Unsupervised machine learning for solar cell materials from the literature

Cite as: J. Appl. Phys. **131**, 064902 (2022); doi: 10.1063/5.0064875
Submitted: 27 July 2021 · Accepted: 23 January 2022 ·
Published Online: 9 February 2022

Lei Zhang[1,2,a)] and Mu He[1]

AFFILIATIONS

[1]Institute of Advanced Materials and Flexible Electronics (IAMFE), School of Chemistry and Materials Science, Nanjing University of Information Science and Technology, 210044 Nanjing, China
[2]Department of Materials Physics, School of Chemistry and Materials Science, Nanjing University of Information Science and Technology, 210044 Nanjing, China

[a)]**Author to whom correspondence should be addressed:** 002699@nuist.edu.cn

ABSTRACT

Machine learning and data-driven methods have been adopted for material science research in recent years; yet, the textual data are not fully embraced by the materials and physics community. In this work, we aim to make the computers unsupervisedly learn the latent information on the solar cell materials based on the textual data with minimal human intervention and perform solar cell materials predictions. An unsupervised machine learning model is constructed by automatically extracting the information from the materials literature database using word embeddings, which successfully establishes the hidden relationships between the materials formulas and their photovoltaic applications. Uncommon solar cell materials predicted by the natural language processing (NLP)-based machine learning method are further evaluated via the first-principles methods to reveal the optoelectronic properties of the predicted candidate, demonstrating the validity of the NLP-assisted machine learning model. This study highlights the text-based machine learning methods for solar cell materials and calls for a wide deployment of the NLP methods for the materials research.

INTRODUCTION

Solar energy is a clean and renewable source that has the potential to replace fossil fuels for sustainable societies. The conversion of the sunlight into electricity is realized by various types of solar cells, including silicon solar cells, copper indium gallium selenide (CIGS) solar cells, perovskite solar cells, dye-sensitized solar cells (DSSCs), and organic solar cells.[1–3] Many solar cells face a series of limitations, such as low power conversion efficiency, poor stability, or higher cost. As a result, new types of solar cell materials should be identified.

The concept of "big data," often accompanied by artificial intelligence, has significantly transformed the way people predict and understand materials.[4,5] The materials science literature includes a large number of images, numbers, and texts, with the Arabic numbers and images readily available via machine learning and image recognition tools.[6] On the other hand, the textual data are less understood by the computer due to the abstract nature of the texts. Nevertheless, with the recent development of the natural language processing (NLP) methods, more information can be extracted from the texts in the materials and chemical literature studies.[7–10] For example, Tshitoyan *et al.* employed the word embeddings to predict the thermoelectric materials.[6] Zhang *et al.* applied the NLP method to analyze bifunctional materials for photo-rechargeable battery applications.[11] Kim *et al.* trained a neural network from the literature to annotate the materials synthesis description and construct a model for synthetic planning.[12] Huang *et al.* prepared a battery database by automatically extracting the materials science literature via ChemDataExtractor.[13,14] The employment of text-based unsupervised machine learning is promising to extract hidden information from materials considering the vast amount of textual data in the published papers and patents.

First-principles calculations starting from the Schrodinger equations have been demonstrated as reliable methods to obtain the electronic and optical properties of materials.[15,16] The first-principles calculation is often adopted in parallel with the experiments to help understand the structure–property relationships of

