Deep learning model for predicting material composition from X-ray photoelectron spectroscopy data

In-Ho Lee

Korea Research Institute of Standards and Science, Daejeon 34113, Korea

J. H. Choi and Yongsup Park*

Department of Physics, Kyung Hee University, Seoul 02447, Korea

(Dated: August 22, 2023)

Abstract

We develop a deep learning model to infer material composition from X-ray photoelectron spectroscopy data and test its performance. We generate a synthetic dataset made of 75000 spectra based on XPS parameter databases and electron scattering theory in the transport approximation. We considered 2, 3, 4, and 5 as the different types of atoms a substance can have. We use synthetic data for the X-ray photoelectron spectroscopy model development and utilize the latest artificial intelligence techniques to build a model that can predict the degree of carbon contamination and composition. The mean absolute error in the carbon contamination prediction for the test dataset is 14.5%. In addition, the mean of the maximum absolute error for the composition ratio prediction is 6.4%.

* corresponding author parky@khu.ac.kr

I. Introduction

X-ray photoelectron spectroscopy(XPS) method is widely used in characterizing materials. XPS analysis is the measurement of the composition and atomistic bonding properties of materials. By irradiating a sample with X-rays we obtain the kinetic energy and intensity of photoelectrons emitted by the photoelectric effect[1]. X-ray energy can be absorbed by one of the core electrons. The energy needed to cause the core electron to be emitted and subsequently detected is characteristic to each element. This characteristic feature allows the use of binding energy to identify the elements present on the surface of the material. Peak intensity at a specific binding energy usually displays the total number of photoelectron counts per second. Sometimes there would be an overlapping peaks, unresolved peaks on the binding energy axis. Specifically, it is important to check the core-levels and Auger-lines for each atom.

XPS can detect all elements except hydrogen and helium with detection limits of approximately 0.1%–1%. Since XPS is extremely surface sensitive, care must be taken to avoid surface contamination. Since each element has its own unique binding energy, comparing the peaks and binding energies in the spectrum can be used to determine the composition of the elements present on the sample surface. Binding energy is characteristic of the chemical environment of the core-excited atom, corresponding structural characterization. Also, when the chemical bonding state of an atom changes, the binding energy usually changes by a few eV, so the chemical bonding state can be inferred from this change. In addition, XPS simultaneously provides chemical information on the chemical structure, degree of carbon contamination, and oxidation state of the sample component atoms. Computational prediction of core-electron binding energies is also developed[2, 3].

XPS results can be challenging to interpret, in general, although there is a data base for binding energy values[4, 5]. Sometimes XPS data are often misinterpreted in the literature. Many factors in the experiment affect binding energies in XPS method. The binding energies of different atomic features can even overlap, further complicating the analysis. In fact, generation of photoelectrons is closely related to processes that result from x-ray bombardment of a surface include emission of a photoelectron, x-ray fluorescence, and emission of an Auger electron.

By obtaining a variety of experimental measurement data, it is possible to understand

the general XPS measurement data. It is also possible to infer XPS spectra of materials with arbitrary compositions from sufficient data. This is known as the traditional XPS analysis and is widely utilized in both materials physics and industrial researches. Similarly, research has recently been conducted on inferring composition ratios using deep learning method[6]. It can be seen that the ability to make inferences from data can be achieved through machine learning.

Recent advances in artificial intelligence have made it possible to make systematic inferences that have never been easily attempted before[7]. In this work, we developed and tested a model for correlating XPS data with chemical composition using probabilistic inference supported by recent machine learning methods. At this point, it would be natural to try to leverage artificial intelligence to build models for predicting carbon contamination and composition ratio that are as accurate as those of long-time experts.

Section II provides machine learning details. Section III presents and discusses model performance. Finally, Section IV concludes the paper.

II. Methods

For training and validation, we use the in silico-generated dataset. It is possible to generate synthetic dataset for XPS by taking into account composition and carbon contamination. The network was trained on the simulated dataset.

A. synthetic data

We generated a synthetic dataset made of 75000 spectra based on XPS parameter databases and electron scattering theory in the transport approximation[8]. Only Al k_α source is used. Each detail of real XPS spectra, including peak position, intensity, inelastic loss backgrounds, chemical shifts, the analyzer transmission function, and the signal-to-noise ratio has been carefully simulated according to available XPS databases and theories. The spectra kinetic energy range was 400–1486 eV on a 2048 energy point grid.

We consider random possible combinations of elements from the list of Li, Be, B, C, ..., and Bi atoms. Each virtual material is composed by a random number(from 2 to 5) of elements, with variable stoichiometry ratios. The carbon contamination leads to an overall

lower XPS intensity[9].

