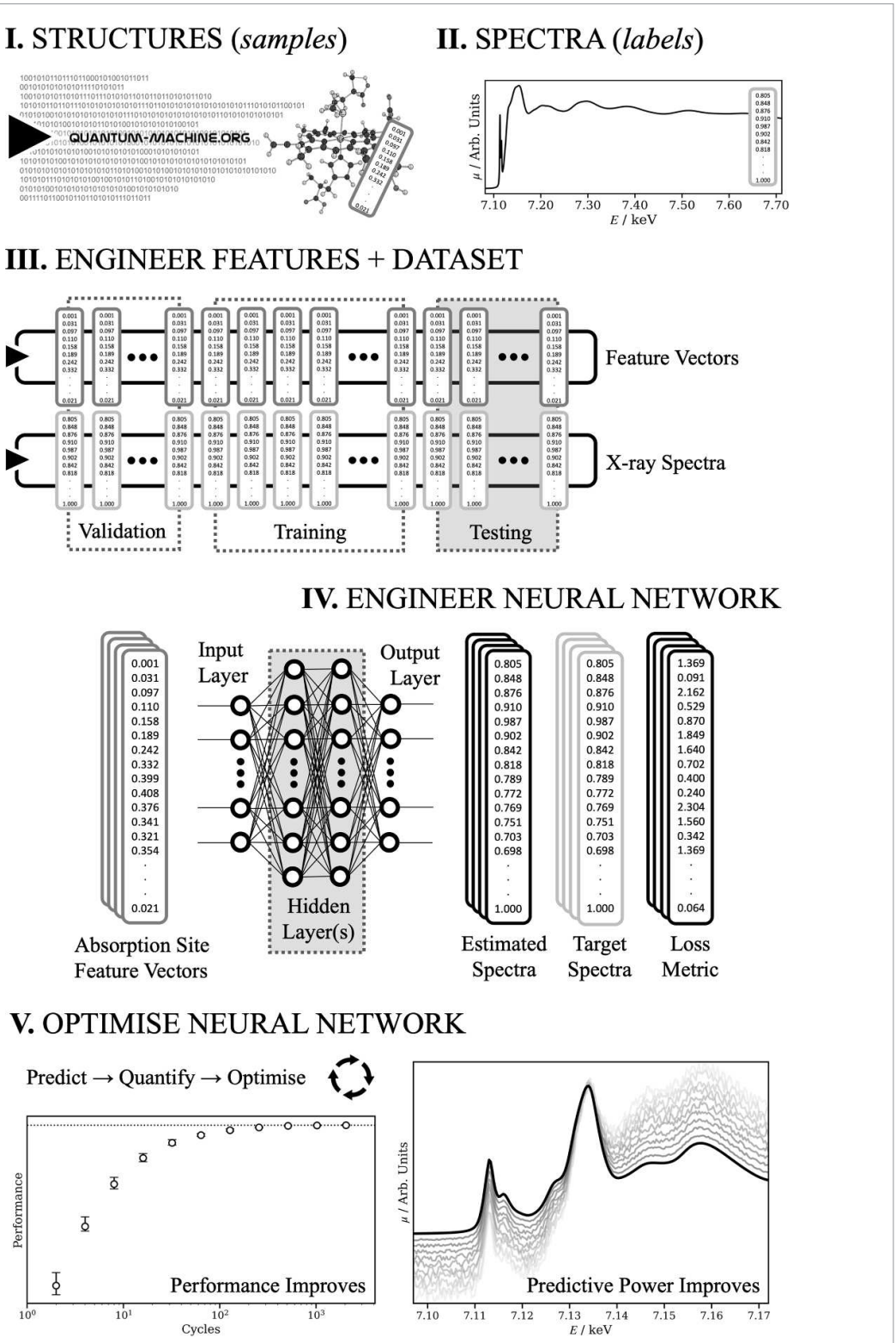

Figure 8. Illustration of the XANESNET workflow as described in [92]. The nuclear geometries are featurised and, alongside their corresponding calculated spectra, form pairs of data used during training. The objective of the DNN is to establish a '*forward*' mapping by optimising the internal weights. Reproduced from [92]. CC BY 3.0.

theoretically-calculated spectra ('*labels*'). This represents the first key step and the method used to simulate the spectra determines the overall accuracy of the model. The samples are encoded as a feature vector, as described in section 3.1, and subsequently fed into the DNN, which attempts to establish a mapping from the feature vector to the spectrum through iterative modification of the network weights.

The influence of structural (section 3.1) and spectral (section 3.2) representations have been discussed above. Table 3 illustrates the effect of the network architecture for three architectures implemented within the XANESNET package [43], namely multilayer perceptron (MLP), convolutional neural network (CNN)

Table 3. Performance of XANESNET for predicting transition metal K-edge using an MLP (428 770 free parameters) CNN (246 626 free parameters) or LSTM (422 376 free parameters), assessed using 250 *held-out* structure-spectrum pairs. The structure-spectrum pairs used in the *held-out* set are the same as those used in [92] and were selected at random from the full training set and never seen by the network. While the nature of the *held-out* set will influence the performance reported, this data which has never been seen by the network provides indicative performance. Structure represented using the wACSF descriptor. Input files and associated data are available at [45].

	Performance/%		
Element	MLP	CNN	LSTM
Ti	4.6 (3.1)	5.8 (3.1)	4.5 (3.7)
V	4.5 (5.0)	6.2 (4.7)	4.9 (4.9)
Cr	3.6 (3.9)	4.2 (3.7)	3.6 (3.7)
Mn	4.3 (2.9)	5.4 (2.5)	3.7 (2.6)
Fe	4.4 (3.6)	5.4 (3.5)	4.4 (3.7)
Co	4.3 (2.9)	7.0 (3.3)	4.2 (3.1)
Ni	4.5 (2.8)	5.3 (3.5)	4.3 (3.0)
Cu	3.6 (2.7)	6.0 (3.4)	3.3 (2.4)
Zn	4.0 (2.6)	5.1 (2.3)	3.3 (2.4)
Mean	**4.2 (3.2)**	**5.8 (3.5)**	**3.8 (3.3)**

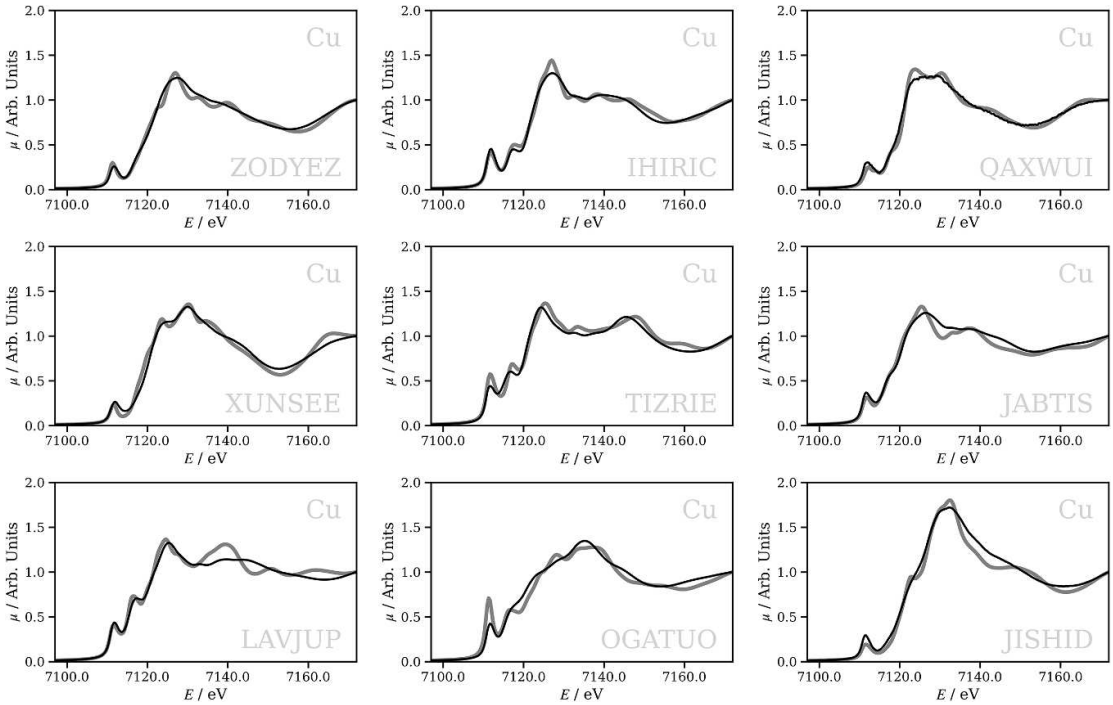

Figure 9. Representative Cu K-edge XANES spectra. The top three show spectra taken from the 20th–30th percentiles, i.e. around the lower quartile. The centre three show spectra from the 45th–55th percentiles, i.e. around the median. The lower three panels show K-edge XANES spectra from the 70th–80th percentiles, i.e. around the upper quartile. The six-character labels in the lower right of each panel are the Cambridge Structural Database (CSD) codes for the samples. Reproduced from [92]. CC BY 3.0.

and long short-term memory (LSTM) network. As representative examples, these have been applied to the transition metal training data described in [92] and openly available at [44]. In all cases, similar performance is observed across all of the first-row transition metal K-edge, with slightly better performance for the Cu and Zn edges, which is associated with the weaker pre-edge in these spectra. This shows that both MLP and LSTM yield overall very similar performance, with the latter yielding a slightly low percentage difference for $\mu_{predicted}$ when compared to μ_{target} of 250 held-out examples. The performance of CNN is slightly worse than the other two. Overall, this is achieved with almost half the internal network weights. To provide context for these numbers, figure 9 illustrates an example of K-edge XANES spectra predicted using the MLP network described above and in [92]. This clearly illustrates that even for the worst performers in the held-out dataset (figure 9 bottom line), the network captures the general spectral shape. The training data and input associated with these simulations can be obtained from [45].

The focus of the work discussed in the previous paragraph has been upon achieving generality, in the sense that networks are aimed at being able to simulate an x-ray spectrum for an arbitrary absorbing atom in any coordination environment for a given absorption edge. We refer to this as a '*Type I*' model, a type which is generally preferable as it avoids the time-consuming requirement to develop a new model for every specific problem. The main challenge associated with developing accurate training sets and achieving generality, as with innumerable ML tasks across all fields, is scale. Indeed, recent DNN models for predicting XAS spectral lineshapes of transition metal K-edges [92] have been trained using molecules from the tmQM training set [157] containing a single geometry of the mono-metallic complexes harvested from the Cambridge structural database (CSD). While—as shown above—this is accurate when used to predict spectral shapes of compounds in a similar chemical space, large uncertainties arise when considering complexes with multiple heavy atoms within the cutoff radius (6 Å) of the absorbing atom or which are strongly distorted from their equilibrium geometry [147, 161]. Consequently, further developments in this field should focus on both the training set and how the structures are represented to optimise performance. However, achieving comprehensive coverage of the chemical space is a significant challenge, especially when seeking to develop a training set using a high-level theory with a large computational burden.

An alternative approach, the so-called '*Type II*' method, is to tailor one's model to a more specialised problem. These models are trained using data for a specific class of systems [162–165]. Indeed, this is the concept behind the originally developed FitIt [154] approach. Kwon *et al* [106] used an MLP in conjunction with the LMBTR, ACSF and SOAP descriptors to directly predict XANES spectra of amorphous carbon. They reported that LMBTR outperforms ACSF and SOAP which the authors attribute to the explicit inclusion of bond lengths and angles that influence XANES spectra. In total, they used 12 528 training samples although did not show how the convergence of the model varies with the size of the training set. These works demonstrate high-level accuracy that can be achieved in the '*Type II*' models, although with the disadvantage that a new model needs to be trained for each new problem addressed. Consequently, it may be beneficial to apply classification models to break inputs into established subgroups, which could then be used to automatically develop individual bespoke models able to achieve generality or their specific class, i.e. use a neural network with a classifier architecture, such as a decision tree, to automatically subdivide chemical space into more manageable groups. Recently, Tefet *et al* [129, 166] have used unsupervised ML methods to classify XAS and XES spectra, distinguishing key properties including oxidation state, bonding, coordination number, and aromaticity. The success of these classification methods could address the challenge of collating sufficent data of sufficient scale to satisfactorily train general '*Type I*' models.

4.2. Reverse mapping: structure ← spectrum

In the previous section, our focus was on the '*forward*' mapping task, i.e. the task of mapping structures and/or structural properties onto the spectroscopic observables (structure → spectrum). This task is analogous to the objective of computational spectroscopy in that a first-principles or density-functional-derived wavefunction is used to compute the spectrum/spectroscopic observable from an (initial) geometry. However, as it provides a direct data-to-interpretation channel, the '*reverse*' mapping task, i.e. the task of translating an (experimental) spectrum into a structure or structural property (structure ← spectrum), is of substantially greater interest to experimentally-focused end-users.

The simplest approach (in terms of conception and implementation) has origins dating back to the inception of x-ray spectroscopic analysis, and involves interpreting experimental x-ray spectra through direct comparison to reference data. While it is possible to carry out such a comparison with a limited subset of domain-specific reference data, general application requires an extensive dataset of reference data and a robust method for the quantitative assessment of the degree of similarity between the recorded and reference x-ray spectra; currently available datasets contain only a small number of experimental x-ray spectral [167] , which greatly limits this approach. The generation of suitably large datasets of x-ray spectral references is presently only practicably possible through theoretical simulation [168]. Zheng and Mathew *et al* [168, 169] for example, have generated such a database (comprising over 800 000 *K*-edge x-ray spectra) through theoretical simulation. These x-ray spectral references can be compared to experimental x-ray spectral data using a diversity of similarity metrics to limit bias. While such comparisons are undoubtedly useful, they are typically only effective for well-defined (e.g. crystalline) molecules and materials [80] and, in addition, as any comparisons are based on a comparison between experiment and theory, they will fail to deliver where the theory does not provide a satisfactory description of the molecules/materials under study.

Clustering and dimensionality-reduction approaches [170, 171] represent appealing methods, widely applied to simplify the problem and provide spectral interpretations. The objective of clustering approaches is to identify a few basis x-ray spectra that can, by their combination, represent satisfactorily a larger dataset; these approaches have been used to great effect for the processing of spatially-resolved x-ray spectra, [130, 172–174] analysis of *in operando* XANES for catalysts and battery materials, [175–178] and for feature

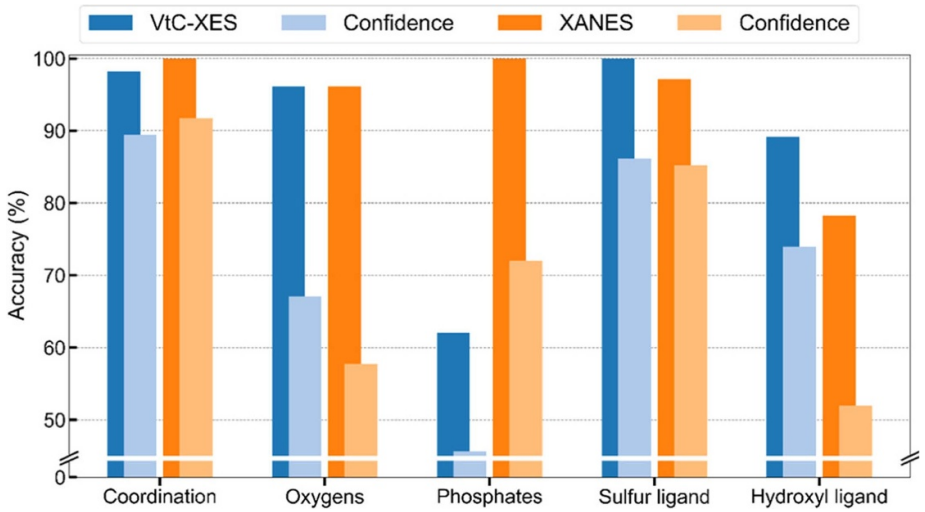

Figure 10. GP classifier prediction accuracy with corresponding average probability for all classifications of x-ray (XANES and valence-to-core XES) spectra. Reprinted with permission from [166]. Copyright (2022) American Chemical Society.

extraction from x-ray spectra [166]. Aarva *et al* [179, 180] have, for example, generated a series of spectroscopic fingerprints which can be compared to—and, crucially, used to interpret—experimental x-ray spectra. The authors used the SOAP descriptor to characterise and cluster based on the chemical environment, providing a direct link between spectrum and (local) structure. Additional unsupervised approaches, e.g. dimensionality reduction [including principle component analysis (PCA)], [181] *t*-distributed stochastic neighbor embeddings (*t*-SNE), [129] uniform manifold approximation and projection (UMAP) operations, [166] multivariate curve resolution, [182–184] and autoencoding [129, 131] have all been applied to x-ray spectroscopy with the objective of finding simplified representations of x-ray spectra which can then be connected directly to the structural/electronic properties of the molecules and materials under study.

An example of the application of decomposition/dimensionality reduction approaches is shown in figure 10 (derived from work by Tetef *et al* [166]). Tetef *et al* [166] carried out a UMAP-embedding-based cluster analysis to investigate the spectral sensitivity of x-ray spectroscopy (P *K*-edge XANES and valence-to-core XES) to structural features of complexes including coordination number and oxidation state. The authors used their cluster analysis to prepare the input for a Gaussian process (GP) classifier to interpret directly the x-ray spectra in the context of a '*reverse*' (structure ← spectrum) mapping task. Figure 10 shows the accuracy of the scheme as a predictor of coordination number, number of oxygen/sulphur/hydroxyl ligands, and phosphate classification. Except for the latter (phosphate classification from valence-to-core XES) the authors were able to achieve accuracy close to or above 80% across all subtasks.

In contrast to mathematical decomposition/dimensionality reduction approaches, Guda *et al* [132] have experimented with the use of physical/chemical intuition to develop compact x-ray spectroscopic (XANES) descriptors. Figure 11(a) illustrates such a descriptor based on x-ray absorption edge position and intensity, and the curvature of post-edge minima and maxima, to give a compact fingerprint of the (local) electronic and geometric structure of the absorbing atom(s). The authors demonstrated (figures 11(b)–(e)) that these compact fingerprints correlate well with the structural properties of interest. In combination with regression and classification machine-learning models, the authors could optimise the exact composition of these descriptors to achieve not only spectral interpretation but also physical/chemical insight.

It is also possible—and perhaps desirable, moreover—to use machine-learned/extracted features (e.g. those derived directly from the data by, for example, a neural network feature extractor) instead of handcrafted features (e.g. those constructed based on physical/chemical intuition or decomposition/dimensionality reduction). Drera *et al* [141] and Pielsticker *et al* [142] for example, have both implemented CNN feature extractors that can be coupled to a regressor/classifier head for the automatic analysis of x-ray (XPS) spectroscopy. Drera *et al* [141] used a dataset of *ca.* 100 000 theoretical x-ray spectra to detect and quantify chemical elements/composition based on the XPS spectrum, while Pielsticker *et al* [142] adopted a similar approach targeting automatic quantification based on transition metal XPS data; the authors also included an uncertainty quantification approach using Monte-Carlo dropout.

