Table of Contents

[1 Chapter 5 – A survey of new developments between X-ray spectroscopy and machine learning 1](#_Toc124859637)

[1.1 Solving the inverse problem 1](#_Toc124859638)

[1.1.1 Supervised machine learning approaches 1](#_Toc124859639)

[1.1.2 Unsupervised machine learning approaches 8](#_Toc124859640)

[1.2 Solving the forward problem 10](#_Toc124859641)

[1.3 References 12](#_Toc124859642)

# Chapter 5 – A survey of new developments between X-ray spectroscopy and machine learning

In this chapter, we will give an overview of the work utilizing machine learning in the context of X-ray absorption spectroscopy, at least at the time that this dissertation was written. We will focus on the inverse problem, where the majority of work is on XANES regression problems.

## Solving the inverse problem

### Supervised machine learning approaches

Seminal work includes the papers by Timoshenko and coworkers, where they predicted coordination from XANES rather than EXAFS.

**Supervised Machine-Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles, Timoshenko, et al., 2017 [Timoshenko, 2017]**

1. The dataset: Pt XANES of nanoparticles
2. What they did: Predict coordination of Pt nanoparticles from spectra using a neural network

Chart

Description automatically generated

Fig. 1 (a) shows the test nanoparticle structures. (b) and (c) show the true versus predicted coordination numbers from the neural network for both the first and fourth coordination shell, respectively, on the test dataset.

**Probing Atomic Distributions in Mono- and Bimetallic Nanoparticles by Supervised Machine Learning, Timoshenko, et al., 2019 [Timoshenko, 2019]**

1. The dataset: EXAFS of Pt and PdAu nanoparticles
2. What they did: Predict partial radial distribution functions (RDF) using a neural network from the wavelet-transformed EXAFS

**“Inverting” X‑ray Absorption Spectra of Catalysts by Machine Learning in Search for Activity Descriptors,** **Timoshenko and Frenkel, 2019 [Timoshenko and Frenkel, 2019]**

1. The dataset: Various
2. What they did: Perspective on an overview of the ML methods one can use to solve the inverse problem for XANES, including decision trees, neural networks, PCA, and MCR-ALS (multivariate curve resolution-alternating least squares).

Another notable paper is the work with the Materials Project database to match an unknown spectrum against all the spectra in the corresponding spectral database for the closest and thus likeliest candidates. Moreover, instead of using a neural network, they utilized ensemble learning, which should (theoretically) be more generalizable than a neural network and has thus seen a rise in popularity, along with gradient boosting algorithms.

**Automated generation and ensemble-learned matching of X-ray absorption spectra,** **Zheng, et al., 2018 [Zheng, 2018]**

1. The dataset: XASdb, a database of over 800,000 k-edge XANES spectra of structures from the Materials Project
2. What they did: Developed the Ensemble-Learned Spectra IdEntification (ELSIE) algorithm, which uses an ensemble of “weak” learners to compare similar spectra and thus identify oxidation state or coordination number.

The following papers explore different ways to generate features of XANES spectra. In each, both features (of the spectra) and properties were selected, and then correlations between those features and the properties of interest were identified.

**Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms, Guda, et al., 2021 [Guda, 2021]**

1. The dataset: XANES of FeSiO2
2. What they did: Correlate XANES features (edge position, intensities, positions, and curvatures of minima and maxima) to predict properties (coordination numbers, bond distances and angles, and oxidation state) using Elastic Net (combining ridge and LASSO regression)

Chart

Description automatically generated

Fig. 2 Training dataset of each descriptor versus property, where the color reflects the CN values in (a), (b), the average Fe-O distance in (c), (d), and iron valence in (e), (f).

**Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships, Torrisi, et al., 2020 [Torrisi, 2020]**

1. The dataset: XANES of several different 3d transition metals
2. What they did: Correlate XANES features to properties (coordination number, Bader charge, and mean nearest neighbor distance), where featurization of XANES spectra include both just pointwise energy values (normal) and third order polynomial fits to different sized energy regions using a random forest (RF)

Diagram

Description automatically generated with medium confidence

Fig. 3 Featurization of XANES spectra included the pointwise energy-intensity values and fitting to third order polynomials for regions with varying energy resolutions.

**How Much Structural Information Could Be Extracted from XANES Spectra for Palladium Hydride and Carbide Nanoparticles, Usoltsev, et al., 2020 [Usoltsev, 2020]**

1. The dataset: Pd K-edge XANES of nanoparticles
2. What they did: Correlate various structural descriptors and their combinations (such as Pd-Pd interatomic distances, hydrogen concentration, and adsorbed hydrocarbons) to the “pure” spectral components obtained from MCR (multivariate curve resolution), where they used PCA to determine number of MCR components

Here are other supervised machine learning applications, but in the context of fitting XANES spectra.

