IOP Publishing Journal of Physics: Condensed Matter
J. Phys.: Condens. Matter **34** (2022) 095902 (10pp) https://doi.org/10.1088/1361-648X/ac3e1e

Prediction of solar cell materials via unsupervised literature learning

Lei Zhang[1,2,*] **and Mu He**[1]

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

E-mail: 002699@nuist.edu.cn

Received 23 August 2021, revised 18 November 2021
Accepted for publication 29 November 2021
Published 15 December 2021

Abstract

Despite the significant advancement of the data-driven studies for physical science, the textual data that are numerous in the literature are not fully embraced by the physics and materials community. In this manuscript, we successfully employ the natural language processing (NLP) technique to unsupervisedly predict the existence of solar cell types including the dye-sensitized solar cells and the perovskite solar cells based on literatures published prior to their first discovery without human annotation. Enlightened by this, we further identify possible solar cell material candidates via NLP starting with a comprehensive training database of 3.2 million paper abstracts published before 2021. The NLP model effectively predicts the existing solar cell materials, while an uncommon solar cell material namely $PtSe_2$ is suggested as an appropriate candidate for the future solar cells. Its optoelectronic properties are comprehensive investigated via first-principles calculations to reveal the decent stability and optoelectronic performance of the NLP-predicted candidate. This study demonstrates the viability of the textual data for the data-driven materials prediction and highlights the NLP method as a powerful tool to reliably predict the solar cell materials.

Keywords: natural language processing, NLP, solar cell, perovskite, machine learning

S Supplementary material for this article is available online

(Some figures may appear in colour only in the online journal)

1. Introduction

Solar cells convert the energy of light into electricity via the photovoltaic effect. Silicon is the predominant material source to fabricate the solar cells, while new types of solar cell materials are emerging to complement the silicon-based counterparts, including those for the dye-sensitized solar cells (DSSCs) and perovskite solar cells [1–9]. In particular, the cutting-edge perovskite solar cells have reached power conversion efficiencies over 25.5% [10], which can potentially disrupt the silicon-based solar cell technologies. These emerging solar cells enjoy cost effectiveness, facile synthesis and decent power conversion efficiencies; yet, many issues should be addressed for their industrial deployment, including the instability and lead contamination of the solar cell materials.

The materials discovery process is typically based on the 'trial-and-error' procedure that causes significantly high costs. Apart from the ineffective 'trial-and-error' processes, new research paradigms are emerging to effectively search for appropriate materials. The data driven method is considered as the fourth paradigm that gives rise to the emergence and popularity of materials informatics, and is of central importance to realize the vision of the materials genome initiative [11].

* Author to whom any correspondence should be addressed.

