The predictive power of NLP models on Perovskite solar cells: BERTforPSC

Naveen Bhatia,\*, Mohammad Khaja Nazeeruddinb, and François Maréchala

aIndustrial Process and Energy Systems Engineering, Ecole Polytechnique Fedérale de Lausanne, Sion, 1951, Switzerland

 bGroup for Molecular Engineering of Functional Materials, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fedérale de Lausanne, Sion, 1951

\*naveen.bhati@epfl.ch

Abstract

With the advent of ChatGPT, natural language processing (NLP) models have gained tremendous interest from the research community and have been applied to a plethora of scientific domains like batteries, pharmaceuticals, recycling plastics, etc., to obtain insights from the existing corpus of literature, and thus making the process of reading, analyzing, interpreting, and reporting the results shorter and faster. However, the applications of such models are still limited to a few fields in the past, and perovskite solar cells (PSCs) are among them. Recently, PSCs power conversion efficiency climbed the mark of 26.1% in a single junction and 33.7% in silicon/perovskite tandem solar cells, putting them in the leading position of next-generation solar cells. However, optimizing decision variables in terms of materials selection and process conditions requires analysis of the huge database of experiments to draw better insights to make them market-competitive in terms of cost and environmental impacts. In this article, authors have used two state-of-the-art NLP models, BERT and SciBERT, to analyze the corpus of stability data based on experimental datasets and further normalised based on storage and testing conditions to visualize the trends and compare their performance with regression-based models. The insights obtained while employing such models with different kinds of datasets where both alpha-numeric keys are presented as model features are also offered, highlighting the limitations of such models. The efficiency and effectiveness of such models in interpreting the causal relationships and predicting the trends will help in utilizing such models for tackling the challenges of optimizing material-process design problems (MPDP) with available data from literature.

**Keywords**: Natural language processing, BERT, Machine learning, Perovskite solar cells

* 1. Introduction

Perovskite solar cells have presented a unique opportunity in the field of PV technologies to have more control over the different aspects of fabrication including the techniques or equipment employed, process parameters used, and materials selection (Huang, 2020) which is not the case with many existing 1st and 2nd generation PV technologies (Singh et al., 2021). However, this paradox of choice also results in using more nuanced tools and techniques to optimize the overall production of PSCs which not only meets the requirements in terms of efficiency and stability but also emphasizes the design of environmental-friendly and sustainable PV technology for the future (Yoo et al., 2022). Given these possibilities, the amount of experimental work carried out has also presented an opportunity to use state-of-the-art techniques for finding better overall solutions from the myriad of existing possibilities (Jacobsson et al., 2022).

Machine learning and artificial intelligence lie at the forefront of these new techniques that are now used in almost all domains (Sarker et al., 2021). Plenty of work has also been carried out in the domain of perovskite solar cells where the focus of application has varied from individual layer-level characteristics (Liu Y. et al, 2022) to cell (Liu Z. et al, 2022) and module-level characteristics (Ramirez et al., 2023). Lu et al. (2021) and Hu et el. (2022) have discussed the trends in efficiency, bandgaps, Voc, Isc, and stability of PSCs using different kinds of machine learning techniques. However, recently with the advent of transformer-based models, many advanced natural language processing-based algorithms have made a mark in the field of ML/AI-based analysis. BERT (Devlin et al., 2018), XLNet (Yang et al, 2019), and SciBERT (Beltagy et al., 2019) are all different variations of the NLP models based on transformer architecture. ChatGPT (OpenAI) and BARD are among the most advanced ones with the number of parameters greater than one billion. Given these tools at hand, it is important to analyze the effectiveness in dealing with scientific data, especially from the field of material science and optoelectronics. However, since the earlier models are mostly trained on web text and general science it is difficult to expect their high performance on domain-specific tasks (Xie et al., 2023) which is partly because of their limited vocabulary and also lack of embeddings for such specific data.

