Data driven calculations of XAS spectra of Grapheneoxides

Official project title

Annika Bande, Kanishka Singh *, Helmholtz-Zentrum f"ur Materialen und Energie GmbH, Freie Universität Berlin*

In Short

- Generation of grapheneoxide structures
- Calculation of XAS spectra using GPAW
- Machine Learning prediction of XAS spectra

Quantum dots made of grapheneoxide (GO) structures are zero-dimensional semiconductor materials that have extensive applications ranging from photocatalysis to photonic qubits.[1] Spectroscopy can be used to reveal the structural features of such materials and with quantum chemistry methods, theoretical spectra can be computed accurately. But since quantum dots can be understood as large, nanosized molecules, these techniques are slow and computationally demanding. Machine learning (ML) methods were demonstrated to be useful in the prediction of properties of molecules[2][3], which is based on the existence of a data set of molecule-property pairs to drive the ML.

Since the computation of XAS spectra is computationally demanding, there are very few, if any, datasets of XAS spectra for large molecules such as GOs. Since ML algorithms require copious amounts of data to train models, this lack of appropriate data has prevented the applications of ML to this problem. The aim of this project is to address this issue, by creating a dataset of representative GO structures and their XAS spectra.

GO layers are single layered oxidised derivatives of graphene, often represented with truncated structures as exemplified in Fig. 1. In the experimental generation of these compounds, the oxidation of graphene results in the replacement of hydrogen atoms with hydroxyl, epoxy and carboxyl functional groups. While the concentration of different atoms in each GO can be determined, the randomness of oxidation and presence of structural defects makes it difficult to determine the exact structure of a GO layer generated experimentally.

X-ray absorption spectroscopy has been crucial in the structural determination of such molecules. Different functional groups appear in the XAS spectra at different peak ranges, thus helping the identification of the structure of each molecule. XAS spectra are either measured experimentally, or they can be computed theoretically using several techniques from quantum chemistry such as TDDFT, Many Body Perturbation Theory, Dynamic Core hole theories etc. These theoretical calculations are used to interpret experimental spectra as they are obtained directly from the structural composition of the underlying molecule. It is therefore crucial that the theoretical XAS spectra are accurate and can be generated in a computationally efficient manner.

A GO layer is theoretically infinite in length, and

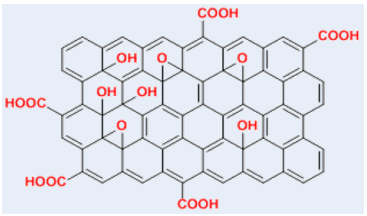

Figure 1: *Representative structure of a GO molecule from the proposed database.*

therefore for calculating their spectra theoretically, one has to use truncated structures. However since XAS computations are expensive, the determination of the spectra of such truncated structures is still an arduous task, especially for large GO structures. The larger a such a theoretical model structure is, the more realistically it can capture the structure of a true, much larger GO quantum dot and therefore, approximate the true structure in a better way. The cost of generating XAS spectra using DFT-based methods increases by the order of N^2, therefore it is prohibitive for large structures. ML methods on the other hand, are not susceptible to such scaling issues, but require training models and large datasets. The most notable work on creating a dataset of GO structures was on two-dimensional GO layers approximated by model structures containing 176-213 atoms and having varying levels of structural defects and oxidation [3]. XAS spectra have then be calculated for these structures using the density-functional theory (DFT) code GPAW [5]. Unsupervised clustering of the XAS spectra was then used to deduce fingerprint spectra for different functional groups present in the structures of the dataset. However, the entire dataset has not been made available publicly, and as such is not useful for supervised ML applications of the project of this proposal.

