Application of Materials Informatics Tools to the Analysis of Combinatorial Libraries of All Metal-Oxides Photovoltaic Cells

Hanoch Senderowitz¹(✉), Abraham Yosipof², and Omer Kaspi¹

¹ Department of Chemistry, Bar Ilan University, 5290002 Ramat Gan, Israel
omerkaspi@gmail.com, hsenderowitz@gmail.com,
hanoch.senderowitz@biu.ac.il

² Department of Information Systems, College of Law and Business,
Ramat-Gan, Israel
avi.yosipof@gmail.com

Abstract. Material informatics is engaged with the application of informatics tools, frequently in the form of machine learning algorithms, to gain insight into structure properties relationships of materials and to design new materials with desired properties. Here we describe the application of such algorithms to the analysis of solar cell (i.e., photovoltaic; PV) libraries made entirely from metal oxides (MOs). MOs-based solar cells hold the potential to provide clean and affordable energy if their power conversion efficiencies are improved. We demonstrate the power of dimensionality reduction methods to visualize the MOs-based solar cell space and the power of several algorithms to develop predictive models for key PV properties. We stress the importance of conducting such studies in collaboration with experimentalists.

Keywords: Materials informatics · QSAR · Solar cells ·
Dimensionality reduction · RANSAC · Genetic programming · kNN

1 Introduction

Materials informatics is an emerging field of research primarily engaged with the application of informatic principles to materials science for the purpose of discovering and developing new materials [1, 2]. Several factors contribute to the continuous development of materials informatics including the rapid growth of available databases that contain experimental and computational information on structures and properties of materials and the multiple similarities between materials informatics and e.g., chemo- and bio-informatics. Thus, this younger field could draw from the experience and large arsenal of tools developed in the more established fields [2].

Central to materials informatics is the application of data mining techniques and in particular machine learning approaches, often referred to as Quantitative Structure Activity Relationship (QSAR) modeling, to derive predictive models for a variety of materials-related “activities”. Such models could accelerate the development of new materials with favorable properties and provide insight into the factors governing these

properties. In this work we present the application of several machine learning tools to the analysis of solar cells.

Growth in energy demands, coupled with the movement towards clean energy, are likely to make solar cells an important part of future energy resources. At present, most solar cells are based on silicon yet new alternatives are constantly emerging with examples including organic photovoltaic (OPV) cells, [3] dye sensitized solar cells (DSSC), [4] and perovskites [5]. Another appealing alternative is presented by solar cells entirely made of metal oxides (MOs). Such cells are reasonably cheap to manufacture due to the abundance of their constituting elements, are environmentally friendly and are stable over time. Yet, their performances in terms of their ability to convert sunlight into electricity need to be improved [6]. Such improvements require the development of new MOs which could benefit from combining combinatorial materials science for producing solar cells libraries with machine learning approaches to analyze the resulting libraries and direct synthesis efforts.

MO-based solar cell libraries are produced through a combinatorial materials science approach involving the non-uniform deposition of two or more layers of different metal oxides on top of a glass support, followed by the introduction of the appropriate metal contacts. A typical library synthesized in this form consists of multiple cells (169 in our case) each of which is typically characterized by multiple composition and photovoltaic (PV) parameters including, for example, the thickness of the different layers that make up the cell and the thickness ratios between them, the band gap of the absorber layer, the short circuit photocurrent density (J_{sc}), the open circuit photovoltage (V_{oc}) and the internal quantum efficiency (IQE) [6, 7].

2 Results and Discussion

In this work we present the application of several machine learning tools to the analysis of single and multiple MO-based solar cell libraries. Some of these tools were recently incorporated into a newly developed MATLAB-based decision support system (DSS) called “PV Analyzer” [8]. PV Analyzer integrates tools common to chemo- and materials-informatics such as simple bi-parametric correlations, heat-maps and principle component analysis (PCA) with a workflow for the development of predictive quantitative structure activity relationship (QSAR) models based on the RANSAC algorithm which was originally developed for computer vision (see below). The PV Analyzer workflow is composed of three blocks: (1) Input data; (2) Data pre-processing and visualization; (3) RANSAC modeling (Fig. 1). Importantly, PV Analyzer is a modular system which allows for the facile incorporation of additional tools.