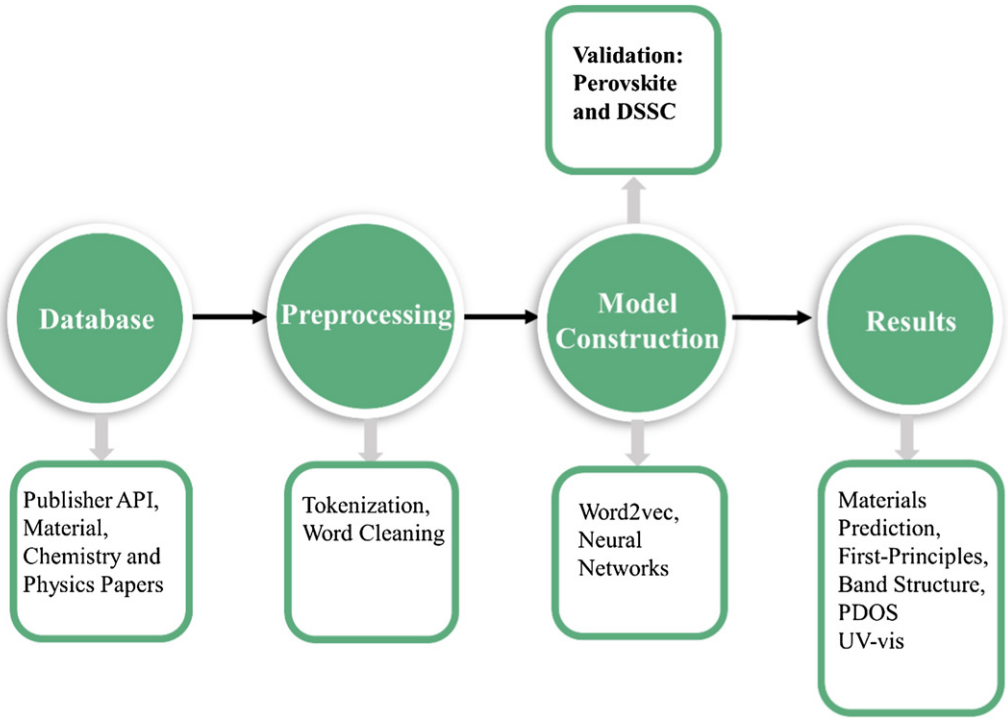

Figure 1. Flowchart of the NLP process including the database preparation, preprocessing, word2vec model construction and results generation. The predictability is validated using the perovskite solar cells and DSSCs as examples.

In the big-data era, the data-driven methods coupled with the machine learning techniques are capable of obtaining the multidimensional virtual design space of the emerging materials via the high-throughput calculations and the high-throughput experiments. While supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately, which detriments the true artificial intelligence concept.

However, compared with other data forms, the textual data are often neglected by the materials community despite of the significant advancement in the natural language processing (NLP) research area. Tshitoyan *et al* demonstrates the NLP method to analyze the materials science texts and predict new thermoelectric materials via the curation of the texts in the paper abstracts [12]. This shows the features of NLP to analyze and predict the thermoelectric materials that complement the other materials prediction methods. In addition, the NLP method has been applied for the materials synthesis plan via the collection of the synthetic texts and the specialized name entity recognition steps [13–21].

First-principles calculations based on the Schrodinger equations are suggested to be effective to evaluate the optoelectronic properties of the solar cell materials and unveil the hidden relationships. Particularly, the density-functional theory is adopted as the primary method to elucidate the atomic pictures and electronic structures of the solar cell materials and the low-dimensional functional materials [22–26]. In parallel with the high-throughput experiments, the high-throughput calculations via the first-principles calculation have been efficaciously demonstrated to design new materials *in silico*. The first-principles calculations are considered as the third paradigm for materials researches and are often combined with experiments and data-driven methods to advance the materials design process [27–29].

In this manuscript, we perform an NLP investigation on solar cell materials in an effort to predict new solar cell materials based on existing knowledge in an unsupervised manner. In this study, the solar cells materials are predicted in an unsupervised manner without any *a priori* correct answer or a teacher, which are significantly different with the traditional machine learning studies for solar cell material predictions. The reliability of the NLP process for the solar cell prediction is evidenced by the successful predictions of DSSCs and perovskite solar cells using the literatures published before their first appearance (1991 for DSSC and 2009 for perovskite solar cell, respectively). The DSSCs and the perovskite solar cells are identified by the NLP technique using the historic textual data published 10 years before their initial reports. Enlightened by this, the materials literatures published before 2021 are collected, which points to many common solar cell materials as well as an uncommon solar cell material $PtSe_2$. The first-principles calculations are performed and the simulated results of the new candidate exhibit appropriate solar cell materials features. As a result, the present study suggests NLP as a reliable tool to predict new solar cell materials for future photovoltaic devices.

