
Article:
Rankine, C. D. orcid.org/0000-0002-7104-847X 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. Journal of Chemical Physics. 164102. ISSN 1089-7690

https://doi.org/10.1063/5.0087255

The Journal of Chemical Physics

Accurate, affordable, and generalizable machine learning simulations of transition metal x-ray absorption spectra using the XANESNET deep neural network

Cite as: J. Chem. Phys. **156**, 164102 (2022); https://doi.org/10.1063/5.0087255
Submitted: 03 February 2022 • Accepted: 25 March 2022 • Accepted Manuscript Online: 25 March 2022 • Published Online: 26 April 2022
Published open access through an agreement with Newcastle University

C. D. Rankine and T. J. Penfold

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Accurate, affordable, and generalizable machine learning simulations of transition metal x-ray absorption spectra using the XANESNET deep neural network

Cite as: J. Chem. Phys. 156, 164102 (2022); doi: 10.1063/5.0087255
Submitted: 3 February 2022 • Accepted: 25 March 2022 •
Published Online: 26 April 2022

C. D. Rankine[a] and T. J. Penfold[b]

AFFILIATIONS

Chemistry–School of Natural and Environmental Sciences, Newcastle University, Newcastle Upon Tyne NE1 7RU, United Kingdom

[a]Electronic mail: conor.rankine@ncl.ac.uk
[b]Author to whom correspondence should be addressed: tom.penfold@ncl.ac.uk

ABSTRACT

The affordable, accurate, and generalizable prediction of spectroscopic observables plays a key role in the analysis of increasingly complex experiments. In this article, we develop and deploy a deep neural network—XANESNET—for predicting the lineshape of first-row transition metal K-edge x-ray absorption near-edge structure (XANES) spectra. XANESNET predicts the spectral intensities using only information about the local coordination geometry of the transition metal complexes encoded in a feature vector of weighted atom-centered symmetry functions. We address in detail the calibration of the feature vector for the particularities of the problem at hand, and we explore the individual feature importance to reveal the physical insight that XANESNET obtains at the Fe K-edge. XANESNET relies on only a few judiciously selected features—radial information on the first and second coordination shells suffices along with angular information sufficient to separate satisfactorily key coordination geometries. The feature importance is found to reflect the XANES spectral window under consideration and is consistent with the expected underlying physics. We subsequently apply XANESNET at nine first-row transition metal (Ti–Zn) K-edges. It can be optimized in as little as a minute, predicts instantaneously, and provides K-edge XANES spectra with an average accuracy of ~±2%–4% in which the positions of prominent peaks are matched with a >90% hit rate to sub-eV (~0.8 eV) error.

I. INTRODUCTION

Wherever there are valuable data to be predicted, processed, labeled, or mined, one is guaranteed to find machine learning models working autonomously and leveraging recent advances in the accessibility of hardware and software optimized for the task at hand. Highly effective machine learning models that are able to extract and learn patterns represented in data without hand-coded heuristics continue to transform what we can do and the way we do it across the physical sciences[1]—as they have in chemistry for quite some time.[2]

The trajectory of machine learning in chemistry inclines steeply upward, and applications continue to grow at pace.[3] In the chemical research and development domain, applications include the design and discovery of new materials,[4–9] catalysts,[10–13] and drugs[14–16] as well as chemical reaction prediction and synthesis planning.[17–25] In the domain of *ab initio* computational chemistry, interest in the disruptive potential of machine learning is surging too.[26–33] Here, there have been significant successes with machine learning models that redress the accuracy/affordability balance of atomistic modeling—from parametric force-fields[34–38] to accurate quantum mechanical properties obtained from low-cost electronic structure calculations[39–43] and accelerated excited-state molecular dynamics.[44–55]

It ought to be of no great surprise that spectroscopy—already in renaissance following fast-paced developments in methodology and instrumentation, especially at high-brilliance light sources[56–60]—should also be simultaneously transformed by

machine learning.[61] Indeed, the two are a natural pairing; machine learning is similarly grounded in linear mathematics (e.g., least-squares regression) and probability (e.g., maximum-likelihood parametric estimation)—concepts that are familiar to experimental spectroscopists. With the popularity of emergent spectroscopies on an upward trajectory, resulting increasingly in situations where new methods and new users are brought together, machine learning offers a route to affordable and accurate "*out-of-the-box*," "*limited-expertise-required*" analyses.

In spectroscopic applications, machine learning models are typically assigned one of the two tasks: either carrying out "*forward*" (structure-to-spectrum) or "*reverse*" (spectrum-to-structure or spectrum-to-property, alternatively "*inverse*") mappings. There are now many examples of machine learning models for "*reverse*" mappings in the literature although comparatively fewer for "*forward*" mappings. These collectively encompass infrared (IR),[62,63] ultraviolet/visible (UV/vis),[49–51,64–67] Raman scattering,[68] neutron scattering,[69] nuclear magnetic resonance (NMR),[70] and x-ray techniques.[71–97] The focus of this article is on a "*forward*" mapping approach in the domain of x-ray absorption spectroscopy (XAS).

The prediction of spectroscopic observables—a paradigmatic "*forward*" mapping—is a central objective of computational chemistry for spectroscopists as it serves as a conduit between experiment and theory. Achieving a detailed understanding of the properties of a molecule/material on the atomic level via simulations is often the key to understanding and explaining experimentally observed phenomena; ultimately, it is also the key to harnessing them in practical applications. The challenge lies in making the calculations capable of capturing satisfactorily the complexity of the phenomena while simultaneously accurate, affordable, and generally applicable enough to appeal to users. It transpires—unsurprisingly—that this is a tall order indeed!

From the perspective of "*forward*" mapping methodologies, there are three distinct approaches: (i) focusing on a spectral window (e.g., a "*fingerprint*" window sensitive to a particular property or observable) and developing a machine learning model to predict directly the resonances within this window;[48,49,98–103] (ii) representing the resonances via a Hamiltonian matrix associated with a closed set of secular equations and developing a machine learning model to predict the Hamiltonian matrix elements;[27,39,41,42,50] and (iii) developing a machine learning model to predict directly the spectral lineshapes.[71–74,104,105] The latter approach, which we adopt in this article and elsewhere where we have worked with machine learning models for XAS in theoretical[71] and practical[73,74] settings, circumvents the formidable challenge of predicting the huge number of resonances around the x-ray absorption edge.[106] Sitting alongside the well-developed theory for XAS (e.g., multiple scattering theory, multiplet theory, and Bethe–Salpeter *k*-space approaches plus extensions of popular *ab initio* quantum chemical strategies),[106] machine learning models for fast "*forward*" XAS mappings are well placed to unlock affordable analyses in particularly challenging cases, e.g., coupling to ultrafast dynamics simulations[107–118] and describing accurately disordered/amorphous materials.[119–124] In these cases, many configurations need to be sampled to simulate XAS with even qualitative accuracy, but the time- and resource-intensiveness of the individual computational calculations presently makes such treatments challenging.[106]

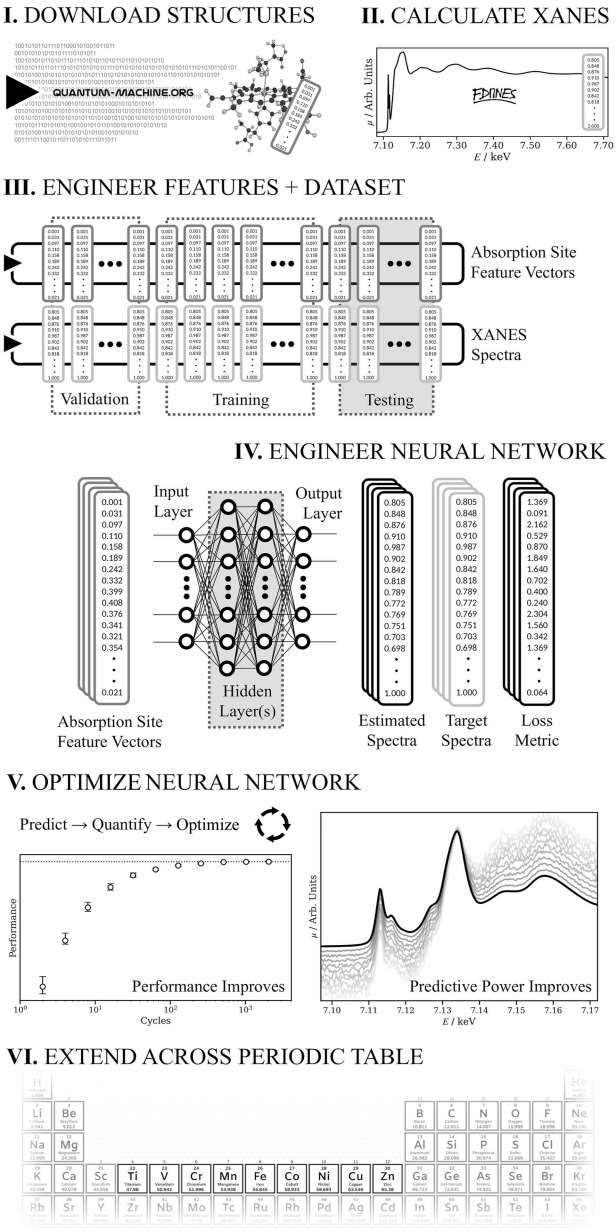

FIG. 1. A schematic of the XANESNET DNN and workflow detailed in this article. The local geometries around first-row transition metal x-ray absorption sites (**I**, "*samples*," Sec. II A) are inputs, and the corresponding theoretically calculated K-edge XANES spectra (**II**, "*labels*," Sec. II C) are outputs. The samples are encoded as descriptive feature vectors (**III**; Sec. II B 2) and associated with their labels to construct reference datasets from which the DNN (**IV**, Sec. II B 1) discovers a "*forward*" structure-to-spectrum mapping via iterative optimization of the internal weights (**V**). We start in the familiar territory at the Fe K-edge and then extend the DNN across the first row of transition metals (Ti–Zn, **VI**).

The K-edge x-ray absorption near-edge structure (XANES) spectra of first-row transition metals—the focus of the present article—are characterized by two distinct regions: the pre- and post-edge. The former contains features corresponding to the promotion

of core electrons into unoccupied orbitals at energies just below the ionization potential. These features provide information about the valence orbital character of the system under study although $3d \leftarrow 1s$ electronic transitions are dipole forbidden, and consequently, these features tend to have low intensity. The latter contains above-ionization-potential resonances that occur due to the interference of the electron wave(s) originating from the x-ray absorption site with the electron wave(s) scattered back from the potential of the neighboring atoms. The low electron kinetic energies (<50 eV) associated with the XANES window of the XAS spectrum mean that these features are dominated by the interference of electron wave scattering pathways involving multiple nearest-neighbor atoms. The XANES window encodes highly detailed geometric (e.g., coordination number and distance) and electronic (e.g., oxidation and spin state) information.

In this article, we build on our earlier proof-of-principle work in Ref. 71 to develop and deploy a deep neural network (DNN)[125]—XANESNET (Fig. 1)—for predicting the lineshape of first-row transition metal K-edge x-ray absorption near-edge structure (XANES) spectra. XANESNET predicts the K-edge XANES spectral intensities using only information about the local coordination geometry of the transition metal complexes. We address in detail the calibration of the feature vector that encodes this information for the particularities of the problem at hand, and we explore the individual feature importance to reveal the physical insight that XANESNET provides at the Fe K-edge. We subsequently transfer XANESNET to nine first-row transition metal (Ti–Zn) K-edges, where we benchmark predictive power and performance.