Timoshenko *et al* [76, 80, 80, 164, 165] have carried out pioneering work in the XAS domain along these lines, demonstrating the predictive power of neural networks to obtain structural insights from both XANES

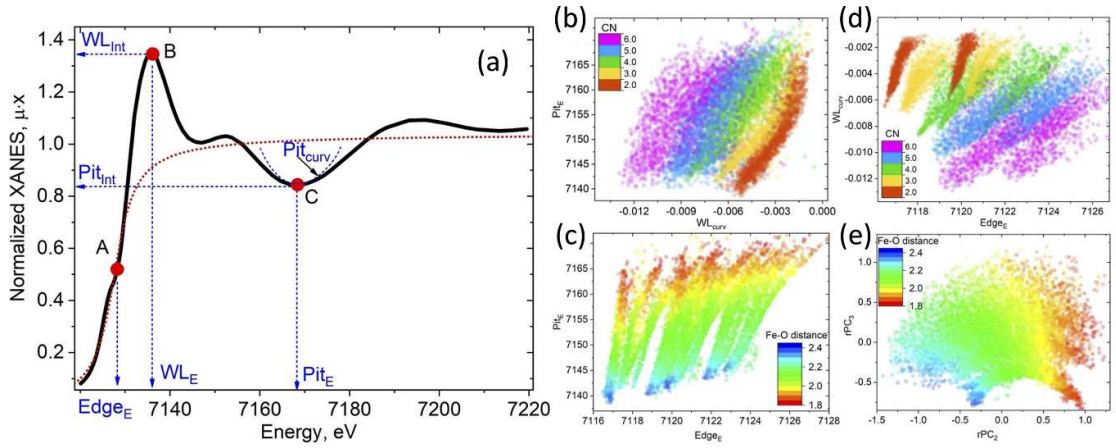

Figure 11. (a) Set of XANES descriptors, including edge energy and slope, white line intensity, curvature, and position, pit intensity, curvature, and position. (b)–(e) show scatter plots of pairs of descriptors against structural parameters, *i.e.* first coordination shell bond lengths and coordination numbers. Reproduced from [132]. CC BY 4.0.

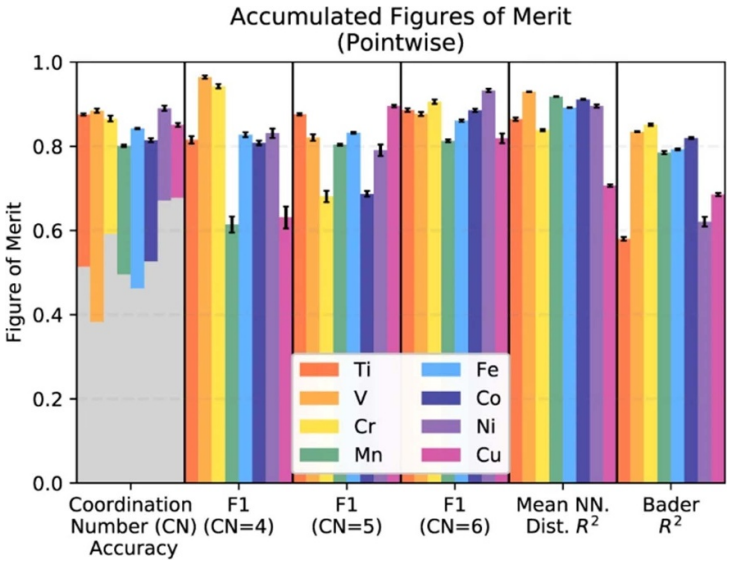

Figure 12. Performance of a random forest (ensemble) ML model reported by Torrisi *et al* [127] evaluated on the task of extracting geometric and electronic parameters from x-ray spectra. Reproduced from [127]. CC BY 4.0.

and EXAFS spectra across and wide variety of systems. This is especially important within the context of the disordered catalytic materials focused upon in their work (for which a satisfactory first-principles analysis is a far-from-trivial task on account of the large number of atomic configurations that have to be considered). While the results are highly encouraging, the authors focus on Type-II, i.e. system/class-specific ML models, and—as such—there remains scope to explore ML models with greater generality in the future.

Carbone *et al* [185] have also carried out work in this space, having developing a framework to classify the symmetry of the coordination environment around an x-ray absorption site. For the first-row transition metal elements, the authors were able to achieve an 86% classification accuracy and were also able to demonstrate that only a small decrease in performance was observed when using only the pre-edge region of the XAS spectrum. These observations are consistent with empirical knowledge which holds that changes in the local coordination environment will modulate most strongly the shape of the resonances in the pre-edge (XANES) region of the XAS spectrum [186]. Torrisi *et al* [127] have also demonstrated that coordination numbers, average first-coordination-shell bond lengths, and the atomic charge of the absorbing atom, in addition to the symmetry of the coordination environment, can also all be learned using a random forest ML model. Their results, reproduced in figure 12, demonstrate >80% classification accuracy against these properties and balanced treatment across the first-row transition metals. Kiyohara *et al* explored a

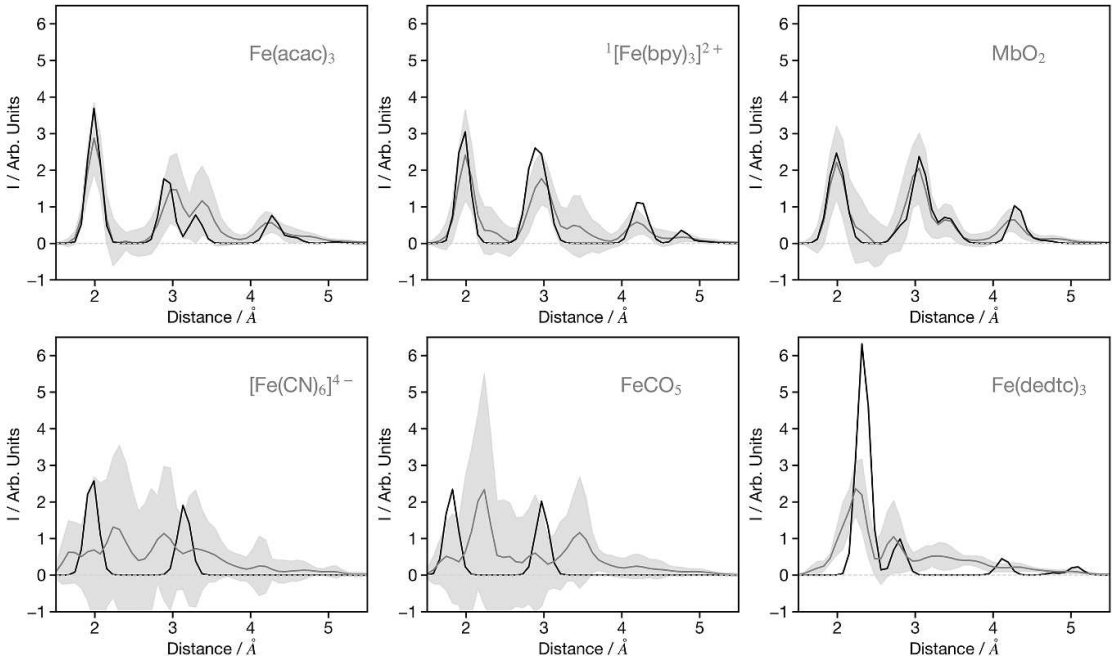

Figure 13. G^2 wACSF derived by the XANESNET CNN applied to experimental spectra, as documented in [190]. The grey lines represent the predicted distributions, while the lighter grey areas indicate the range of variation within $\pm 2\sigma$, determined using bootstrap resampling. The black traces depict the expected distributions based on structures reported in experimental studies The top three panels are the best predictions, while the bottom three panels are the worst predictions. Reproduced from [190]. CC BY 4.0.

combination clustering and decision-tree ML model where, for a selection of oxygen and carbon K-edge XAS spectra, categories of spectra were first clustered together and a decision-tree model was used to derive subsequently the correspondence between the distinctive x-ray spectral features characterising each cluster of x-ray spectra and the geometric properties. of interest [187]. In addition to these classification ML models, an earlier study by Kiyohara and Mizoguchi [188] and a study from Higashi and Ikeno [189] both reported regression ML models for mapping x-ray spectra onto two-body PDFs and applied these to the analysis of oxygen K-edge XAS spectra. The authors were able to use the PDFs to extract geometric parameters, such as the expected first-coordination-shell bond lengths, to high accuracy with relative errors <0.2 Å.

However, although these previous works showcase highly effective ML models, to date they have largely been developed using—and evaluated against—theoretical x-ray spectra. This fails to align with the proposed practical intent of these methodologies, which is to extract information from experimental x-ray spectra. David *et al* [190] have recently implemented a CNN that maps Fe K-edge XAS spectra into a pseudo-PDF based on the two-body wACSF (G^2) terms. Although David *et al* trained their CNN using theoretically-generated x-ray spectra, the authors evaluated the performance of their CNN against experimental x-ray spectra. Figure 13 shows six G^2 wACSF predicted for the experimental Fe K-edge XAS spectra of $Fe(acac)_3$; [191] $[Fe(bpy)_3]^{2+}$; [192] MbO_2; [193] $[Fe(CN)_6]^{4-}$; [194] $FeCO_5$; [195] and $Fe(dedtc)_3$ [196].

$Fe(acac)_3$, $[Fe(bpy)_3]^{2+}$, and MbO_2) show strong performance for the first two coordination shells. However, in the spirit of improving the performance of these approaches, it is more instructive to understand the examples for which the CNN delivers poor performance, i.e. $[Fe(CN)_6]^{4-}$, $FeCO_5$, and $Fe(dedtc)_3$. For the former two transition metal complexes, previous work has highlighted the challenge of the network for describing systems containing linear bonds like carbonyls and cyanides, [161] owing to an x-ray '*focusing effect*' that exerts a strong influence on the appearance of the x-ray spectrum. However, while the predictions for $[Fe(CN)_6]^{4-}$ and $FeCO_5$ are poor, the uncertainty is also large, demonstrating that the ML model is aware of its limitations. In contrast, $Fe(dedtc)_3$ not only yields an inaccurate set of predicted G^2 wACSF but—judged by a low uncertainty—exhibits over-confidence. This arises from the challenge of transferring a network trained on theoretical x-ray spectra to experimental x-ray spectra. The long Fe-S bonds (*ca.* 2.3 Å) in $Fe(dedtc)_3$ lead to a breakdown of the '*muffin-tin*' approximation used to simulate the Fe K-edge XAS spectra under the MS approach. Hence, even though the network is trained on molecules sharing a similar structure, leading to a high level of confidence, this confidence is misplaced because the training data fails to coincide with experimental spectra for such scenarios.

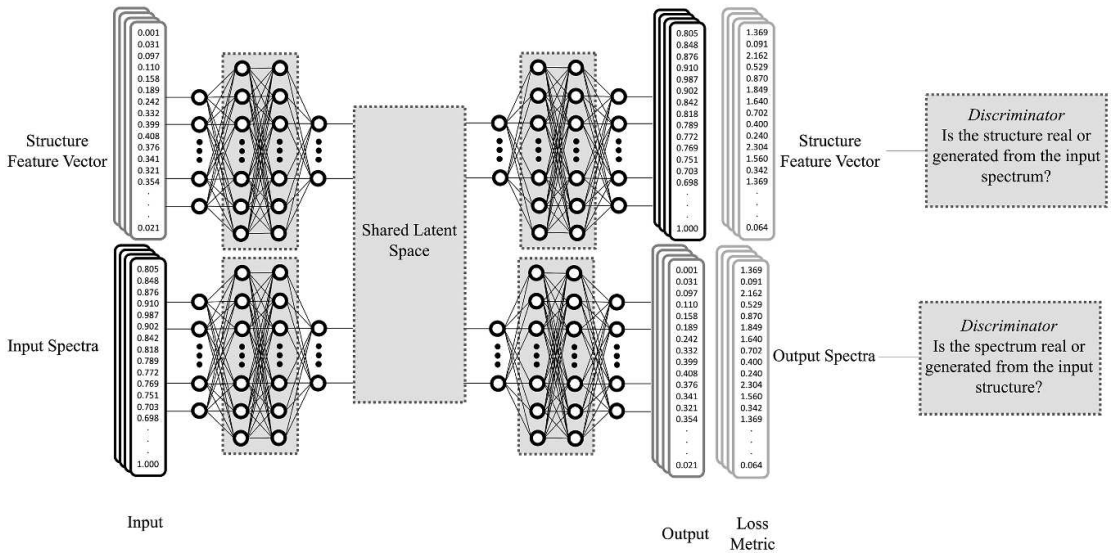

Figure 14. Schematic representation of the Auto-Encoder Generative Adversarial Network (AEGAN) implemented within the XANESNET package [43] The dual learning cycle-consistent model uses two coupled autoencoders with a shared latent space to ensure consistent *forward* and *reverse* interpretations of the structure and spectra.

This could be solved by training on well-characterised experimental data. Nonetheless, despite advancements like laboratory-based x-ray spectroscopy [197, 198], which have improved our capacity to obtain experimental x-ray spectra, it remains a formidable challenge to gather the quantity and quality of x-ray spectra necessary for ML model training. This is not to say that it is not possible; indeed, Chen *et al* [128] recently used experimental data during the training of their network to predict properties such as oxidation state. The authors demonstrated that when representing the spectra as a continuous distribution function, they were able to classify the changing oxidation state of a battery material during cycling. However, despite the promising results, the authors highlighted the challenge associated with cases where a mismatch between experimental and computational spectra emerges. An alternative approach, recently applied to inelastic neutron scattering (INS) data [199] is to use generative adversarial networks to translate theoretical spectra into those which mimic their experimental counterparts. In [199], their Exp2SimGAN, based upon dual contrastive learning GAN (DCLGAN) [200] was designed to convert a simulated dataset into one that resembles an experiment and was applied to convert between convolved and unconvolved INS spectra. In this area, cycleGANs have received attention, owing to their ability to translate information between two domains within an unsupervised framework [201, 202]. This approach, to date, has been used to translate from one domain to another but exploits a cycle consistency loss to ensure that the data can be trained without the need for paired and transformations are kept as close to the original as possible. Consequently, considering the results presented in [199], it should be considered that a similar approach could be used to overcome the absence of experimental data for training *reverse* networks, where networks such as cycleGANs are used to translate calculated spectra to appear more like their experimental counterparts. Here, theoretically derived spectra could be passed through a cycleGAN to generate pseudo-experimental data, with the potential to improve the performance of *reverse* models. However, this still requires the development of a database of well-characterised experimental x-ray spectra, which should be a key focus of future work.

4.3. Self-consistency: bidirectional networks

Sections 4.1 and 4.2 outlined methods capable of addressing the structure/property to spectrum and spectrum to structure/property mapping problems. However, one of the potential limitations of this approach is the independent nature of the networks and therefore there is no way of guaranteeing self-consistency i.e. the forward and reverse predictions give the consistent with each other. This could be enforced using cycle consistency as discussed in the previous section. To address this, figure 14 shows an Auto-Encoder Generative Adversarial Network (AEGAN) implemented within the XANESNET package [43]. This model adopts two coupled autoencoders with a shared latent space and cycle consistency loss to ensure consistent *forward* and *reverse* interpretations of the structure and spectra. Consequently, the network incorporates 6 loss functions, which must be carefully balanced to optimise network performance. This highlights the challenge associated with optimising the performance of more complicated networks.

Table 4. Performance of the XANESNET AEGAN network for all of the transition metal K-edge spectra, assessed using 250 *held-out* structure-spectrum pairs. The structure-spectrum pairs used in the *held-out* set is the same as those used in [92] and were selected at random from the full training set and never seen by the network. While the nature of the *held-out* set will influence the performance reported, this data which has never been seen by the network provides indicative performance. Spectra are represented as discretised energy points and the structure is represented using 32 G^2 wACSFs and 64 G^4 wACSFs. Input files and associated data are available at [45].

	Performance/%			
	Predict		Reconstruct	
Element	Forward	Reverse	Forward	Reverse
Ti	9.3 (5.6)	4.3 (3.7)	6.9 (3.6)	4.1 (2.7)
V	7.6 (9.8)	5.1 (4.9)	6.8 (7.3)	4.5 (3.2)
Cr	4.8 (5.4)	2.7 (3.6)	3.8 (4.2)	2.3 (2.9)
Mn	17.4 (11.9)	6.8 (3.2)	10.7 (6.5)	7.2 (4.6)
Fe	7.6 (4.6)	3.3 (2.7)	6.1 (3.0)	2.5 (1.7)
Co	9.5 (5.6)	3.6 (2.6)	8.2 (4.7)	3.4 (2.5)
Ni	7.9 (4.6)	4.4 (3.0)	7.4 (3.8)	5.1 (3.3)
Cu	14.5 (8.3)	4.1 (3.3)	13.1 (8.0)	4.4 (2.7)
Zn	7.1 (4.9)	3.8 (2.9)	6.7 (3.8)	3.6 (2.5)
Mean	**9.5 (6.7)**	**4.2 (3.3)**	**7.7 (5.0)**	**4.1 (2.9)**

Table 4 shows the performance of this model across the first-row transition metal training set [44]. Overall the performance on predicting the spectra and indeed reconstruction is slightly worse than presented in section 4.1. This difference in performance is likely to be linked to the complexity of the network, which in contrast to independent networks is more sensitive to variations in the hyperparameters of the network. This is especially true for the description of the loss functions, indeed overall this network has 6 independent loss functions that are combined and the relative weighting between them can influence performance. In addition, while appealing due to their cycle consistency, these dual-learning models exhibit larger networks which usually necessitate more free parameters. For example, the model used in table 4 contains just over 1300 000 free parameters. Consequently, while the present performance is non-optimal, the ability to ensure cycle consistency is an appealing property, and further work should be invested in the development of such networks.

4.4. Δ-learning

The objectives of the ML techniques explained thus far have been to transform and translate between structural and spectral representations without a need for first-principles calculations. While these have been successful, a substantial obstacle in creating models which are both precise and broadly applicable is the scale of training data required. Achieving comprehensive coverage of the chemical space remains a formidable challenge, particularly when attempting to create a training dataset using a high-level theory that demands substantial computational resources.

However, it is important to recognise that while the most accurate and costly calculations provide precise, quantitative agreement between experiment and theory, much simpler calculations can still provide qualitative/semi-quantitative interpretations of experiments [23]. Indeed, the general spectral shape and most of the relevant physics are often captured through computationally inexpensive methods (*e.g. multiple scattering theory [69]*), while the outstanding small corrections to the spectral shape required to achieve quantitative agreement are usually by far the most computationally demanding. Consequently, one approach to reducing the computational expense and the requirement to develop large training sets is to adopt the composite strategy, Δ-learning as introduced by Ramakrishnan *et al* [204].