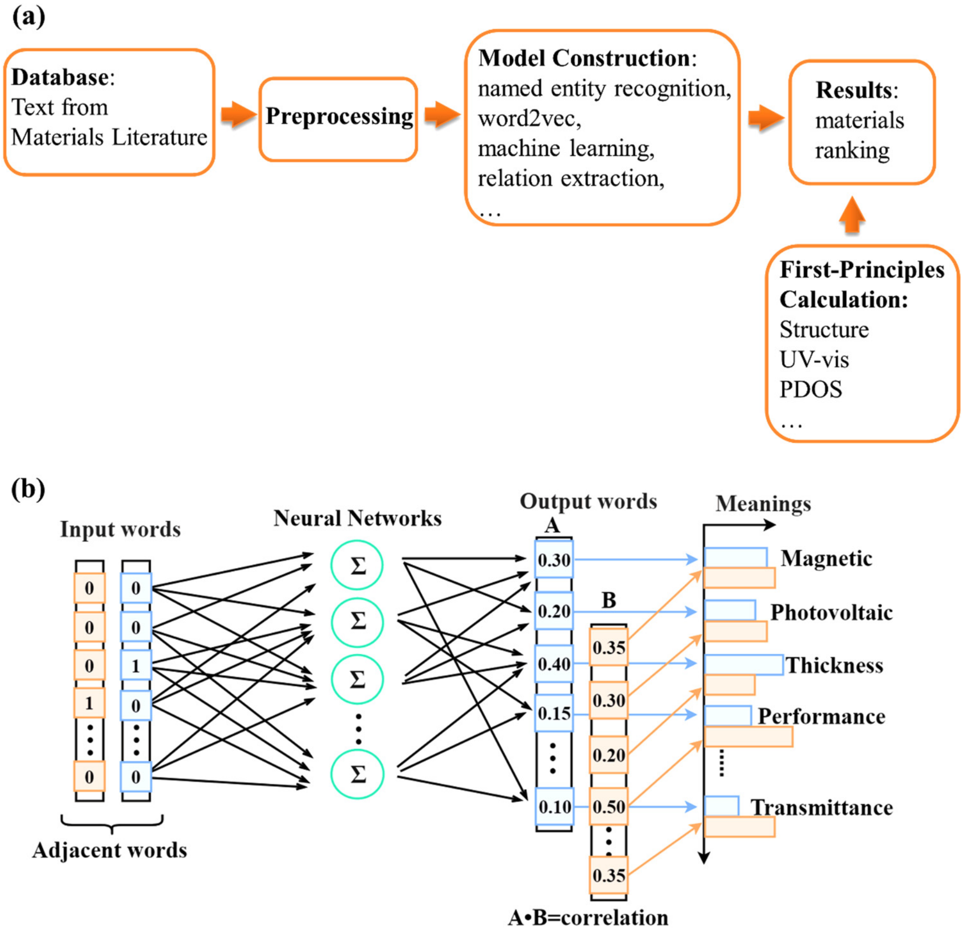
(a)
Database: Text from Materials Literature
Preprocessing
Model Construction: named entity recognition, word2vec, machine learning, relation extraction, …
Results: materials ranking
First-Principles Calculation: Structure UV-vis PDOS …
(b)
Input words
Neural Networks
Output words
A
B
Meanings
Magnetic
Photovoltaic
Thickness
Performance
Transmittance
0.30
0.20
0.40
0.15
0.10
0.35
0.30
0.20
0.50
0.35
Adjacent words
A•B=correlation

FIG. 1. (a) Flowchart of the text-based machine learning process for the solar cell materials. First, the material-domain database is prepared by downloading the abstracts of materials papers. Second, the preprocessing (text cleaning, sentence splitting, tokenization, spell checker, pos tagging, lemmatization, stemming, etc.) is performed. Third, the names of the materials are recognized in the name entity recognition step, and a machine learning model is constructed via word2vec to establish the relationships between the materials names and their applications. Finally, first-principles calculations are performed to verify the predicted candidates. (b) The explanation of the NLP process with the word2vec technique that codified the texts and construct neural networks for the word embedding. The prediction is based on the calculation of the cosine similarity between the word vectors such that the relationships of the materials and properties/applications can be extracted.

materials and their hidden mechanisms and can be employed to obtain materials databases in a high-throughput manner with lower cost and reduced time, facilitating the data-driven investigations on materials.

In this paper, we employ the NLP-based machine learning methods to predict solar cell materials, which are analyzed via first-principles calculations. We aim to make the computers unsupervisedly learn the latent information on the solar cell materials based on the textual data with minimal human intervention and perform solar cell materials predictions. The unsupervised machine learning process successfully identifies common solar cells materials in an unsupervised manner. In addition, an unconventional solar cell material As_2O_5 is suggested by the machine learning model. The density of states, UVvis absorption spectra, and band structures of the predicted material are obtained via first-principles calculations to evaluate its feasibility for the photovoltaic applications

COMPUTATIONAL DETAILS

The machine learning process includes textual database preparation, preprocessing, model construction, results generation, and first-principles calculations [Fig. 1(a)]. First, the material-domain database is prepared by downloading the abstracts of materials papers, which are considered accurate sources that summarize the major scientific concepts of the articles. The abstract database consists of 50 000 materials science papers from SpringerLink (https://link.springer.com/). Second, preprocessing (text cleaning, sentence splitting, tokenization, spell checker, pos tagging, lemmatization, stemming, etc.) is performed and ChemDataExtractor is employed to perform the word segmentation on the abstract database. The abstracts related to announcement, book review, and correction are deleted. In the NLP process, the proper tokenization (Fig. S1, supplementary material) in the preprocessing step is critical,[9,17] and the stop words and the customized dictionary are included such that the redundant stop words can be deleted and the technical phrases in the materials and chemical domain can be correctly recognized. Third, the names of the materials are recognized in the name entity recognition step, and a machine learning model is constructed via the word2vec method[18,19] to establish relationships between the materials names and their applications. The skip-gram model is employed, with the context window of 8, the number of iterations of 5, and