The PV libraries studied in this work were gathered over time from multiple literature reports [6, 7, 9, 10 and references cited therein]. Importantly, all were manufactured by the same lab using largely similar methods. First, looking at individual libraries, we demonstrated the ability of standard PCA to identify outliers in the PV space and together with self organizing maps (SOM) to unveil local and global effects of a Molybdenum Oxide layer on PV properties [9]. Next, five libraries were compiled into a unified dataset for a total of 1165 solar cells where each solar cell was characterized by seven experimentally measured PV properties including short circuit

Fig. 1. Screenshots describing the application of the DSS to a specific solar cell library. (A) PV Analyzer's main screen. From this screen the user can control the tasks available in the DSS. (B) Correlation matrix screen showing the correlations between all descriptors and all activities (C) Activities/descriptors assignment and data curation screen. (D) Heat map screen showing the heat-map obtained for *IQE*. (E) Test set/Training set selection screen. (F) PCA screen showing the loading plot (left) and scoring plot (right) resulting from the PCA performed on the descriptors. (G) RANSAC control screen. Taken from reference [8].

photocurrent density (J_{sc}), open circuit photovoltage (V_{oc}), internal quantum efficiency (IQE), maximum photovoltaic power (P_{max}), fill factor (FF), series resistance (R_s) and Shunt resistance (R_{sh}). This database was subjected to several linear and non linear dimensionality reduction methods (PCA, kernel-PCA, Isomap and Diffusion map) to allow for a facile visualization of the resulting “PV space”. Furthermore, we studied the relative performances of the different methods in terms of their ability to separate, in the reduced space, the five original libraries which were used to make up the space and maintain their original neighborhood [10]. We found that all methods were able to segregate the different libraries into unique regions of the reduced space (Fig. 2) but that PCA performed best in terms of the ability to correctly maintain the local environment of samples whereas Isomap did the best job in assigning class membership based on the identity of nearest neighbors, i.e., it was the best classifier. In addition, diffusion map identified the smallest number of outliers. We also found that many of the outliers identified by all methods could be rationalized.

Fig. 2. 3D representation of the integrated database after different dimensionality reduction methods. Subfigures A-D represent the different libraries plotted in the reduced space following PCA, Kernel PCA, Isomap and Diffusion map, respectively. Taken from reference [10].

Next we used several machine learning algorithms such as k nearest neighbors (kNN), genetic programming (GP) [11] and RANSAC to develop predictive QSAR models for key PV properties including J_{sc} , V_{oc} and IQE for several libraries [7, 12]. In particular, the Random Sample Consensus (RANSAC) algorithm is a predictive modeling tool widely used in the image processing field for cleaning datasets from noise. In this work we demonstrated that RANSAC could be used as a “one stop shop” algorithm for the development and validation of QSAR models, performing multiple tasks including outlier removal, descriptors selection, model development and predictions of test set samples while using the concept of applicability domain [13]. Contrary to most conventional QSAR models, the input descriptors used in these studies were experimentally measured rather than theoretically computed and included

the thicknesses of the different MO layers used to construct the cells and the thickness ratios between them. The need to rely on measured rather than on calculated descriptors is because when the (well defined) different MOs are combined (see above for a short description of the manufacturing process), the exact composition/structure of the resulting solar cell is not well defined and consequently is not amenable to the calculation of structure-based descriptors. Our results (Table 1) demonstrate that QSAR models with good prediction statistics could be developed by any of these methods for most PV properties (except for several poor models developed for *Voc*) and that these models highlight important factors affecting these properties in accord with experimental findings. Specifically, the thicknesses of the different MO layers play a pivotal role in determining the PV properties. The resulting models are therefore suitable for designing better solar cells. Interestingly, the success of developing *k*NN-based models demonstrates that the similar property principle which is well established in the small molecules/pharmaceuticals hyperspace also holds in the PV space.

Table 1. Performances of models derived by different algorithms for external test sets of three MOs-based solar cells libraries. Results are presented for three key PV properties, namely, *Jsc*, *Voc* and *IQE*. Columns marked with “AD” provide the results following the application of an applicability domain.