In the Δ-learning framework, models are engineered to correct characteristics acquired from a less computationally demanding calculation to align with those associated with a more advanced yet computationally intensive methodology, effectively performing a correction from low-level to high-level theory without entailing the costs of high-level methods. This approach has been widely used across quantum chemistry [205–207]. For x-ray spectroscopy, one can deploy an ansatz:

$$\mu(E)^{\mathrm{H}} = \mu(E)^{\mathrm{L}} + \Delta(E)^{\mathrm{ML}} \tag{19}$$

where $\mu(E)^{\mathrm{H}}$ is the spectrum calculated at a high level of theory, $\mu(E)^{\mathrm{L}}$ is the spectrum computed at the lower level of theory and $\Delta(E)^{\mathrm{ML}}$ is the correction learned (see figure 15).

Figure 16 shows recent results obtained using the Δ-learning strategy by Watson *et al* [203] applied to the Rh L_3-edge. This work demonstrates that the Δ-learning strategy can quickly learn the difference between

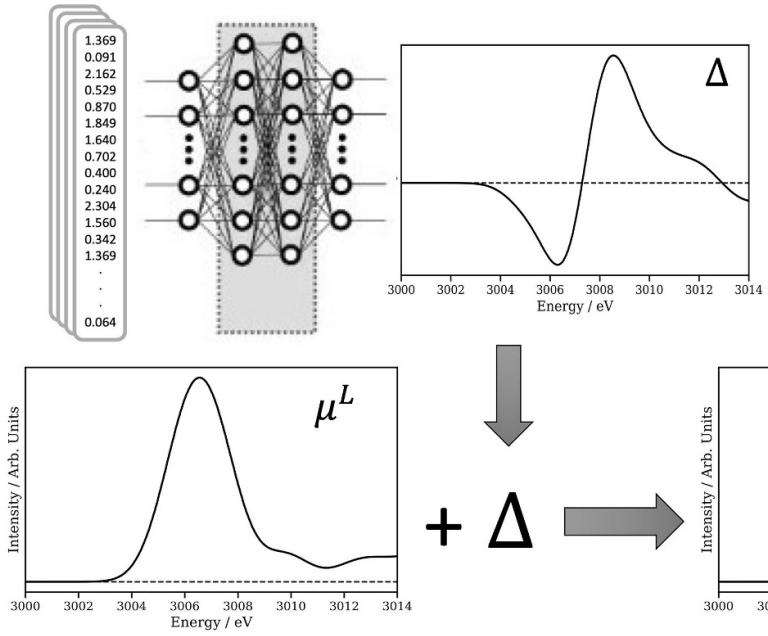

Figure 15. Schematic of the Δ-learning approach. The featurised local geometries are used in conjunction with the difference between their TDDFT(BLYP) and TDDFT(B3LYP) calculated spectra are outputs. Once optimised, the predicted difference is added to the TDDFT(BLYP) spectrum to recreate a spectrum equivalent to TDDFT(B3LYP). Reproduced from [203]. CC BY 4.0.

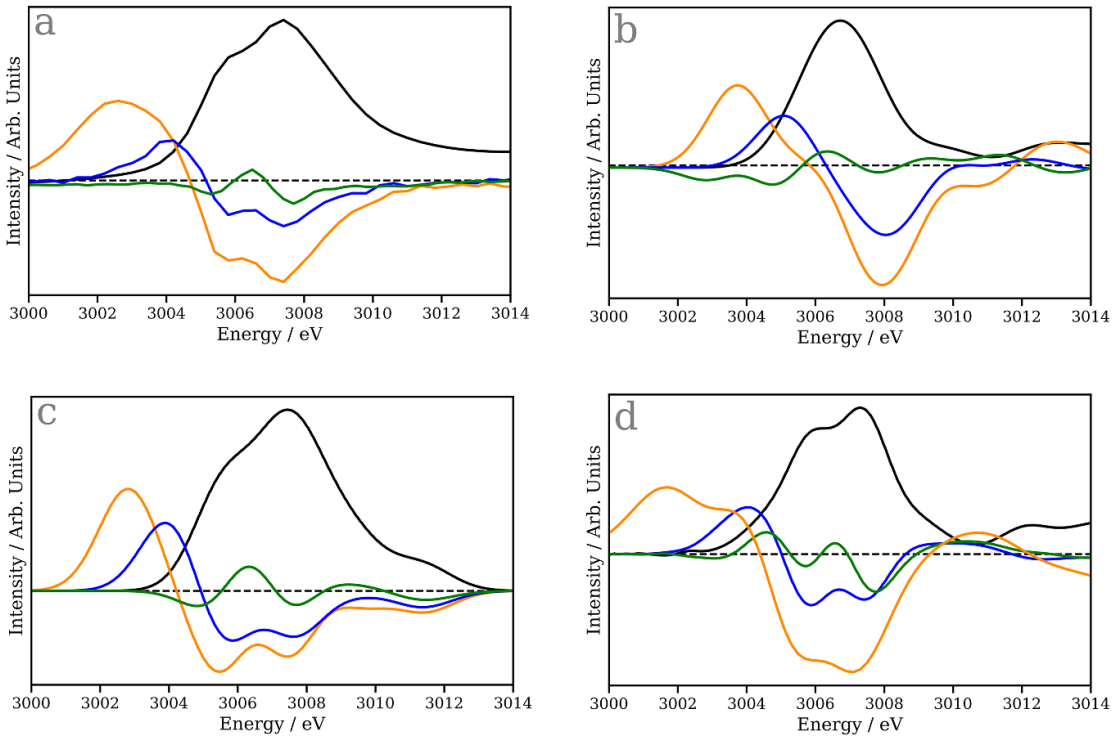

Figure 16. Ground state (black) and transient Rh L_3 spectra at 250 fs (orange), 10 ps (blue) and >190 ns (green) of cyclopentadienyl rhodium carbonyl, $CpRh(CO)_2$. (a) Experimental spectra reproduced from [208]. (b) Spectra calculated using TDDFT(BLYP), (c) TDDFT(B3LYP) and (d) Δ-Learning model. The dashed line represents zero intensity. Reproduced from [203]. CC BY 4.0.

TDDFT(BLYP) and TDDFT(B3LYP) computed spectra, providing a composite method for obtaining accurate core-hole spectra at reduced computational cost, as $\mu(E)^{\mathrm{H}}$ can be achieved using $\mu(E)^{\mathrm{L}}$ and the predicted $\Delta(E)^{\mathrm{ML}}$ from the developed model. The accuracy of this approach, shown in figure 16, is demonstrated by simulating Rh L_3-edge spectra tracking the C-H activation of octane by a cyclopentadienyl rhodium carbonyl complex [208], where we demonstrate the Δ-learning model can accurately reproduce the TDDFT(B3LYP) spectra at TDDFT(BLYP) cost.

Future developments in this area should prioritise the expansion of this approach, with particular attention to enlarging both the training dataset and the Δ, i.e. the disparity in quality between the lower and higher-level quantum chemistry methods. Indeed, in this respect, the p-DOS representation developed by Middleton *et al* [116] could be classed as a Δ-learning scheme as it involves translating an input closely related to the single particle spectrum within the dipole approximation to a higher-level of theory. This approach shows significant promise, although requires further testing across a broader profile of applications. Furthermore, considering that the Δ values aim to address deficiencies in the underlying theory linked to the lower-level methods, it may be feasible to discern patterns. In such instances, as exemplified in section 3.2, representing the Δs as a reduced number of principal components could streamline the network's operations.

5. Developing accurate training sets

The performance of any model will only ever be as good as the training data with which it is developed. Indeed, to replicate spectral features accurately, high-quality training data must be supplied which covers a representative proportion of the *feature space* of interest. As identified in the previous section, a significant challenge associated with developing accurate training sets is scale. The quantity of experimental data available is simply insufficient and, therefore, datasets are presently generated using theoretical calculations.

To date, ML models in x-ray spectroscopy have used available structures and spectra. However, to continue progress in this area, an increasing focus needs to be placed on developing accurate training sets. For those based on computational spectra, there are three main considerations: (i) the computational level of theory used, (ii) the sampling approach for choosing additional systems to include in the training set, and (iii) the training strategy. For the computational level of theory, the field has witnessed significant progress across the last decade [23], so it is now possible to calculate x-ray spectra using a hierarchy of methods meaning the principal bottleneck is scale, which could be addressed using efficient sampling and training strategies. There are myriad methods and examples in the literature to sub-sample and train the ML models [209–224], yet few have explored this avenue for developing training sets for ML in x-ray spectroscopy.

Figure 17 (grey points) shows the performance of the XANESNET MLP model developed for the Pt L_3-edge [110] as a function of training set size when using four different sampling techniques, namely random sampling (RS), furthest point sampling (FPS) [211], similarity-based sampling (SBL) [210] and uncertainty based active learning (AL) sampling [213]. The datasets and input files for this data are available at [45]. Starting from an initial training set of 1000 randomly selected samples, the first approach increases the training set size based upon randomly selecting additional samples, while the furthest-based sampling and similarity-based sampling calculate the Euclidean distance between the samples in the training set and add new samples based upon those furthest away or closest to the existing samples. Finally, uncertainty-based learning uses the bootstrapping technique (section 7) to establish and add samples that exhibit a large uncertainty and therefore are likely to be poorly represented within the training set. All examples exhibit a rapid decay, followed by a slower progress after ~4000 structure-spectra pairs. The furthest point sampling provides the lowest mean squared error when 16 000 samples have been added to the training set.

Beyond simply the size of the training set, the training strategy can influence the performance of the models. For the data in figure 17 (grey points), each point is essentially independent in the sense that it is derived from a model trained with that number of samples, without knowledge of previous models for smaller training sets. In contrast, figure 17 (black) shows the effect of using curriculum learning (CL) [215, 216] to train the models. CL is a strategy that aims at training an ML model from easier data to more complex data, which imitates the meaningful learning order in human curricula. For x-ray spectroscopy, it is not immediately apparent what constitutes an easy or difficult x-ray spectrum to learn [203]. Consequently, the curriculum is developed based on the sampling strategies discussed in the previous paragraph. Here each model is initiated using the optimal weights of the previous model and consequently, aligned with the fundamental idea behind CL, the complexity of the model, defined as the size of the training set, is gradually increased throughout training. By building upon the existing model, the subsequent models inherit the benefits and insights gained from the previous training iterations. The results of figure 17 (black points) show that all sampling methods have a distinct advantage with the CL approach, for samples greater than 4000, with the largest influence being observed for the uncertainty-based sampling. In addition to the significant improvement, this learning curve for the uncertainty-based sampling also retains a significant gradient at 16 000 samples suggesting that increasing the size of the training sets can still yield sizeable increases in performance.

Overall, this section has highlighted some of the strategies used in the literature to develop and refine training sets. While this is highlighted for a specific example, i.e. the Pt L_3-edge, these have been rarely

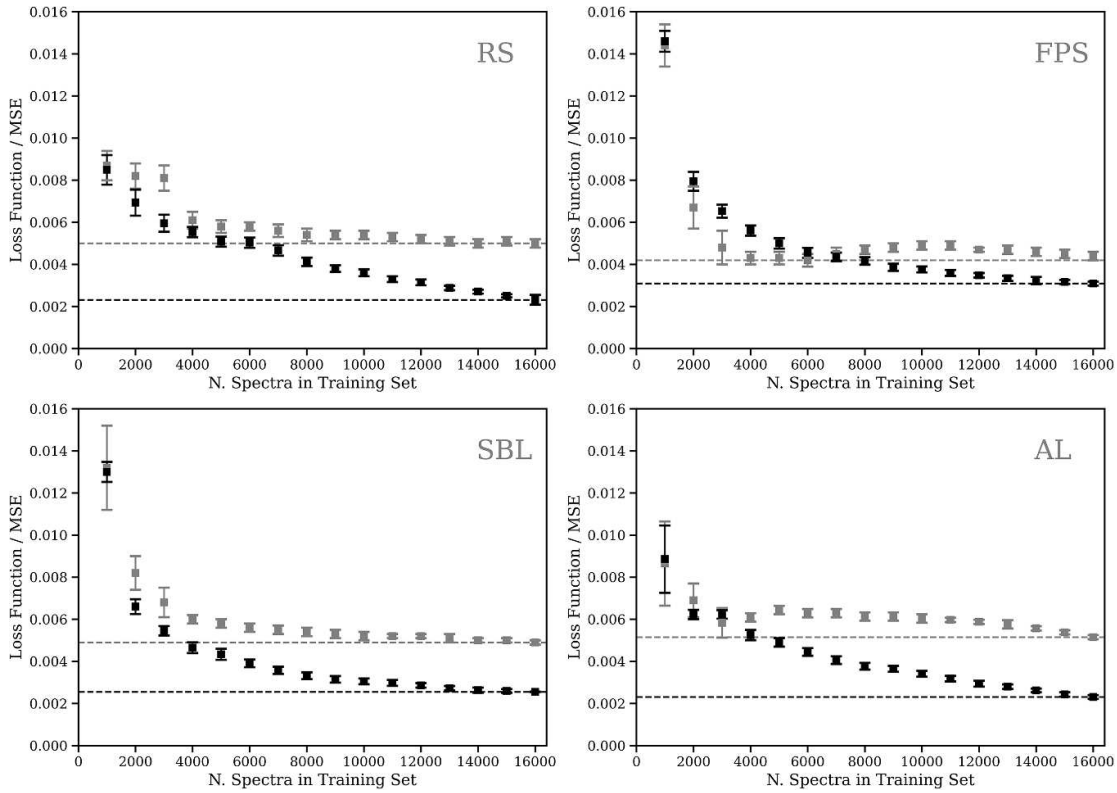

Figure 17. Performance of the DNN in terms of the loss function, MSE, of the DNN model as a function of training set size for the Pt L_2-edge training dataset with (black) and without (grey) using CL in the training process. The upper left, upper right, lower left, and lower right panels show the performance after sampling with Random Sampling (RS), Furthest point sampling (FPS), Similarity-based learning (SBL), and active learning (AL) respectively. All results presented are obtained from five-times-repeated five-fold cross-validation.

applied and/or investigated in detail for x-ray spectroscopy, and consequently this represents an area for development in this field.

6. Interpreting model behaviour

A key limitation associated with the use of ML models is that they are often used in a black-box manner and therefore the rationale behind predictions, i.e. spectral interpretation, is not obtained. As the fundamental, principal power and draw of ML algorithms, particularly deep networks, is that they are able to extract and encode higher-dimensional patterns and relationships within datasets which are non-trivial for human beings to perceive and interpret (i.e. they derive connections which naturally resist perception via human intuition), creating any digestible, rationalisable interpretation of the algorithm's behaviour, during either training or application, naturally presents a non-trivial challenge. The philosophical question of how to define trustworthy, cogent metrics of interpretability for any given machine algorithm is extant within ML generally, and remains a lively topic of general discussion [225]. Nevertheless, in the field of computational spectroscopy, one of the key objectives is not simply to provide a calculation that agrees with the experiment, but to permit a detailed interpretation of the peak assignments or the physical origin of changes observed between samples. Consequently, understanding and explaining the performance of a network, without the use of additional first-principles theoretical calculations is a key challenge. Indeed, for end users interpretability is important for the contextualisation of results and for developers, it provides a means to interrogate whether the models are getting the correct prediction for the 'right reasons'. Therefore, ML researchers for spectroscopy have implemented several such techniques in order to better enable informed decision-making and effectual application of ML models for developers and users.

Several such strategies have been created to make models interpretable, as discussed in [226, 227]. These can be divided into two groups, model-specific and model-agnostic (or model-independent) strategies. Methods can also provide local explanations, i.e. inform why a model has made a specific prediction. Alternatively, global explanations can inform, in a general sense, why a model behaves as it does. However, despite its importance, there are relatively few applications of interpretability applied within the context of x-ray spectroscopy [126, 127, 130, 132, 228].

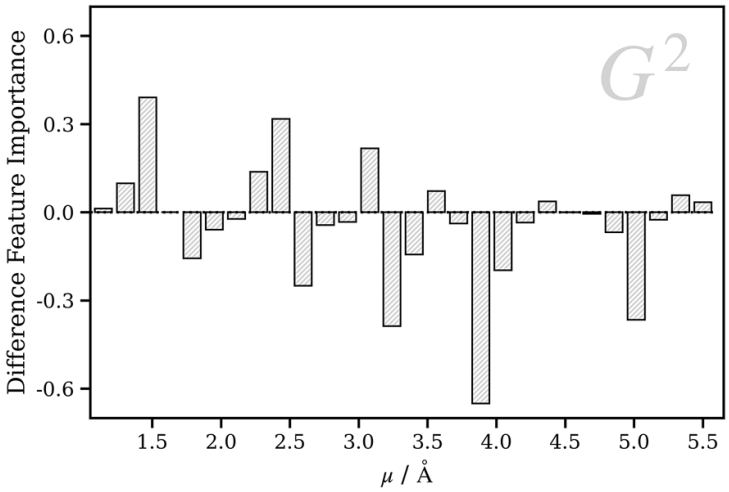

Figure 18. Relative importance of the G^2 wACSF terms as a function of distance from the absorbing atom for high- and low-energy spectral regions at the Fe K-edge. Positive differences indicate a stronger importance of these distances to high-energy spectral features, while negative differences indicate a stronger importance of low-energy spectral features. Reproduced from [92]. CC BY 4.0.

The simplest insight into performance can be measured by reducing the number of input features and assessing their influence on the performance of the model. For x-ray absorption this has been achieved using the action of a variance threshold filter, i.e. removing the features in order of which demonstrates the least variance when averaged over the whole training set [92]. This will provide a global insight into the importance of specific input features but provide limited insight for individual predictions. A similar global perspective can be obtained using relative feature importance, assessed via scrambling the values of each input feature over the reference dataset and assessing the performance penalty. Using this approach, figure 18 assesses the difference feature importance as a function of distance from the absorbing atom for high- and low-energy spectral regions at the Fe K-edge [92]. Indeed, if the difference feature importance is positive, it indicates that this region is more important for the high-energy spectral region. In contrast, if the difference feature importance is negative the distance is more important for the low-energy spectral region. Figure 18 displays a general shift in the difference from positive to negative values as the distance from the absorbing atom is increased illustrating that atoms closest to the absorbing atom are more important at the high energy region, while the low energy region has a larger *field-of-view*. Crucially, this aligns with the underlying physics: i.e. when core photoelectrons have low energy near the absorption edge they exhibit longer wavelengths, and consequently this spectral region is more responsive to structural features that are located farther from the x-ray absorption site. Conversely, in the higher-energy region, photoelectrons have greater kinetic energy, leading to shorter wavelengths, and reducing the range of structural information they can yield [10].