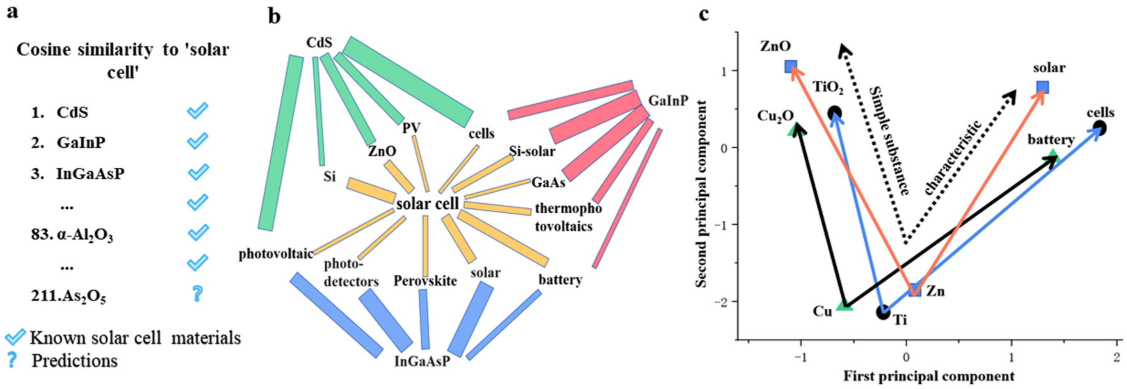

FIG. 2. (a) The output list of solar cell materials predicted by the machine learning method. The ranking is based on the absolute value of cosine similarity between the word vectors of the chemical formula and the solar cell. (b) The predicted correlations of the solar cell materials, their context words, and the target word (solar cell); the higher the correlation, the thicker the width of the connection line. (c) Two-dimensional projection of some keyword vectors. The 200-dimensional word vectors are visualized after the dimension reduction, demonstrating domain knowledge and model validity.

the vector dimension size of 200. The machine-learned model gradually matures through continuous iterative training and, finally, outputs each word in the form of a vector. The skip-gram method is employed in this study. There are other methods such as a continuous bag of words (CBOWs) (Fig. S2, supplementary material), which trains faster than the skip-gram counterpart.[20,21] In this study, the skip-gram is employed because it is generally more accurate and can better extract the information from the word embedding. In addition, the following figure is provided in the revised manuscript to demonstrate the cosine similarity and the NLP process. The NLP model is constructed based on the word2vec technique that codified the texts and neural networks are established for the word embedding [Fig. 1(b)]. The prediction is based on the calculation of the cosine similarity between the word vectors such that the relationships of the materials and the properties/applications can be extracted. The cosine similarity is calculated according to the following formula:

$$\text{similarity} = \cos\theta = \frac{\bar{x}\cdot\bar{y}}{||\bar{x}||\,||\bar{y}||}, \tag{1}$$

where $\bar{x}\cdot\bar{y}$ is the vector dot product from x and y, $||\bar{x}||$ is the long vector x, $||\bar{y}||$ is the long vector y.

Finally, first-principles calculations are performed to obtain the optimized structure, simulated UV–vis spectra, and electronic properties, which verify the validity of the machine learning model. The first-principles calculations are carried out in CASTEP,[22] employing the Perdue–Burke–Ernzerhof (PBE) functional and 430 eV cut-off energy. The van der Waal effects are accounted for using the Tkatchenko–Scheffler (TS) scheme.[23] The convergence tolerances for the energy, force, and displacement are 2×10^{-6} eV/atom, 0.03 eV/Å, and 0.002 Å. A k-point set of $6\times3\times3$ is employed for the geometrical optimization and energy calculation.

RESULTS AND DISCUSSION

The machine learning model outputs the well-known solar cell materials (Fig. 2 and Table I), including Si, GaAs, ZnO, CIGS, InP, c-Si, CdS, GaInP, and InGaAsP, in an unsupervised manner by reading the materials literature with minimal human intervention. Their correlation coefficients with the solar cell (the absolute value of the cosine product of the two word vectors) exceed 0.5; for example, the correlation coefficients of CdS, GaInP, and InGaAsP are 0.574 723 06, 0.572 785 44, and 0.569 808 90, respectively, which are suggested to be highly correlated with the solar cell. In fact, the materials up to the top 210 are all known solar cell materials [Fig. 2(a)]. Apart from the common solar cells materials that have been widely reported in the literature, several uncommon materials are predicted by the NLP-based method, such as As_2O_5 that ranks 211th; as a result, the optoelectronic properties of As_2O_5 are further examined via the first-principles calculations to evaluate the viability of the machine learning prediction (*vide infra*).