Library	Activity	<i>k</i> NN		RANSAC		GP
		Q_{ext}^2	Q_{ext}^2 (AD)	Q_{ext}^2	Q_{ext}^2 (AD)	Q_{ext}^2
<i>TiO</i> ₂ <i>Cu</i> ₂ <i>O</i> (Ag)	<i>Jsc</i>	0.92	0.92	0.69	0.87	0.76
	<i>Voc</i>	0.89	0.89	0.80	0.80	0.78
	<i>IQE</i>	0.87	0.87	0.69	0.83	0.72
<i>TiO</i> ₂ <i>Cu</i> ₂ <i>O</i> (Ag Cu)	<i>Jsc</i>	0.89	0.88	0.76	0.84	0.74
	<i>Voc</i>	0.56	0.55	0.62	0.73	0.50
	<i>IQE</i>	0.91	0.89	0.78	0.82	0.72
<i>TiO</i> ₂ <i>Co</i> ₃ <i>O</i> ₄ <i>MoO</i> ₃	<i>Jsc</i>			0.82	0.82	
	<i>Voc</i>			0.00	0.33	
	<i>IQE</i>			0.79	0.79	

3 Conclusions

In this work we demonstrated the power of materials informatics methods to visualize the PV space of MO-based solar cell libraries and to develop predictive QSAR models for key PV properties. Importantly, we wish to emphasize the need to perform studies of the type described in this work in close collaboration with experimentalists in order to both provide physics/chemistry based explanation to the observed trends/effects/correlations and to capitalize on the results.

References

1. Rajan, K.: *Mater. Today* **8**, 38–45 (2005). [https://doi.org/10.1016/S1369-7021\(05\)71123-8](https://doi.org/10.1016/S1369-7021(05)71123-8)
2. Senderowitz, H., Tropsha, A.: *J. Chem. Inf. Model.* **58**, 1313–1314 (2018). <https://doi.org/10.1021/acs.jcim.8b00016>
3. Scharber, M.C., Sariciftci, N.S.: *Prog. Polym. Sci.* **38**, 1929–1940 (2013). <https://doi.org/10.1016/j.progpolymsci.2013.05.001>
4. Faccio, R., Fernandez-Werner, L., Pardo, H., Mombre, A.W.: *Recent Pat. Nanotechnol.* **5**, 46–61 (2011). <https://doi.org/10.2174/187221011794474930>
5. Rong, Y., et al.: *Science* **361**(6408) (2018). <https://doi.org/10.1126/science.aat8235>
6. Rühle, S., et al.: *Phys. Chem. Chem. Phys.* **16**, 7066–7073 (2014). <https://doi.org/10.1039/C4CP00532E>
7. Yosipof, A., Nahum, O.E., Anderson, A.Y., Barad, H., Zaban, A., Senderowitz, H.: *Mol. Inf.* **34**, 367–379 (2015). <https://doi.org/10.1002/minf.201400174>
8. Kaspi, O., Yosipof, A., Senderowitz, H.: *Mol. Inf.* **37**(9–10), e1800067 (2018). <https://doi.org/10.1002/minf.201800067>
9. Yosipof, A., Kaspi, O., Majhi, K., Senderowitz, H.: *Mol. Inf.* **35**, 622–628 (2016). <https://doi.org/10.1002/minf.201600050>
10. Kaspi, O., Yosipof, A., Senderowitz, H.: *J. Chem. Inf. Model.* **58**, 2428–2439 (2018). <https://doi.org/10.1021/acs.jcim.8b00552>
11. Mitchell, M.: *An Introduction to Genetic Algorithms*. MIT Press, London (1998). ISBN 0-262-13316-4
12. Kaspi, O., Yosipof, A., Senderowitz, H.: *J. Chemoinf.* **9**, 34 (2017). <https://doi.org/10.1186/s13321-017-0224-0>
13. Fischler, M.A., Bolles, R.C.: *Commun. ACM* **24**, 381–395 (1981). <https://doi.org/10.1145/358669.358692>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.