II. TECHNICAL DETAILS

A. Datasets

Our reference datasets comprise x-ray absorption site geometries ("*samples*") of first-row transition metal (Ti–Zn) complexes harvested from the transition metal Quantum Machine (tmQM) dataset.[126,127] The dataset for each first-row transition metal comprised all of the structures from the tmQM dataset containing that element, as extracted from the 2020 release of the Cambridge Structural Database (CSD) and subsequently optimized at the GFN2-xTB level of theory. The tmQM dataset was initially generated by applying seven filters to exclude: (i) all structures except those containing a single transition metal; (ii) all structures not containing a minimum of one C and one H atom (allowing only these other elements: B, Si, N, P, As, O, S, Se, F, Cl, Br, and I); (iii) the structure of all extraneous molecular components beyond that of the transition metal complex (e.g., counter-ions); (iv) all polymeric structures; (v) all structures without three-dimensional coordinates; (vi) all structures with disordered atoms; and (vii) all structures with charges greater than +1 and less than −1. Full details of the construction and composition of the tmQM dataset can be found in Ref. 127.

K-edge XANES spectra ("*labels*") for these structures were calculated using multiple scattering theory (MST) as implemented in the FDMNES[128,129] package (Sec. II C). We have developed nine independent reference datasets, one for each first-row transition metal (Ti–Zn) x-ray absorption edge; the number of samples contained in the reference datasets scales from as few as ~1100 (V) to ~8660 (Ni). A summary of the number of samples contained in the reference datasets is given in the supplementary material (Table S1). We have made the reference datasets publicly available (see our Data Availability Statement for details).

250 samples from each reference dataset were isolated at random to form "*held-out*" testing datasets (evaluated post-optimization only, Sec. III D). The remaining samples comprised the training and validation datasets used during optimization (Secs. III A–III C). The training and validation subsets were constructed "*on-the-fly*" throughout via repeated K-fold cross-validation with five repeats and five fold, i.e., a five-times-repeated 80:20 split.

B. Deep neural network

1. Architecture

The architecture of the XANESNET DNN used in this article is based on the deep multilayer perceptron (MLP) model and comprises an input layer, two hidden layers, and an output layer. All layers are dense, i.e., fully connected, and each hidden layer performs a nonlinear transformation using the rectified linear unit (*relu*) activation function. The input layer comprises N neurons (to accept a feature vector of length N encoding the local environment around an x-ray absorption site, Sec. II B 2), the hidden layers each comprise 512 neurons, and the output layer comprises 226 neurons from which the discretized K-edge XANES spectrum is retrieved after regression, i.e., XANESNET is a multi-output MLP with each output neuron corresponding to the spectral intensity at a given energy grid point. The architecture of the XANESNET DNN is $[N \times 512 \times 512 \times 226]$.

The internal weights, $\mathbf{W}$, are optimized via iterative feed-forward and backpropagation cycles to minimize the empirical loss, $J(\mathbf{W})$, defined here as the mean-squared error (MSE) between the predicted, $\mu_{predict}$, and target, μ_{target}, K-edge XANES spectra over the reference dataset, i.e., an optimal set of internal weights, $\mathbf{W}^*$, is sought that satisfies $\underset{\mathbf{W}}{argmin}(J(\mathbf{W}))$.

Gradients of the empirical loss with respect to the internal weights, $\delta J(\mathbf{W})/\delta \mathbf{W}$, were estimated over minibatches of 32 samples and updated iteratively according to the Adaptive Moment Estimation (ADAM)[130] algorithm. The learning rate for the ADAM algorithm was set to 1×10^{-4}. The internal weights were initially set according to the He[131] uniform distribution. Unless explicitly stated in this article, optimization was carried out over 512 iterative epochs.

Regularization was implemented to minimize the propensity of overfitting; batch standardization and dropout were applied at each hidden layer. The probability, p, of dropout was set to 0.25.

The XANESNET DNN is programmed in Python 3 with the TensorFlow[132]/Keras[133] application programming interface (API) and integrated into a scikit-learn[134] (*sklearn*) data pre- and post-processing pipeline via the KerasRegressor wrapper for scikit-learn. The Atomic Simulation Environment[135] (ASE) API is used to handle and manipulate molecular structures. The code is publicly available under the GNU Public License (GPLv3) on GitLab.[136]

2. Featurization

The local environments around x-ray absorption sites are encoded via dimensionality reduction using the weighted atom-centered symmetry function (wACSF) descriptor of Gastegger

et al.,[137] which builds on top of the generalized ACSF descriptor introduced by Behler[138,139] to overcome the unfavorable scaling as the number of atom types in the dataset grows. The recent review by Behler in Ref. 140 is strongly recommended to the unfamiliar reader. The wACSF descriptor (or "*feature vector*," $\mathbf{G}_i$) for an arbitrary absorption site, i, is constructed via concatenation of a "*global*" (G^1) wACSF, n radial (G^2, two-body) wACSFs, and m angular (G^4, three-body) wACSFs, i.e., it takes the form

$$\boldsymbol{G}_i = \{G_i^1, G_{i,1}^2, G_{i,2}^2, \ldots, G_{i,n}^2, G_{i,1}^4, G_{i,2}^4, \ldots, G_{i,m}^4\}, \tag{1}$$

where n and m are chosen to cover satisfactorily the radial and angular space of the reference dataset and discriminate different atomic environments.

The G^1, G^2, and G^4 wACSF each takes the forms

$$G_i^1 = \sum_{j \neq i} f_c(r_{ij}), \tag{2}$$

$$G_i^2 = \sum_{j \neq i} Z_j \cdot f_c(r_{ij}) \cdot \exp^{-\eta(r_{ij}-\mu)^2}, \tag{3}$$

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

where i, j, and k are the index atomic sites, Z_i is the atomic number of the atom at the site i, r_{ij} is the interatomic distance between sites i and j, and θ_{jik} is the interatomic angle between sites j, i, and k. f_c is a radial cutoff function (the cutoff set at some radial distance, r_c) that ensures that the wACSFs vary smoothly and, ultimately, go to zero where $r_{ij} \geq r_c$; it take the form

$$f_c(r_{ij}) = \begin{cases} 0.5 \times \left(\cos\left(\frac{\pi r_{ij}}{r_c}\right) + 1\right) & \text{for } r_{ij} \leq r_c, \\ 0 & \text{for } r_{ij} > r_c. \end{cases} \tag{5}$$

The radial distance, r_c, supplied to f_c has to be sufficiently large to include an appropriate number of nearest neighbors. From the perspective of an absorbing atom in x-ray spectroscopy, r_c has to reflect the "*field of view*" (i.e., the maximum cutoff distance to which XANES is sensitive); for this reason, r_c = 6.0 Å throughout.

η, μ, λ, and ζ are parameters that have to be calibrated. The effects of η and μ on the radial resolution and extent and of λ and ζ on the angular resolution and extent are illustrated in Fig. 2. The calibration of these parameters can be achieved manually or automatically—in the latter case, e.g., via an intelligent sampling/Bayesian approach, decomposition, or principle component analysis (PCA),[141] or using a genetic algorithm.[137] An alternative approach designed to work "*out-of-the-box*" is given by the suggested parameterization strategy of Gastegger *et al.*, described in Ref. 137. Here, one first defines an auxiliary radial grid, **R**, as a linearly interpolated space of k points, r, between $r_{\text{min.}}$ and $r_{\text{max.}}$ and then

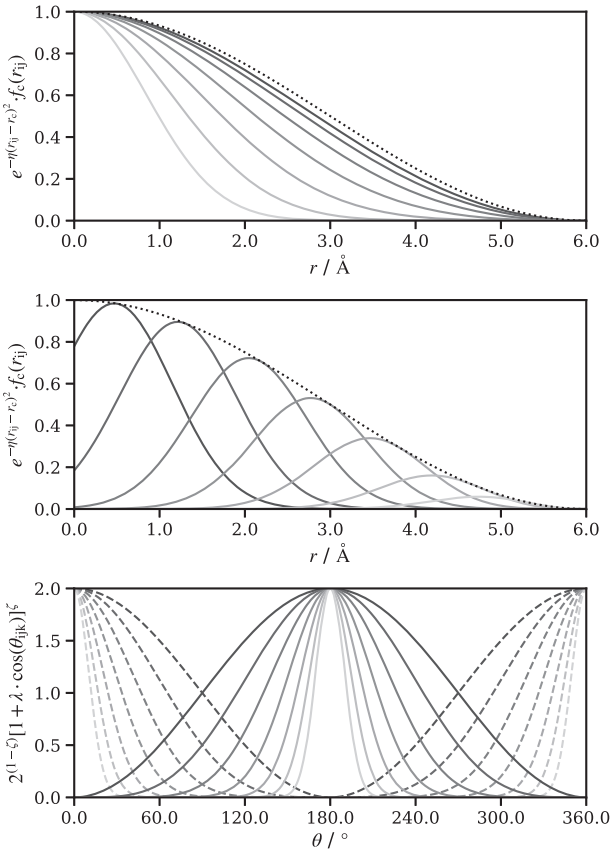

FIG. 2. Schematic of the effect of the η, μ, λ, and ζ parameters on the symmetry function forms. Upper panel: a "*centered*" parameterization scheme where μ = 0.0 and η is varied; lighter-colored lines correspond to higher values of η. Center panel: a "*shifted*" parameterization scheme where η is fixed and μ is varied; lighter-colored lines correspond to higher values of μ. Lower panel: the effect of the λ and ζ parameters on the angular component of a G^4 symmetry function; the solid and dashed lines correspond to λ = +1.0 and λ = −1.0, respectively, and lighter-colored lines correspond to higher values of ζ.

obtains either "*centered*" (Fig. 2, upper panel) pairs of η and μ parameters via setting μ to zero in all cases and setting η as

$$\eta_i = \frac{1}{2r_i^2} \tag{6}$$

or "*shifted*" (Fig. 2, center panel) pairs of η and μ parameters via setting μ to each point on the auxiliary radial grid and setting η as

$$\eta = \frac{1}{2(\Delta r)^2}. \tag{7}$$

In the former case [Eq. (6)], the wACSFs are centered at the x-ray absorption site and differ in their radial extent. In the latter case [Eq. (7)], their radial extent is constant, and their center shifts away

from the x-ray absorption site, profiling the local environment in a series of concentric "*shells*."

G^4 wACSFs additionally need to have λ and ζ parameters defined. Every pair of η and μ parameters is typically repeated for $\lambda = \pm 1.0$ to obtain a full 360° angular view, and each triple of η, μ, and λ parameters can optionally be repeated for a series of values of ζ to refine the angular resolution (Fig. 2, lower panel).

Unless explicitly stated in this article, all G^2 wACSFs were constructed according to the "*shifted*" scheme and all G^4 wACSFs were constructed according to the "*centered*" scheme.

