André Orth*, Hawo Höfer, Alexei Nefedov, Mehrdad Jalali, Christof Wöll, and Markus Reischl

ML-Based XPS Quantification Supported by Synthetic Dataset Generation

https://doi.org/10.1515/cdbme-2024-2118

Abstract: With growing interest in laboratory automation and high-throughput systems, the amount of generated experimental data is rapidly increasing while analysis methods still require many manual work hours from experts. This is prevalent in X-ray photoelectron spectroscopy (XPS), where quantification is a complex, time-consuming, and error-prone task. We therefore propose a neural network-based workflow to make this process more approachable. As training data availability ranges from insufficient to non-existent, our workflow creates a synthetic dataset containing XPS signals and corresponding area percentages based on binding energies supplied by the user. As a result, no previous measurements are needed. After training on the synthetic data, the neural network can predict area percentages of the known binding energies with high confidence. This workflow can therefore be adapted for XPS quantification tasks to filter significant data and supervise processes. Moreover, this enables non-experts to analyze spectra and can help experts to reduce focus on important spectra.

Keywords: XPS, Quantification, Machine Learning, Synthetic Dataset

1 Introduction

X-ray photoelectron spectroscopy (XPS) is a state-of-the-art method in material science to gain information about elemental compositions as well as chemical and electronic states of elements. By exciting photoelectrons with X-rays, the surface of samples is measured as spectra that show intensities of said photoelectrons over their respective binding energies [1, 2]. XPS is used in different research fields like development of batteries or thin-films [3, 4]. Another important field is medical science, where analysis of implants, medical equipment and alloys regarding aging processes, biocompatibility and antimicrobiality are important [5–7].

To obtain the desired information about occurring elements and their atomic percentages $at.\%$, each spectra is analyzed by hand. This task is very time demanding, requires highly skilled experts and is quite error prone due to its complexity [8, 9]. With the rise of high-throughput systems and lab automation gaining importance, the amount of measurements and resulting data is drastically increasing. To prevent bottlenecks in new processes, analyzing methods need to be scaled accordingly. Especially for XPS, where each measurement needs to be thoroughly examined by hand, tools for easier access for non-experts and preliminary decisions based on complex systems without the need for experts are vital.

There has been limited research into the quantification of XPS data using neural networks, although this has been focused on survey spectra over a larger energy range. The work by Drera et al. [10] uses training data based on binding energies from a database, which gives users a possibility to select contained elements. This is a limiting factor, as binding energies can vary and a more detailed quantification of for example orbitals of one element is not considered. Pielsticker et al. [11] use linear combinations of real measurements to create training data. Therefore, a user needs measurements of each element in its pure form and without impurities to not introduce any bias into the dataset. These approaches suffice for broader quantification of given elements. The breakdown of signals into peaks associated with orbits need more precise data in the form of narrow or detail scans as well as specific binding energies that can change between setups [12]. As these are the major works concerning XPS, looking at other spectroscopy methods, Li et al. [13] have presented tools for analyzing 1D-Nuclear Magnetic Resonance (NMR) spectra, that include peak fitting solutions using neural networks. This can give insights on improving the network approach for XPS analysis, although differing data generation is a limiting factor.

Therefore, we introduce an approach based on synthetic data that focuses on detail scans. We only use given binding energies and peak constraints to create datasets on which our models are then trained and tested. This is our first application of neural networks on XPS, while previous research focused on X-ray Diffraction (XRD) with similarities in data generation and network architecture [14]. Consequently, this is a proof-of-concept and validated on synthetic data.

***Corresponding author: André Orth,** Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology (KIT), Eggenstein-Leopoldshafen, Germany, e-mail: andre.orth@kit.edu
Hawo Höfer, Markus Reischl, Institute for Automation and Applied Informatics, KIT, Eggenstein-Leopoldshafen, Germany
Alexei Nefedov, Mehrdad Jalali, Christof Wöll, Institute of Functional Interfaces, KIT, Eggenstein-Leopoldshafen, Germany

Open Access. © 2024 The Author(s), published by De Gruyter. (cc) BY This work is licensed under the Creative Commons Attribution 4.0 International License.