The relationships between the chemical formula and the characteristic words related to the target word solar cell can be visualized intuitively [Fig. 2(b)]. "CdS" is closely related to the word "cells" and "photovoltaic," which is revealed by the thick connection lines between the pairs (high correlation coefficients). In addition, "Si" and "ZnO" are closely related to "solar cell" evidenced by the high correlation coefficients (thick lines), which agrees with the domain knowledge that the two materials are largely employed for solar cell applications. The transition words between CdS and solar cells, such as Si and ZnO, helps establish materials–application relationships, representing the knowledge graphs.[24,25] Furthermore, the most relevant transition word that connects "GaInP" and "solar cell" is "Si," while the word "battery" demonstrates negligible relevance with "GaInP," demonstrating that GaInP is a solar cell material rather than a battery material. The same situation occurs for "InGaAsP," which is closely related to the word "solar" and negligibly related to "battery" (narrow line), suggesting the applicability of InGaAsP in solar cells rather than batteries. Last but not least,

TABLE I. Ten possible solar cell materials are suggested by the NLP-based machine learning model. Ranking is according to the correlation coefficients (cosine similarity between the two word vectors) between the materials and the solar cell. Common solar cell materials are successfully predicted by the unsupervised machine learning model.

Chemical formula	Relevance coefficient	Chemical formula	Relevance coefficient
1. CdS	0.574 723 06	6. ZnO	0.550 606 55
2. GaInP	0.572 785 44	7. c-Si	0.550 409 80
3. InGaAsP	0.569 808 90	8. InGaN-based	0.549 664 26
4. InP-based	0.566 466 10	9. GaAs	0.549 351 30
5. Si	0.563 009 13	10. Cu(In,Ga)Se_2	0.546 454 85

"CdS" is closely related to "cells," "PV," and "photovoltaic," while "GaInP" is closely related to "cells," "Si solar," and "thermophotovoltaics"; these demonstrate the applicability of these materials in these three areas that validate the machine learning process. To sum up, the NLP model successfully extracts the domain knowledge of the solar cell materials in an unsupervised way.

In order to further demonstrate the validity of the machine learning model, two-dimensional vector projection of the chemical formula, the elements, and their attribute word vectors is plotted to demonstrate the relationships between the word vectors [Fig. 2(c)]. After the dimensional reduction operation, the original 200-dimensional word vectors are displayed in a two-dimensional coordinate system, and the textual words can be added and subtracted like vectors. The visualization of word vectors is important to justify the NLP-based machine learning model, with the addition and subtraction of word vectors reflecting the domain knowledge. The metal elements and their oxides tend to gather in different regions, and the vectors from elements to oxides are consistent. Similarly, the elemental metallic elements and their oxides gather in different areas, and the vectors from the elements to the oxides are almost parallel with each other. In addition, the word vector difference between Cu_2O and Cu is similar to the vector difference between TiO_2 and Ti, and the vector difference between the exemplar applications (solar, battery, and cells) and their corresponding elements is consistent. As a result, the model after the NLP training successfully extracts the relationships and concepts in materials science (the relationships between metals and their oxides) in a quantitative manner and correctly, distinguishes the characteristics of the chemical formula and the corresponding properties.

TABLE II. Correlation (cosine similarity between the two word vectors) between common synonyms predicted by the model, with the small correlation coefficients between the antonyms verifying the NLP-based machine learning model.

Test word 1	Test word 2	Relevance coefficient
1. HOMO	LUMO	0.870 259 46
2. PV	Photovoltaic	0.597 266 73
3. Solar	Light	0.420 134 00
4. Molecular	Orbitals	0.585 716 50
5. InP	GaAs	0.698 754 90
6. Photovoltaic	Solar	0.645 049 33
7. Li-ion	Battery	0.685 017 70
8. Cells	Batteries	0.333 488 67

TABLE III. Correlations (cosine similarity between the two word vectors) between several unrelated technical words predicted by the model, with the small correlation coefficients between the antonyms verifying the NLP-based machine learning model.

Test word 1	Test word 2	Relevance coefficient
1. PV	Electromagnetic	0.143 457 52
2. Battery	Orbital	0.169 476 54
3. LUMO	Thermoelectric	0.171 693 97
4. Water	Piezoelectric	0.175 615 88
5. Photolysis	Plating	0.093 706 19

The NLP model was examined using common synonyms and antonyms in the solar cell material domain. The highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) have the largest correlation coefficient (0.87) (Table II), which is associated with the frequent appearance of the two important energy levels that are extensively employed for the bandgap calculations and charge transport mechanism explanations. In addition, the synonym pairs (PV and photovoltaic, solar and light, molecular and orbitals, InP and GaAs, photovoltaic and solar, and Li-ion and battery) exhibit high correlation coefficients from 0.42 to 0.70, which agree with the domain knowledge. Several unrelated words are examined to further evaluate the NLP model, which successfully predicts the low correlation coefficients and identifies the irrelevant relationships between the technical words. For example, PV and electromagnetic have a correlation coefficient of 0.14; battery and orbital have a correlation coefficient of 0.17; LUMO and thermoelectric have a correlation coefficient of 0.17; water and piezoelectric have a correlation coefficient of 0.18; and photolysis and plating have a correlation coefficient of 0.09 (Table III). The material maps of various solar cells are plotted to examine the model, which is based on the principle component analysis (PCA) with dimension reduction projection of the word vectors of the solar cell materials, with each point representing a material name [Fig. 3(a)]. The distance between different materials is related to the cosine similarity of the two word vectors, and similar materials tend to cluster together in the diagram, such as the dye-sensitized solar cells (DSSCs). The number of occurrences of the eight most common elements in the diagram demonstrates that Si, Ti, O, C, Ga, S, In, and Cu are important ingredients for the solar cell materials, which corresponds to the common solar cell types (such as silicon, TiO_2, GaAs, and CIGS solar cells). As a result, the unsupervised learning model successfully establishes the relationships between the technical words in the materials domain and is promising for the material prediction task.