C. XANES simulation

All first-row transition metal (Ti–Zn) K-edge XANES spectra were calculated using MST as implemented in the FDMNES[128,129] package. The spectral windows were set between −15.0 and +60.0 eV (relative to the x-ray absorption edges; see Table S1), and the absorption cross sections were calculated in steps of 0.2 eV (i.e., 376 points). A self-consistent muffin-tin potential with a cutoff radius of 6.0 Å around the x-ray absorption site was used. The interaction with the x-ray field was described by the electric quadrupole approximation, and scalar relativistic effects were included.

The calculated absorption cross sections were preprocessed via convolution with a fixed-width Lorentzian function (the width, Γ_i, depending on the x-ray absorption edge, see Table S1) and resampled via interpolation into 226 points.

III. RESULTS AND DISCUSSION

We turn to the results and discussion here, which are broken down as follows. In the first place, we parameterize a suitable $\mathbf{G}_i$ feature vector (Sec. III A) and, subsequently, explore elements of the data preprocessing pipeline (Sec. III B), assessing the performance of the XANESNET DNN at the Fe K-edge. In the second place, we explore what the XANESNET DNN takes into consideration when predicting Fe K-edge XANES spectra (i.e., which features matter and to what extent, Sec. III C). We subsequently apply the XANESNET DNN to all nine of the first-row transition metal databases (Sec. III D), showcasing its broad applicability across systems demonstrating a large variation of oxidation states and chemistries.

A. Featurization and parameterization

In this section, we address the way in which the local environments around the transition metal x-ray absorption sites are introduced into the XANESNET DNN, i.e., we address the encoding or "*featurization*," of the Cartesian coordinates as parameterized $\mathbf{G}_i$ vectors (Sec. II B 2). We initially focus on the Fe K-edge reference dataset; results for the other eight reference datasets are, however, included in the supplementary material.

In the first instance, we assess the performance of the "*centered*" and "*shifted*" parameterization schemes (Sec. II B 2) for the G^2 and G^4 wACSFs. Figure 3 displays the relative performance of the XANESNET DNN at the Fe K-edge where the local environments around the x-ray absorption sites are featurized as $\mathbf{G}_i$ vectors of length 97, i.e., containing a single G^1 wACSF and either 96 G^2 (Fig. 3, left panel) or 96 G^4 (Fig. 3, right panel) wACSFs.

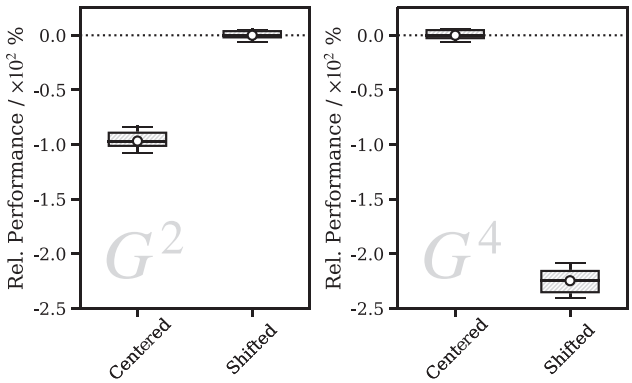

FIG. 3. Performance at the Fe K-edge for the "*centered*" and "*shifted*" parameterization schemes. Performance is plot relative (in %) to the best performance in the panel. Validation results; five-times-repeated fivefold cross-validation. Left Panel: 96 G^2 wACSFs. Right Panel: 96 G^4 wACSFs.

Reflecting the results presented in Ref. 137, we verify that the G^2 and G^4 wACSFs benefit from a "*shifted*" and "*centered*" parameterization scheme, respectively. However, the performance penalty for following the less-suitable of the two parameterization schemes is much greater for the G^4 wACSF in this work (−225%) compared to Ref. 137 (−20%). In contrast, the performance penalty for the G^2 wACSF in this work (−100%) is in line with the aforementioned results (−75%). Acknowledging differences in the $\mathbf{G}_i$ vector length and machine-learning model architecture, this result nonetheless evidences that the extent to which the G^4 wACSFs are parameterized optimally is of comparably greater importance in this work as they communicate comparably more information in the context of the present problem. This reflects either (i) a more "direct" physical relationship between the inputs and outputs [i.e., a stronger link between the local (angular) environment and the transition metal K-edge XANES spectrum (cf. enthalpies in Ref. 137), which could be expected as resonances in the post-edge are, after all, geometric in origin] or (ii) the greater importance of the G^4 wACSF, generally, in discriminating between the diverse coordination geometries of the transition metal complexes in the reference dataset(s). We return to the latter point throughout this article.

Performance is predictably improved via mixing G^2 and G^4 wACSFs. Figure 4 displays the relative performance of the XANESNET DNN at the Fe K-edge as a function of the $G^2 : G^4$ composition of the (length 97) $\mathbf{G}_i$ vector. These data are displayed for the other eight transition metal K-edge reference datasets in the supplementary material (Fig. S1) and exhibit similar trends to those shown in Fig. 4. Performance is optimal with 32 G^2 and 64 G^4 wACSFs and displays a heavy skew toward the inclusion of angular information in a 2:1 $G^4 : G^2$ ratio.

Performance is modestly improved further via the inclusion of higher values of ζ into the G^4 wACSF. In order to keep the length and composition (32 G^2 and 64 G^4 wACSFs) of the $\mathbf{G}_i$ vector constant and considering that each triple of η, μ, and λ parameters is repeated for each additional value of ζ by construction, sets of one {1}, two {1, 2}, four {1, 2, 4, 8}, and eight {1, 2, 4, 8, 16, 32, 64, 128} additional values of ζ were trialed. Figure 5 displays the relative

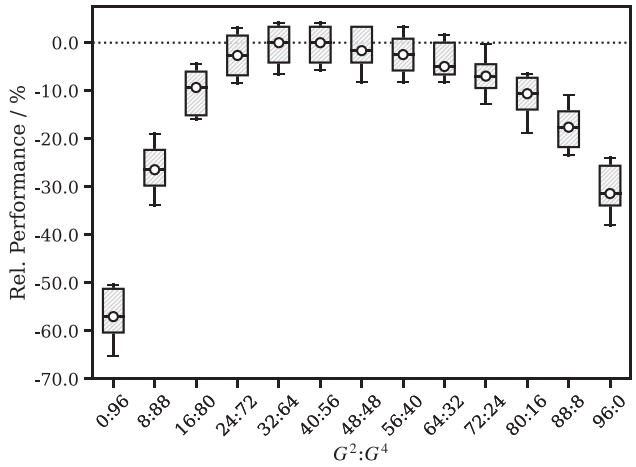

FIG. 4. Performance at the Fe K-edge as a function of the G^2 : G^4 composition of the $\mathbf{G}_i$ vector. Performance is plot relative (in %) to the best performance in the panel. Validation results; five-times-repeated fivefold cross-validation. 96 $G^{2/4}$ wACSFs.

performance of the XANESNET DNN at the Fe K-edge as a function of the greatest value of ζ, ζ_{max}, included. These data are displayed for the other eight transition metal K-edge reference datasets in the supplementary material (Fig. S2). Figure 5 shows an improvement in performance up to $\zeta_{max} = 128$ compared to $\zeta_{max} = 1$ (−10%). The inclusion of higher values of ζ focuses the angular extent of the G^4 wACSF around 180° (Fig. 2). This perhaps has limited utility in machine learning applications using popular databases of small organic systems (e.g., QM7 and QM9) where linear and right-angled triples of atoms are infrequently encountered but is of considerable utility here where it apparently improves the ability of the XANESNET DNN to discriminate between local transition metal coordination environments as these angles are commonplace in canonical coordination geometries, e.g., octahedral, square-planar, square-base, and trigonal-(bi)pyramidal.

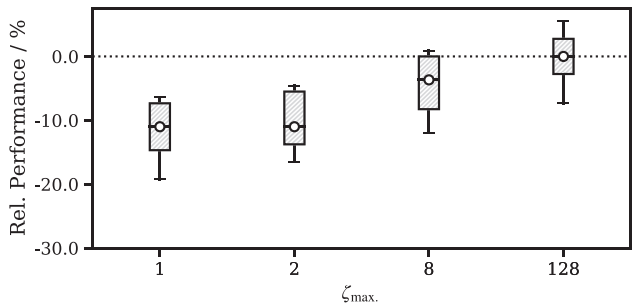

FIG. 5. Performance at the Fe K-edge as a function of the maximum value of ζ, ζ_{max}, used in the G^4 wACSF. Values of ζ used are {1}, {1, 2}, {1, 2, 4, 8}, and {1, 2, 4, 8, 16, 32, 64, 128}. Performance is plot relative (in %) to the best performance in the panel. Validation results; five-times-repeated fivefold cross-validation. 32 G^2 wACSFs and 64 G^4 wACSFs.

We will consequently carry forward a (length 97) $\mathbf{G}_i$ vector comprising the G^1 wACSF and 32 and 64 G^2 and G^4 wACSF, respectively, with G^4 wACSFs up to $\zeta_{max} = 8$ to balance the performance gain attainable by adding higher values of ζ against the cost of sacrificing pairs of μ and η parameters expressly and, consequently, limiting flexibility.

B. Optimization and performance

The $\mathbf{G}_i$ vector parameterized in Sec. III A now delivers strong performance at the Fe K-edge, yet it is still—in a sense—suboptimal as it is likely to contain low-variance features and feature-to-feature correlations as a by-product of its construction that are (in the best case) redundant or (in the worst case) an obstacle to noise-free learning. Using variance and correlation threshold filters in the data preprocessing pipeline, redundant (low-variance and/or highly correlated) features in the $\mathbf{G}_i$ vectors are able to be eliminated.

Figure 6 displays the relative performance of the XANESNET DNN at the Fe K-edge as a function of the percentage of features eliminated via action of a variance threshold filter. These data are displayed for the other eight transition metal K-edge reference datasets in the supplementary material (Fig. S3). It is possible to eliminate up to 25% of features (performance penalty $< -1\%$) from the $\mathbf{G}_i$ vector without consequence and, potentially, up to 50% of features without incurring a wholly unacceptable performance penalty (−10%), should exceptionally compact $\mathbf{G}_i$ vectors be required.

Erring on the side of caution and eliminating 25% of features from the $\mathbf{G}_i$ vector yield a truncated $\mathbf{G}_i$ vector of length 71 (with the G^1 wACSF retained and otherwise comprising 28 G^2 and 42 G^4 wACSFs). The reduced dimensions of the truncated $\mathbf{G}_i$ vector coupled with the compact $[N \times 512 \times 512 \times 226]$ architecture (Secs. II B 1 and II B 2) reduce the number of internal weights in the XANESNET DNN to 414 208 (cf. >3 000 000 in our earlier work, Ref. 71), lowering the propensity for overfitting, accelerating

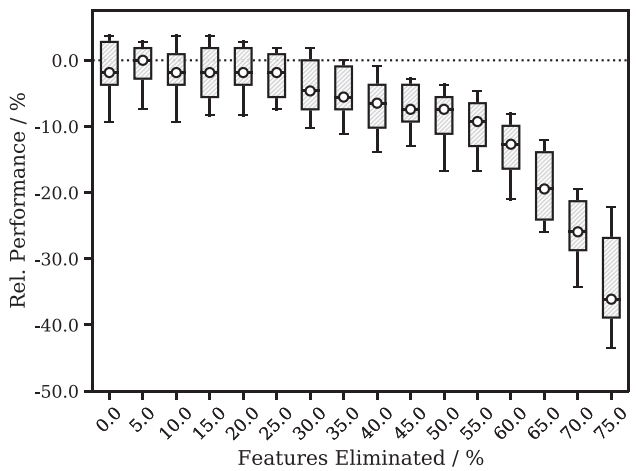

FIG. 6. Performance at the Fe K-edge as a function of the percentage of features eliminated via action of a variance threshold filter. Performance is plot relative (in %) to the best performance in the panel. Validation results; five-times-repeated fivefold cross-validation. 32 G^2 wACSFs and 64 G^4 wACSFs.