Although feature importance can offer insights, they can be misleading. A key challenge is that if high levels of correlation exist between input features, if a feature is removed from the model, it may be compensated for by a correlated feature, thus masking the true level of importance of the feature. An alternative approach is Shapley analysis based on the SHAP method [30], which can also provide local explanations, i.e. explain each prediction from the model. However, it should be stressed that this approach does not remove the challenge of correlation between features. To illustrate this approach, figure 19 shows the absolute SHAP feature importance (black), for predictions of the *held-out* Fe K-edge training set [161]. These are compared to the curved waved amplitude of 2 body multiple scattering pathways extracted from the FEFF software (grey). Overall, there is broad agreement between the two, consistent with a model mimicking the correct physics. This analysis is promising, but presently only includes the two body terms and therefore future extensions should incorporate the influence of high-body MS expansions, which are known to be important in the XANES spectra. In addition, a detailed analysis of how this interpretation is correlated to the quality of spectral prediction should be established.

The aforementioned approaches have focused on structural representations. Recently Kotobi *et al* [126] used a graph neural network combined with feature attribution to deliver interpretation in terms of a linear combination of core-to-valence orbital transitions, comparable to information that arises from a quantum chemistry calculation. Figure 20 shows an example of these attributions, which inform the interpretation of each peak in terms of both the core (especially important when multiple absorbers contribute to the same

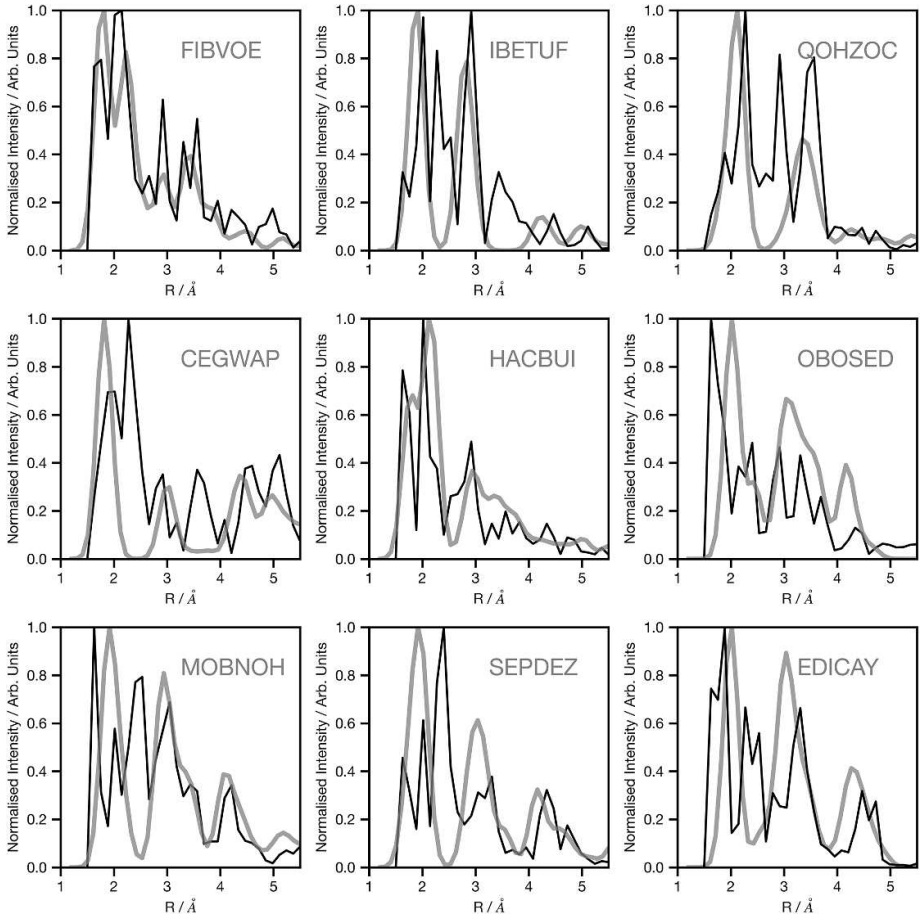

Figure 19. Normalised G^2 wACSF feature importance arising from the Shapley analysis (grey) and curved waved amplitude of 2 body multiple scattering pathways extracted from FEFF simulations. The upper three panels show K-edge XANES spectra from the 0th–10th percentiles, i.e. the best performers when the held-out set is ranked by MSE. The center 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.

space) and valence orbitals from which the transition derives. This work demonstrated excellent agreement between core orbitals and spectral peaks, and although the performance slightly declined with valence orbital assignment, the results remain highly promising for incorporating explainability into ML models, which enables end-users to access insight into the physical origin of spectroscopic predictions.

Finally, it has been proposed that the attention mechanism [229], increasingly populated in modern ML architectures, could potentially be used to provide interpretations visualisation of the attention weights that have been used to interpret the performance of the model [230, 231]. However, some studies argue this is not the case [232, 233] and therefore further work is required, especially in x-ray spectroscopy, where such interpretation has not yet been applied.

7. Quantifying uncertainty

Accurate ML models are beginning to open up new possibilities to accelerate analyses in x-ray spectroscopy while, through taking advantage of the recent developments outlined in section 6, simultaneously also providing insight into interpretation. ML model performance remains nonetheless dependent on the quality of the data that the ML models are trained with and, consequently—unless the training data cover as completely as possible the relevant chemical space. Poor performance inevitably arises in some cases. The ability to quantify accurately the uncertainty in ML model predictions is valuable, especially when provided as a metric for (non-expert) users who may not be familiar with the (limited) coverage of the training data. Fortunately, a number of approaches and metrics are available for quantifying uncertainty; in the domain of chemical ML, examples can be found for, e.g. the design of experiments used to synthesise nanoparticles, [234–236] the optimisation of the mechanical properties of materials, [237, 238] and, more generally, in the space of molecular property prediction [239].

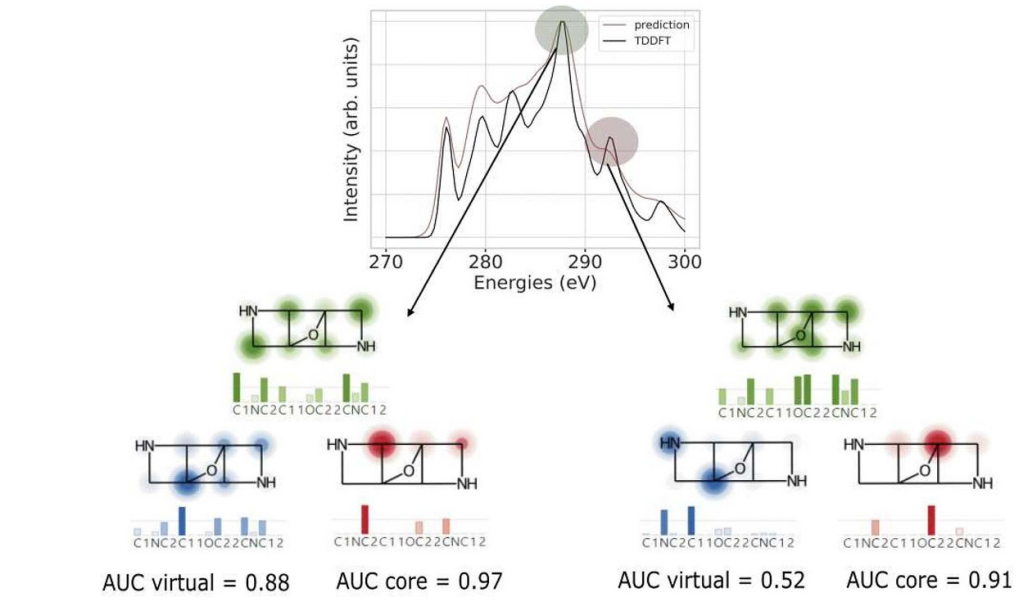

Figure 20. Attributions (green) are compared with the ground truth of core (red) and virtual (blue) orbitals via area under the curve (AUC) values for two peaks of an XAS spectrum predicted by the GATv2model in [126]. The model has higher AUC values when a peak in the predicted spectrum follows the TDDFT result. Reproduced from [126]. CC BY 4.0.

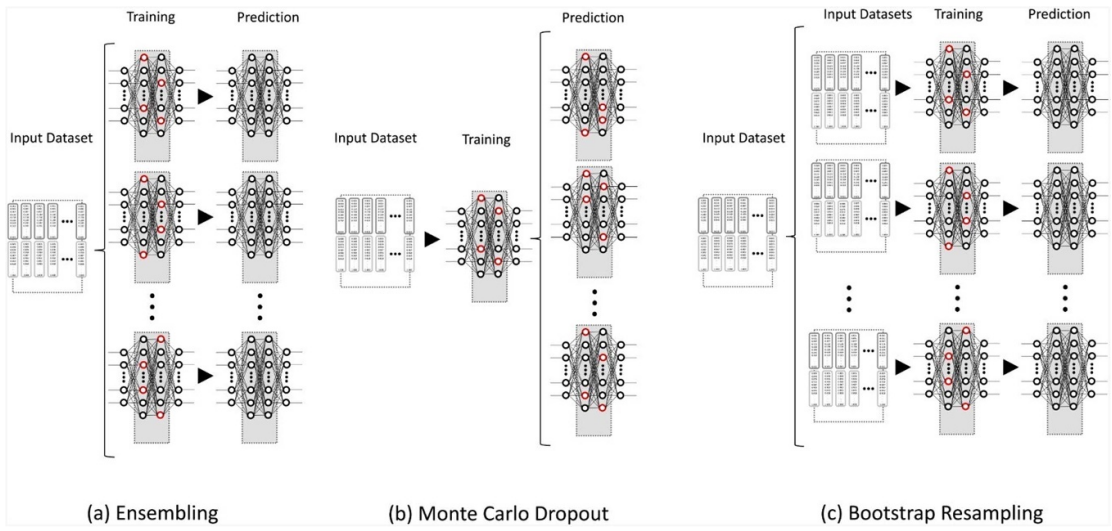

Figure 21. Schematic of simple approaches for quantifying uncertainty in ML model predictions: (a) ensembling, (b) Monte-Carlo dropout, and (c) bootstrap resampling as applied to a neural network. The red circles indicate neurons that are dropped out of the neural network. Reproduced from [108]. CC BY 4.0.

Uncertainty in ML model predictions arises principally in two forms: *aleatoric* and *epistemic* [240]. The former (aleatoric uncertainty) arises from incomplete training data, i.e. an ML model is used to produce a prediction for an input outside the scope of the training dataset; the latter (epistemic uncertainty) arises from model variability in the sense that there are multiple (similar) solutions to the task of optimising the ML model weights and this introduces a degree of variability into the ML model that is built even when exposed to the same training dataset. To attempt to address and quantify uncertainty, three approaches have been applied in the domain of x-ray spectroscopy: (i) ensembling, [108, 147], (ii) Monte-Carlo dropout, [108] and (iii) bootstrap resampling [108, 161]. All of these approaches are shown schematically in figure 21.

Ensembling (figure 21(a)) is discussed in the context of x-ray spectroscopy in [108, 147]; principally, it involves the optimisation of multiple ML models using the same training dataset. Although each ML model in the ensemble learns from the same data, each is instantiated probabilistically with a different set of initial internal weights before learning, and the outcome is that the optimal internal weights of the trained ML models all vary slightly. From the ensemble of probabilistically-instantiated ML models, the mean prediction

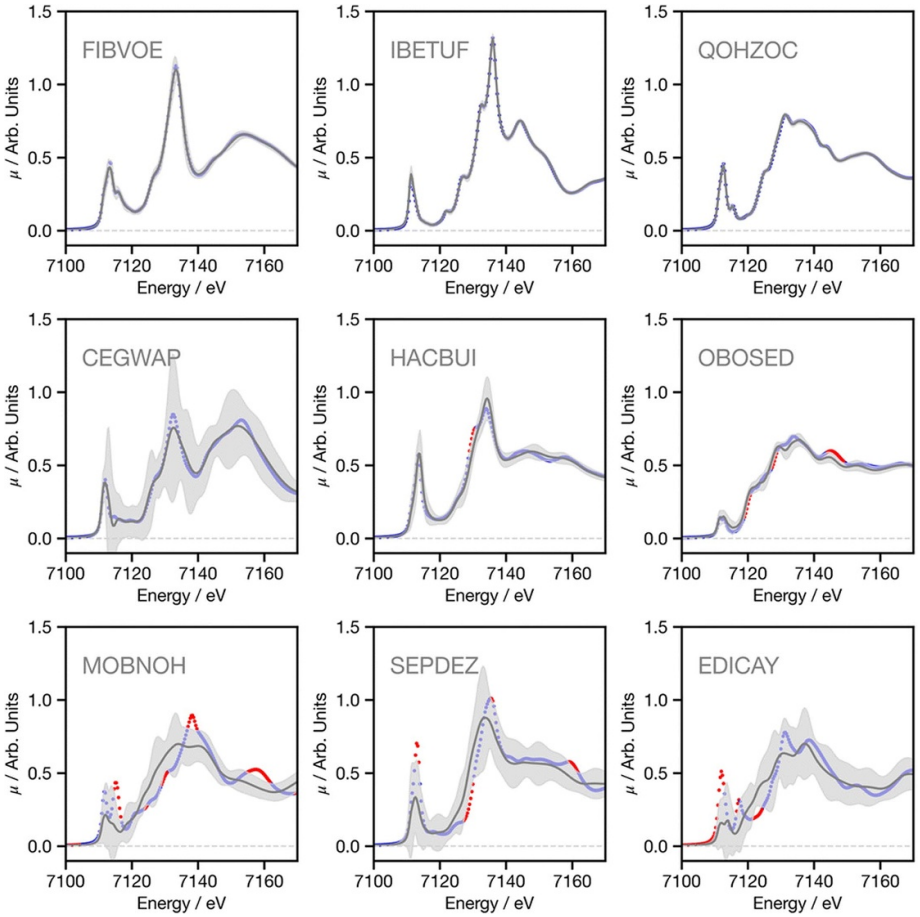

Figure 22. Example K-edge XANES spectra for Fe-containing samples, where the solid grey is the average prediction and the light grey shaded region represents $\pm 3\sigma$. The ground truth spectrum is shown using the blue points, which become red when the ground truth is outside the $\pm 3\sigma$ of the predicted spectrum. Reproduced from [161]. CC BY 4.0.

and standard deviation can be derived. Consequently, the ensembling method supplies metrics which are able to quantify the uncertainty arising from intrinsic *model uncertainty*, and therefore also quantify *epistemic* uncertainty. Monte-Carlo dropout (figure 21(b)) exploits probabilistic dropout during prediction, where the model variability derives from the use of dropout during prediction in addition to during training [108]. Finally, bootstrap resampling [108, 161] (figure 21(c)) serves as a method for estimating statistics on a population by repeatedly drawing samples from a dataset, with replacement of samples at each repetition. The advantages of this approach are most clearly observed when characterising the uncertainty associated with incomplete training data. The bootstrap-sampled reference datasets, which are of the same size as the original and therefore will contain duplicates, introduce dataset diversity to each instance of the ML model and consequently, the multiple models used can again be used to predict the mean prediction and standard deviation of spectral predictions.

Figure 22 exemplifies the performance of the uncertainty quantification at the Fe K-edge [161]. This clearly shows that uncertainty increases as the quality of the predictions decreases, especially for the lowest three panels. Indeed, in [161] the authors showed that $\pm 3\sigma$ from the predicted spectrum covered >90% of the points in the truly calculated spectra and therefore could be reliably used to assess the quality of any prediction. Importantly, consistent with previous work [147], the model also exhibited a slight underconfidence, in that it was more likely to provide a large uncertainty for the good prediction than vice versa. Underconfidence was most commonly observed when linear bonds, such as CO or CN were present in the sample. This clearly highlights a limitation of the model for capturing the well established focusing effect on x-ray spectra.

8. Applications: interpretation of disorder and time-resolved experiments

In this section, we will explore the performance of the ML methodologies discussed above through a curated selection of case studies. The advantage of ML methodologies in x-ray spectroscopy is most obvious when a large number of computational x-ray spectroscopic simulations are required to describe satisfactorily the

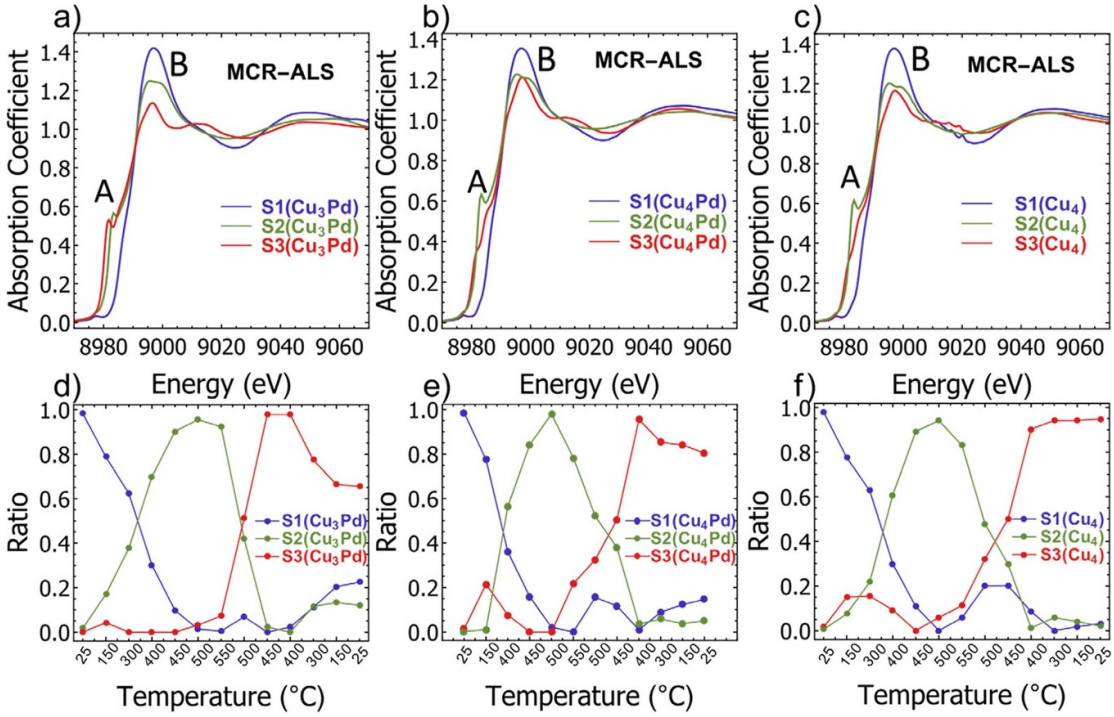

Figure 23. MCR analyses of (a) Cu_3Pd, (b) Cu_4Pd, and (c) Cu_4 resolved into pure species. Temperature-dependent concentration profiles obtained by MCR for (d) Cu_3Pd, (e) Cu_4Pd, and (f) Cu_4. Reprinted with permission from [258]. Copyright (2021) American Chemical Society.

system under study: the clearest examples of such cases are when the x-ray spectra contain dynamical information, either as a consequence of the intrinsic disorder of the system under study or during time-resolved studies in which dynamics are (photochemically) induced. Indeed, such studies often require a large number of configurational 'snapshots' [241–253] of the system to be sampled [e.g. from a molecular dynamics (MD) simulation] to describe adequately the x-ray spectrum. This is a time- and resource-intensive task for first-principles simulations but can be addressed using ML algorithms [130, 145, 146, 156, 162, 174, 254–257] that alleviate the bottleneck associated with computing the x-ray spectra for all of the sampled configurations.