In this project, the approach is to develop a database that contains structures of different size to understand the local and global effects of XAS spectra for ML applications. Rather than using a large representative GO layer with defects, the structures we use vary in number of hexagonal rings, as well as

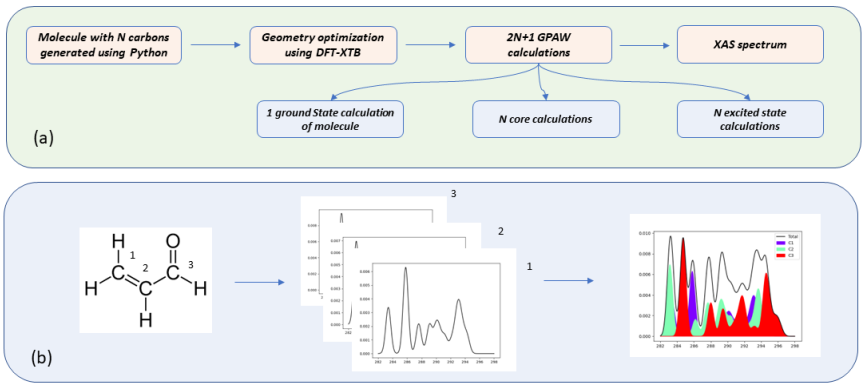

Figure 2: *(a) The process of generating a XAS spectrum for a molecule in the proposed database using GPAW. (b) The XAS spectrum of acrolein generated using the GPAW core hole method, shows how information from local sites appears in the total spectrum. Individual core hole spectra can be an important source of data for ML applications.*

degree of oxidation. The proposed dataset will have molecules with number of atoms ranging from 36 to more than 120 atoms. The large variation of size and composition will cover a large chemical space consisting of GO-like structures as well as representative structures that can be used to train ML models that can approximate XAS of large structures from these small representative molecule.

The DFT code GPAW uses the core-hole method to generated carbon K-edge XAS spectra. This means that for each C site in the molecule, N, two DFT calculations have to be performed one with the electron configuration of the ground state, the other of the excited state which is the core-hole electron configuration. Therefore, the number of calculations is in the order of $2N$ per molecule. Then the individual spectra are combined to generate into the XAS spectrum of the entire structure. In GPAW the calculations on each C atom are independent from each other. Therefore, massively parallel architectures can be used to generate these spectra, which is why GPAW can be used to generate extensive datasets for ML applications.

In one of our previous works, we were able to demonstrate that several Graph Neural Networks can be used to predict spectra of small molecules using a database [4]. Graph ML approaches can be used to train models on both graph (molecule) and node (atom) level data, and the large number of site-level spectra that will be created in this database, can be used to train an ML model which will predict site level spectra depending on the node environment. These site level spectra can then be aggregated to generate molecular spectra.

Our previous experiences make us confident that the targeted database of this proposal will be crucial to the acceleration of XAS spectra prediction and, furthermore, useful to the development of ML algorithms in molecular property prediction and graph learning. We have performed initial calculations on the testing account setup at HLRN and have been able to generate spectra for some of representative structures of the database. Through the HLRN resources we will be able to extend our present calculations to a larger number of structures, and therefore be able to create this dataset.

WWW

https://www.helmholtz-berlin.de/theochem

More Information

[1] D. Dreyer, S. Park, C. Bielawski, R. Ruoff *Chemical Society Reviews* **39**, 1 (2009). doi: 10.1039/B917103G

[2] J. Gilmer, S. Schoenholz, P. Riley, O. Vinyals, G. Dahl *arXiv*, doi:10.48550/arXiv.1704.01212

[3] A. Aarva, S. Sainio, V. Deringer, M. Caro, T. Laurila *The Journal of Physical Chemistry C* **125**, 33 (2021). doi: 10.1021/10.1021/acs.jpcc.1c03238

[4] K. Singh, J. Munchmeyer, L. Weber, U. Leser, A. Bande *Journal of Chemical Theory and Computation* **18**, 7 (2022). doi: 10.1021/acs.jctc.2c00255

[5] J. Enkovaara, N. Romero, S. Shende, J. Mortensen *Procedia Computer Science* **4**, 1 (2011). doi:10.1016/j.procs.2011.04.003

Project Partners

Sunel De Kock Michael Walter, University of Freiburg

Funding

Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS)

DFG Subject Area

327-02