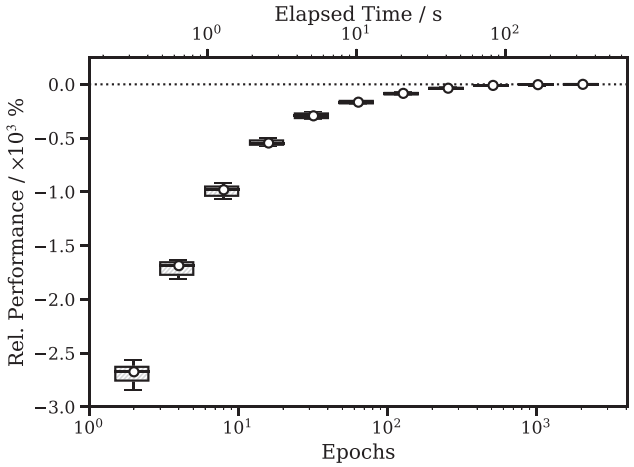

FIG. 7. Performance at the Fe K-edge as a function of the number of feedforward/backpropagation epochs and the elapsed time in seconds (optimized using an NVIDIA RTX 3070). Performance is plot relative (in %) to the best performance in the panel. Validation results; five-times-repeated fivefold cross-validation. 28 G^2 wACSFs and 42 G^4 wACSFs.

optimization, and opening up the opportunity to investigate computationally intensive feature selection algorithms (Sec. III C).

Figure 7 displays the relative performance of the XANESNET DNN at the Fe K-edge as a function of the number of feedforward/backpropagation epochs and the elapsed time in seconds taken to carry out the optimization. These data are displayed for the other eight transition metal K-edge reference datasets in the supplementary material (Fig. S4). With the reference datasets used in this article, the XANESNET DNN takes advantage of its simple and compact MLP architecture; it can be optimized to convergence in ~512–1024 feedforward/backpropagation epochs—a process that can be completed in as little as a minute using an off-the-shelf commercial-grade central processing unit (CPU) (AMD Ryzen Threadripper 3970X, 3.7–4.5 GHz) or graphics processing unit (GPU) (NVIDIA RTX 3070, 5888 CUDA cores, 1.5–1.7 GHz).

C. Feature importance and selection

In this section, we carry forward the $\mathbf{G}_i$ vector parameterized in Sec. III A with 25% of the features eliminated through the action of the variance filter as in Sec. III B. We turn our attention toward addressing a different question: what is the XANESNET DNN taking into consideration when predicting K-edge XANES spectra (i.e., which features matter, and to what extent?) and can it be considered physical?

The relative inference feature importance of each of the features comprising the $\mathbf{G}_i$ vector has been assessed via scrambling the values of the $\mathbf{G}_i$ vectors featurewise over the reference dataset and assessing the performance penalty in each instance at inference time. The objective of this feature importance experiment is to identify how reliant the XANESNET DNN is on each feature for the purpose of producing accurate predictions: the greater the performance penalty when the feature is scrambled, the greater the reliance on that feature the model expresses. Figure 8 displays the results of the feature importance experiment on the XANESNET DNN at the Fe K-edge, evaluated on the validation datasets constructed "*on-the-fly*" via five-times-repeated fivefold cross-validation. The feature importance of each of the G^2 (Fig. 8, center panel) and G^4 (Fig. 8, lower panel) wACSFs, using the relative performance as a proxy, is plot relative to the optimal baseline performance. These data are displayed for the other eight transition metal K-edge reference datasets in the supplementary material [Figs. S5 (G^2) and S6 (G^4)].

In the first place, we focus on the feature importance of the G^2 wACSF (Fig. 8, center panel); these mirror the radial distribution of atomic sites around the x-ray absorption site (Fig. 8, upper panel). The greatest feature importance is found for the first coordination shell around the x-ray absorption site [windows I, II (coordination with light, first-row elements, e.g., C, N, O, and F) and III (coordination with heavier, second-row-and-above elements, e.g., Si, P, S, Cl, Br, and I), Fig. 8, upper panel] with decreasing feature importance found for the second (windows IV and V) and third (window VI and beyond) coordination shells. The feature importance approximately reflects the density of atomic sites at the distance at which the G^2 wACSF is centered on the radial distribution, i.e., at the associated value of the μ parameter (Sec. II B 2) although this is not without exception. For example, the G^2 wACSFs centered around 1.5–1.6 Å (μ = 1.47 and 1.63 Å) have among the highest feature importance in the $\mathbf{G}_i$ vector, yet there are very few atomic sites located at this distance in the radial distribution (window I). Leakage of feature importance from the most important G^2 wACSF (μ = 1.8 Å, window II, which encodes the first coordination shell) is a contributing factor as the Gaussians centered here overlap on account of their full-widths-at-half-maxima (FWHM ~0.3 Å), and if one feature is scrambled, the radial information lost can be recovered partially from neighboring features. However, the values of the G^2 wACSFs centered around 1.5–1.6 Å are also strongly indicative of a particular class of the coordination complex in the reference dataset—the transition metal hydride—as no other atomic sites are as close to the x-ray absorption site as H in these coordination complexes. In this sense, these G^2 wACSFs act as useful yet rudimentary "classifiers" and are allocated a higher feature importance than one would otherwise expect, given the low density of atomic sites at this distance in the radial distribution.

In the second place, we focus on the feature importance of the G^4 wACSF (Fig. 8, lower panel). Each white/shaded block represents G^4 wACSF constructed with a fixed value of ζ_{max} (Sec. II B 2) from the set employed ({1, 2, 8, 128}, Sec. III A), and the trend of increasing feature importance (i.e., increasing performance) with increasing value(s) of ζ supports our earlier results. Within each white/shaded block, the same trend or pattern recurs. There are two peaks in feature importance that appear as if merged into a single peak where ζ_{max} = 1.0 and that separate as ζ_{max} is increased and the angular resolution is refined (Fig. 2). These correspond to the two key types of local angular environment around x-ray absorption sites: the linear (180°) and right-angled (90°) coordination geometries, e.g., octahedral and square-planar, among others, and the tetrahedral (105°–115°) coordination geometries. It is interesting to note that while the feature importance of the G^4 wACSF for the other eight transition metal K-edge reference datasets (Fig. S6) shows similar trends, Ni and Zn have comparably greater G^4 feature importance than one would otherwise expect. We associate this with the greater number of four-coordinate transition metal

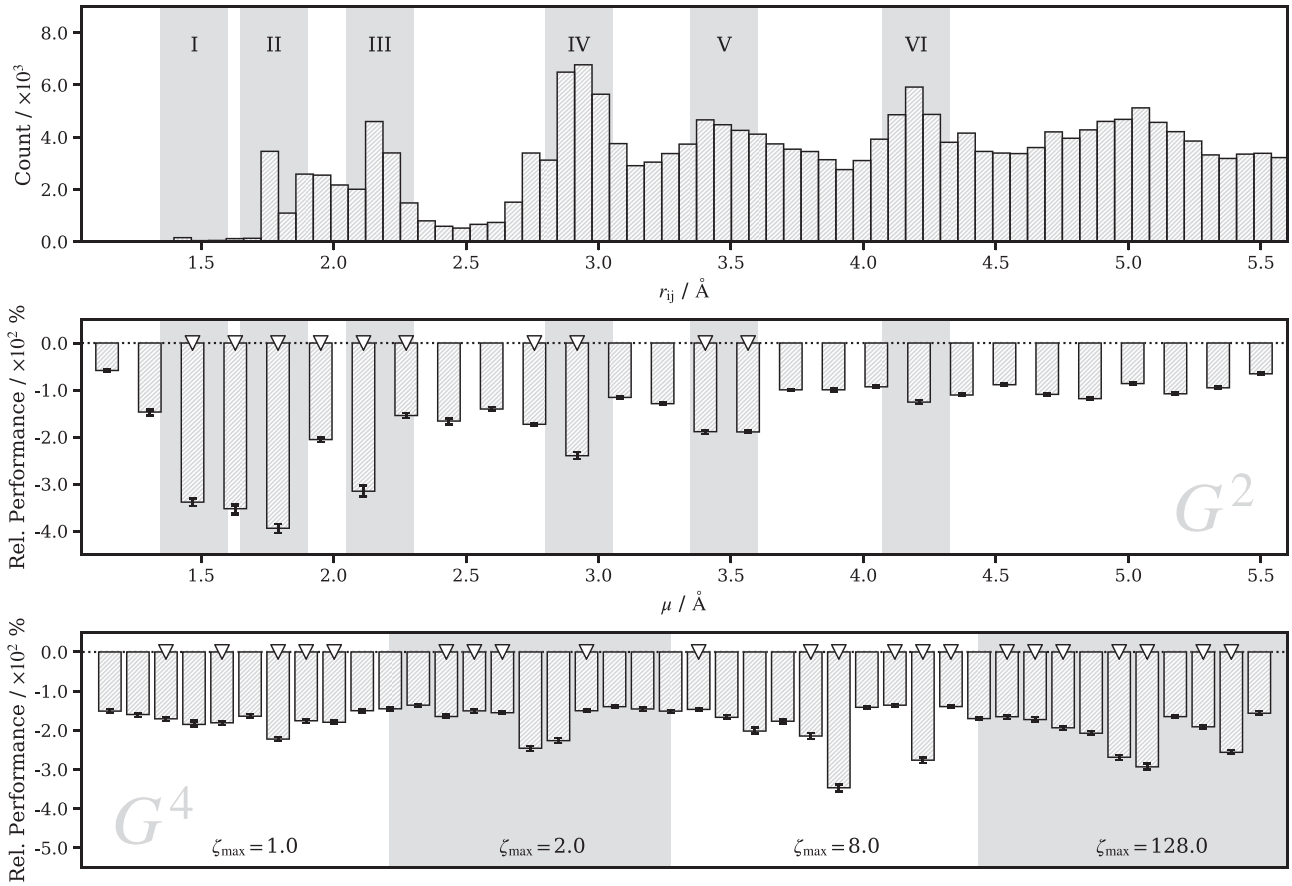

FIG. 8. Feature importance for G^2 and G^4 wACSFs at the Fe K-edge. Upper panel: histogram of the radial distribution of atomic sites around the x-ray absorption site in the Fe K-edge reference dataset. Center panel: feature importance for G^2 wACSF. Performance is plot relative (in %) to the baseline. Triangular markers indicate G^2 wACSFs selected via sequential feature selection (SFS). Lower panel: feature importance for G^4 wACSF. Performance is plot relative (in %) to the baseline. Triangular markers indicate G^4 wACSFs selected via SFS. 28 G^2 wACSFs and 42 G^4 wACSFs.

complexes contained in the Ni and Zn reference datasets[127]—in particular, the prevalence of tetrahedral and square-planar coordination geometries—and the utility of the G^4 wACSF for discriminating between them.