8.1. Dynamics of size-selected Cu_xPd_y clusters during catalysis

Size-selected clusters are important model catalysts and establishing structure-activity relationships for such species is a key step towards mechanistic understanding. In [258], Liu *et al* studied propane oxidation reactions using size-selected Cu_xPd_y clusters. Interpretation of experiments like those carried out by Liu *et al* is often challenging owing in part to the small sizes of the clusters and in part to the continuous structural changes occurring under reaction (*e.g. operando*) conditions. In this work, the authors used multivariate curve resolution (MCR) analysis to identify the different phases (figures 23(a)–(c)) of each cluster and to quantify their concentration under *operando* conditions as a function of temperature (figures 23(d)–(f)).

Liu *et al* [258] further developed a CNN to predict the coordination numbers of the clusters which, given their small sizes, can be conveniently connected to their structure [80]. Their CNN was trained upon calculated spectra, obtained using MS calculations as implemented within the FEFF package. For this specific case, the authors were able to demonstrate a strong agreement between calculated and experimental spectra which enhances the accuracy of the network. Based upon this approach, the authors were able to extract the chemical states and compositions of the clusters, along with information about their structures, which could be correlated to their catalytic activity and selectivity.

8.2. Structural changes during reduction of polyoxometalates

Owing to their ability to store reversibly multiple electrons, polyoxometalates (POMs) [259] are appealing materials for the electrochemical storage of energy and, consequently, have been both employed in redox flow batteries [260] and as an alternative to carbon-based cathodes in molecular cluster batteries [261]. To improve the performance of energy storage materials like these, it is crucial to understand the electronic and geometric structural properties that govern their redox behavior. Figure 24 shows a comparison between an experimental [262] and DNN-predicted Mo K-edge XANES spectrum of $\mathbf{PMo_{12}}^{3-}$, [145] and also presents

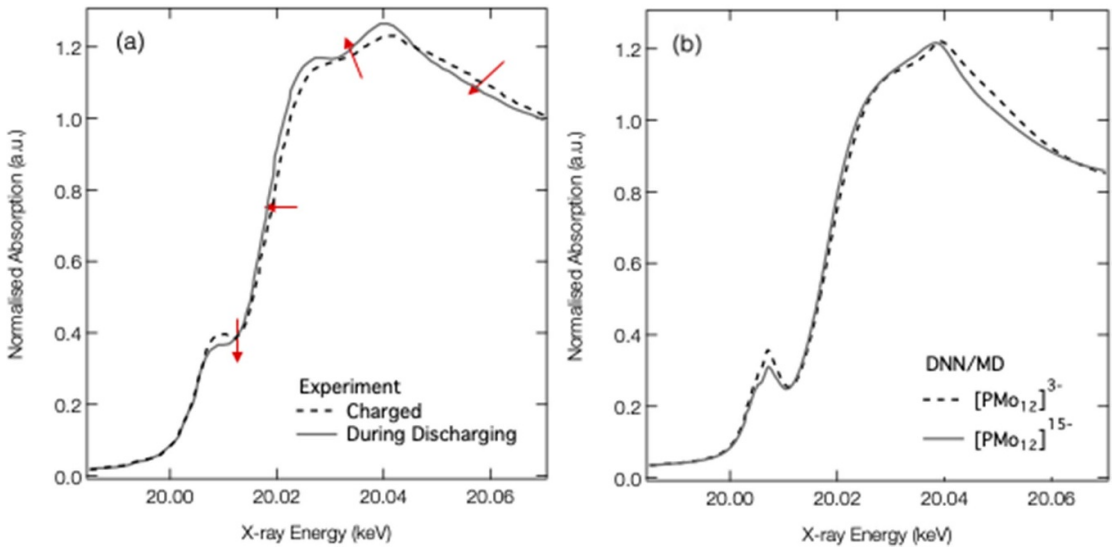

Figure 24. (a) Experimental (Wang *et al* [262]) Mo K-edge XANES spectrum of the charged $\mathbf{PMo_{12}}^{3-}$ POM (dashed) and *operando* (Wang *et al* [262]) Mo K-edge XANES spectrum recorded during discharge with a cell voltage of 2.5 V (solid). (b) DNN-predicted ensemble-averaged Mo K-edge XANES spectra for $\mathbf{PMo_{12}}^{3-}$ (dashed) and $\mathbf{PMo_{12}}^{15-}$ (solid) obtained using a nuclear ensemble of 65 000 QM/MM MD snapshots. Reprinted from [262], Copyright (2021), with permission from Elsevier.

the *in-operando*-acquired Mo K-edge XANES spectrum during active discharging [262]. These x-ray spectra are challenging to model computationally for a number of reasons, not least because of the dynamic nature of POMs, coupled with their strong interaction with the solvent environment [263]. As such, configurational sampling via MD simulation provides the most statistically-reliable insight into their (ensemble-averaged) properties and their x-ray spectra are most appropriately calculated computationally by sampling configurations obtained through these MD simulations (i.e. under the nuclear ensemble approximation).

The key spectral changes accompanying discharge of $\mathbf{PMo_{12}}^{3-}$ (figure 24(a)) can be summarised in four points: (i) a loss in pre-edge intensity associated with the elongation of the Mo-O bond distances; (ii) a red shift of the x-ray absorption edge; (iii) an increase in the white-line intensity; (iv) and a loss of intensity in the spectral feature around *ca.* 20.06 keV. As Falbo *et al* discuss in [145], the DNN/MD ensembling approach reproduces all of the key features observed in the experimental Mo K-edge XANES spectrum; this is not the case if one computes the Mo K-edge XANES spectrum using only a single indicative equilibrium structure (i.e. without configurational averaging via the nuclear ensemble approximation). The red shift of the x-ray absorption edge is associated with the reduction of the Mo sites in $\mathbf{PMo_{12}}^{3-}$, the consequence of which is a lowering of the binding energy of the core electrons. The decrease in pre-edge intensity is a response to the lengthening of the Mo-O bonds in $\mathbf{PMo_{12}}^{3-}$, and the tendency of the O-Mo-O angles to adopt a more right-angular geometry brings the (local) coordination environment around each Mo x-ray absorption site closer to C_{4v} symmetry, leading to a commensurate decrease in $4d/5p$ orbital mixing. Surprisingly, despite strong solute-solvent interactions, the explicit modelling of the environment (e.g. the presence of Li^+ and the solvent) has no great effect on the Mo K-edge XANES spectra.

8.3. NN-EXAFS reveals oxygen evolution reaction (OER) mechanism of $Co_xFe_{3-x}O_4$ materials

Bimetallic transition metal oxides such as spinel-like $Co_xFe_{3-x}O_4$ materials are attractive catalysts for the OER in alkaline electrolytes. However, there remains work to be done towards understanding the catalytically-active state of these $Co_xFe_{3-x}O_4$ materials; such information is crucial to guide the design and development of further-improved catalysts. In [256], Timoshenko *et al* applied *operando* quick EXAFS (QEXAFS) at the Co K-edge to study the structural changes and phase transitions taking place in these $Co_xFe_{3-x}O_4$ catalysts under operational conditions. The authors performed PCA analysis of the Co K-edge x-ray spectra over the whole time domain and reported that only four principal component vectors were sufficient to describe the entire dataset. Using the distinct differences between the principal component vectors, the authors were able to propose structural/chemical changes consistent with their observation.

To support their analysis, the authors also used the NN-EXAFS [80, 264] approach (figure 25) to investigate the evolution of the (local) structure (including concentration, coordination numbers, and bond lengths) around the Co x-ray absorption sites during active catalysis. This NN is developed using calculated data and applied to experimental data to extract the aforementioned spectral details. Their NN-EXAFS indicated that the local structural environment around the tetrahedral Co^{2+} sites could be characterised as a

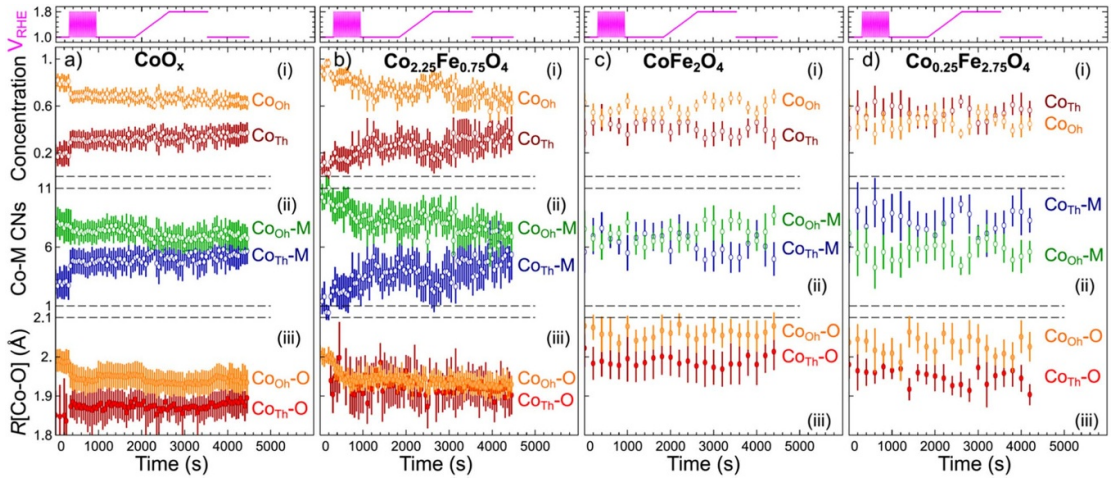

Figure 25. Evolution of the (local) structure parameters associated with Co x-ray absorption sites in (a) CoO_x, (b) $Co_{2.25}Fe_{0.75}O_4$, (c) $CoFe_2O_4$, and (d) $Co_{0.25}Fe_{2.75}O_4$ under oxygen evolution reaction conditions (*e.g. in operando*) as obtained using QEXAFS. The (local) structural parameters were extracted by integrating PDFs obtained via an NN-EXAFS analysis. Reproduced from [256]. CC BY 4.0.

disordered spinel-like structure. Alongside their PCA analysis, the authors concluded that these catalysts exhibit a segregated structure in which an Fe-rich but electrochemically-passive phase coexists with a catalytically-active Co-rich phase. For the latter, NN-EXAFS demonstrated the formation of active sites exhibiting Co^{3+} octahedra. Besides the significant catalytic insight, this work highlighted the strength of the NN-EXAFS approach and its suitability for aiding the interpretation of dynamical data from disordered samples *in operando*: indeed, the QEXAFS experiment provides a large quantity of data, presenting a particular challenge for traditional analyses, yet a NN-EXAFS analysis can be carried out within seconds (after the neural network has been trained).

8.4. On-the-Fly deep neural networks for simulating time-resolved spectroscopy

ML can also be applied to dynamical data acquired on a much faster timescale, opening up the possibility for ML-aided interpretation of experiments using ultrashort and ultrabright x-ray pulses at X-FELs. Figure 26 shows the performance of a DNN applied in such a case to a proposed time-resolved x-ray experiment [162]. Middleton *et al* [162] used a DNN to simulate the experimental S K-edge XANES signal using excited-state MD simulations of the ring-opening mechanism of the small cyclic disulphide 1,2-dithiane [265]. The DNN was trained on-the-fly from first-principles computational data with a train-test process that was repeated through the timesteps of the excited-state MD simulation until such a time as the predicted S K-edge XANES spectra could be produced with sufficient accuracy to replace the computationally-intensive quantum chemical calculations. Middleton *et al* demonstrated that *ca.* 100 fs of excited-state MD simulation provided sufficient first-principles computational data to train the DNN which was then able to predict accurately and affordably the S K-edge XANES spectra at future (i.e. unseen) times.

Figures 26(a) and (b) show a comparison between the calculated and DNN-predicted S K-edge XAS spectra from 110 fs (i.e. the time that the DNN was trained up until) to 900 fs. There is good agreement between the two, and the DNN captures the periodic behavior observed in both spectral bands which are associated with changes in electronic state populations and changes in the S-S internuclear distance. Figures 26(c) and (d) present a more detailed evaluation of performance by illustrating the ability of the DNN to produce predictions reliably for an individual trajectory. This scenario exhibits more pronounced shifts in the predicted spectra compared to predictions on the entire ensemble of trajectories (figures 26(a) and (b)), which is attributed to the incoherent nature of the extended temporal dynamics of 1,2-dithiane. Despite the higher resolution of the spectra for the single trajectory, there is still a notable accord between the computed and predicted spectra.

For this test case, it is clear that the use of structural and spectral data up until 110 fs is sufficient to train the DNN as the (photo)dynamics of dithiane after S-S bond fission are principally expressed within this period. Through analysis of the magnitude and positions of the spectral features alongside the geometric distortions, the authors showed that the majority of the geometric and spectral (i.e. input and output) space has been traversed by the 110 fs mark, facilitating the DNN to predict the x-ray spectra for future times, where most geometries fall within this space. However, for future applications where the selection of training data required to achieve convergence may be less intuitive, it has been demonstrated that the ensemble

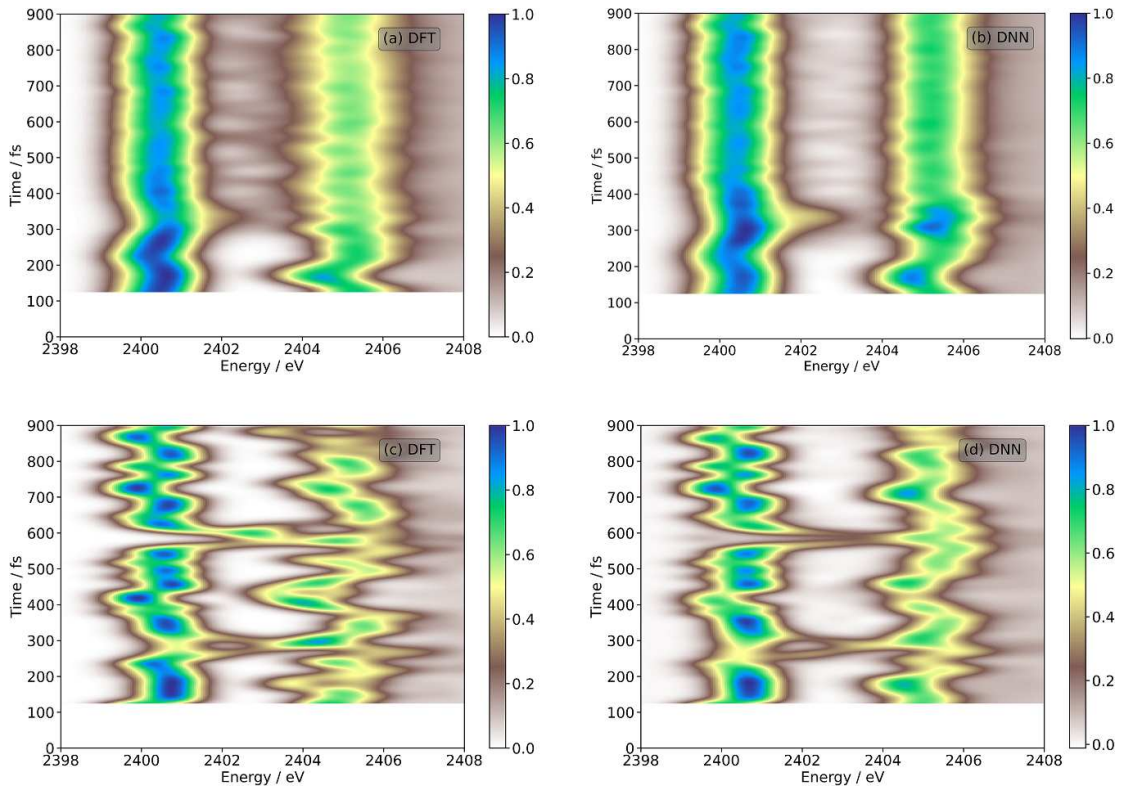

Figure 26. Time-resolved S K-edge XANES spectra associated with the ultrafast excited-state ring-opening dynamics of 1,2-dithiane. (a) Calculated using REW-TDDFT at the DKH-BP86/def2-TZVP level. (b) Predicted using the XANESNET DNN. Plots are shown for unseen data from 120 fs onwards; the XANESNET DNN was trained on the data from all timesteps up until 120 fs. Example spectra for a specific trajectory calculated using REW-TDDFT at the DKH-BP86/def2-TZVP level and predicted using the DNN are shown in (c) and (d), respectively. Reproduced from [162]. CC BY 4.0.

approach to quantify DNN uncertainty (as described in section 7) can be employed to gauge the performance. The ensembling technique presents a robust method for determining whether the DNN is trained satisfactorily to produce reliably the key features of the x-ray spectra.

9. Conclusions

Rapid advances in instrumentation and experimental methodology, coupled with increasing data acquisition rates and ever-improving spectral and spatiotemporal resolution, have pushed the envelope considerably in x-ray spectroscopy, transforming the technique beyond recognition. These developments have not only widened the accessibility and applicability of x-ray spectroscopy but have enabled novel experiments utilising the ultrashort and ultrabright x-ray pulses available at X-FELs. Underpinning the qualitative and quantitative interpretations of the experimental data, computational spectroscopy has become an increasingly important tool to complement these experiments and has, in itself, been driven forward in response to the challenges presented by experimental developments. While computational x-ray spectroscopy has, to date, focused primarily on the development of ever-faster and ever-better first-principles computational chemical techniques, machine-learning methods are beginning to emerge and expand the scope and reach of data analysis.

In this Topical Review, we have detailed recent developments in machine-learning methods for computational x-ray spectroscopy, exploring each step of the workflow from the underpinning theory which the machine-learning models are tasked with replicating to the preparation of the datasets and optimisation of the models, the interpretation of their outputs, and the quantification of their uncertainties. It is clear that recent research efforts in this space have led to significant progress; machine-learning approaches are now capable of '*forward*' (structure → spectrum) and '*reverse*' (structure ← spectrum) mappings between structure and x-ray spectroscopic observables across multiple x-ray absorption edges, elements, and experiments.