First-principles calculations are then performed on As_2O_5 (ranked 211th) that is predicted by the NLP-based machine learning model, which evaluates the optoelectronic properties of the material and the model validity. The simulation reveals an optimized structure with a unit cell size of $4.72 \times 8.69 \times 8.82$ Å^3, with

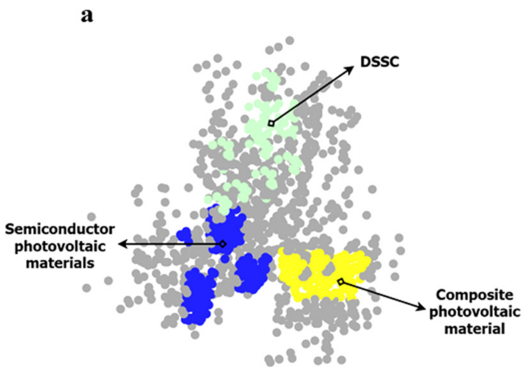

FIG. 3. Material maps. (a) PCA dimension reduction projection of solar cell materials based on the word vectors, with each point representing a material name. The distance between different materials is related to cosine similarity. Similar materials tend to cluster together in the diagram, such as the dye-sensitized solar cells (DSSCs). (b) The number of occurrences of the eight most common elements in the diagram.

every two arsenic atoms sharing five oxygen atoms [Fig. 4(a)]. A direct bandgap of 1.6 eV at the Gamma point is predicted [Fig. 4(b)]. In the projected density of states (PDOS) spectra, the valence band is predominately contributed by the p orbital of the oxygen atoms, while the conduction band is predominately contributed by the s orbital of the arsenic atoms [Fig. 4(c)]. The absorption coefficients of As_2O_5 in the UV–vis region range from 1200 cm^{-1} (780 nm) to 14 000 cm^{-1} (400 nm) [Fig. 4(d)], which demonstrates decent light absorption in the visible region. The structure and properties of $CH_3NH_3PbI_3$ calculated using the same level of theory are also shown for comparison purposes [Figs. 4(e)–4(h)]. The two materials exhibit similar bandgap values and demonstrate well-connected networks that are appropriate for the charge transfer purpose. The valence band of As_2O_5 is mainly contributed by O-2p orbitals while that of $CH_3NH_3PbI_3$ is mainly contributed by the I-5p orbitals; the conduction band of As_2O_5 is mainly contributed by As-4 s orbitals while that of $CH_3NH_3PbI_3$ is mainly contributed by the Pb-6p orbitals. As_2O_5 has weaker absorption intensities in the UV–vis region than $CH_3NH_3PbI_3$; nevertheless, the difference is not significant and As_2O_5 demonstrates proper band structures for solar cell applications. To sum up, the electronic and optical properties of As_2O_5 predicted by the first-principles calculations suggest that the unsupervised machine learning model is feasible for solar cell material prediction.