In Fig. 9, we alternatively assess the feature importance of the G^2 wACSF in two different regions of the XANES spectrum; a lower-energy region in the neighborhood of the x-ray absorption edge spanning −3.0 → +3.0 eV and a higher-energy region in the post-edge spanning +50.0 → +56.0 eV (relative to the x-ray absorption edge). Figure 9 displays the difference feature importance obtained by subtracting the relative feature importance in the latter from the former.

The first coordination shell is of approximately equal importance to the accurate prediction of the XANES spectrum in each of the two regions. However, G^2 wACSFs with lower and higher values of μ (encoding atomic sites closer to and further from, respectively, the x-ray absorption site) are relatively more and less important, respectively, in the higher-energy region. Figure 9 indicates a shift from a balanced reliance on all of the G^2 wACSFs in the lower-energy region near the x-ray absorption edge to increased reliance on only those G^2 wACSFs with lower values of μ that encode atomic sites in the first coordination shell as the energy is increased. Importantly, this mirrors the expected physics: core photoelectrons excited close to the x-ray absorption edge (i.e., in the lower-energy region) have low kinetic energy and, by extension, longer wavelengths—consequently, this region of the x-ray absorption spectrum is more sensitive to the structure further away from the x-ray absorption site. However, in the higher-energy region, the greater kinetic energy of the core photoelectrons—which, consequently, have shorter wavelengths —results in a reduced "*field of view*," limiting the structural sensitivity to the immediate locality of the x-ray absorption site. Indeed, resonances with energy >50 eV above the x-ray absorption edge are usually classified as belonging to the extended x-ray absorption fine structure (EXAFS) region, which is well understood to exhibit structural sensitivity only to the first coordination shell around the x-ray absorption site.[142]

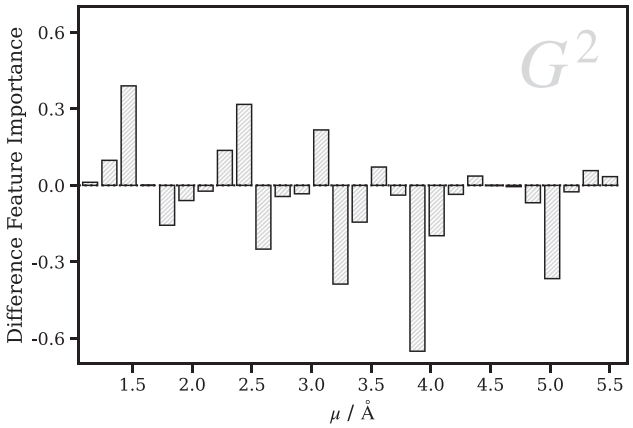

FIG. 9. Difference (high-energy region – low-energy region) feature importance for G^2 wACSF. The high-energy region of the XANES spectrum spans +50.0 → +56.0 eV, and the low-energy region of the XANES spectrum spans −3.0 → +3.0 eV (relative to the x-ray absorption edge). Fe K-edge. Validation results; five-times-repeated fivefold cross-validation. 28 G^2 wACSFs and 42 G^4 wACSFs.

Armed with what we now know about feature importance, we can use the carried-forward $\mathbf{G}_i$ vector to construct a further-truncated $\mathbf{G}_i$ vector from the ground up including only the most important features, i.e., following a "*select-from-model*" strategy. Figure 10 displays the performance of the XANESNET DNN as a function of the percentage of features included in this further-truncated $\mathbf{G}_i$ vector. Only about 60% of the features from the original carried-forward $\mathbf{G}_i$ vector are required to obtain performance that converges to the baseline. Including only these features yields a compact $\mathbf{G}_i$ vector of length 43 containing only the most important information: the G^1 wACSF and 12 and 30 G^2 and G^4 wACSFs, respectively. The composition is displayed pictorially in the inset pie chart on Fig. 10—again, the G^4 wACSFs are overweighted

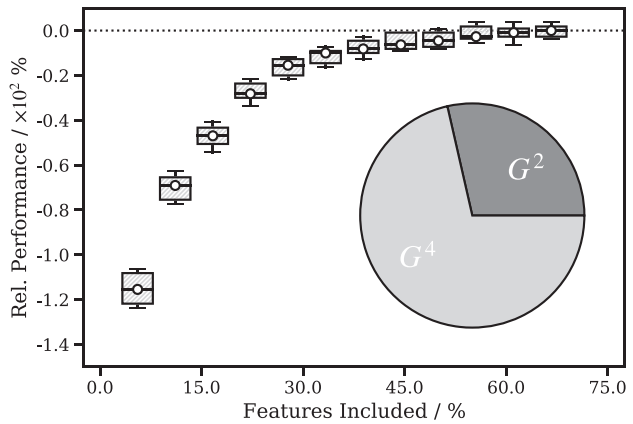

FIG. 10. Performance at the Fe K-edge as a function of the percentage of features included via a "*select-from-model*" strategy targeting high feature importance. Performance is plot relative (in %) to the baseline. Validation results; five-times-repeated fivefold cross-validation. 28 G^2 wACSFs and 42 G^4 wACSFs.

compared to the G^2 wACSF in an approximate 1:2 ratio, indicative of their importance in discriminating between the diverse coordination geometries of the transition metal complexes in the reference dataset.

To demonstrate that this ground-up construction based on feature importance is not biased by including only the features with high evaluated feature importance when taken together, i.e., from the feature importance experiment with the whole carried-forward $\mathbf{G}_i$ vector exposed to the XANESNET DNN, we have also carried out another ground-up construction and top-down deconstruction using "*forward*" and "*backward*" sequential feature selection (SFS), respectively. The SFS experiment involves adding (in the "*forward*" formulation) or eliminating (in the "*backward*" formulation) features sequentially to/from the $\mathbf{G}_i$ vector; the choice of feature to add or eliminate from the pool of available features is made to maximize the performance of the machine-learning model, and each feature addition or elimination is trialed independently. SFS is consequently a computationally intensive feature selection algorithm and can require hundreds to thousands of iterations for a DNN, depending on the target length of the desired $\mathbf{G}_i$ vector.

The plots displaying the feature importance of the G^2 (Fig. 8, center panel) and G^4 (Fig. 8, lower panel) wACSF are decorated with triangular markers above the features that were selected via "*forward*" SFS (the "*backward*" SFS result was not materially different) to obtain a further-truncated $\mathbf{G}_i$ vector of length 33. All of the G^2 wACSFs covering the first coordination shell (windows I, II, and III, Fig. 8, upper panel) were selected as were G^2 wACSFs with high feature importance in the second coordination shell (windows IV and V). Of the G^4 wACSFs, those with highest feature importance were not all selected although high-importance features were still selected more often than not, and more features were selected from high-ζ blocks.

The $\mathbf{G}_i$ vector constructed via "*forward*" SFS comprised the G^1 wACSF and 10 G^2 wACSFs and 22 G^4 wACSFs, i.e., it converged toward a similar composition and, incidentally, toward similar performance by comparison with the longer $\mathbf{G}_i$ vector constructed via the "*select-from-model*" strategy.

D. Extension to transition metal K-Edges

The XANESNET DNN demonstrably needs very little judiciously selected information to deliver accurate and affordable predictions of Fe K-edge XANES spectra for arbitrary Fe x-ray absorption sites; radial information on the first (and to a lesser extent, the second) coordination shell suffices with angular information sufficient to separate satisfactorily key coordination geometries (Sec. III C). Although the exact composition of the $\mathbf{G}_i$ vector is dataset-dependent (one of the themes we have explored in this article with respect to the coordination complexes in the tmQM dataset and the particularities of the problem at hand), the calibration carried out here is extensible across the first-row transition metal (Ti–Zn) reference datasets as coordination distances are not greatly different on average and canonical coordination geometries are found consistently. In this section, we demonstrate the performance of the XANESNET DNN at predicting the K-edge XANES spectra of the nine "*held-out*" transition metal test datasets (Ti–Zn, 250 samples each, Sec. II A).

Figure 11 displays histograms of the median percentage error, $\Delta\mu$, between target, μ_{target}, and predicted, $\mu_{predict}$, first-row transition metal K-edge XANES spectra; key properties of these distributions (medians, upper and lower quartiles, and skewness coefficients) are tabulated in Table I. Across the nine first-row transition metal reference datasets, the median $\Delta\mu$ is typically sub–5% (~4.3%, on average) with the lower and upper quartiles situated symmetrically ~2%–3% under and above, respectively, presenting a tight interquartile range of ~3%–5% that testifies to the balanced performance of the XANESNET DNN. Coupled with the high positive skewness coefficients (>1.0) across the reference datasets that place predictions squarely toward the higher-performance end of these figures, we are confident that the XANESNET DNN delivers accurate and affordable predictions that generalize well across this block of the Periodic Table. Figure S7 shows the median percentage error as a function of energy for each of the reference datasets with the lower and upper quartiles and the 5th and 95th percentiles indicated. Figures S8–S16 show illustrative example K-edge XANES spectra for each of the reference datasets, presenting examples drawn from around the median (45th–55th percentile) and the lower (20th–30th percentile) and upper (70th–80th percentile) quartiles to showcase the performance that one would expect from the XANESNET DNN.