2 Method

For the deployment of our approach, we have created a workflow and split it into two paths. An exemplary overview of this workflow is shown in Figure 1. The first path (solid arrows) concerns the creation of synthetic data and training of the neural network. The second path (dashdotted arrows) is the application of the neural network on supplied spectra.

An important difference to previous research is the different prediction goal of our network. Experts are commonly interested in atomic percentages ($at.\%$) for quantification of contained elements. However, we train our network to predict peak area percentages ($pa.\%$). This is because atomic percentages are calculated from peak areas using Relative Sensitivity Factors (RSF). These differ for each element and instrument manufacturers, who therefore often include RSF values with their product or accompanying software. As the aim of this approach is variability, we therefore do not include the RSF conversion to atomic percentages within the data and network workflow. If needed by the user, peak area percentages can be converted to the actual peak area within the measurement and using their respective RSF to the atomic percentages with

$$X_A = \frac{I_{Am}/I_A^{Av}}{\sum_i I_{im}/I_i^{Av}}. \quad (1)$$

Here, X_A represents the atomic percentage of element A, I_{im} the peak areas of each element i and I_i^{Av} the corresponding RSF [15]. This preliminary exclusion allows for a broader usecase and adaptability. In the following, we will discuss the synthetic data generation and network architecture in more detail.

As a large amount of synthetic spectra is needed to train our network, we generate synthetic peaks following the structure shown in Figure 1. First off, we require the input of binding energies of the contained elements. There are additional parameters that can be tuned to a specific problem if there are known constraints. For each synthetic signal and contained element a Full Width Half Maximum (FWHM) is randomly chosen from a range (default $[0.5 - 3]$ eV). Pseudo-Voigt profiles are calculated with a given η, which shifts the profile towards a pure Gaussian ($\eta = 0$) or pure Lorentzian ($\eta = 1$) shape. Each profile is then normalized by its area, multiplied with a predefined $pa.\%$ and combined to generate a synthetic signal.

First steps during XPS analysis is to calculate and remove the underlying background. Three types are most common. For very simple examples, fitting a linear background can be sufficient. Shirley and Tougaard are more common and accurate, but more complex to achieve a better fit [16]. Shirley is an iterative process based on the assumption that the background is proportional to the total number of photoelectrons starting at the current binding energy. Tougaard introduces further complexity by including effects of inelastically scattered electrons. There are many complex effects to consider when calculating backgrounds from real measurements, which are also highly dependent on the contained elements. As this approach aims to not rely on specific element information, Tougaard can not be applied and we chose a combination of Shirley and linear for our approach to create synthetic backgrounds. The Shirley background $S(E)$ is calculated according to Vegh et al. [17], using a simplified approximation of an integral over the measurement $y(E)$ based on the binding energy E

$$S(E) = Y_0 * \int_E^{+\infty} y(E')dE. \quad (2)$$

This does not take into account or simulate effects like inelastically scattered electrons. As we are only considering detail scans, this effect has less of an impact [18]. We approximate the background in these smaller binding energy ranges by a combination of a linear and the calculated Shirley background. The gradient of the linear background over the measurement length is chosen from a parameter range that defaults to $[-0.2, 0.2]$.

As last addition, we add Gaussian noise. Its maximum can be changed as an input parameter of the workflow and defaults to 10% of the standard deviation of the measurement. During data generation, a uniform random value is created with the given maximum for each signal and added onto our data. The signal is then normalized to a range of $[0, 1]$ and added to the dataset with its $pa.\%$ and other necessary parameters. The next step of our workflow is to build the neural network.