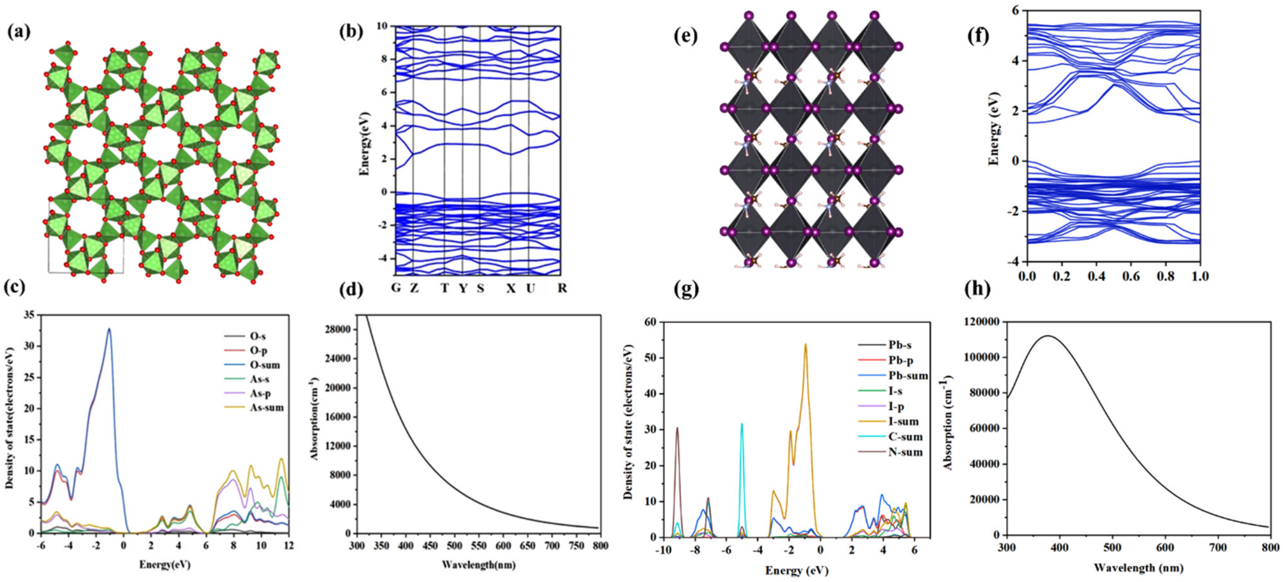

FIG. 4. Structure and properties of As_2O_5. For comparison purposes, the structure and properties of an established photovoltaic material $CH_3NH_3PbI_3$ calculated using the same level of theory are also shown. (a) The optimized crystal structure of As_2O_5. (b) The band structure of As_2O_5, demonstrating a direct bandgap of 1.5 eV. (c) PDOS spectra of As_2O_5. (d) Simulated UV–vis absorption spectra of As_2O_5. (e) The optimized crystal structure of $CH_3NH_3PbI_3$. (f) The band structure of $CH_3NH_3PbI_3$, demonstrating a direct bandgap of 1.5 eV. (g) PDOS spectra of $CH_3NH_3PbI_3$. (h) Simulated UV–vis absorption spectra of $CH_3NH_3PbI_3$.

CONCLUSIONS

An unsupervised machine learning model is constructed for the solar cell materials using texts in the literature. From the data science point of view, the text-based machine learning method will benefit from the wide availability of textual data in the materials literature. In the machine learning process, the chemical formula and the solar cell target words are represented as vectors, which lead to the successful relationship extraction of the materials and their applications. The common solar cell materials are effectively identified by the unsupervised machine learning model. The uncommon solar cell material appearing in the list is then examined via the first-principles calculations to reveal the validity of the machine learning model, and the predicted candidates such as As_2O_5 exhibit decent electronic and optical properties that are favorable for the solar cell applications. The present study calls for more NLP-based machine learning studies for materials research, considering the notorious data scarcity and sparsity problems in materials science.

SUPPLEMENTARY MATERIAL

See the supplementary material for the description of the NLP model, the tokenization process, the skip-gram, and CBOW architectures in word2vec, the cosine similarity definition, and the first-principles calculations.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NNSFC, No. 51702165). The authors acknowledge computational support from NSCCSZ Shenzhen, China.

AUTHOR DECLARATIONS

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The authors have approved the final version of the manuscript.

DATA AVAILABILITY

The data that support the findings of the study are openly available in GitHub, Ref. 26.

REFERENCES

[1] R. Prasanna, T. Leijtens, S. P. Dunfield, J. A. Raiford, E. J. Wolf, S. A. Swifter, J. Werner, G. E. Eperon, C. de Paula, A. F. Palmstrom, C. C. Boyd, M. F. A. M. van Hest, S. F. Bent, G. Teeter, J. J. Berry, and M. D. McGehee, Nat. Energy **4**, 939 (2019).

[2] A. Agresti, A. Pazniak, S. Pescetelli, A. Di Vito, D. Rossi, A. Pecchia, M. Auf der Maur, A. Liedl, R. Larciprete, D. V. Kuznetsov, D. Saranin, and A. Di Carlo, Nat. Mater. **18**, 1228 (2019).

[3] Q. Chen, J. Wu, X. Ou, B. Huang, J. Almutlaq, A. A. Zhumekenov, X. Guan, S. Han, L. Liang, Z. Yi, J. Li, X. Xie, Y. Wang, Y. Li, D. Fan, D. B. L. Teh, A. H. All, O. F. Mohammed, O. M. Bakr, T. Wu, M. Bettinelli, H. Yang, W. Huang, and X. Liu, Nature **561**, 88 (2018).

[4] M. Park, J. Ryu, W. Wang, and J. Cho, Nat. Rev. Mater. **2**, 16080 (2017).

[5] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nat. Mater. **12**, 191 (2013).

[6] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, Nature **571**, 95 (2019).

[7] J. Pennington, R. Socher, and C. Manning, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014), pp. 1532–1543.