The training dataset size is 75000. The validation dataset size and test dataset size are 600 and 3000, respectively. The present network takes as input the 2048 spectral points ($\{x_j\}, j = 1, \ldots, 2048.$) and produces two outputs, the normalized carbon contamination level $c(0 \leq c \leq 1)$ and the normalized intensity ($\{y_e\}, 0 \leq y_e \leq 1, e = 1, \ldots, 81.$) of the 81 element.

B. deep learning model

The single neural network we used predicts each of the two by itself. The first is the degree of carbon contamination($0 \leq c \leq 1$) and the second is the composition ratio($0 \leq y_e \leq 1,\ e = 1, 2, 3, \ldots, 81$). As explained earlier, H and He are excluded from the study. We built a model that simultaneously optimizes the regression(contamination degree) and regression(composition ratio) loss functions using multiple one-dimensional convolutional nueral network layers, multihead attention layer, dense layers, and dropout layers. Our model is based on TensorFlow. A 'softmax' activation function for the final layer is used to normalize the outputs so that $\sum_{e=1}^{81} y_e = 1$. An 'adam' optimizer[10] and the tuned loss function allowed for a robust training. A 'sigmoid' activation, is used to identify the level of carbon contamination c. Total number of trainable parameters is 8765675.

C. loss functions

Mean squared logarithmic error($MSLE$) is considered to be an improvement over using percentage based errors for training because its numerical properties are better.

$$MSLE(\{y_e\}, \{y_e^t\}) = \frac{1}{81} \sum_{e=1}^{81} \left\{ log(1 + y_e) - log(1 + y_e^t) \right\}^2, \tag{1}$$

where $\{y_e\}$ are the network outputs and $\{y_e^t\}$ the target values. It is less sensitive to outliers than mean squared error since the logarithmic transformation compresses the error values. The loss function is scale independent as it is a difference of two log values which is the same as log of the ratio of the values. Due to the loss being log it penalizes underestimates more than overestimates.

D. training and validation

We plotted the loss functions over the training epoch in the Fig. 1. We also plotted the loss functions for the validation dataset on the model at the same time. The loss functions are related to the prediction of the degree of carbon contamination and the prediction of the composition ratio, respectively.

The mean absolute error($| c - c^t |$, $0 \leq c \leq 1$. c, c^t are predicted carbon contamination value and true carbon contamination value, respectively.) in the carbon contamination prediction for the test dataset(sample size: 1000) is 0.145. For the same test dataset, the mean of the maximum absolute error($max \mid y_e - y_e^t \mid$, $0 \leq y_e \leq 1$) for the composition ratio prediction is 0.064.

III. Results and discussion

IV. Conclusions

To summarize, we developed an accurate and efficient deep learning model to infer material composition from XPS data. We have shown that the model we built has expert-level predictive capabilities and can be used to infer composition and carbon contamination as the simplest application. We generated a synthetic dataset made of 75000 spectra based on XPS parameter databases and electron scattering theory in the transport approximation. We considered 2, 3, 4, and 5 as the different types of atoms a substance can have. The mean absolute error in the carbon contamination prediction for the test dataset is 14.5%. In addition, the mean of the maximum absolute error for the composition ratio prediction is 6.4%.

This research was supported by Enhancement of Measurement Standards and Technologies in Physics funded by Korea Research Institute of Standards and Science (No. KRISS-2021-

GP2021-0002).

---

[1] C. S. Fadley, X-ray photoelectron spectroscopy: Progress and perspectives, Journal of Electron Spectroscopy and Related Phenomena **178**, 2 (2010).

[2] D. Golze, M. Hirvensalo, P. Hernández-León, A. Aarva, J. Etula, T. Susi, P. Rinke, T. Laurila, and M. A. Caro, Accurate computational prediction of core-electron binding energies in carbon-based materials: A machine-learning model combining density-functional theory and gw, Chemistry of Materials **34**, 6240 (2022).

[3] Q. Sun, Y. Xiang, Y. Liu, L. Xu, T. Leng, Y. Ye, A. Fortunelli, W. A. Goddard III, and T. Cheng, Machine learning predicts the x-ray photoelectron spectroscopy of the solid electrolyte interface of lithium metal battery, The Journal of Physical Chemistry Letters **13**, 8047 (2022).

[4] J. Chastain and R. C. King Jr, Handbook of x-ray photoelectron spectroscopy, Perkin-Elmer Corporation **40**, 221 (1992).