While these works illustrate the significant progress achieved, they also highlight extensive opportunities to enhance the application of ML techniques for x-ray spectroscopy. In particular, for forward-mapping ML models, a need remains to develop accurate training sets that cover chemical space satisfactorily to enable the

ML models that make use of them to be applicable with genuine generality across a broad range of practical problems. However, these training sets also need to be able to capture spectral trends associated with minor structural changes if these ML models are also to be used for the fitting of experimental x-ray spectra. For reverse-mapping ML models, the key challenge relates to identifying and handling appropriately the mismatches (arising from limitations of the underlying theory) between the theoretical x-ray spectra used during training and the experimental x-ray spectra to which these ML models are most usefully applied. Progress in each of these domains will significantly increase the quantity and quality of information that can be extracted from experimental spectra using forward- and reverse-mapping ML models, providing unparalleled support for direct experimental data analysis in x-ray spectroscopy.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://gitlab.com/team-xnet.

Acknowledgments

The research described in this paper was funded by the Leverhulme Trust (Project No. RPG-2020-268) and EPSRC (Grant Nos. EP/S022058/1, EP/R021503/1, EP/R51309X/1, EP/X035514/1, EP/X026973/1 and EP/W008009/1).

Conflict of interest

The authors declare no competing financial interests.

Code details

Accompanying materials including software, training and testing sets, and tutorials are freely available at: https://gitlab.com/team-xnet.

ORCID iDs

Thomas Penfold https://orcid.org/0000-0003-4490-5672
Luke Watson https://orcid.org/0000-0003-1255-8055
Thomas Pope https://orcid.org/0000-0001-7552-9812
Julia Kaczmarek https://orcid.org/0009-0004-6922-0447
Conor Rankine https://orcid.org/0000-0002-7104-847X

References

[1] Greczynski G and Hultman L 2020 X-ray photoelectron spectroscopy: towards reliable binding energy referencing *Prog. Mater. Sci.* **107** 100591
[2] Hess C 2021 New advances in using Raman spectroscopy for the characterization of catalysts and catalytic reactions *Chem. Soc. Rev.* **50** 3519–64
[3] Sá J 2014 *High-Resolution XAS/XES: Analyzing Electronic Structures of Catalysts* (CRC Press)
[4] Mukamel S *et al* 2020 Roadmap on quantum light spectroscopy *J. Phys. B: At. Mol. Opt. Phys.* **53** 072002
[5] Barone V, Alessandrini S, Biczysko M, Cheeseman J R, Clary D C, McCoy A B, DiRisio R J, Neese F, Melosso M and Puzzarini C 2021 Computational molecular spectroscopy *Nat. Rev. Methods Primers* **1** 38
[6] Puzzarini C, Bloino J, Tasinato N and Barone V 2019 Accuracy and interpretability: The devil and the holy grail. new routes across old boundaries in computational spectroscopy *Chem. Rev.* **119** 8131–91
[7] Van Bokhoven J A and Lamberti C 2016 *X-ray Absorption and X-ray Emission Spectroscopy: Theory and Applications* vol 1 (Wiley)
[8] Rehr J J, Kas J J, Vila F D, Prange M P and Jorissen K 2010 Parameter-free calculations of x-ray spectra with FEFF9 *Phys. Chem. Chem. Phys.* **12** 5503–13
[9] Milne C J, Penfold T J and Chergui M 2014 Recent experimental and theoretical developments in time-resolved x-ray spectroscopies *Coord. Chem. Rev.* **277** 44–68
[10] Penfold T J, Milne C J and Chergui M 2013 Recent advances in ultrafast x-ray absorption spectroscopy of solutions *Adv. Chem. Phys.* **153** 1–41
[11] Hwu Y and Margaritondo G 2021 Synchrotron radiation and x-ray free-electron lasers (x-fels) explained to all users, active and potential *J. Synchrotron Radiat.* **28** 1014–29
[12] Hastings J, Pellegrini C and Marinelli A 2020 *Physics of and Science with X-ray Free-Electron Lasers* vol 199 (IOS Press)
[13] Yang F, Xuefei Feng Y-S L, Li Cheng Kao P-A G, Yang W and Guo J 2021 In situ/operando (soft) x-ray spectroscopy study of beyond lithium-ion batteries *Energy Environ. Mater.* **4** 139–57
[14] Liu X, Wang D, Liu G, Srinivasan V, Liu Z, Hussain Z and Yang W 2013 Distinct charge dynamics in battery electrodes revealed by in situ and operando soft x-ray spectroscopy *Nat. Commun.* **4** 2568

[15] Xue Z, Jizhou Li, Pianetta P and Liu Y 2022 Data-driven lithium-ion battery cathode research with state-of-the-art synchrotron x-ray techniques *Acc. Mater. Res.* **3** 854–65
[16] Lin F *et al* 2017 Synchrotron x-ray analytical techniques for studying materials electrochemistry in rechargeable batteries *Chem. Rev.* **117** 13123–86
[17] Chen M, Chou S-L and Dou S-X 2019 Understanding challenges of cathode materials for sodium-ion batteries using synchrotron-based x-ray absorption spectroscopy *Batter. Supercaps* **2** 842–51
[18] Wernet P 2019 Chemical interactions and dynamics with femtosecond x-ray spectroscopy and the role of x-ray free-electron lasers *Phil. Trans. R. Soc.* A **377** 20170464
[19] Katayama T *et al* 2023 Atomic-scale observation of solvent reorganization influencing photoinduced structural dynamics in a copper complex photosensitizer *Chem. Sci.* **14** 2572–84
[20] Britz A *et al* 2020 Resolving structures of transition metal complex reaction intermediates with femtosecond exafs *Phys. Chem. Chem. Phys.* **22** 2660–6
[21] Attar A R, Aditi Bhattacherjee C D P, Schnorr K, Closser K D, Prendergast D and Leone S R 2017 Femtosecond x-ray spectroscopy of an electrocyclic ring-opening reaction *Science* **356** 54–59
[22] Garratt D *et al* 2022 Direct observation of ultrafast exciton localization in an organic semiconductor with soft x-ray transient absorption spectroscopy *Nat. Commun.* **13** 3414
[23] Rankine C D and Penfold T J 2021 Progress in the theory of x-ray spectroscopy: From quantum chemistry to machine learning and ultrafast dynamics *J. Phys. Chem.* A **125** 4276–93
[24] Wenzel J, Holzer A, Wormit M and Dreuw A 2015 Analysis and comparison of CVS-ADC approaches up to third order for the calculation of core-excited states *J. Chem. Phys.* **142** 214104
[25] Wenzel J, Wormit M and Dreuw A 2014 Calculating core-level excitations and x-ray absorption spectra of medium-sized closed-shell molecules with the algebraic-diagrammatic construction scheme for the polarization propagator *J. Comput. Chem.* **35** 1900–15
[26] Wenzel J, Wormit M and Dreuw A 2014 Calculating x-ray absorption spectra of open-shell molecules with the unrestricted algebraic-diagrammatic construction scheme for the polarization propagator *J. Chem. Theory Comput.* **10** 4583–98
[27] Yu Sokolov A 2018 Multi-reference algebraic diagrammatic construction theory for excited states: general formulation and first-order implementation *J. Chem. Phys.* **149** 204113
[28] Coriani S and Koch H 2015 Communication: x-ray absorption spectra and core-ionization potentials within a core-valence separated coupled cluster framework *J. Chem. Phys.* **143** 181103
[29] Coriani S and Koch H 2016 Erratum: Communication: x-ray absorption spectra and core-ionization potentials within a core-valence separated coupled cluster framework *J. Chem. Phys.* **145** 149901
[30] Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Advances in Neural Information Processing Systems* p 30
[31] Besley N A 2021 Modeling of the spectroscopy of core electrons with density functional theory *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **11** e1527
[32] Besley N A 2020 Density functional theory based methods for the calculation of x-ray spectroscopy *Acc. Chem. Res.* **53** 1306–15
[33] Chen Z *et al* 2021 Machine learning on neutron and x-ray scattering and spectroscopies *Chem. Phys. Rev.* **2** 031301
[34] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 Recent advances and applications of machine learning in solid-state materials science *npj Comput. Mater.* **5** 83
[35] Jha D, Ward L, Paul A, Liao W K, Choudhary A, Wolverton C and Agrawal A 2018 ElemNet: deep learning the chemistry of materials from only elemental composition *Sci. Rep.* **8** 17593
[36] Zhou Z, Kearnes S, Li L, Zare R N and Riley P 2019 Optimization of molecules via deep reinforcement learning *Sci. Rep.* **9** 10752
[37] Antono E, Matsuzawa N N, Ling J, Edward Saal J, Arai H, Sasago M and Fujii E 2020 Machine-learning guided quantum chemical and molecular dynamics calculations to design novel hole-conducting organic materials *J. Phys. Chem.* A **124** 8330–40
[38] de Almeida A F, Moreira R and Rodrigues T 2019 Synthetic organic chemistry driven by artificial intelligence *Nat. Rev. Chem.* **3** 589–604
[39] Dral P O 2020 Quantum chemistry in the age of machine learning *J. Phys. Chem. Lett.* **11** 2336–47
[40] Chen W K, Liu X Y, Fang W H, Dral P O and Cui G 2018 Deep learning for nonadiabatic excited-state dynamics *J. Phys. Chem. Lett.* **9** 6702–8
[41] Schütt K T, Gastegger M, Tkatchenko A, Müller K R and Maurer R J 2019 Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions *Nat. Commun.* **10** 5024
[42] Schütt K T, Kessel P, Gastegger M, Nicoli K A, Tkatchenko A and Müller K R 2019 SchNetPack: a deep learning toolbox for atomistic systems *J. Chem. Theory Comput.* **15** 448–55
[43] XANESNET 2023 (available at: https://gitlab.com/team-xnet/xanesnet)
[44] XANESNET Training Data 2023 (available at: https://gitlab.com/team-xnet/training-sets)
[45] XANESNET Tutorials 2023 (available at: https://gitlab.com/team-xnet/tutorials)
[46] Gallo E and Glatzel P 2014 Valence to core x-ray emission spectroscopy *Adv. Mater.* **26** 7730–46
[47] Andrea Cannizzo C J M, Consani C, Gawelda W, Ch Bressler F V M and Chergui M 2010 Light-induced spin crossover in Fe (ii)-based complexes: the full photocycle unraveled by ultrafast optical and x-ray spectroscopies *Coord. Chem. Rev.* **254** 2677–86
[48] Van der Veen R M, Kas J J, Milne C J, Pham V-T, El Nahhas A, Lima F A, Vithanage D A, Rehr J J, Abela R and Chergui M 2010 L-edge xanes analysis of photoexcited metal complexes in solution *Phys. Chem. Chem. Phys.* **12** 5551–61
[49] George G N, Hackett M J, Sansone M, Gorbaty M L, Kelemen S R, Prince R C, Harris H H and Pickering I J 2014 Long-range chemical sensitivity in the sulfur k-edge x-ray absorption spectra of substituted thiophenes *J. Phys. Chem.* A **118** 7796–802
[50] Matthias Kahk J, Michelitsch G S, Maurer R J, Reuter K and Lischner J 2021 Core electron binding energies in solids from periodic all-electron δ-self-consistent-field calculations *J. Phys. Chem. Lett.* **12** 9353–9
[51] Matthias Kahk J and Lischner J 2019 Accurate absolute core-electron binding energies of molecules, solids and surfaces from first-principles calculations *Phys. Rev. Mater.* **3** 100801
[52] Annegarn M, Matthias Kahk J and Lischner J 2022 Combining time-dependent density functional theory and the δscf approach for accurate core-electron spectra *J. Chem. Theory Comput.* **18** 7620–9
[53] Besley N A 2021 Density functional theory calculations of core–electron binding energies at the k-edge of heavier elements *J. Chem. Theory Comput.* **17** 3644–51
[54] Smolentsev G, Soldatov A V, Messinger J, Merz K, Weyhermüller T, Bergmann U, Pushkar Y, Yano J, Yachandra V K and Glatzel P 2009 x-ray emission spectroscopy to study ligand valence orbitals in Mn coordination complexes *J. Am. Chem. Soc.* **131** 13161–7

[55] Lee N, Petrenko T, Bergmann U, Neese F and DeBeer S 2010 Probing valence orbital composition with iron kβ x-ray emission spectroscopy *J. Am. Chem. Soc.* **132** 9715–27
[56] De Groot F 2005 Multiplet effects in x-ray spectroscopy *Coord. Chem. Rev.* **249** 31–63
[57] Rehr J J 2006 Theory and calculations of x-ray spectra: XAS, XES, XRS, and NRIXS *Radiat. Phys. Chem.* **75** 1547–58
[58] Wang X, de Groot F M F and Cramer S P 1997 Spin-polarized x-ray emission of 3d transition-metal ions: a comparison via kα and kβ detection *Phys. Rev.* B **56** 4553
[59] Kotani A 1998 Theory of x-ray emission spectra in *f* and *d* electron systems *J. Electron Spectrosc. Relat. Phenom.* **92** 171–9
[60] Stavitski E and De Groot F M F 2010 The CTM4XAS program for EELS and XAS spectral shape analysis of transition metal L edges *Micron* **41** 687–94
[61] de Groot F M F *et al* 2021 2p x-ray absorption spectroscopy of 3d transition metal systems *J. Electron Spectrosc. Relat. Phenom.* **249** 147061
[62] De Groot F 2001 High-resolution x-ray emission and x-ray absorption spectroscopy *Chem. Rev.* **101** 1779–808
[63] Josefsson I, Kunnus K, Schreck S, Föhlisch A, de Groot F, Wernet P and Odelius M 2012 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** 3565–70
[64] Pinjari R V, Delcey M G, Guo M, Odelius M and Lundberg M 2014 Restricted active space calculations of L-edge x-ray absorption spectra: from molecular orbitals to multiplet states *J. Chem. Phys.* **141** 124116
[65] Delcey M G, Kragh Sørensen L, Vacher M, Couto R C and Lundberg M 2019 Efficient calculations of a large number of highly excited states for multiconfigurational wavefunctions *J. Comput. Chem.* **40** 1789–99
[66] Maganas D, Kristiansen P, Duda L-C, Knop-Gericke A, DeBeer S, Schlögl R and Neese F 2014 Combined experimental and ab initio multireference configuration interaction study of the resonant inelastic x-ray scattering spectrum of Co_2 *J. Phys. Chem.* C **118** 20163–75
[67] Pollock C J, Ulises Delgado-Jaime M, Atanasov M, Neese F and DeBeer S 2014 Kβ mainline x-ray emission spectroscopy as an experimental probe of metal–ligand covalency *J. Am. Chem. Soc.* **136** 9453–63
[68] Guo M, Kallman E, Kragh Sørensen L, Delcey M G, Pinjari R V and Lundberg M 2016 Molecular orbital simulations of metal 1s2p resonant inelastic x-ray scattering *J. Phys. Chem.* A **120** 5848–55
[69] Rehr J J and Albers R C 2000 Theoretical approaches to x-ray absorption fine structure *Rev. Mod. Phys.* **72** 621
[70] Fay M J, Proctor A, Hoffmann D P and Hercules D M 1988 Unraveling exafs spectroscopy *Anal. Chem.* **60** 1225A–43A
[71] Koningsberger D C, Mojet B L, Van Dorssen G E and Ramaker D E 2000 XAFS spectroscopy; fundamental principles and data analysis *Top. Catal.* **10** 143–55
[72] Funke H, Chukalina M and Scheinost A C 2007 A new *FEFF*-based wavelet for EXAFS data analysis *J. Synchrotron Radiat.* **14** 426–32
[73] Penfold T J, Tavernelli I, Milne C J, Reinhard M, El Nahhas A, Rafael Abela U R and Chergui M 2013 A wavelet analysis for the x-ray absorption spectra of molecules *J. Chem. Phys.* **138** 014104
[74] Jeong E-S and Han S-W 2023 Comparison of fourier-transformed and wavelet-transformed EXAFS *J. Korean Phys. Soc.* **84** 1–10
[75] Rossi T, Penfold T J, Rittmann-Frank M H, Reinhard M, Rittmann J, Borca C N, Grolimund D, Milne C J and Chergui M 2014 Characterizing the structure and defect concentration of ZnO nanoparticles in a colloidal solution *J. Phys. Chem.* C **118** 19422–30
[76] Timoshenko J, Wrasman C J, Luneau M, Shirman T, Cargnello M, Bare S R, Aizenberg J, Friend C M and Frenkel A I 2019 Probing atomic distributions in mono- and bimetallic nanoparticles by supervised machine learning *Nano Lett.* **19** 520–9
[77] Timoshenko J, Sang Jeon H, Sinev I, Haase F T, Herzog A and Roldan Cuenya B 2020 Linking the evolution of catalytic properties and structural changes in copper–zinc nanocatalysts using operando EXAFS and neural-networks *Chem. Sci.* **11** 3727–36
[78] Martini A, Negri C, Bugarin L, Deplano G, Abasabadi R K, Lomachenko K A, Janssens T V W, Bordiga S, Berlier G and Borfecchia E 2022 Assessing the influence of zeolite composition on oxygen-bridged diamino dicopper (ii) complexes in cu-cha deno x catalysts by machine learning-assisted x-ray absorption spectroscopy *J. Phys. Chem. Lett.* **13** 6164–70
[79] Martini A, Bugaev A L, Guda S A, Guda A A, Priola E, Borfecchia E, Smolders S, Janssens K, De Vos D and Soldatov A V 2021 Revisiting the extended x-ray absorption fine structure fitting procedure through a machine learning-based approach *J. Phys. Chem.* A **125** 7080–91
[80] Timoshenko J and Frenkel A I 2019 "Inverting" x-ray absorption spectra of catalysts by machine learning in search for activity descriptors *ACS Catal.* **9** 10192–211
[81] Terry J *et al* 2021 Analysis of extended x-ray absorption fine structure (EXAFS) data using artificial intelligence techniques *Appl. Surf. Sci.* **547** 149059
[82] Prange M P, Govind N, Stinis P, Ilton E S and Howard A A 2023 A multifidelity and multimodal machine learning approach for extracting bonding environments of impurities and dopants from x-ray spectroscopies *Technical Report* (Pacific Northwest National Laboratory (PNNL))
[83] Rehr J J and Albers R A 1990 Scattering-matrix formulation of curved-wave multiple-scattering theory: application to x-ray-absorption fine structure *Phys. Rev.* B **41** 8139–49
[84] Natoli C R, Benfatto M, Brouder C, Ruiz Lopez M F and Foulis D L 1990 Multichannel multiple-scattering theory with general potentials *Phys. Rev.* B **42** 1–25
[85] Rehr J J, Albers R A and Zabinsky S 1992 High-order multiple-scattering calculations of x-ray-absorption fine structure *Phys. Rev. Lett.* **69** 3397–400
[86] Rehr J J and Ankudinov A L 2005 Progress in the theory and interpretation of XANES *Coord. Chem. Rev.* **249** 131–40
[87] Briois V, Sainctavit P, Long G J and Grandjean F 2001 Importance of photoelectron multiple scattering in the iron k-edge x-ray absorption spectra of spin-crossover complexes: full multiple scattering calculations for several iron (ii) trispyrazolylborate and trispyrazolylmethane complexes *Inorg. Chem.* **40** 912–8
[88] El Nahhas A *et al* 2013 x-ray absorption spectroscopy of ground and excited rhenium–carbonyl–diimine complexes: evidence for a two-center electron transfer *J. Phys. Chem.* A **117** 361–9
[89] James Penfold T, Reinhard M, Hannelore Rittmann-Frank M, Tavernelli I, Rothlisberger U, Milne C J, Glatzel P and Chergui M 2014 x-ray spectroscopic study of solvent effects on the ferrous and ferric hexacyanide anions *J. Phys. Chem.* A **118** 9411–8
[90] Zabinsky S I, Rehr J J, Ankudinov A, Albers R C and Eller M J 1995 Multiple-scattering calculations of x-ray-absorption spectra *Phys. Rev.* B **52** 2995
[91] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 Physics-inspired structural representations for molecules and materials *Chem. Rev.* **121** 9759–815