Here, the NLP models were fine-tuned for the regression tasks of predicting the stability of PSCs by enhancing the vocabulary size of these models with existing perovskite solar cells FAIR (findable, accessible, interoperable, and reusable) dataset (Jacobsson et al., 2022) and are compared with existing regression models. Finally, hyperparameter optimization for all models was carried out to realize the best performance on the dataset.

* 1. Data and models

In the first part of the article, existing machine learning techniques both linear and non-linear are explored to set the base level of prediction capabilities of these techniques and provide a baseline for comparing the NLP models like BERT. Also, the objective is to predict the continuous values of stability and therefore regression-based analysis was carried out using these models.

* + 1. Data cleaning and processing

The data for PSC stability was used based on the analysis presented by Zhang et al. (2022) for the FAIR database which has more than 42000 data points from the experiments and contain data related to the active materials (embedded materials), assisting materials, additives and techniques used for fabricating different layers along with the process parameters for certain cases. Thus, it can be assumed that data for the present analysis is a simulated data based on experiment with physics-based modeling to account for the variations for the storage conditions like temperature, humidity and light exposure. This original dataset contains around 7400 datapoints from 2013 to 2021. Afterwards, the data is cleaned for missing stability values and only data after 2017 is chosen to avoid the bias from the earlier experiments in the field. Also, the data was filtered to represent the most commonly used perovskite thus avoiding sparse data depending on the light intensity, simulator class, perovskite-inspired structure, perovskite ABX3 structure, single crystal, and cell architecture. In the next step, cleaning of data was based on different layers of the stack, their deposition procedure, the solvent used, synthesis atmosphere, ions composition for perovskite, and additives used such that they represent most of the data. This cleaning procedure is adapted to improve the generalization of models as in other cases it will lead to very high dimensional data. Based on this finally we had 2783 data points which were then cleaned for JV measurement data (missing values for PCE, Voc, Isc, and FF). However, this dataset still has some missing values for certain columns and different amputation strategies were used to treat them while building ML models on top of them. Before that, one-hot encoding strategy is used to treat the categorical data as most of the data exists in terms of the name of elements, compounds, or techniques. In terms of data engineering, a few new columns were also created taking into account the annealing process characteristics like different annealing temperatures, maximum annealing temperature, and total energy absorbed during annealing per unit mass and specific heat as it is related to the phase conversion and might have a good effect on stability. In the end, we have 444 features for our dataset including JV measurement characteristics and stability (TS80 (based on stabilized efficiency at the end of the burn-in-region, Zhang et al. (2022))).

While checking the quality of the dataset, it was found that there were several outliers in terms of stability which were cleaned using the z-score criteria of 3. Even after outliers removal, the data of stability was highly skewed on the right (positive skewed, skewness =9.12) and therefore log transformation was applied to make the data look more normally distributed (skewness=-0.07) which performs better with ML techniques. Also, the cell area values were imputed with the median values. Moreover, the perovskite additives and HTL additives columns were imputed with ‘undoped’ value. The columns with no variance are removed at this point. Afterward, stratification splitting of data was carried out using the log-transformed stability values so that training (80%) and test data (20%) have values from all kinds of stability regions. Figure 1 shows the variation in stability values with different kind of layers i.e., ETL and HTL in the overall stack.

 

Figure 1: Range of stability values with different a) ETLs and b) HTLs

It shows the range of values possible for any given type of material/stack depending on the other features of the overall cell. During the Pearson correlation check between the features of the dataset, it was found that log\_TS80m is highly correlated to TS80 which is basically some kind of direct indication of stability. However, in real-life scenarios, we would not have access to this data as this is something we would like to predict and therefore all features giving a direct measure of stability are removed. All other features have a lower than 0.22 correlation with the target variable i.e., log\_TS80m. A pipeline with an imputation strategy for numerical values using median and standard scaler and one-hot encoding transformation for categorical attributes was used to treat the dataset at this point. Based on these transformations, the final data has 425 columns or features.

