The Journal of Open Source Software

spectrapepper: A Python toolbox for advanced analysis of spectroscopic data for materials and devices.

Enric Grau-Luque[1], Fabien Atlan[1], Ignacio Becerril-Romero[1], Alejandro Perez-Rodriguez[1, 2], Maxim Guc[1], and Victor Izquierdo-Roca[1]

1 Catalonia Institute for Energy Research (IREC), Jardins de les Dones de Negre 1, 08930 Sant Adrià de Besòs, Spain **2** Departament d'Enginyeria Electrònica i Biomèdica, IN2UB, Universitat de Barcelona, C/ Martí i Franqués 1, 08028 Barcelona, Spain

DOI: 10.21105/joss.03781

Software

- Review
- Repository
- Archive

Editor: Rachel Kurchin

Reviewers:

- @stuartcampbell
- @ksunden

Submitted: 23 September 2021
Published: 17 November 2021

Statement of need

In recent years, the complexity of novel high-tech materials and devices has increased considerably. This complexity is primarily in the form of increasing numbers of components and broader ranges of applications. An example of the latter is the last generation of thin-film solar cells, which comprise several functional micro- and nano- layers including back contact, absorber, buffer, and transparent front contact. Most of these layers are complex multicomponent compounds (Cu(In,Ga)Se2, Sb2Se3, CdTe, CdS, Zn(O,S), ZnO:Al, etc.) that require fine-tuning of their physicochemical properties to ensure functionality and high peformance (Chopra et al., 2004; Powalla et al., 2018). This embedded complexity means that further development of such devices requires advanced characterization and methodologies that allow correlating the physicochemical data of the different layers (chemical composition, structural properties, defect concentration, etc.) with the performance of the final devices in a fast, precise, and reliable way. In this regard, non-destructive methodologies based on spectroscopic characterization techniques (Raman, photoluminescence, X-ray fluorescence, reflectance, transmittance, etc.) have already been demonstrated to possess a high versatility and potential for this type of analyses (Dimitrievska et al., 2019; Guc et al., 2017; Oliva et al., 2016). These spectroscopy-based methodologies can provide deep information that encompasses the complexity of novel materials and devices in a non-destructive way, providing a profound understanding of their properties, failure mechanisms, and possible improvements (Grau-Luque et al., 2021). The latest advances in the application of spectroscopic methodologies for complex materials and devices include the implementation of combinatorial analysis (CA), artificial intelligence (AI) and machine learning (ML), that have been already used in few studies and are slowly becoming more common (Chen et al., 2020). Furthermore, the widespread use of this kind of tools in both laboratory environments and on-line/in-line monitoring of production lines is predicted to shorten development times by a factor of 10, from 10 to 20 years to just a few years (Aspuru-Guzik & Persson, 2018; Correa-Baena et al., 2018; Maine & Garnsey, 2006; Mueller et al., 2016). Unfortunately, several barriers for researchers to implement CA, AI, and ML remain (Gu et al., 2019; Mahmood & Wang, 2021). One of them is the proper pre-processing of spectroscopic data that allows not only to emphasize the relevant changes in the spectra, but also to combine data obtained from different techniques and instruments. Also, the use of ML requires substantial amounts of high-quality data for a precise analysis of the physicochemical parameters of new materials and devices, which necessitates the use of automated systems for massive characterization measurements. In other words, the implementation of automated high-throughput experiments and the capability to perform big-data pre-processing to enhance features of spectroscopic data for ML, and subsequent CA, requires deep theoretical, statistical, analytical, and programming knowledge. Therefore, simple and practical platforms that help researchers to apply such tools are paramount to accelerate their

universal adoption and ultimately shorten the development times of new materials and devices (Butler et al., 2018).

Overview

spectrapepper is a Python package that aims to ease and accelerate the use of advanced tools such as machine learning and combinatorial analysis, through simple, straightforward, and intuitive code and functions. This library includes a wide range of tools for spectroscopic data analysis in every step, including data acquisition, processing, analysis, and visualization. Ultimately, **spectrapepper** enables the design of automated measurement systems for spectroscopy and the combinatorial analysis of big data through statistics, artificial intelligence, and machine learning. **spectrapepper** is built in Python 3 (Van Rossum & Drake, 2009), and also uses third-party packages including `numpy` (Harris et al., 2020), `pandas` (Reback et al., 2021), `scipy` (Virtanen et al., 2020), and `matpotlib` (Hunter, 2007), and encourages the user to use `scikit-learn` (Pedregosa et al., 2011) for machine learning applications. **spectrapepper** comes with full documentation, including quick start, examples, and contribution guidelines. Source code and documentation can be downloaded from https://github.com/spectrapepper/spectrapepper.

Features

A brief and non-exhaustive list of features includes:

- Baseline removal functions.
- Normalization methods.
- Noise filters, trimming tools, and despiking methods (Barton & Hennelly, 2019; Whitaker & Hayes, 2018).
- Chemometrics algorithms to find peaks, fit curves, and deconvolve spectra.
- Combinatorial analysis tools, such as Spearman, Pearson, and n-dimensional correlation coefficients.
- Tools for ML applications, such as data merging, randomization, and decision boundaries.
- Sample data and examples.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 952982 (Custom-Art project) and Fast Track to Innovation Programme under grant agreement no. 870004 (Solar-Win project). Authors from IREC belong to the SEMS (Solar Energy Materials and Systems) Consolidated Research Group of the "Generalitat de Catalunya" (ref. 2017 SGR 862) and are grateful to European Regional Development Funds (ERDF, FEDER Programa Competitivitat de Catalunya 2007–2013). MG acknowledges the financial support from Spanish Ministry of Science, Innovation and Universities within the Juan de la Cierva fellowship (IJC2018-038199-I).