**PyFitit: The software for quantitative analysis of XANES spectra using machine-learning algorithms, Martini, et al., 2019 [Martini, 2019]**

1. The dataset: Fitting XANES spectra package with Ce L3 edge and Fe(terpy)2 as examples
2. What they did: Array of ML algorithms (Gradient Boosting of Random Trees, Radial Basis Functions and Neural Networks) to fit spectra. They also use Latin hypercube sampling (LHS) to generate molecular deformations, which is generally most effective for sampling a high-dimensional parameter space .

**Assessing arsenic species in foods using regularized linear regression of the arsenic K-edge X-ray absorption near edge structure, Jahrman, et al., 2022 [Jahrman, 2022]**

1. The dataset: As K-edge XANES
2. What they did: LASSO regression to perform linear combination fitting onto a reference library from the spectra

**Random Forest Models for Accurate Identification of Coordination Environments from X-Ray Absorption Near-Edge Structure, Zheng, et al., 2020 [Zheng, 2020]**

1. The dataset: K-edge XANES
2. What they did: Predict coordination number from XANES spectra using a random forest, which was trained on a database of 190,000 spectra. They analyze feature importance using the drop-variable technique.

**Solving the structure of ‘‘single-atom’’ catalysts using machine learning – assisted XANES analysis, Xiang, et al., 2022 [Xiang, 2022]**

1. The dataset: Co XANES of “single-atom” catalysts
2. What they did: PCA to determine number of species for linear combination fitting. They also used a neural network to predict the distance between Co and C from the carbonyl, dC, and the distance between Co and bottom O, dO.

The following paper is unique in that it performs classification, rather regression.

**Classification of local chemical environments from x-ray absorption spectra**

**using supervised machine learning, Carbone, et al., 2019 [Carbone, 2019]**

1. The dataset: Simulated K-edge XANES of eight 3d transition metals (Ti, V, Cr, Mn, Fe, Co, Ni, and Cu)
2. What they did: Classify local coordination environment using a neural network

Diagram

Description automatically generated

Fig. 4 Their workflow for classifying spectra into the three coordination environments: tetrahedral (*T*4), and square pyramidal (*S*5), and octahedral (*O*6). Of note, their structural database was the Materials Project.

### Unsupervised machine learning approaches

**Latent Representation Learning for Structural Characterization of Catalysts, Routh, et al., 2021 [Routh, 2021]**

1. The dataset: Pd K-edge XANES
2. What they did: Correlated the dimensions of the latent space of an autoencoder to physical properties like coordination number (N), interatomic distance (R), and hydrogen fraction (H). They found applying PCA to the latent space produced stronger correlations. Then they trained a neural network to predict N, R, and H from the transformed latent space (the PCA-on-latent space representation).

**Machine learning approaches for ELNES/XANES,** **Mizoguchi and Kiyohara, 2020 [Mizoguchi and Kiyohara, 2020]**

1. The dataset: 39 electron energy loss near edge structure (ELNES), a.k.a. O K-edge XANES spectra, 14 of mono-metal oxides and 25 polymorphous SiO2
2. What they did: Used a decision tree to explain and predict hierarchical clustering of spectra, and then used a neural network to predict classes from decision tree.

Diagram

Description automatically generated

Fig. 5 (a) Structure to spectra correlation (b) Clustering spectra using hierarchical clustering (Wasserstein distance as a similarity metric) (c) Decision tree to determine the underlying properties distinguishing the three clusters.

## Solving the forward problem

The forward problem is instead predicting XANES spectra from structural parameters. While the inverse problem dominates the applications of machine learning in the X-ray spectroscopy community, there is some notable work to solve the forward problem and thus replace time-consuming DFT calculations.

**Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy, Carbone, et al., 2020**

1. What they did: Used graph based neural network to predict XANES from molecular structures in the QM9 database.

**Accurate, Affordable, and Generalisable Machine Learning Simulations of Transition Metal X-ray Absorption Spectra using the XANESNET Deep Neural Network, Rankine and Penfold, 2022**

1. What they did: Use a deep neural network to predict K-edge XANES for nine first-row transition metals (Ti-Zn) from weighted atom-centered symmetry functions (wACSF), a featurization of the local coordination geometry.

Diagram

Description automatically generated

Fig. 6 Mean percentage error between the target (from simulations) and predicted spectra on the test set.

The following paper is unique in that it predicts Valence-to-Core X-ray emission spectroscopy (VtC-XES) spectra rather than XANES.

**A deep neural network for valence-to-core X-ray emission spectroscopy, Penfold and Rankine, 2022**

1. What they did: They extended their neural net to predict row transition metal K-edge VtC-XES from weighted atom-centered symmetry functions (wASF) as input.

## References