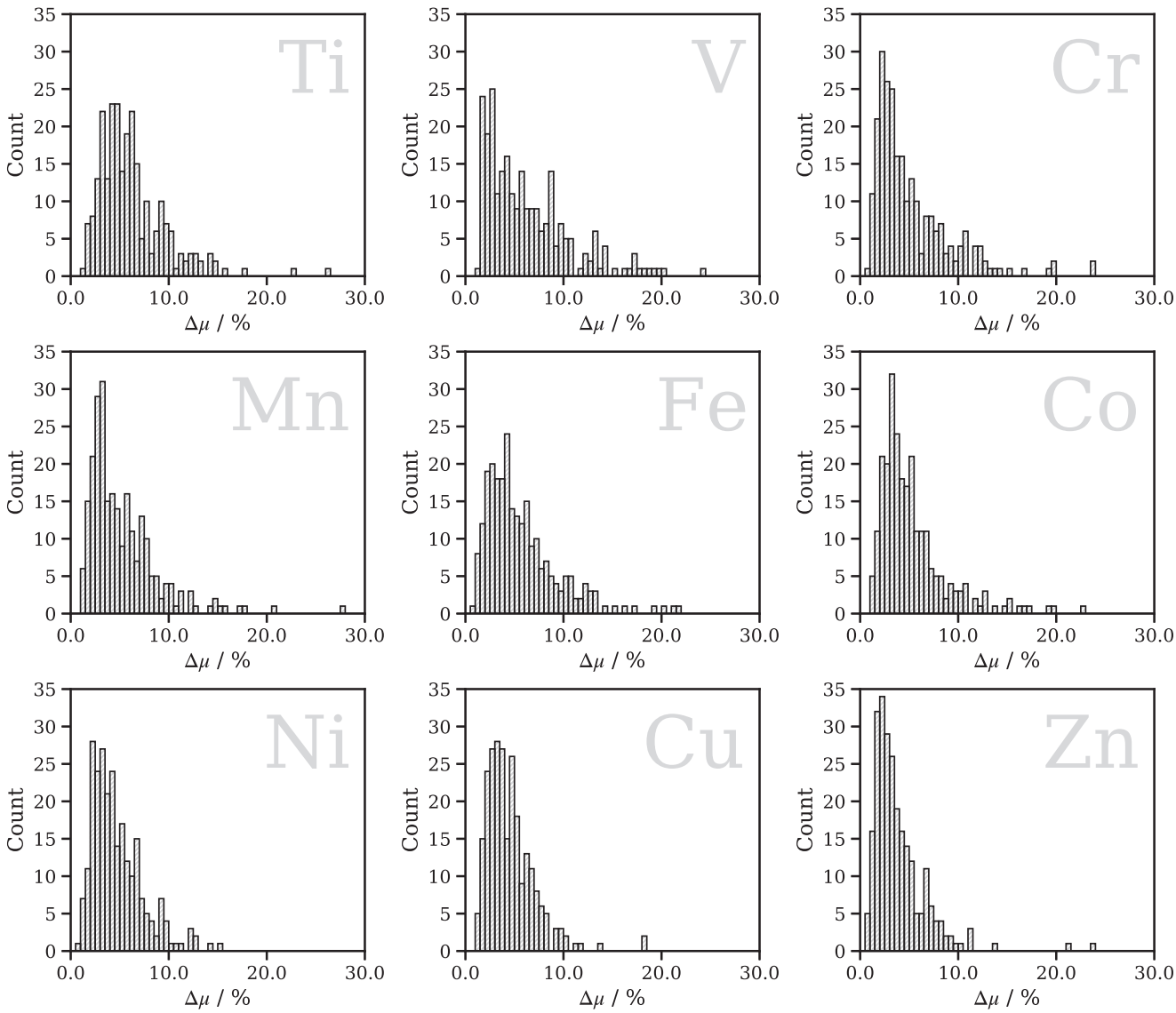

FIG. 11. Histograms of the median percentage error, $\Delta\mu$, between target, μ_{target}, and predicted, $\mu_{predict}$, first-row transition metal K-edge XANES spectra. Evaluated on nine "*held-out*" transition metal test datasets (Ti–Zn) containing 250 randomly selected samples each (Sec. II A). 28 G^2 wACSFs and 42 G^4 wACSFs.

TABLE I. Summary[a] of the median percentage errors, $\Delta\mu_{median}$ (%), upper and lower quartiles, and skewness coefficients for the $\Delta\mu$ distribution histograms (Fig. 11).

Edge	$\Delta\mu_{median}$	Upper quart.	Lower quart.	Skew.
Ti	5.5 (3.8)	7.7 (5.7)	4.0 (2.3)	1.898
V	5.2 (3.2)	8.6 (6.0)	2.9 (1.9)	1.625
Cr	3.8 (2.5)	6.9 (4.7)	2.5 (1.5)	1.926
Mn	4.3 (2.8)	6.7 (4.8)	2.9 (1.9)	2.242
Fe	4.7 (3.1)	7.2 (4.8)	3.1 (2.0)	1.607
Co	4.3 (2.8)	6.3 (4.3)	3.1 (1.9)	2.058
Ni	4.1 (2.6)	6.0 (4.0)	2.8 (1.7)	1.286
Cu	4.0 (2.7)	5.6 (4.2)	2.8 (1.7)	2.007
Zn	3.2 (2.2)	4.9 (3.5)	2.2 (1.5)	3.005

[a]Values in parenthesis are after arctangent broadening; Table S1.

The predicted K-edge XANES spectra can optionally be broadened via an additional postprocessing step to account for diverse effects on the spectral resolution including, although not limited to, core-hole lifetime broadening, instrument response, and many-body effects, e.g., inelastic losses. If this postprocessing step is carried out [as is routine and typically with an energy-dependent arctangent function, see Eq. (2) in Ref. 71], performance is improved appreciably (see the values in parentheses in Table I, arctangent broadening parameters are tabulated in Table S1). Across the nine first-row transition metal reference datasets, the median $\Delta\mu$ is reduced to ~3% (2.8%, on average) and the interquartile range tightens further to ~2%–3% post-broadening with the greatest improvements in the finely structured edge region of the K-edge XANES spectra.

Figure 12 displays parity plots of the error in energy, ΔE, between target, E_{target}, and predicted, $E_{predict}$, peak positions in the

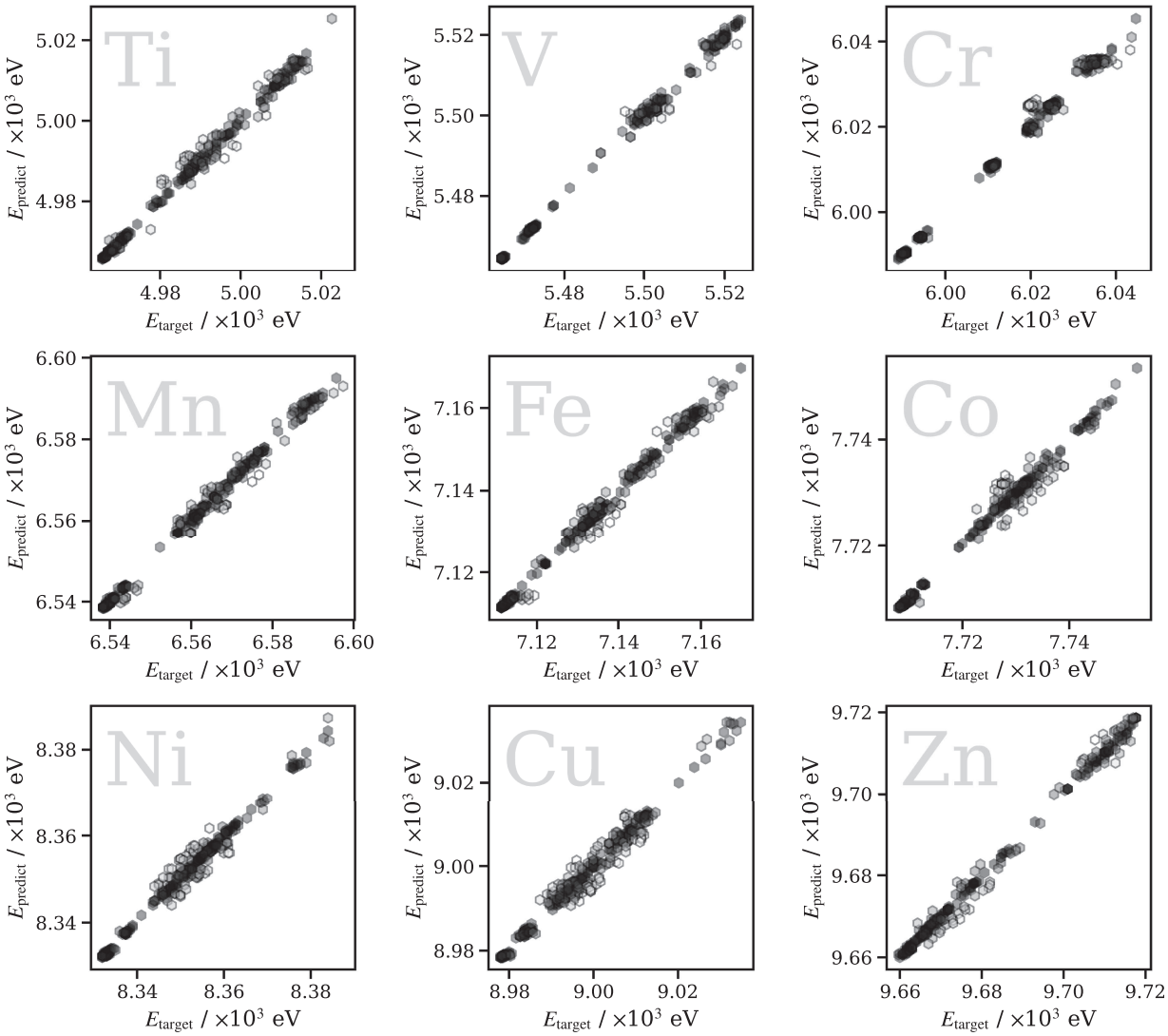

FIG. 12. Parity plots of target, E_{target}, and predicted, $E_{predict}$, peak positions. Evaluated on nine "*held-out*" transition metal test datasets (Ti–Zn) containing 250 randomly selected samples each (Sec. II A). 28 G^2 wACSFs and 42 G^4 wACSFs.

TABLE II. Summary of the mean peak position errors, ΔE_{mean} (eV), maximum peak position errors, ΔE_{max} (eV), standard deviations, σ (eV), and R^2 coefficients for the peak position parity plots (Fig. 12).

Edge	ΔE_{mean}	ΔE_{max}	σ	R^2
Ti	0.86	4.01	1.12	0.996
V	0.54	3.96	0.81	0.999
Cr	0.65	3.55	1.08	0.997
Mn	0.76	3.91	1.04	0.997
Fe	0.83	3.81	1.11	0.996
Co	0.74	5.33	1.15	0.993
Ni	0.88	4.98	1.19	0.993
Cu	0.99	4.60	1.26	0.991
Zn	0.95	4.18	1.22	0.997

first-row transition metal K-edge XANES spectra (a key metric for the experimental spectroscopist); key properties (means, maxima, standard deviations, and R^2 coefficients) are tabulated in Table II. For completeness, Fig. S17 displays parity plots of the error in intensity, $\Delta\mu$, between target, μ_{target}, and predicted, $\mu_{predict}$, peak intensities. The XANESNET DNN consistently predicts the positions of prominent peaks in the target K-edge XANES spectra to sub-eV (~0.80 eV, on average) accuracy across the nine first-row transition metal reference datasets, reproducing >90% of identified targets. The coefficients of determination, R^2—which are for all reference datasets, >0.99—evidence encouragingly strong linear relationships between E_{target} and $E_{predict}$.

IV. CONCLUSION

In this article, we have built on our earlier proof-of-principle work in Ref. 71 and practical applications in Refs. 72 and 74 to develop and deploy a new compact neural network—the XANESNET DNN—for predicting the line shape of transition metal K-edge XANES spectra. The XANESNET DNN is >80% smaller, an order of magnitude faster to optimize, and yet nonetheless displays improved predictive power and an encouraging potential for generality across the Periodic Table. We have extended the scope of our study beyond the familiar Fe K-edge to the nine first-row transition metal (Ti–Zn) K-edges and assessed the predictive power and generality of the XANESNET DNN here. Our model is able to predict K-edge XANES spectral intensities with an average accuracy of ~±2%–4% across the selected spectral windows (−15.0 → +60 eV relative to each x-ray absorption edge) and to predict the positions of prominent peaks with a >90% hit rate and sub-eV (~0.80 eV) accuracy.

We have addressed in detail the calibration of the feature vector ($\mathbf{G}_i$) that encodes the information on the local environment around the x-ray absorption site and carried out an assessment of the relative importance of the individual features—particularly the radial (G^2) and angular (G^4) components. We found that very little judiciously selected geometric information is actually needed or, indeed, used to map feature vectors onto the lineshape of the corresponding K-edge XANES spectrum; radial information on the first (and to a lesser extent, the second) coordination shells suffices alongside a quantity of angular information sufficient to separate satisfactorily key classes of coordination geometry. We found in addition that the relative importance of the individual features differs depending on the spectral window under consideration. In low-energy windows near the x-ray absorption edge, all features are taken into account in a balanced way, while in higher-energy windows in the post-edge, features encoding radial information closer to the x-ray absorption site are ascribed higher importance, mirroring the expected physics in the shift from multiple scattering to single scattering with increasing energy.