* + 1. Classical Machine learning models

For ML modeling, initially, a screening is made on a set of 18 models without going for hyperparameter optimization using their default settings. Once models with relatively better performance were found, complete hyperparameter optimization was carried out either using GridSearchCV or RandomSearchCV depending upon the size of the search space to avoid excessive time consumption. For evaluating the models, two metrics were used, RMSE (root mean squared error) and R2 values (coefficient of determination). The models included in the first check included linear regression, decision tree regressor (DTR), random forest regressor (RF), support vector regression (SVM), KNeighbours regressor (KNN), extreme gradient boost regressor (XGB), ridge regressor (RR), lasso regressor, Bayesian regressor (BR), elastic net regressor (ELN), gradient boost regressor (GBR), Adaboost regressor (ADA), multi-layer perceptron regressor (MLP, neural networks), gaussian process regressor (GPR), extra trees regressor (ETR), kernel ridge regressor (KRR), and light gradient boost regressor (LGBM). Data was adapted accordingly to use these different kinds of algorithms/models. Finally, out of the 18 models only 7 were selected for hyperparameter tuning. These models include RF, SVM, XGB, ELN, GPR, ETR, and LGBM. The cross-validation was inherently performed using 5-fold CV in both these hyperparameter tuning methods. The performance of the models on the training and test dataset is shown in Figure 2 for both RMSE and R2 scoring metrics where RF model shows (the best performance in terms of both metrics without overfitting the training dataset (i.e., difference between the performance on 2 datasets is almost same). Here, the RMSE values are based on scaled log-transformed stability data.

 

Figure 2: Classical models performance (R2 and RMSE) on a) training dataset and b) test dataset

Based on the best model, the importance of various features was visualized using the recursive elimination method which recursively eliminates the least important features in the dataset. Figure 3 b) shows the relative importance of features where the highest contribution are from the JV measurements followed by energy absorbed, cell area, PSC thickness and maximum annealing temperature. Figure 3 c) shows SHAP analysis to see the nature of effect of these features where it was found that Voc, FF, PCE, Jsc are all positively impacting the stability. Also, the descriptors introduced after feature engineering like energy absorbed and maximum temperature during perovskite annealing are found to have positive impacts on the PSC stability performance.

  

Figure 3: a) RF model performance on test data, b) relative feature importance, and c) SHAP analysis for RF model

* + 1. NLP Machine learning models

In this section, two state-of-the-art models, BERT (both cased and uncased) and its variant SciBERT were used to see the performance on the task of stability prediction. In this case, since the models are based on transformer architecture, the starting dataset needs to be treated such that it can be used with these models. As most of the features in the dataset are categorical in nature, they can be directly used without any transformations as they are from the natural language domain. However, the numerical features due to a huge range of scale have to be treated in a manner that it doesn’t explode the vocabulary of the models while at the same time providing enough context for creating meaningful embeddings during the training process. To overcome this challenge, the numerical data was converted to scientific form with one decimal precision which significantly reduced the vocab size for all the numbers involved in the dataset. Afterward, the training data and test data (80:20) were created in a similar manner as for above-discussed regression models with log transformation of the target variable. During training both R2 and RMSE metrics were evaluated, however, improvements were made using R2 metrics. Hyperparameter optimization was carried out for learning rates (lr) and batch sizes (B) in grid search manner for BERT variants and random manner for SciBERT. Finally, the best models were selected based on their performance of the test data. With SciBERT, the best performance was found with uncased variant with R2 value as 0.22 (B=64, lr=2e-5) on test data. The tokenization based on SciBERT is shown in the figure below where color of the words are indicative of the relative scores assigned to them using integrated gradient method. For cased and uncased BERT, the best test performances (R2 value) were 0.225 (B=64, lr=2e-5) and 0.23 (B=64, lr=2e-5) which are lower a bit lower than





Figure 4: BERT results for relative scores of a) first 15 words b) last 15 words c) test data fitting and d) entire sentence

classical regression models (Figure 3 a)). Moreover, for analyses of the features, scores for each word based on the integrated gradients method are used as shown in Figure 4 a) and b) where the color of words in the sentence corresponds to their relative scores in the final prediction (Fig. 5).