[5] B. V. Crist, Xps in industry—problems with binding energies in journals and binding energy databases, Journal of Electron Spectroscopy and Related Phenomena **231**, 75 (2019).

[6] G. Drera, C. M. Kropf, and L. Sangaletti, Deep neural network for x-ray photoelectron spectroscopy data analysis, Machine Learning: Science and Technology **1**, 015008 (2020).

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).

[8] W. S. Werner, Electron transport in solids for quantitative surface analysis, Surface and Interface Analysis: An International Journal devoted to the development and application of techniques for the analysis of surfaces, interfaces and thin films **31**, 141 (2001).

[9] S. Evans, Correction for the effects of adventitious carbon overlayers in quantitative xps analysis, Surface and Interface Analysis: An International Journal devoted to the development and application of techniques for the analysis of surfaces, interfaces and thin films **25**, 924 (1997).

[10] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

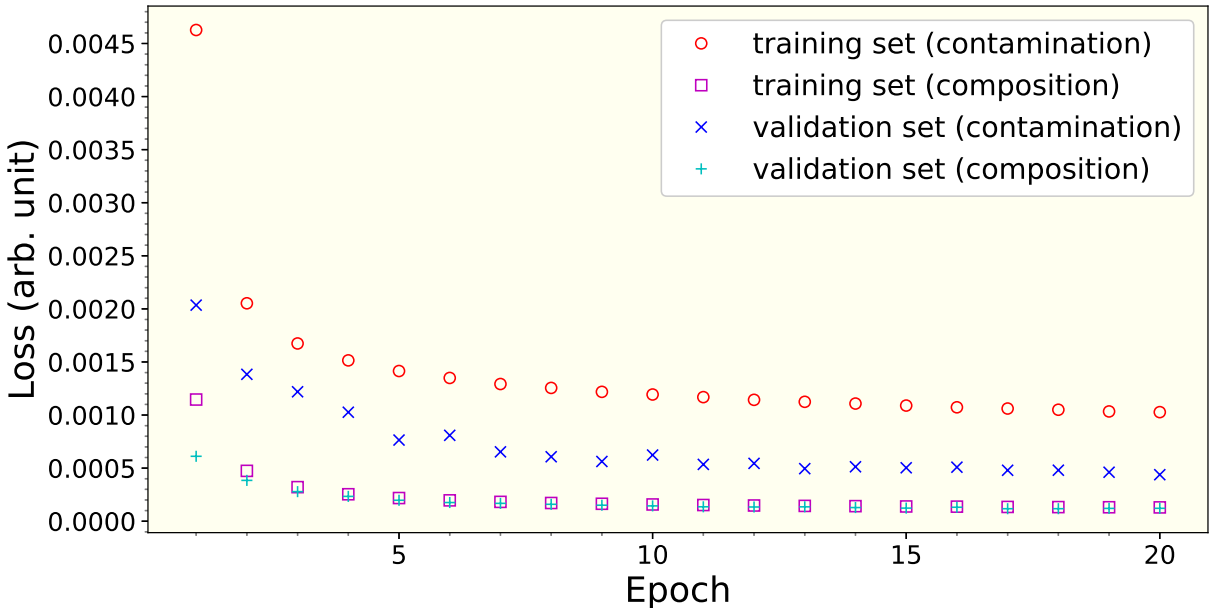

FIG. 1. The loss functions are related to the prediction of the degree of carbon contamination and the prediction of the composition ratio, respectively. Mean squared logarithmic error($MSLE$) is used for both contamination prediction and composition prediction.