Although the exact composition of the feature vector is dataset-dependent (one of the themes explored in this article with respect to the coordination complexes in the tmQM dataset and the particularities of our problem), the calibration carried out here has, nonetheless, proved extensible across our first-row transition metal (Ti–Zn) reference datasets.

While accuracy, affordability, and generality (with respect to the identity of the absorption site) are no longer cardinal challenges, there are, of course, new challenges to tackle and opportunities to embrace which, most pressingly, include (i) the incorporation of electronic information and (ii) dataset curation. On the topic of (i), the XANESNET DNN currently considers only the local geometric environment around the x-ray absorption site of interest—consequently, its ability to describe charge-state-dependent spectral features remains uncertain. For (ii), high-quality balanced training sets rivaling popular molecular organic datasets have to be curated/constructed. Here, there is a great potential for intelligent (guided) and/or combinatorial strategies, which we expect to work well alongside advances in high-throughput computing.

SUPPLEMENTARY MATERIAL

The supplementary material contains a summary of reference datasets used for each of the nine first-row transition metals. It also includes data analyzing the parameterization of the feature vector associated with the eight (not Fe) first-row transition metal systems not shown in the main text. Finally, percentage errors as a function of energy for predicted XANES spectra, example spectra, and parity plots for each transition metal are shown.

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, and EP/R51309X/1). C.D.R. is supported by a Doctoral Prize Fellowship (Grant No. EP/R51309X/1). This research made use of the Rocket High Performance Computing (HPC) service at Newcastle University. C.D.R. additionally acknowledges the Alan Turing Institute via which access to the EPSRC-supported (Grant No. EP/T022205/1) Joint Academic Data Science Endeavor (JADE) HPC cluster was provided under Project No. JAD029.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Open Data Commons Open Database License at http://dx.doi.org/10.25405/data.ncl.19087478.

REFERENCES

[1]G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Rev. Mod. Phys. **91**, 045002 (2019).
[2]J. Gasteiger and J. Zupan, Angew. Chem., Int. Ed. **32**, 503 (1993).
[3]A. C. Mater and M. L. Coote, J. Chem. Inf. Model. **59**, 2545 (2019).
[4]K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Nature **559**, 547 (2018).
[5]W. Sun, Y. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao *et al.*, Sci. Adv. **5**, eaay4275 (2019).
[6]N. Artrith, J. Phys: Energy **1**, 032002 (2019).
[7]S. Chibani and F.-X. Coudert, APL Mater. **8**, 080701 (2020).
[8]R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, npj Comput. Mater. **3**, 54 (2017).
[9]Q. Tao, P. Xu, M. Li, and W. Lu, npj Comput. Mater. **7**, 23 (2021).
[10]Z. Li, X. Ma, and H. Xin, Catal. Today **280**, 232 (2017).
[11]Z. Li, S. Wang, W. S. Chin, L. E. Achenie, and H. Xin, J. Mater. Chem. A **5**, 24131 (2017).
[12]J. R. Kitchin, Nat. Catal. **1**, 230 (2018).
[13]A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, and S. E. Denmark, Science **363**, eaau5631 (2019).
[14]R. Mercado, T. Rastemo, E. Lindelöf, G. Klambauer, O. Engkvist, H. Chen, and E. Jannik Bjerrum, Mach. Learn.: Sci. Technol. **2**, 025023 (2021).
[15]V. D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A. G. Papadiamantis, V. Aidinis, I. Lynch, D. Greco, and G. Melagraki, Int. J. Mol. Sci. **22**, 1676 (2021).
[16]H. Achdout, A. Aimon, E. Bar-David, H. Barr, A. Ben-Shmuel, J. Bennett, M. L. Bobby, J. Brun, S. Bvnbs, M. Calmiano *et al.*, bioRxiv:10.1101/2020.10.29.339317 (2020).
[17]M. H. S. Segler, M. Preuss, and M. P. Waller, Nature **555**, 604 (2018).
[18]C. W. Coley, W. H. Green, and K. F. Jensen, Acc. Chem. Res. **51**, 1281 (2018).
[19]C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, Chem. Sci. **10**, 370 (2019).
[20]P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, and T. Laino, Chem. Sci. **11**, 3316 (2020).
[21]J. S. Schreck, C. W. Coley, and K. J. M. Bishop, ACS Cent. Sci. **5**, 970 (2019).
[22]V. H. Nair, P. Schwaller, and T. Laino, CHIMIA **73**, 997 (2019).
[23]D. P. Kovács, W. McCorkindale, and A. A. Lee, Nat. Commun. **12**, 1695 (2021).
[24]P. Schwaller, A. C. Vaucher, T. Laino, and J.-L. Reymond, Mach. Learn.: Sci. Technol. **2**, 015016 (2021).
[25]W. Gao, R. Mercado, and C. W. Coley, arXiv:2110.06389 (2021).
[26]G. B. Goh, N. O. Hodas, and A. Vishnu, J. Comput. Chem. **38**, 1291 (2017).
[27]K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, Nat. Commun. **10**, 5024 (2019).
[28]M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, and K. Burke, Nat. Commun. **11**, 5223 (2020).
[29]F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, Annu. Rev. Phys. Chem. **71**, 361 (2020).
[30]P. O. Dral, J. Phys. Chem. Lett. **11**, 2336 (2020).
[31]O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, Nat. Rev. Chem. **4**, 347 (2020).
[32]B. Huang and O. A. Von Lilienfeld, Chem. Rev. **121**, 10001 (2021).
[33]J. Westermayr, M. Gastegger, K. T. Schütt, and R. J. Maurer, J. Chem. Phys. **154**, 230903 (2021).
[34]S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017).
[35]O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, Chem. Rev. **121**, 10142 (2021).
[36]V. Vassilev-Galindo, G. Fonseca, I. Poltavsky, and A. Tkatchenko, J. Chem. Phys. **154**, 094119 (2021).
[37]G. Fonseca, I. Poltavsky, V. Vassilev-Galindo, and A. Tkatchenko, J. Chem. Phys. **154**, 124102 (2021).
[38]I. Poltavsky and A. Tkatchenko, J. Phys. Chem. Lett. **12**, 6551 (2021).
[39]Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, and T. F. Miller III, J. Chem. Phys. **153**, 124111 (2020).
[40]A. S. Christensen, S. K. Sirumalla, Z. Qiao, M. B. O'Connor, D. G. A. Smith, F. Ding, P. J. Bygrave, A. Anandkumar, M. Welborn, F. R. Manby *et al.*, J. Chem. Phys. **155**, 204103 (2021).
[41]M. Welborn, L. Cheng, and T. F. Miller III, J. Chem. Theory Comput. **14**, 4772 (2018).
[42]L. Cheng, N. B. Kovachki, M. Welborn, and T. F. Miller III, J. Chem. Theory Comput. **15**, 6668 (2019).
[43]S. Dick and M. Fernandez-Serra, Nat. Commun. **11**, 3509 (2020).
[44]W.-K. Chen, X.-Y. Liu, W.-H. Fang, P. O. Dral, and G. Cui, J. Phys. Chem. Lett. **9**, 6702 (2018).
[45]P. O. Dral, M. Barbatti, and W. Thiel, J. Phys. Chem. Lett. **9**, 5660 (2018).
[46]J. Westermayr, M. Gastegger, M. F. S. J. Menger, S. Mai, L. González, and P. Marquetand, Chem. Sci. **10**, 8100 (2019).
[47]J. Westermayr and P. Marquetand, Mach. Learn.: Sci. Technol. **1**, 043001 (2020).
[48]J. Westermayr, F. A. Faber, A. S. Christensen, O. A. von Lilienfeld, and P. Marquetand, Mach. Learn.: Sci. Technol. **1**, 025009 (2020).
[49]J. Westermayr, P. Marquetand, and P. Marquetand, J. Chem. Phys. **153**, 154112 (2020).
[50]J. Westermayr and R. J. Maurer, Chem. Sci. **12**, 10755 (2021).
[51]J. Westermayr and P. Marquetand, Chem. Rev. **121**, 9873 (2021).
[52]W. B. How, B. Wang, W. Chu, A. Tkatchenko, and O. V. Prezhdo, J. Phys. Chem. Lett. **12**, 12026 (2021).
[53]B. Wang, W. Chu, A. Tkatchenko, and O. V. Prezhdo, J. Phys. Chem. Lett. **12**, 6070 (2021).
[54]P. O. Dral and M. Barbatti, Nat. Rev. Chem. **5**, 388 (2021).
[55]A. Ullah and P. O. Dral, New J. Phys. **23**, 113019 (2021).
[56]P. Emma, R. Akre, J. Arthur, R. Bionta, C. Bostedt, J. Bozek, A. Brachmann, P. Bucksbaum, R. Coffee, F.-J. Decker *et al.*, Nat. Photonics **4**, 641 (2010).
[57]E. Allaria, R. Appio, L. Badano, W. A. Barletta, S. Bassanese, S. G. Biedron, A. Borga, E. Busetto, D. Castronovo, P. Cinquegrana *et al.*, Nat. Photonics **6**, 699 (2012).
[58]T. Ishikawa, H. Aoyagi, T. Asaka, Y. Asano, N. Azumi, T. Bizen, H. Ego, K. Fukami, T. Fukui, Y. Furukawa *et al.*, Nat. Photonics **6**, 540 (2012).
[59]D. Khakhulin, F. Otte, M. Biednov, C. Bömer, T.-K. Choi, M. Diez, A. Galler, Y. Jiang, K. Kubicek, F. A. Lima *et al.*, Appl. Sci. **10**, 995 (2020).
[60]M. Maiuri, M. Garavelli, and G. Cerullo, J. Am. Chem. Soc. **142**, 3 (2019).
[61]C. A. Meza Ramirez, M. Greenop, L. Ashton, and I. ur Rehman, Appl. Spectrosc. Rev. **56**, 733 (2021).
[62]M. Gastegger, J. Behler, and P. Marquetand, Chem. Sci. **8**, 6924 (2017).
[63]J. L. Lansford and D. G. Vlachos, Nat. Commun. **11**, 1513 (2020).
[64]K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, and P. Rinke, Adv. Sci. **6**, 1801367 (2019).
[65]Y. Zhang, S. Ye, J. Zhang, C. Hu, J. Jiang, and B. Jiang, J. Phys. Chem. B **124**, 7284 (2020).
[66]R. P. Xian, V. Stimper, M. Zacharias, S. Dong, M. Dendzik, S. Beaulieu, B. Schölkopf, M. Wolf, L. Rettig, C. Carbogno, S. Bauer, and R. Ernstorfer, arXiv:2005.10210 (2020).
[67]B.-X. Xue, M. Barbatti, and P. O. Dral, J. Phys. Chem. A **124**, 7199 (2020).
[68]L. Pan, P. Zhang, C. Daengngam, S. Peng, and M. Chongcheawchamnan, J. Raman Spectrosc. **53**, 6 (2022).
[69]Z. Chen, N. Andrejevic, N. C. Drucker, T. Nguyen, R. P. Xian, T. Smidt, Y. Wang, R. Ernstorfer, D. A. Tennant, M. Chan, and M. Li, Chem. Phys. Rev. **2**, 031301 (2021).
[70]F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, Nat. Commun. **9**, 4501 (2018).