We use a model architecture based on Conv1D and MaxPooling1D layers to predict $pa.\%$. This is partly derived from current publications [10, 11] and previous findings while working with XRD [14]. The dataset is split into three subsets for training, validation and testing using a split of $80/10/10\%$. The neural network is built and compiled using the Python package TensorFlow. Its architecture is displayed and described in Table 1. The input size of the first layer equals our

Tab. 1: Network architecture with input size of 200 signal points (20 eV, 0.1 eV/step) and normalized output for 5 labels.

Layer	Output	Parameters #
Input	200, 1	0
Conv1D	200, 12	192
MaxPooling1D	199, 12	0
Conv1D	199, 12	2127
MaxPooling1D	198, 12	0
Conv1D	198, 12	2127
MaxPooling1D	197, 12	0
Flatten	2364	0
Dense	200	473000
Dense	5	1005
Output_norm	5	0

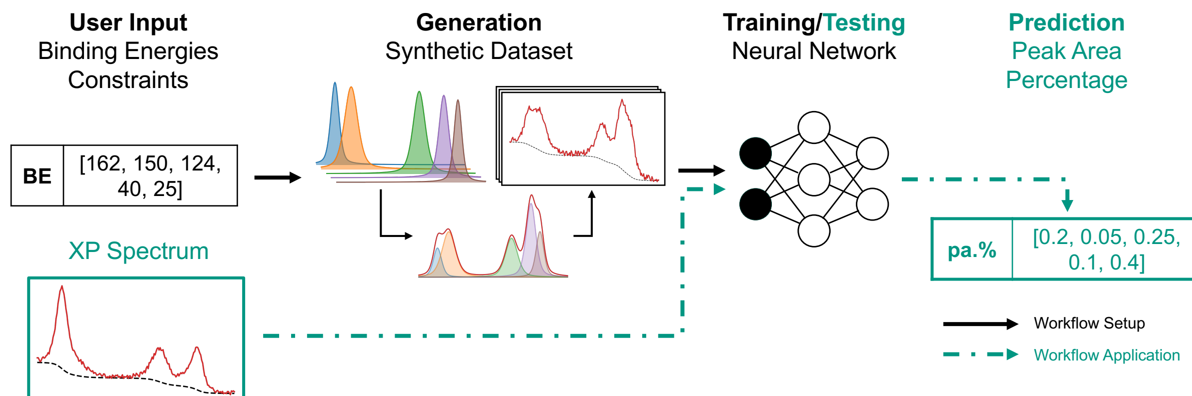

Fig. 1: A workflow overview with user input of binding energies, additional constraints and X-ray photoelectron (XP) spectra for quantification. Based on binding energies and constraints, single peaks per binding energy are created and combined. Additional augmentations (backgrounds and noise) are added to create the final synthetic signal. The resulting dataset is used to train the neural network. The supplied XP spectra is then analyzed by the network (dashdotted arrows) to receive a prediction of the containing $pa.\%$.

measurement size and output the amount of elements and their given binding energies. The network consists of 478.541 trainable parameters. We use Mean Absolute Error (MAE) as loss and metric to supervise the training and validation as well as comparing our test results. The Conv1D and first Dense layers all use a ReLu activation. All Conv1D layers use a filter size of 12 and kernel size of 15. The last Dense layer uses a linear activation and is followed by a normalization layer as it is used by Drera et al. [10]. However, we do not use the same multilevel convolution sub-net header [19]. This has shown good results for survey scans but is not suitable for our approach on detail scans. Two additional tools are used during training. To prevent plateauing, the TensorFlow callback ReduceLROnPlateau reduces the learning rate after 20 epochs of no improvement. Moreover, if there is no further improvement over 50 epochs, the training is terminated by EarlyStopping.