[8] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, Bioinformatics **17**, 574–582 (2001).

[9] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain, J. Chem. Inf. Model. **59**, 3692 (2019).

[10] Z. Jensen, E. Kim, S. Kwon, T. Z. H. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, and E. Olivetti, ACS Cent. Sci. **5**, 892 (2019).

[11] M. He and L. Zhang, Int. J. Energy Res. **45**, 15521 (2021).

[12] E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum, S. Jegelka, and E. Olivetti, J. Chem. Inf. Model. **60**, 1194 (2020).

[13] M. C. Swain and J. M. Cole, J. Chem. Inf. Model. **56**, 1894 (2016).

[14] S. Huang and J. M. Cole, Sci. Data **7**, 260 (2020).

[15] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Sci. Adv. **3**, e1701816 (2017).

[16] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, and M. Scheffler, Sci. Adv. **5**, eaav0693 (2019).

[17] D. M. Wilary and J. M. Cole, J. Chem. Inf. Model. **61**, 4962 (2021).

[18] K. W. CHURCH, Nat. Lang. Eng. **23**, 155 (2017).

[19] L. Ma and Y. Zhang, in *2015 IEEE International Conference on Big Data (Big Data)* (IEEE, 2015), pp. 2895–2897.

[20] B. Bansal and S. Srivastava, Proc. Comput. Sci. **132**, 1147 (2018).

[21] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, Expert Syst. Appl. **69**, 214 (2017).

[22] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, and M. C. Payne, J. Phys.: Condens. Matter **14**, 2717 (2002).

[23] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 073005 (2009).

[24] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," in *IEEE Transactions on Neural Networks and Learning Systems* (IEEE, 2021).

[25] X. Zou, J. Phys.: Conf. Ser. **1487**, 012016 (2020).

[26] M. He and L. Zhang (2021). "NLP-As2O5", GitHub. https://github.com/Zhang-NJ-Lab/MuHe/tree/main/As2O5

Unsupervised Machine Learning for Solar Cell Materials from Literature

Lei Zhang[ab*] *and Mu He*[a]

[a] *Institute of Advanced Materials and Flexible Electronics (IAMFE), School of Chemistry and Materials Science, Nanjing University of Information Science & Technology, 210044, Nanjing, China. Email: 002699@nuist.edu.cn*

[b] *Department of Materials Physics, School of Chemistry and Materials Science, Nanjing University of Information Science & Technology, 210044, Nanjing, China.*

KEYWORDS Natural Language Processing; Machine Learning; NLP; Solar Cell

Supporting Information

NLP Model

The NLP model is based on the word2vec technique that produces word embedding for better word representation and captures a large number of precise syntactic and semantic word relationships. It is a shallow two-layered neural network that can detect synonymous words and suggest additional words for partial sentences once it is trained, where words are represented in the form of vectors and similar meaning words appear together and dissimilar words are far away from each other. Since neural networks do not understand the texts and instead they only understand numbers, the word embedding is employed to provide an effective way to convert text to a numeric vector.[1] Word2vec is a widely used algorithm based on neural networks, which is commonly referred to as "deep learning" (though word2vec itself is rather shallow). Using large amounts of unannotated plain text, word2vec learns relationships between words automatically. The output are vectors, and the outputs demonstrate h linear relationships such as:

vector (king) – vector (man) + vector (woman) ≈ vector (queen)

vector (Paris) ≈ vector (France) – vector (Italy) + vector (Rome)

The model can be constructed via genism using codes such as:

```
import gensim.models

sentences = MyCorpus()

model = gensim.models.Word2Vec(sentences=sentences)
```

In the NLP process, the proper tokenization (Figure S1) in the preprocessing step is critical, and the stop words and the customized dictionary should be included, such that the redundant stop words can be deleted and the technical phrases in the materials and chemical domain can be recognized.

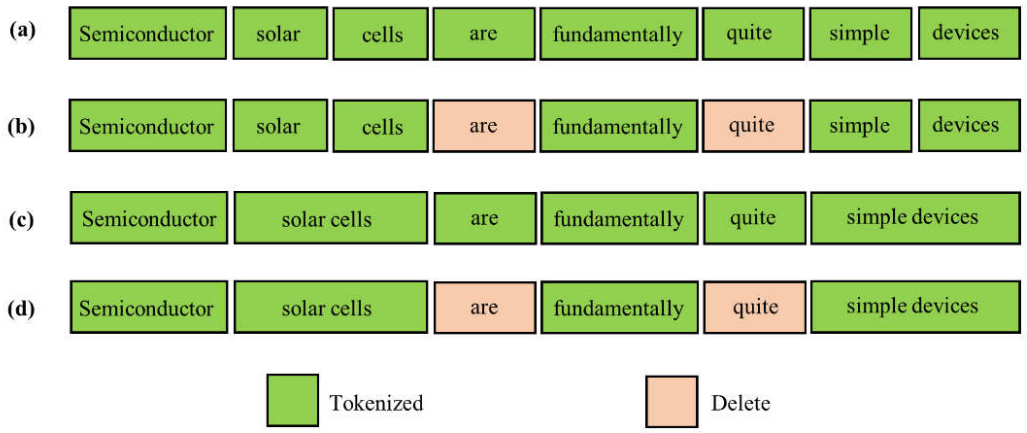

Figure S1. Tokenization process. (a) Tokenization with neither stop words nor customized dictionary. (b) Tokenization with stop words (the stop word are deleted). (c) Tokenization with customized dictionary. (d) Tokenization with both stop words and customized dictionary (the stop word are deleted).