[71] C. D. Rankine, M. M. M. Madkhali, and T. J. Penfold, J. Phys. Chem. A **124**, 4263 (2020).
[72] M. M. M. Madkhali, C. D. Rankine, and T. J. Penfold, Molecules **25**, 2715 (2020).
[73] M. M. M. Madkhali, C. D. Rankine, and T. J. Penfold, Phys. Chem. Chem. Phys. **23**, 9259 (2021).
[74] E. Falbo, C. D. Rankine, and T. J. Penfold, Chem. Phys. Lett. **780**, 138893 (2021).
[75] M. R. Carbone, S. Yoo, M. Topsakal, and D. Lu, Phys. Rev. Mater. **3**, 033604 (2019).
[76] M. R. Carbone, M. Topsakal, D. Lu, and S. Yoo, Phys. Rev. Lett. **124**, 156401 (2020).
[77] K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J. J. Rehr, S. P. Ong, and K. A. Persson, Sci. Data **5**, 180151 (2018).
[78] C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson, and S. P. Ong, npj Comput. Mater. **4**, 12 (2018).
[79] C. Zheng, C. Chen, Y. Chen, and S. P. Ong, Patterns **1**, 100013 (2020).
[80] J. Timoshenko, D. Lu, Y. Lin, and A. I. Frenkel, J. Phys. Chem. Lett. **8**, 5091 (2017).
[81] J. Timoshenko, A. Halder, B. Yang, S. Seifert, M. J. Pellin, S. Vajda, and A. I. Frenkel, J. Phys. Chem. C **122**, 21686 (2018).
[82] J. Timoshenko, M. Ahmadi, and B. Roldan Cuenya, J. Phys. Chem. C **123**, 20594 (2019).
[83] M. Ahmadi, J. Timoshenko, F. Behafarid, and B. Roldan Cuenya, J. Phys. Chem. C **123**, 10666 (2019).
[84] J. Timoshenko, C. J. Wrasman, M. Luneau, T. Shirman, M. Cargnello, S. R. Bare, J. Aizenberg, C. M. Friend, and A. I. Frenkel, Nano Lett. **19**, 520 (2019).
[85] J. Timoshenko and A. I. Frenkel, ACS Catal. **9**, 10192 (2019).
[86] I. Miyazato, L. Takahashi, and K. Takahashi, Mol. Syst. Des. Eng. **4**, 1014 (2019).
[87] S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram, and L. Hung, npj Comput. Mater. **6**, 109 (2020).
[88] S. Kiyohara and T. Mizoguchi, J. Phys. Soc. Jpn. **89**, 103001 (2020).
[89] A. A. Guda, S. A. Guda, A. Martini, A. L. Bugaev, M. A. Soldatov, A. V. Soldatov, and C. Lamberti, Radiat. Phys. Chem. **175**, 108430 (2020).
[90] S. A. Guda, A. S. Algasov, A. A. Guda, A. Martini, A. N. Kravtsova, A. L. Bugaev, L. V. Guda, and A. V. Soldatov, J. Surf. Invest.: X-Ray, Synchrotron Neutron Tech. **15**, 934 (2021).
[91] D. Y. Kirsanova, M. A. Soldatov, Z. M. Gadzhimagomedova, D. M. Pashkov, A. V. Chernov, M. A. Butakova, and A. V. Soldatov, J. Surf. Invest.: X-Ray, Synchrotron Neutron Tech. **15**, 485 (2021).
[92] E. G. Kozyr, A. L. Bugaev, S. A. Guda, A. A. Guda, K. A. Lomachenko, K. Janssens, S. Smolders, D. De Vos, and A. V. Soldatov, J. Phys. Chem. C **125**, 27844 (2021).
[93] D. M. Pashkov, A. A. Guda, M. V. Kirichkov, S. A. Guda, A. Martini, S. A. Soldatov, and A. V. Soldatov, J. Phys. Chem. C **125**, 8656 (2021).
[94] A. Martini, A. L. Bugaev, S. A. Guda, A. A. Guda, E. Priola, E. Borfecchia, S. Smolders, K. Janssens, D. De Vos, and A. V. Soldatov, J. Phys. Chem. A **125**, 7080 (2021).
[95] A. Martini, A. A. Guda, S. A. Guda, A. L. Bugaev, O. V. Safonova, and A. V. Soldatov, Phys. Chem. Chem. Phys. **23**, 17873 (2021).
[96] A. Tereshchenko, D. Pashkov, A. Guda, S. Guda, Y. Rusalev, and A. Soldatov, Molecules **27**, 357 (2022).
[97] S. Tetef, N. Govind, and G. T. Seidler, Phys. Chem. Chem. Phys. **23**, 23586 (2021).
[98] P. M. Mishra, L. Avaldi, P. Bolognesi, K. C. Prince, R. Richter, and U. R. Kadhane, J. Phys. Chem. A **118**, 3128 (2014).
[99] Y. Mei, C. Li, N. Q. Su, and W. Yang, J. Phys. Chem. A **123**, 666 (2018).
[100] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, Phys. Rev. B **89**, 205118 (2014).
[101] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).
[102] A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen, and P. Rinke, J. Chem. Phys. **150**, 204121 (2019).
[103] O. Rahaman and A. Gagliardi, J. Chem. Inf. Model. **60**, 5971 (2020).
[104] A. Sanchez-Gonzalez, P. Micaelli, C. Olivier, T. Barillot, M. Ilchen, A. Lutman, A. Marinelli, T. Maxwell, A. Achner, M. Agåker *et al.*, Nat. Commun. **8**, 15461 (2017).
[105] A. A. Kananenka, K. Yao, S. A. Corcelli, and J. L. Skinner, J. Chem. Theory Comput. **15**, 6850 (2019).
[106] C. D. Rankine and T. J. Penfold, J. Phys. Chem. A **125**, 4276 (2021).
[107] G. Capano, M. Chergui, U. Rothlisberger, I. Tavernelli, and T. J. Penfold, J. Phys. Chem. A **118**, 9861 (2014).
[108] G. Capano, T. J. Penfold, U. Röthlisberger, and I. Tavernelli, CHIMIA **68**, 227 (2014).
[109] G. Capano, C. J. Milne, M. Chergui, U. Rothlisberger, I. Tavernelli, and T. T. J. Penfold, J. Phys. B: At., Mol. Opt. Phys. **48**, 214001 (2015).
[110] T. Katayama, T. Northey, W. Gawelda, C. J. Milne, G. Vankó, F. A. Lima, R. Bohinc, Z. Németh, S. Nozawa, T. Sato *et al.*, Nat. Commun. **10**, 3606 (2019).
[111] N. H. List, A. L. Dempwolff, A. Dreuw, P. Norman, and T. J. Martínez, Chem. Sci. **11**, 4180 (2020).
[112] T. Northey, J. Norell, A. E. A. Fouda, N. A. Besley, M. Odelius, and T. J. Penfold, Phys. Chem. Chem. Phys. **22**, 2667 (2020).
[113] T. J. Penfold, M. Pápai, T. Rozgonyi, K. B. Møller, and G. Vankó, Faraday Discuss. **194**, 731 (2016).
[114] S. P. Neville, V. Averbukh, S. Patchkovskii, M. Ruberti, R. Yun, M. Chergui, A. Stolow, and M. S. Schuurman, Faraday Discuss. **194**, 117 (2016).
[115] S. P. Neville, V. Averbukh, M. Ruberti, R. Yun, S. Patchkovskii, M. Chergui, A. Stolow, and M. S. Schuurman, J. Chem. Phys. **145**, 144307 (2016).
[116] S. P. Neville, M. Chergui, A. Stolow, and M. S. Schuurman, Phys. Rev. Lett. **120**, 243001 (2018).
[117] I. Seidu, S. P. Neville, R. J. MacDonell, and M. S. Schuurman, Phys. Chem. Chem. Phys. **24**, 1345 (2022).
[118] T. J. Penfold, J. Szlachetko, F. G. Santomauro, A. Britz, W. Gawelda, G. Doumy, A. M. March, S. H. Southworth, J. Rittmann, R. Abela *et al.*, Nat. Commun. **9**, 478 (2018).
[119] N. Huse, H. Wen, D. Nordlund, E. Szilagyi, D. Daranciang, T. A. Miller, A. Nilsson, R. W. Schoenlein, and A. M. Lindenberg, Phys. Chem. Chem. Phys. **11**, 3951 (2009).
[120] G. Gavrila, K. Godehusen, C. Weniger, E. T. J. Nibbering, T. Elsaesser, W. Eberhardt, and P. Wernet, Appl. Phys. A **96**, 11 (2009).
[121] M. Reinhard, T. J. Penfold, F. A. Lima, J. Rittmann, M. H. Rittmann-Frank, R. Abela, I. Tavernelli, U. Rothlisberger, C. J. Milne, and M. Chergui, Struct. Dyn. **1**, 024901 (2014).
[122] V.-T. Pham, T. J. Penfold, R. M. van der Veen, F. Lima, A. El Nahhas, S. L. Johnson, P. Beaud, R. Abela, C. Bressler, I. Tavernelli *et al.*, J. Am. Chem. Soc. **133**, 12740 (2011).
[123] J. Szlachetko, J. Sá, M. Nachtegaal, U. Hartfelder, J.-C. Dousse, J. Hoszowska, D. L. Abreu Fernandes, H. Shi, and C. Stampfl, J. Phys. Chem. Lett. **5**, 80 (2014).
[124] O. Cannelli, C. Bacellar, R. A. Ingle, R. Bohinc, D. Kinschel, B. Bauer, D. S. Ferreira, D. Grolimund, G. F. Mancini, and M. Chergui, Struct. Dyn. **6**, 064303 (2019).
[125] Y. Lecun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).
[126] Quantum Machine, 2021, quantum-machine.org/datasets.
[127] D. Balcells and B. B. Skjelstad, J. Chem. Inf. Model. **60**, 6135 (2020).
[128] O. Bunău and Y. Joly, J. Phys.: Condens. Matter **21**, 345501 (2009).
[129] O. Bunău, A. Y. Ramos, and Y. Joly, *International Tables for Crystallography, Vol. I: X-ray Absorption Spectroscopy and Related Techniques* (Wiley, 2021).
[130] D. P. Kingma and J. L. Ba, arXiv:1412.6980 (2014).
[131] K. He, X. Zhang, S. Ren, and J. Sun, arXiv:1502.01852 (2015).
[132] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," (2015); available at tensorflow.org.
[133] Keras, 2015, github.com/keras-team/keras.
[134] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, J. Mach. Learn. Res. **12**, 2825 (2011).

[135]A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus *et al.*, J. Phys.: Condens. Matter **29**, 273002 (2017).
[136]XANESNET, 2021, gitlab.com/conor.rankine/xanesnet.
[137]M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, J. Chem. Phys. **148**, 241709 (2018).
[138]J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[139]J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[140]J. Behler, Chem. Rev. **121**, 10037 (2021).
[141]G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).
[142]T. J. Penfold, C. J. Milne, and M. Chergui, Adv. Chem. Phys. **153**, 1 (2013).