3 Results

For an exemplary test of the data pipeline as well as the network architecture, we chose five binding energies within a 20 eV range, a Voigt shape with a set ratio of $\eta = 0.5$, the default FWHM range of $[0.5, 3]$ eV and 10% noise. The background consists of the Shirley adaption with an added linear component in the default range. The binding energies were chosen randomly under the following constraints. The distance between two peaks must be at least 1 eV. The distance to both binding energy limits, in this case 0 and 20 eV, must be at least 2.5 eV. Of 500.000 created signals, we trained the network using 80% and validated on 10%. The training took 17 minutes on an NVIDIA TITAN RTX and was ended by EarlyStopping after 115 epochs. Testing the trained network on the remaining 10% took 3.8 seconds and resulted in an MAE of 0.003319. (Without learning rate reduction and early stopping, a similar result was be achieved. However, this took over 400 epochs during training and was therefore not a viable option.) A distribution of the maximum absolute errors is shown in Figure 2.

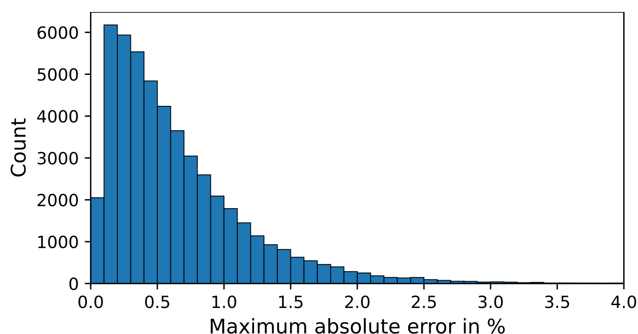

Fig. 2: Distribution of maximum absolute errors in % of network prediction

Here we can see that over 80% of the signals are predicted with a maximum error of less than 1%, while 49% even fall below 0.5%. For further comparison, reproductions of the two worst and two best predictions are displayed in Figure 3. The signals are calculated from true and predicted $pa.\%$ as well as FWHM and binding energies from the synthetic dataset. Therefore, they are shown without background or noise for easier comparison. Worse predictions occurred for signals with noise levels above 8% while noise levels of less than 1% resulted in more accurate predictions. The errors mainly occur for peaks that are closer together, which is also an error source during manual analysis.

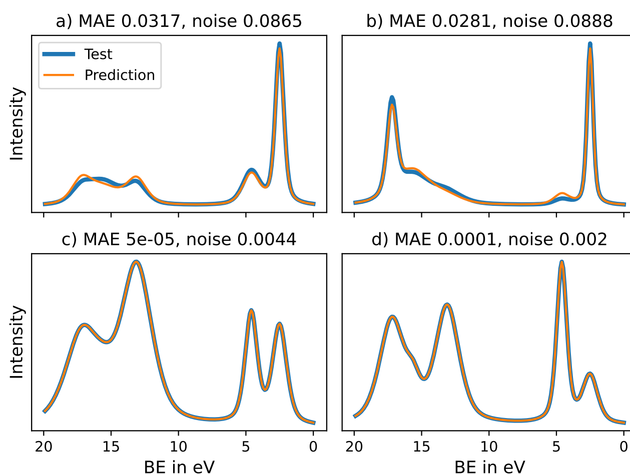

Fig. 3: Reproduction of the worst (a, b) and best (c, d) predictions

4 Conclusion

XPS analysis is a highly complex process that requires a high level of experience and expertise. In parallel with other referenced research, the aim of our approach is to ease the hurdle of this complexity to non-experts and make tools accessible to a wider audience. This can be used in the medical field to improve accuracy for upcoming high-throughput and lab automated systems and simplify research and supervision of for example aging processes. Therefore, also experts can benefit by not needing to analyze every measurement.

We have presented a workflow that creates synthetic spectra based on a wide array of changeable parameters and a model architecture that is able to predict area percentages of given elements with very low errors in short time. These tools are still subject to approximations and simplifications due to the complexity of the issue where an expert may outperform the workflow.