Skip-gram vs. CBOW

The word2vec skip-gram model takes in pairs (word1, word2) generated by moving a window across text data, and trains a 1-hidden-layer neural network based on the synthetic task of given an input word, giving us a predicted probability distribution of nearby words to the input. A virtual one-hot encoding of words goes through a 'projection layer' to the hidden layer; these projection weights are later interpreted as the word embeddings. So if the hidden layer has 200 neurons, this network will give us 200-dimensional word embeddings. The CBOW word2vec model is very similar to the skip-gram model and also has a 1-hidden-layer neural network. However, the synthetic training task now uses the average of multiple input context words to predict the centre word rather than a single word as in skip-gram (Figure S2).

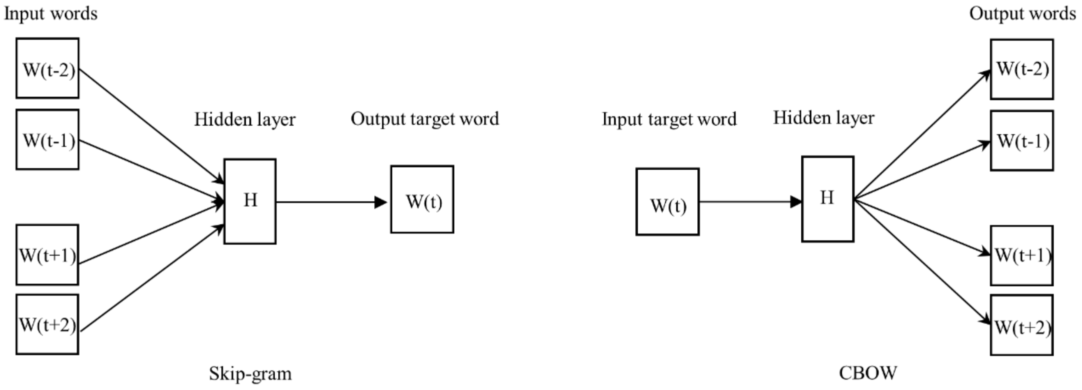

Figure S2. The CBOW and skip-gram models architecture in word2vec

Cosine Similarity and Pearson Coefficients

In this study, correlation coefficients are defined by the cosine similarity values between the two 200-dimensional word vectors, rather than the Pearson correlation coefficients (r). The word2vec model can be calculated using the value of word vectors obtained using the cosine similarity equation, which is the calculation of the similarity between two n-dimensional vectors by looking for a cosine value from the angle between the two and is often used to compare documents in text mining. For Pearson correlation coefficients, 0.5 is a threshold value for the weak correlation and 0.8 is generally considered for strong correlation. For example, there is no correlation if $r=0$; there is weak correlation if $0<r<0.5$; there is moderate correlation if $0.5<r<0.8$; there is strong correlation if $0.8<r<1$.[1]

First-Principles Calculations

The first-principles calculations, also known as *ab-initio* from Latin, is commonly employed in condensed matter physics, quantum chemistry and materials science to obtain the structures and properties of molecules and solids. It widely uses Hartree-Fock (HF) method and density functional theory (DFT) to solve the Schrodinger equation and obtain the physical and chemical properties of matter. As long as the wave function of the Schrodinger equation of the real multi-particle system can be solved, any properties of the particle system can be obtained. However, it should be noted that since there are too many particles in the system and the existence of cross-terms, the exact Schrodinger equation cannot be accurately solved. Therefore, the density functional theory (DFT) method based on the overall electron density as the input variable is employed to make the approximation. In a narrower sense, first-principles calculation only refers to the calculation based on DFT, in which the electron densities are solved to estimate the physical properties of the materials. The first-principles calculation is now widely used in physical sciences to generally refer to the physical property calculation using quantum mechanics:

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},t)=\hat{H}\Psi(\mathbf{r},t) \tag{1}$$

In this manuscript, the CASTEP software designed by the TCM group in the University of Cambridge is used to perform the first-principles calculations.

References

[1] D. Jatnika, M.A. Bijaksana, and A.A. Suryani, Procedia Comput. Sci. **157**, 160 (2019).