* 1. Conclusions

Here, the authors have shown how new ML techniques can be used for solving the most pressing challenges of the PSCs and performed an exhaustive study on the different classical regression-based models and state-of-the-art BERT model performance on the dataset of PSC efficiency. It is shown that NLP models can be used to infer information and trends from existing datasets that are already curated rather than training them on a huge corpus of raw data from the literature which can be time-consuming and might have plenty of unnecessary information. Moreover, training on such structured data also allows us to understand and leverage their abilities on datasets created based on critical insights and domain expertise. However, more efforts have to be made on the training part of such models to improve their overall performance in comparison to classical models.

References

OpenAI. (2023). *ChatGPT* [Large language model]. <https://chat.openai.com>

BARD (2023). Google AI, <https://bard.google.com/chat>

Zhang, Z., Wang, H., Jacobsson, T.J. and Luo, J., 2022. Big data driven perovskite solar cell stability analysis.

Huang, Y., Liu, T., Liang, C., Xia, J., Li, D., Zhang, H., Amini, A., Xing, G. and Cheng, C., 2020. Towards simplifying the device structure of high‐performance perovskite solar cells.

Singh, B.P., Goyal, S.K. and Kumar, P., 2021. Solar PV cell materials and technologies: Analyzing the recent developments. Materials Today: Proceedings, 43, pp.2843-2849

Yoo, J.J., Shin, S.S. and Seo, J., 2022. Toward efficient perovskite solar cells: progress, strategies, and perspectives.

Jacobsson, T.J., Hultqvist, A., García-Fernández, A., Anand, A., Al-Ashouri, A., Hagfeldt, A., Crovetto, A., Abate, A., Ricciardulli, A.G., Vijayan, A. and Kulkarni, A., 2022. "An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles."

Sarker, I.H., 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions.

Lu, Y., Wei, D., Liu, W., Meng, J., Huo, X., Zhang, Y., Liang, Z., Qiao, B., Zhao, S., Song, D. and Xu, Z., 2023. Predicting the device performance of the perovskite solar cells from the experimental parameters through machine learning of existing experimental results.

Liu, Z., Rolston, N., Flick, A.C., Colburn, T.W., Ren, Z., Dauskardt, R.H. and Buonassisi, T., 2022. Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing.

Liu, Y., Yan, W., Han, S., Zhu, H., Tu, Y., Guan, L. and Tan, X., 2022. How machine learning predicts and explains the performance of perovskite solar cells.

Hu, Y., H., X., Z., L., Z., T., Y., J., J., B., M., Y., D., X., Z., L., W., J. and C., B., 2022. Machine‐Learning Modeling for Ultra‐Stable High‐Efficiency Perovskite Solar Cells.

Ramírez, E.A., V., J.P., F., A., M., J.F., B., R. and J., F., 2023. Blade‐Coated Solar Minimodules of Homogeneous Perovskite Films Achieved by an Air Knife Design and a Machine Learning‐Based Optimization.

Xie, T., Wa, Y., H. W., Z., Y., L., Y., L., Q., Wang, S., Kit, C., Grazian, C. and Hoex, B., 2023. Large Language Models as Master Key: Unlocking the Secrets of Materials Science with GPT.

Beltagy, I., L, K. and C., A., 2019. SciBERT: A pretrained language model for scientific text.

Yang, Z., D., Z., Y., Y., C., J., S., R.R. and L., Q.V., 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Devlin, J., C., M.W., L., K. and T., K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.