A sensible step for future work is modeling backgrounds closer to real measurements while at best staying independent of element-specific information. As data availability is scarce, validation of the model using a variety of real measurements is another important goal which could not be included in this work. A comparison between human and ML analysis results can lead to further insights as to where current errors are made and new technology is able to assist. For this, additional benchmarks need to be created in collaboration with experts.

References

[1] G. Greczynski et al. “X-ray photoelectron spectroscopy: towards reliable binding energy referencing.” In: *Progress in Materials Science* 107 (2020), p. 100591.

[2] C. S. Fadley. “X-ray photoelectron spectroscopy: Progress and perspectives.” In: *Journal of Electron Spectroscopy and Related Phenomena* 178 (2010), pp. 2–32.

[3] V. Shutthanandan et al. “Applications of XPS in the characterization of Battery materials.” In: *Journal of Electron Spectroscopy and Related Phenomena* 231 (2019).

[4] P. T. Hsieh et al. “Luminescence mechanism of ZnO thin film investigated by XPS measurement.” In: *Applied Physics A* 90 (2008), pp. 317–321.

[5] B.-S. Kang et al. “XPS, AES and SEM analysis of recent dental implants.” In: *Acta Biomaterialia* 5.6 (2009), p. 2222.

[6] M. Szczesna-Antczak et al. “Biomodification and biodeterioration of carbon coatings by fungal strains.” In: *International Biodeterioration & Biodegradation* 88 (2014), pp. 106–117.

[7] S. W. Park et al. “Mesoporous TiO2 implants for loading high dosage of antibacterial agent.” In: *Applied Surface Science* 303 (2014), pp. 140–146.

[8] G. H. Major et al. “Assessment of the frequency and nature of erroneous x-ray photoelectron spectroscopy analyses in the scientific literature.” In: *Journal of Vacuum Science & Technology A* 38.6 (2020). 061204.

[9] P. Sherwood. “The use and misuse of curve fitting in the analysis of core X-ray photoelectron spectroscopic data.” In: *Surface and Interface Analysis* 51.6 (2019), pp. 589–610.

[10] G. Drera et al. “Deep neural network for x-ray photoelectron spectroscopy data analysis.” In: *Machine Learning: Science and Technology* 1.1 (2020), p. 015008.

[11] L. Pielsticker et al. “Convolutional neural network framework for the automated analysis of transition metal X-ray photoelectron spectra.” In: *Analytica Chimica Acta* 1271 (2023), p. 341433.

[12] B. Singh et al. “Comparison of the equivalent width, the autocorrelation width, and the variance as figures of merit for XPS narrow scans.” In: *Journal of Electron Spectroscopy and Related Phenomena* 197 (2014), pp. 112–117.

[13] D.-W. Li et al. “DEEP Picker1D and Voigt Fitter1D: a versatile tool set for the automated quantitative spectral deconvolution of complex 1D-NMR spectra.” In: *Magnetic Resonance* 4.1 (2023), pp. 19–26.

[14] J. Schuetzke et al. “Accelerating Materials Discovery: Automated Identification of Prospects from X-Ray Diffraction Data in Fast Screening Experiments.” In: *Advanced Intelligent Systems* 6.3 (2024), p. 2300501.

[15] M. Seah et al. “Quantitative AES. VIII: Analysis of Auger electron intensities from elemental data in a digital auger database.” In: *Surface and Interface Analysis* 26.12 (1998), pp. 908–929.

[16] M. H. Engelhard et al. “Introductory guide to backgrounds in XPS spectra and their impact on determining peak intensities.” In: *Journal of Vacuum Science & Technology A* 38.6 (2020), p. 063203.

[17] J. Végh. “The analytical form of the Shirley-type background.” In: *Journal of Electron Spectroscopy and Related Phenomena* 46.2 (1988), pp. 411–417.

[18] B. Fultz et al. “Inelastic electron scattering and spectroscopy.” In: *Transmission Electron Microscopy and Diffractometry of Materials* (2008), pp. 163–221.

[19] C. Szegedy et al. “Going Deeper With Convolutions.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2015.