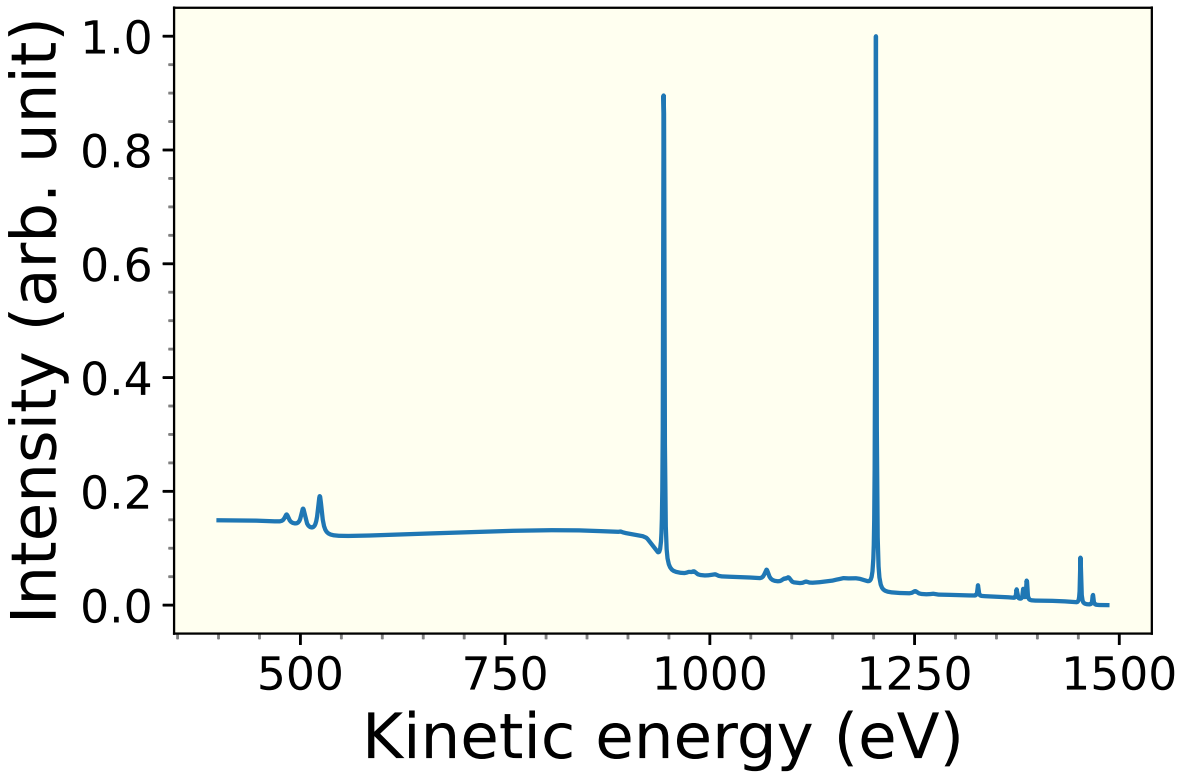

FIG. 2. Deep learning model.

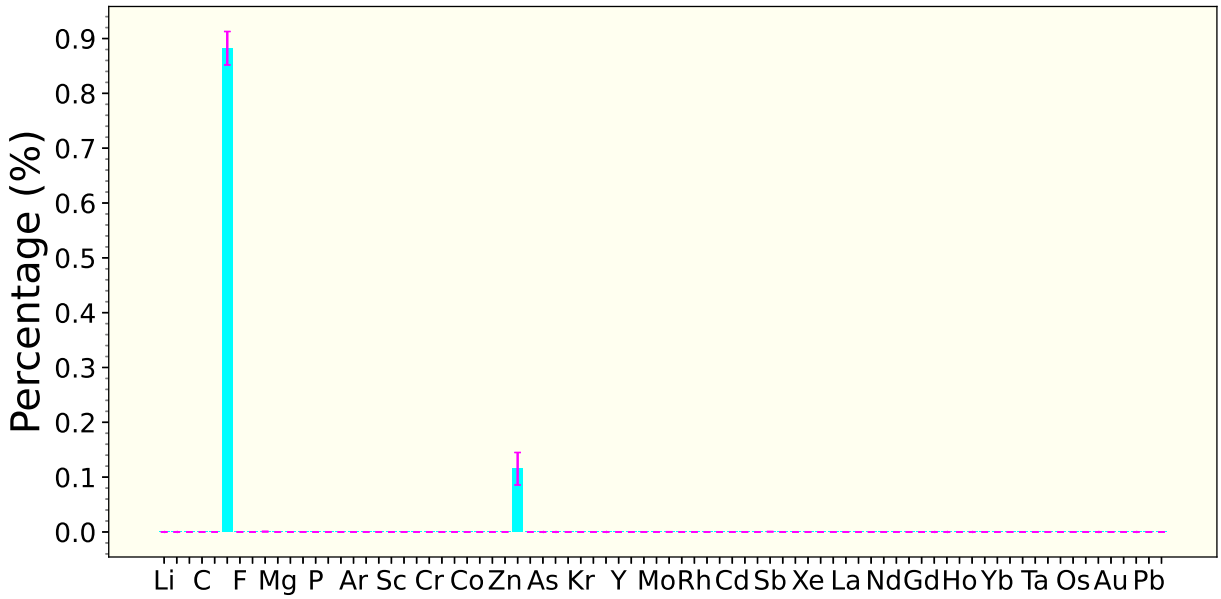

FIG. 3. Deep learning model.