[92] Douglas Rankine C and Penfold T J 2022 Accurate, affordable and generalizable machine learning simulations of transition metal x-ray absorption spectra using the XANESNET deep neural network *J. Chem. Phys.* **156** 164102
[93] Fernandez M, Trefiak N R and Woo T K 2013 Atomic property weighted radial distribution functions descriptors of metal–organic frameworks for the prediction of gas uptake capacity *J. Phys. Chem.* C **117** 14095–105
[94] Krykunov M and Woo T K 2018 Bond type restricted property weighted radial distribution functions for accurate machine learning prediction of atomization energies *J. Chem. Theory Comput.* **14** 5229–37
[95] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces *Phys. Rev. Lett.* **98** 146401
[96] Behler J 2011 Atom-centered symmetry functions for constructing high-dimensional neural network potentials *J. Chem. Phys.* **134** 074106
[97] Imbalzano G, Anelli A, Giofré D, Klees S, Behler J and Ceriotti M 2018 Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials *J. Chem. Phys.* **148** 241730
[98] Gastegger M, Schwiedrzik L, Bittermann M, Berzsenyi F and Marquetand P 2018 WACSF—weighted atom-centered symmetry functions as descriptors in machine learning potentials *J. Chem. Phys.* **148** 241709
[99] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115
[100] Sandip D, Bartók A P, Csányi G and Ceriotti M 2016 Comparing molecules and solids across structural and alchemical space *Phys. Chem. Chem. Phys.* **18** 13754–69
[101] Barnard T, Tseng S, Darby J P, Bartók A P, Broo A and Sosso G C 2023 Leveraging genetic algorithms to maximise the predictive capabilities of the soap descriptor *Mol. Syst. Design Eng.* **8** 300–15
[102] Darby J P, Kermode J R and Csányi G 2022 Compressing local atomic neighbourhood descriptors *npj Comput. Mater.* **8** 166
[103] Huo H and Rupp M 2022 Unified representation of molecules and crystals for machine learning *Mach. Learn.: Sci. Technol.* **3** 045017
[104] Hansen K, Biegler F, Ramakrishnan R, Wiktor Pronobis O A V L, Muller K-R and Tkatchenko A 2015 Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space *J. Phys. Chem. Lett.* **6** 2326–31
[105] Rupp M, Tkatchenko A, Müller K-R and Anatole Von Lilienfeld O 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[106] Kwon H *et al* 2023 Harnessing neural networks for elucidating x-ray absorption structure–spectrum relationships in amorphous carbon *J. Phys. Chem.* C **127** 16473–84
[107] Hirai H, Iizawa T, Tamura T, Karasuyama M, Kobayashi R and Hirose T 2022 Machine-learning-based prediction of first-principles XANES spectra for amorphous materials *Phys. Rev. Mater.* **6** 115601
[108] Penfold T J and Rankine C D 2022 A deep neural network for valence-to-core x-ray emission spectroscopy *Mol. Phys.* **121** e2123406
[109] Vladyka A, Sahle C J and Niskanen J 2023 Towards structural reconstruction from x-ray spectra *Phys. Chem. Chem. Phys.* **25** 6707–13
[110] Watson L, Rankine C D and Penfold T J 2022 Beyond structural insight: a deep neural network for the prediction of Pt $L_{2/3}$-edge x-ray absorption spectra *Phys. Chem. Chem. Phys.* **24** 9156–67
[111] Welborn M, Cheng L and Miller T F 2018 Transferability in machine learning for electronic structure via the molecular orbital basis *J. Chem. Theory Comput.* **14** 4772–9
[112] Karandashev K and von Lilienfeld O A 2022 An orbital-based representation for accurate quantum machine learning *J. Chem. Phys.* **156** 114101
[113] Fabrizio A, Briling K R and Corminboeuf C 2022 $SPA^{H}M$: the spectrum of approximated hamiltonian matrices representations *Digit. Discovery* **1** 286–94
[114] Llenga S and Gryn'ova G 2023 Matrix of orthogonalized atomic orbital coefficients representation for radicals and ions *J. Chem. Phys.* **158** 214116
[115] Lüder J 2021 Determining electronic properties from l-edge x-ray absorption spectra of transition metal compounds with artificial neural networks *Phys. Rev.* B **103** 045140
[116] Middleton C, Curchod B and Penfold T 2024 Partial density of states representation for accurate deep neural network predictions of x-ray spectra *ChemrXiv* (https://doi.org/10.26434/chemrxiv-2024-bbrgt)
[117] Chen C, Weike Y, Zuo Y, Zheng C and Ping Ong S 2019 Graph networks as a universal machine learning framework for molecules and crystals *Chem. Mater.* **31** 3564–72
[118] Xie T and Grossman J C 2018 Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties *Phys. Rev. Lett.* **120** 145301
[119] Wieder O, Kohlbacher S, Kuenemann M, Garon A, Ducrot P, Seidel T and Langer T 2020 A compact review of molecular property prediction with graph neural networks *Drug Discovery Today* **37** 1–12
[120] Kearnes S, McCloskey K, Berndl M, Pande V and Riley P 2016 Molecular graph convolutions: moving beyond fingerprints *J. Comput.-Aided Mol. Design* **30** 595–608
[121] Batatia I, Kovacs D P, Simm G, Ortner C and Csányi G 2022 Mace: higher order equivariant message passing neural networks for fast and accurate force fields *Advances in Neural Information Processing Systems* vol 35 pp 11423–36
[122] Drautz R 2019 Atomic cluster expansion for accurate and transferable interatomic potentials *Phys. Rev.* B **99** 014104
[123] Batatia I, Péter Kovács D, Simm G N C, Ortner C and Csányi G 2023 Mace: Higher order equivariant message passing neural networks for fast and accurate force fields *NeurIPS Proceedings* vol 1050 p 26
[124] Kapil V, Péter Kovács D, Csányi G and Michaelides A 2024 First-principles spectroscopy of aqueous interfaces using machine-learned electronic and quantum nuclear effects *Faraday Discuss.* **249** 50–68
[125] Carbone M R, Topsakal M, Deyu L and Yoo S 2020 Machine-Learning x-ray absorption spectra to quantitative accuracy *Phys. Rev. Lett.* **124** 156401
[126] Kotobi A, Singh K, Höche D, Bari S, Meißner R H and Bande A 2023 Integrating explainability into graph neural network models for the prediction of x-ray absorption spectra *J. Am. Chem. Soc.* **145** 22584–98
[127] Torrisi S B, Carbone M R, Rohr B A, Montoya J H, Yang H, Yano J, Suram S K and Hung L 2020 Random forest machine learning models for interpretable x-ray absorption near-edge structure spectrum-property relationships *npj Comput. Mater.* **6** 109
[128] Chen Y, Chen C, Hwang I, Davis M J, Yang W, Sun C, Ping Ong S and Chan M K Y 2023 Robust machine learning inference from x-ray absorption near edge spectra through featurization (arXiv:2310.07049)

[129] Tetef S, Govind N and Seidler G T 2021 Unsupervised machine learning for unbiased chemical classification in x-ray absorption spectroscopy and x-ray emission spectroscopy *Phys. Chem. Chem. Phys.* **23** 23586–601
[130] Tetef S, Pattammattel A, Chu Y S, Chan M K Y and Seidler G T 2023 Manifold projection image segmentation for nano-XANES imaging *APL Mach. Learn.* **1** 046119
[131] Routh P K, Liu Y, Marcella N, Kozinsky B and Frenkel A I 2021 Latent representation learning for structural characterization of catalysts *J. Phys. Chem. Lett.* **12** 2086–94
[132] Guda A A *et al* 2021 Understanding x-ray absorption spectra by means of descriptors and machine learning algorithms *npj Comput. Mater.* **7** 203
[133] Madkhali M M M, Rankine C D and Penfold T J 2020 The role of structural representation in the performance of a deep neural network for x-ray spectroscopy *Molecules* **25** 2715
[134] Yves Joly O B, Lorenzo J-E, Galera R-M, Grenier S and Thompson B 2009 Self-consistency, spin-orbit and other advances in the FDMNES code to simulate XANES and RXD experiments *J. Phys.: Conf. Ser.* **190** 012007
[135] Bauer M 2014 HERFD-XAS and valence-to-core-XES new tools to push the limits in research with hard x-rays? *Phys. Chem. Chem. Phys.* **16** 13827–37
[136] Nash B *et al* 2022 Combining diagnostics, modeling and control systems for automated alignment of the TES beamline *J. Phys.: Conf. Ser.* **2380** 012103
[137] Campbell S I, Allan D B, Barbour A M, Olds D, Rakitin M S, Smith R and Wilkins S B 2021 Outlook for artificial intelligence and machine learning at the NSLS-II *Mach. Learn.: Sci. Technol.* **2** 013001
[138] Nash B *et al* 2021 X-ray beamline control with machine learning and an online model *Proc. ICALEPCS* vol 21 pp 695–9
[139] Edelen A, Nicole Neveu C M, Emma C and Ratner D 2019 Machine learning models for optimization and control of x-ray free electron lasers *NeurIPS Machine Learning for the Physical Sciences Workshop*
[140] Sanchez-Gonzalez A *et al* 2017 Accurate prediction of x-ray pulse properties from a free-electron laser using machine learning *Nat. Commun.* **8** 15461
[141] Drera G, Kropf C M and Sangaletti L 2020 Deep neural network for x-ray photoelectron spectroscopy data analysis *Mach. Learn.: Sci. Technol.* **1** 015008
[142] Pielsticker L, Nicholls R L, DeBeer S and Greiner M 2023 Convolutional neural network framework for the automated analysis of transition metal x-ray photoelectron spectra *Anal. Chim. Acta* **1271** 341433
[143] Westermayr J and Maurer R J 2021 Physically inspired deep learning of molecular excitations and photoemission spectra *Chem. Sci.* **12** 10755–64
[144] Rankine C D, Madkhali M M M and Penfold T J 2020 A deep neural network for the rapid prediction of x-ray absorption spectra *J. Phys. Chem.* A **124** 4263–70
[145] Falbo E, Rankine C D and Penfold T J 2021 On the analysis of x-ray absorption spectra for polyoxometallates *Chem. Phys. Lett.* **780** 138893
[146] Madkhali M M M, Rankine C D and Penfold T J 2021 Enhancing the analysis of disorder in x-ray absorption spectra: application of deep neural networks to t-jump-x-ray probe experiments *Phys. Chem. Chem. Phys.* **23** 9259–69
[147] Ghose A, Segal M, Meng F, Liang Z, Hybertsen M S, Xiaohui Q, Stavitski E, Yoo S, Deyu L and Carbone M R 2023 Uncertainty-aware predictions of molecular x-ray absorption spectra using neural network ensembles *Phys. Rev. Res.* **5** 013180
[148] Andrea Martini S A G *et al* 2020 Pyfitit: The software for quantitative analysis of xanes spectra using machine-learning algorithms *Comput. Phys. Commun.* **250** 107064
[149] Hwang I-H, Solovyev M A, Han S-W, Chan M K Y, Hammonds J P, Heald S M, Kelly S D, Schwarz N, Zhang X and Sun C-J 2022 AXEAP: a software package for x-ray emission data analysis using unsupervised machine learning *J. Synchrotron Radiat.* **29** 1309–17
[150] Sun Q, Xiang Y, Liu Y, Liang X, Leng T, Yifan Y, Fortunelli A, Goddard W A and Cheng T 2022 Machine learning predicts the x-ray photoelectron spectroscopy of the solid electrolyte interface of lithium metal battery *J. Phys. Chem. Lett.* **13** 8047–54
[151] Golze D, Hirvensalo M, Hernández-León P, Aarva A, Etula J, Susi T, Rinke P, Laurila T and Caro M A 2022 Accurate computational prediction of core-electron binding energies in carbon-based materials: a machine-learning model combining density-functional theory and GW *Chem. Mater.* **34** 6240–54
[152] Capano G *et al* 2013 The role of hartree–fock exchange in the simulation of x-ray absorption spectra: a study of photoexcited $[Fe(bpy)_3]^{2+}$ *Chem. Phys. Lett.* **580** 179–84
[153] Kozyr E G, Bugaev A L, Guda S A, Guda A A, Lomachenko K A, Janssens K, Smolders S, De Vos D and Soldatov A V 2021 Speciation of ru molecular complexes in a homogeneous catalytic system: fingerprint xanes analysis guided by machine learning *J. Phys. Chem.* C **125** 27844–52
[154] Smolentsev G and Soldatov A V 2007 FitIt: new software to extract structural information on the basis of XANES fitting *Comput. Mater. Sci.* **39** 569–74
[155] Martini A, Hursán D, Timoshenko J, Rüscher M, Haase F, Rettenmaier C, Ortega E, Etxebarria A and Roldan Cuenya B 2023 Tracking the evolution of single-atom catalysts for the CO_2 electrocatalytic reduction using operando x-ray absorption spectroscopy and machine learning *J. Am. Chem. Soc.* **145** 17351–66
[156] Trummer D, Searles K, Algasov A, Guda S A, Soldatov A V, Ramanantoanina H, Safonova O V, Guda A A and Coperet C 2021 Deciphering the phillips catalyst by orbital analysis and supervised machine learning from cr pre-edge xanes of molecular libraries *J. Am. Chem. Soc.* **143** 7326–41
[157] Balcells D and Bjerkem Skjelstad B 2020 tmqm dataset—quantum geometries and properties of 86k transition metal complexes *J. Chem. Inform. Model.* **60** 6135–46
[158] Ruddigkeit L, Van Deursen R, Blum L C and Reymond J-L 2012 Enumeration of 166 billion organic small molecules in the chemical Universe database GDB-17 *J. Chem. Inf. Model.* **52** 2864–75
[159] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules *Sci. Data* **1** 140022
[160] Mathew K, Zheng C, Winston D, Chen C, Dozier A, Rehr J J, Ping Ong S and Persson K A 2018 High-throughput computational x-ray absorption spectroscopy *Sci. Data* **5** 1–8
[161] Verma S, Khadijah Nik Aznan N, Garside K and Penfold T J 2023 Uncertainty quantification of spectral predictions using deep neural networks *Chem. Commun.* **59** 7100–3
[162] Middleton C, Rankine C D and Penfold T J 2023 An on-the-fly deep neural network for simulating time-resolved spectroscopy: predicting the ultrafast ring opening dynamics of 1,2-dithiane *Phys. Chem. Chem. Phys.* **25** 13325–34