References

Aspuru-Guzik, A., & Persson, K. (2018). Materials Acceleration Platform - Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods with Ar-

tificial Intelligence. *Report of the Clean Energy Materials Innovation Challenge Expert Workshop*, *January*, 1–108. http://nrs.harvard.edu/urn-3:HUL.InstRepos:35164974

Barton, S. J., & Hennelly, B. M. (2019). An Algorithm for the Removal of Cosmic Ray Artifacts in Spectral Data Sets. *Applied Spectroscopy*, *73*(8), 893–901. https://doi.org/10.1177/0003702819839098

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2

Chen, C., Zuo, Y., Ye, W., Li, X., Deng, Z., & Ong, S. P. (2020). A Critical Review of Machine Learning of Energy Materials. *Advanced Energy Materials*, *10*(8), 1903242. https://doi.org/10.1002/aenm.201903242

Chopra, K. L., Paulson, P. D., & Dutta, V. (2004). Thin-film solar cells: an overview. *Progress in Photovoltaics: Research and Applications*, *12*(2-3), 69–92. https://doi.org/10.1002/PIP.541

Correa-Baena, J.-P., Hippalgaonkar, K., Van Duren, J., Jaffer, S., Chandrasekhar, V. R., Stevanovic, V., Wadia, C., Guha, S., & Buonassisi, T. (2018). Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing. *Joule*. https://doi.org/10.1016/j.joule.2018.05.009

Dimitrievska, M., Oliva, F., Guc, M., Giraldo, S., Saucedo, E., Pérez-Rodríguez, A., & Izquierdo-Roca, V. (2019). Defect characterisation in Cu2ZnSnSe4 kesterites: Via resonance Raman spectroscopy and the impact on optoelectronic solar cell properties. *Journal of Materials Chemistry A*, *7*(21), 13293–13304. https://doi.org/10.1039/c9ta03625c

Grau-Luque, E., Anefnaf, I., Benhaddou, N., Fonoll-Rubio, R., Becerril-Romero, I., Aazou, S., Saucedo, E., Sekkat, Z., Perez-Rodriguez, A., Izquierdo-Roca, V., & Guc, M. (2021). Combinatorial and machine learning approaches for the analysis of Cu2ZnGeSe4: influence of the off-stoichiometry on defect formation and solar cell performance. *Journal of Materials Chemistry A*, *9*(16), 10466–10476. https://doi.org/10.1039/d1ta01299a

Gu, G. H., Noh, J., Kim, I., & Jung, Y. (2019). Machine learning for renewable energy materials. *Journal of Materials Chemistry A*, *7*(29), 17096–17117. https://doi.org/10.1039/c9ta02356a

Guc, M., Hariskos, D., Calvo-Barrio, L., Jackson, P., Oliva, F., Pistor, P., Perez-Rodriguez, A., & Izquierdo-Roca, V. (2017). Resonant Raman scattering based approaches for the quantitative assessment of nanometric ZnMgO layers in high efficiency chalcogenide solar cells. *Scientific Reports 2017 7:1*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-01381-4

Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). *Array programming with NumPy* (No. 7825; Vol. 585, pp. 357–362). Nature Research. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Mahmood, A., & Wang, J.-L. (2021). Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy & Environmental Science*, *14*(1), 90–105. https://doi.org/10.1039/d0ee02838j

Maine, E., & Garnsey, E. (2006). Commercializing generic technology: The case of advanced materials ventures. *Research Policy*, *35*, 375–393. https://doi.org/10.1016/j.respol.2005.12.006

Mueller, T., Kusne, A. G., & Ramprasad, R. (2016). Machine Learning in Materials Science: Recent Progress and Emerging Applications. In *Reviews in computational chemistry* (Vol. 29, pp. 186–273). wiley. https://doi.org/10.1002/9781119148739.ch4

Oliva, F., Kretzschmar, S., Colombara, D., Tombolato, S., Ruiz, C. M., Redinger, A., Saucedo, E., Broussillou, C., Monsabert, T. G. de, Unold, T., Dale, P. J., Izquierdo-Roca, V., & Pérez-Rodríguez, A. (2016). Optical methodology for process monitoring of chalcopyrite photovoltaic technologies: Application to low cost Cu(In,Ga)(S,Se)2 electrodeposition based processes. *Solar Energy Materials and Solar Cells*, *158*, 168–183. https://doi.org/10.1016/j.solmat.2015.12.036

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python* (Vol. 12, pp. 2825–2830). http://scikit-learn.sourceforge.net.

Powalla, M., Paetel, S., Ahlswede, E., Wuerz, R., Wessendorf, C. D., & Friedlmeier, T. M. (2018). *Applied Physics Reviews*, *5*(4), 041602. https://doi.org/10.1063/1.5061809

Reback, J., McKinney, W., Jbrockmendel, Bossche, J. V. den, Augspurger, T., Cloud, P., Hawkins, S., Gfyoung, Sinhrks, Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., Patrick, Garcia, M., Schendel, J., … H-vetinari. (2021). *pandas-dev/pandas: Pandas 1.2.4*. https://doi.org/10.5281/ZENODO.4681666

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace. ISBN: 1441412697

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Walt, S. J. van der, Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Whitaker, D. A., & Hayes, K. (2018). A simple algorithm for despiking Raman spectra. *Chemometrics and Intelligent Laboratory Systems*, *179*, 82–84. https://doi.org/10.1016/j.chemolab.2018.06.009