[163] Timoshenko J, Lu D, Lin Y and Frenkel A I 2017 Supervised machine-learning-based determination of three-dimensional structure of metallic nanoparticles *J. Phys. Chem. Lett.* **8** 5091–8
[164] Timoshenko J, Halder A, Yang B, Seifert S, Pellin M J, Vajda S and Frenkel A I 2018 Subnanometer substructures in nanoassemblies formed from clusters under a reactive atmosphere revealed using machine learning *J. Phys. Chem.* C **122** 21686–93
[165] Timoshenko J, Ahmadi M and Cuenya B R 2019 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 **123** 20594–604
[166] Tetef S, Kashyap V, Holden W M, Velian A, Govind N and Seidler G T 2022 Informed chemical classification of organophosphorus compounds via unsupervised machine learning of x-ray absorption spectroscopy and x-ray emission spectroscopy *J. Phys. Chem.* A **126** 4862–72
[167] Schmitt B *et al* 2014 The sshade project: an european database infrastructure in solid spectroscopy *European Planetary Science Congress* vol 9
[168] Mathew K, Zheng C, Winston D, Chen C, Dozier A, Rehr J J, Ong S P and Persson K A 2018 Data descriptor: high-throughput computational x-ray absorption spectroscopy *Sci. Data* **5** 108151
[169] Zheng C *et al* 2018 Automated Generation and ensemble-learned matching of x-ray absorption spectra *npj Comput. Mater.* **4** 12
[170] Swann E, Sun B, Cleland D M and Barnard A S 2018 Representing molecular and materials data for unsupervised machine learning *Mol. Simul.* **44** 905–20
[171] Martini A and Borfecchia E 2020 Spectral decomposition of x-ray absorption spectroscopy datasets: methods and applications *Crystals* **10** 664
[172] Lerotic M, Jacobsen C, Schäfer T and Vogt S 2004 Cluster analysis of soft x-ray spectromicroscopy data *Ultramicroscopy* **100** 35–57
[173] Lerotic M, Jacobsen C, Gillow J B, Francis A J, Wirick S Vogt S and Maser J 2005 Cluster analysis in soft x-ray spectromicroscopy: finding the patterns in complex specimens *J. Electron Spectrosc. Relat. Phenom.* **144** 1137–43
[174] Tetef S, Pattammattel A, Chu Y S, Chan M K Y and Seidler G T 2024 Accelerating nano-XANES imaging via feature selection *Digit. Discovery* **3** 201–9
[175] Schmidt J E, Xinwei Y, van Ravenhorst I K, Oord R, Shapiro D A, Young-Sang Y, Bare S R, Meirer F, Poplawsky J D and Weckhuysen B M 2019 Probing the location and speciation of elements in zeolites with correlated atom probe tomography and scanning transmission x-ray microscopy *ChemCatChem* **11** 488–94
[176] Beale A M, Jacques S D M, Marco Di Michiel J F W M, Price S W T, Senecal P, Vamvakeros A and Paterson J 2018 x-ray physico-chemical imaging during activation of cobalt-based fischer–tropsch synthesis catalysts *Phil. Trans. R. Soc.* A **376** 20170057
[177] Price S W T, Ignatyev K Geraki K, Basham M, Filik J, Vo N T, Witte P T, Beale A M and Mosselmans J F W 2015 Chemical imaging of single catalyst particles with scanning μ-XANES-CT and μ-XRF-CT *Phys. Chem. Chem. Phys.* **17** 521–9
[178] Boesenberg U *et al* 2013 Mesoscale phase distribution in single particles of LIFEPO4 following lithium deintercalation *Chem. Mater.* **25** 1664–72
[179] Aarva A, Deringer V L, Sainio S, Laurila T and Caro M A 2019 Understanding x-ray spectroscopy of carbonaceous materials by combining experiments, density functional theory and machine learning. Part I: Fingerprint spectra *Chem. Mater.* **31** 9243–55
[180] Aarva A, Sainio S, Deringer V L, Caro M A and Laurila T 2021 x-ray spectroscopy fingerprints of pristine and functionalized graphene *J. Phys. Chem.* C **125** 18234–46
[181] Xiang S, Huang P, Li J, Liu Y, Marcella N, Routh P K, Gonghu Li and Frenkel A I 2022 Solving the structure of "single-atom" catalysts using machine learning–assisted xanes analysis *Phys. Chem. Chem. Phys.* **24** 5116–24
[182] Usoltsev O A, Bugaev A L, Guda A A, Guda S A and Soldatov A V 2022 How much structural information could be extracted from xanes spectra for palladium hydride and carbide nanoparticles *J. Phys. Chem.* C **126** 4921–8
[183] Martini A, Guda A A, Guda S A, Dulina A, Tavani F, D'Angelo P, Borfecchia E and Soldatov A V 2021 Estimating a set of pure xanes spectra from multicomponent chemical mixtures using a transformation matrix-based approach *Synchrotron Radiation Science and Applications: Proc. 2019 Meeting of the Italian Synchrotron Radiation Society-Dedicated to Carlo Lamberti* (Springer) pp 65–84
[184] Voronov A, Urakawa A, van Beek W, Tsakoumis N E, Emerich H and Rønning M 2014 Multivariate curve resolution applied to in situ x-ray absorption spectroscopy data: An efficient tool for data processing and analysis *Anal. Chim. Acta* **840** 20–27
[185] Carbone M R, Yoo S, Topsakal M and Deyu L 2019 Classification of local chemical environments from x-ray absorption spectra using supervised machine learning *Phys. Rev. Mater.* **3** 033604
[186] Farges F *et al* 1997 Ti k-edge xanes studies of ti coordination and disorder in oxide compounds: comparison between theory and experiment *Phys. Rev.* B **56** 1809
[187] Kiyohara S, Kikumasa K, Shibata K and Mizoguchi T 2022 Automatic determination of the spectrum–structure relationship by tree structure-based unsupervised and supervised learning *Ultramicroscopy* **233** 113438
[188] Kiyohara S and Mizoguchi T 2020 Radial distribution function from x-ray absorption near edge structure with an artificial neural network *J. Phys. Soc. Japan* **89** 103001
[189] Higashi M and Ikeno H 2023 Extraction of local structure information from x-ray absorption near-edge structure: a machine learning approach *Mater. Trans.* **64** MT–MG2022028
[190] David T, Khadijah Nik Aznan N, Garside K and James Penfold T 2023 Towards the automated extraction of structural information from x-ray absorption spectra *Digit. Discovery* **2** 1461–70
[191] Deb A and Cairns E J 2006 In situ x-ray absorption spectroscopy—a probe of cathode materials for li-ion cells *Fluid Phase Equilib.* **241** 4–19
[192] Bressler C M C *et al* 2009 Femtosecond xanes study of the light-induced spin crossover dynamics in an iron (ii) complex *Science* **323** 489–92
[193] Lima F A, Penfold T J, Van Der Veen R M, Reinhard M, Abela R, Tavernelli I, Rothlisberger U, Benfatto M, Milne C J and Chergui M 2014 Probing the electronic and geometric structure of ferric and ferrous myoglobins in physiological solutions by Fe k-edge absorption spectroscopy *Phys. Chem. Chem. Phys.* **16** 1617–31
[194] Atkins A J, Bauer M and Jacob C R 2015 High-resolution x-ray absorption spectroscopy of iron carbonyl complexes *Phys. Chem. Chem. Phys.* **17** 13937–48
[195] Chen W-T, Hsu C-W, Lee J-F, Pao C-W and Hsu I-J 2020 Theoretical analysis of Fe k-edge xanes on iron pentacarbonyl *ACS Omega* **5** 4991–5000
[196] Mebs S, Braun B, Kositzki R, Limberg C and Haumann M 2015 Abrupt versus gradual spin-crossover in Fe^{ii} $(phen)_2$ $(NCS)_2$ and Fe^{iii} $(dedtc)_3$ compared by x-ray absorption and emission spectroscopy and quantum-chemical calculations *Inorg. Chem.* **54** 11606–24

[197] Németh Z, Szlachetko J, Bajnóczi Eva G and Vankó G 2016 Laboratory von Hámos x-ray spectroscopy for routine sample characterization *Rev. Sci. Instrum.* **87** 103105
[198] Seidler G T, Mortensen D R, Remesnik A J, Pacold J I, Ball N A, Barry N, Styczinski M and Hoidn O R 2014 A laboratory-based hard x-ray monochromator for high-resolution x-ray emission spectroscopy and x-ray absorption near edge structure measurements *Rev. Sci. Instrum.* **85** 113906
[199] Anker A S, Butler K T, Manh Duc L, Perring T G and Thiyagalingam J 2023 Using generative adversarial networks to match experimental and simulated inelastic neutron scattering data *Digit. Discovery* **2** 578–90
[200] Han J, Shoeiby M, Petersson L and Ali Armin M 2021 Dual contrastive learning for unsupervised image-to-image translation *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 746–55
[201] Zhu J-Y, Park T, Isola P and Efros A A 2017 Unpaired image-to-image translation using cycle-consistent adversarial networks *Proc. IEEE Int. Conf. on Computer Vision* pp 2223–32
[202] Almahairi A, Rajeshwar S, Sordoni A, Bachman P and Courville A 2018 Augmented cyclegan: Learning many-to-many mappings from unpaired data *Int. Conf. on Machine Learning* (PMLR) pp 195–204
[203] Watson L, Pope T, Jay R M, Banerjee A, Wernet P and Penfold T 2023 A δ-learning strategy for interpretation of spectroscopic observables *Struct. Dyn.* **10** 064101
[204] Ramakrishnan R, Dral P O, Rupp M and Von Lilienfeld O A 2015 Big data meets quantum chemistry approximations: the δ-machine learning approach *J. Chem. Theory Comput.* **11** 2087–96
[205] Bogojeski M, Vogt-Maranto L, Tuckerman M E, Müller K-R and Burke K 2020 Quantum chemical accuracy from density functional approximations via machine learning *Nat. Commun.* **11** 5223
[206] Qiao Z, Welborn M, Anandkumar A, Manby F R and Miller T F 2020 Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features *J. Chem. Phys.* **153** 124111
[207] Qiao Z, Christensen A S, Welborn M, Manby F R, Anandkumar A and Miller T F 2022 Informing geometric deep learning with electronic interactions to accelerate quantum chemistry *Proc. Natl Acad. Sci.* **119** e2205221119
[208] Jay R M *et al* 2023 Tracking C–H activation with orbital resolution *Science* **380** 955–60
[209] José M V, Galván I M and Isasi P 2006 Lazy training of radial basis neural networks *Artificial Neural Networks (ICANN 2006)* ed S D Kollias, A Stafylopatis, W Duch and E Oja (Springer) pp 198–207
[210] Lemm D, von Rudorff G F and von Lilienfeld O A 2023 Improved decision making with similarity based machine learning *Mach. Learn.: Sci. Technol.* **4** 045043
[211] Eldar Y, Lindenbaum M, Porat M and Zeevi Y 1997 The farthest point strategy for progressive image sampling *IEEE Trans. Image Process.* **6** 1305–15
[212] Cordova M, Engel E A, Stefaniuk A, Paruzzo F, Hofstetter A, Ceriotti M and Emsley L 2022 A machine learning model of chemical shifts for chemically and structurally diverse molecular solids *J. Phys. Chem.* C **126** 16710–20
[213] Smith J S, Nebgen B, Lubbers N, Isayev O and Roitberg A E 2018 Less is more: sampling chemical space with active learning *J. Chem. Phys.* **148** 241733
[214] Feng S, Wang K, Wang F, Zhang Y and Zhao H 2022 Less is more: a new machine-learning methodology for spatiotemporal systems *Commun. Theor. Phys.* **74** 055601
[215] Bengio Y, Louradour J, Collobert R and Weston J 2009 Curriculum learning *ICML '09: Proc. 26th Annual Int. Conf. on Machine Learning* (*Montreal Quebec, Canada*, *14–18 June 2009*) pp 41–48
[216] Jeffrey L E 1993 Learning and development in neural networks: the importance of starting small *Cognition* **48** 71–99
[217] Sanger T D 1994 Neural network learning control of robot manipulators using gradually increasing task difficulty *IEEE Trans. Robot. Autom.* **10** 323–33
[218] Borgeaud S *et al* 2022 Improving language models by retrieving from trillions of tokens *Proc. 39th Int. Conf. on Machine Learning* (*Proc. of Machine Learning Research*) (PMLR) pp 2206–40
[219] Westermayr J, Gilkes J, Barrett R and Maurer R J 2023 High-throughput property-driven generative design of functional organic molecules *Nat. Comput. Sci.* **3** 139–48
[220] Reker D and Schneider G 2015 Active-learning strategies in computer-assisted drug discovery *Drug Discovery Today* **20** 458–65
[221] Gastegger M, Behler J and Marquetand P 2017 Machine learning molecular dynamics for the simulation of infrared spectra *Chem. Sci.* **8** 6924–35
[222] Podryabinkin E V and Shapeev A V 2017 Active learning of linearly parametrized interatomic potentials *Comput. Mater. Sci.* **140** 171–80
[223] Browning N J, Raghunathan Ramakrishnan O A von L and Roethlisberger U 2017 Genetic optimization of training sets for improved machine learning models of molecular properties *J. Phys. Chem. Lett.* **8** 1351–9
[224] Dral P O, Owens A, Yurchenko S N and Thiel W 2017 Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels *J. Chem. Phys.* **146** 244108
[225] Zachary C L 2018 The Mythos of Model Interpretability: in machine learning, the concept of interpretability is both important and slippery *Queue* **16** 31–57
[226] Polishchuk P 2017 Interpretation of quantitative structure–activity relationship models: past, present and future *J. Chem. Inform. Model.* **57** 2618–39
[227] Oviedo F, Lavista Ferres J, Buonassisi T and Butler K T 2022 Interpretable and explainable machine learning for materials science and chemistry *Acc. Mater. Res.* **3** 597–607
[228] Wang S and Jiang J 2023 Interpretable catalysis models using machine learning with spectroscopic descriptors *ACS Catal.* **13** 7428–36
[229] Niu Z, Zhong G and Hui Y 2021 A review on the attention mechanism of deep learning *Neurocomputing* **452** 48–62
[230] Bahdanau D, Cho K and Bengio Y 2014 Neural machine translation by jointly learning to align and translate (arXiv:1409.0473)
[231] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Łukasz and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* vol 30
[232] Jain S and Wallace B C 2019 Attention is not explanation (arXiv:1902.10186)
[233] Serrano S and Smith N A 2019 Is attention interpretable? (arXiv:1906.03731)
[234] Krishnadasan S, Brown R J C, Demello A J and Demello J C 2007 Intelligent routes to the controlled synthesis of nanoparticles *Lab Chip* **7** 1434–41
[235] Fitzpatrick D E, Battilocchio C and Ley S V 2016 A novel internet-based reaction monitoring, control and autonomous self-optimization platform for chemical synthesis *Org. Process Res. Dev.* **20** 386–94

[236] Epps R W, Bowen M S, Volk A A, Abdel-Latif K, Han S, Reyes K G, Amassian A and Abolhasani M 2020 Artificial chemist: an autonomous quantum dot synthesis bot *Adv. Mater.* **32** 2001626
[237] Gongora A E, Bowen X, Perry W, Okoye C, Riley P, Reyes K G, Morgan E F and Brown K A 2020 A bayesian experimental autonomous researcher for mechanical design *Sci. Adv.* **6** eaaz1708
[238] Gongora A E, Snapp K L, Whiting E, Riley P, Reyes K G, Morgan E F and Brown K A 2021 Using simulation to accelerate autonomous experimentation: a case study using mechanics *iScience* **24** 102262
[239] Gubaev K, Podryabinkin E V and Shapeev A V 2018 Machine learning of molecular properties: Locality and active learning *J. Chem. Phys.* **148** 241727
[240] Scalia G, Grambow C A, Pernici B, Yi-Pei Li and Green W H 2020 Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction *J. Chem. Inform. Model.* **60** 2697–717
[241] Gloria Capano C J M, Majed Chergui U R, Tavernelli I and Penfold T J 2015 Probing wavepacket dynamics using ultrafast x-ray spectroscopy *J. Phys. B: At. Mol. Opt. Phys.* **48** 214001
[242] Katayama T *et al* 2019 Tracking multiple components of a nuclear wavepacket in photoexcited cu (i)-phenanthroline complex using ultrafast x-ray spectroscopy *Nat. Commun.* **10** 3606
[243] Penfold T J *et al* 2013 Solvent-induced luminescence quenching: static and time-resolved x-ray absorption spectroscopy of a copper (i) phenanthroline complex *J. Phys. Chem.* A **117** 4591–601
[244] D'Angelo P, Zitolo A, Migliorati V, Mancini G, Persson I and Chillemi G 2009 Structural investigation of lanthanoid coordination: a combined xanes and molecular dynamics study *Inorg. Chem.* **48** 10239–48
[245] Pham V-T *et al* 2011 Probing the transition from hydrophilic to hydrophobic solvation with atomic scale resolution *J. Am. Chem. Soc.* **133** 12740–8
[246] D'Angelo P, Maria Roscioni O, Chillemi G, Della Longa S and Benfatto M 2006 Detection of second hydration shells in ionic solutions by XANES: computed spectra for ni2+ in water based on molecular dynamics *J. Am. Chem. Soc.* **128** 1853–8
[247] Migliorati V, Zitolo A, Chillemi G and D'Angelo P 2012 Influence of the second coordination shell on the XANES spectra of the Zn2+ ion in water and methanol *ChemPlusChem* **77** 234–9
[248] Qing Y, Zhou J, Zhao T, Zhao H, Chu W, Sheng Z, Chen X, Marcelli A, Luo Y and Ziyu W 2012 Identification of 13-and 14-coordinated structures of first hydrated shell of $[AuCl]^-$ acid aqueous solution by combination of MD and XANES *J. Phys. Chem.* B **116** 7866–73
[249] Wang J, Hsu C-S, Tai-Sing W, Chan T-S, Suen N-T, Lee J-F and Ming Chen H 2023 In situ x-ray spectroscopies beyond conventional x-ray absorption spectroscopy on deciphering dynamic configuration of electrocatalysts *Nat. Commun.* **14** 6576
[250] Routh P K, Marcella N and Frenkel A I 2023 Speciation of nanocatalysts using x-ray absorption spectroscopy assisted by machine learning *J. Phys. Chem.* C **127** 5653–62
[251] Milne C J *et al* Disentangling the evolution of electrons and holes in photoexcited ZnO nanoparticles *Struct. Dyn.* **10** 2023
[252] Penfold T J *et al* 2018 Revealing hole trapping in zinc oxide nanoparticles by time-resolved x-ray spectroscopy *Nat. Commun.* **9** 478
[253] Hannelore Rittmann-Frank M, Milne C J, Rittmann J, Reinhard M, Penfold T J and Chergui M 2014 Mapping of the photoinduced electron traps in TiO2 by picosecond x-ray absorption spectroscopy *Angew. Chem., Int. Edn.* **53** 5858–62
[254] Merkling P J, Muñoz-Páez A, Pappalardo R R and Sánchez Marcos E 2001 Combination of xanes spectroscopy and molecular dynamics to probe the local structure in disordered systems *Phys. Rev.* B **64** 092201
[255] Andrejevic N, Andrejevic J, Bernevig B A, Regnault N, Han F, Fabbris G, Nguyen T, Drucker N C, Rycroft C H and Mingda Li 2022 Machine-learning spectral indicators of topology *Adv. Mater.* **34** 2204113
[256] Timoshenko J, Haase F T, Saddeler S, Rüscher M, Sang Jeon H, Herzog A, Hejral U, Bergmann A, Schulz S and Roldan Cuenya B 2023 Deciphering the structural and chemical transformations of oxide catalysts during oxygen evolution reaction using quick x-ray absorption spectroscopy and machine learning *J. Am. Chem. Soc.* **145** 4065–80
[257] Darby Dyar M *et al* 2016 Use of multivariate analysis for synchrotron micro-xanes analysis of iron valence state in amphiboles *Am. Mineral.* **101** 1171–89
[258] Liu Y, Halder A, Seifert S, Marcella N, Vajda S and Frenkel A I 2021 Probing active sites in cu x pd y cluster catalysts by machine-learning-assisted x-ray absorption spectroscopy *ACS Appl. Mater. Interfaces* **13** 53363–74
[259] Miras H N, Yan J, Long D-L and Cronin L 2012 Engineering polyoxometalates with emergent properties *Chem. Soc. Rev.* **41** 7403–30
[260] Friedl J, Holland-Cunz M V, Cording F, Pfanschilling F L, Wills C, McFarlane W, Schricker B, Fleck R, Wolfschmidt H and Stimming U 2018 Asymmetric polyoxometalate electrolytes for advanced redox flow batteries *Energy Environ. Sci.* **11** 3010–8
[261] Chen H-Y *et al* 2017 In situ x-ray absorption near edge structure studies and charge transfer kinetics of $Na_6[V_{10}O_{28}]$ electrodes *Phys. Chem. Chem. Phys.* **19** 3358–65
[262] Wang H, Hamanaka S, Nishimoto Y, Irle S, Yokoyama T, Yoshikawa H and Awaga K 2012 In operando x-ray absorption fine structure studies of polyoxometalate molecular cluster batteries: polyoxometalates as electron sponges *J. Am. Chem. Soc.* **134** 4918–24
[263] Falbo E and Penfold T J 2020 Redox potentials of polyoxometalates from an implicit solvent model and QM/MM molecular dynamics *J. Phys. Chem.* C **124** 15045–56
[264] Rüscher M, Herzog A, Timoshenko J, Sang Jeon H, Frandsen W, Kühl S and Cuenya B R 2022 Tracking heterogeneous structural motifs and the redox behaviour of copper–zinc nanocatalysts for the electrocatalytic CO_2 reduction using operando time resolved spectroscopy and machine learning *Catal. Sci. Technol.* **12** 3028–43
[265] Rankine C D, Nunes J ao P F, Robinson M S, Lane P D and Wann D A 2016 A theoretical investigation of internal conversion in 1, 2-dithiane using non-adiabatic multiconfigurational molecular dynamics *Phys. Chem. Chem. Phys.* **18** 27170–4