2. Experimental details

A dataset consisting of 3.2 million paper abstracts is prepared via the Springer Nature API to train the word embedding model. The literature abstracts include the physics, chemistry

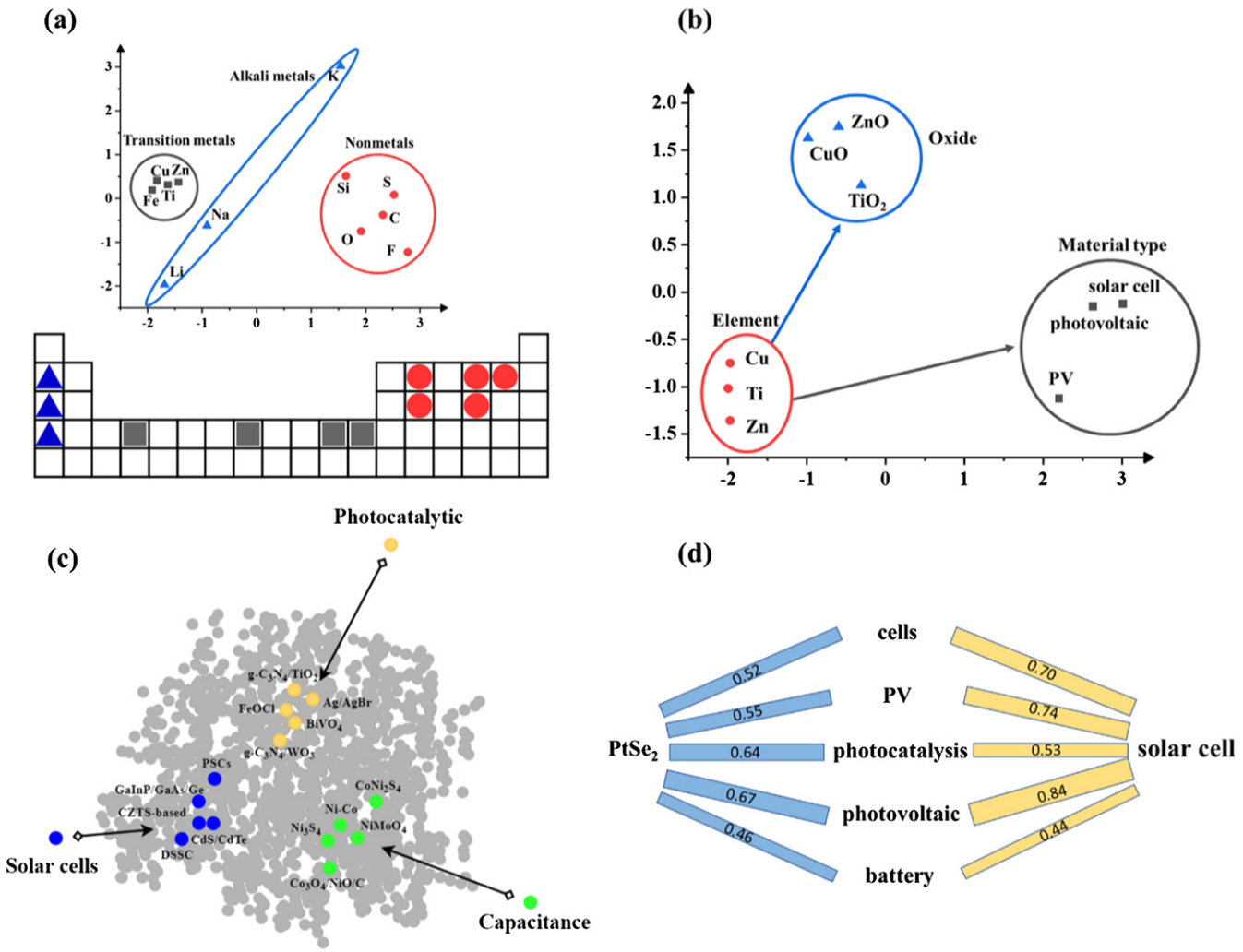

Figure 2. Hidden knowledge in materials science extracted by the present unsupervised machine learning model with minimal human intervention, which demonstrates the significance of the unsupervised machine learning model. (a) Periodic table is automatically constructed via the NLP process without any human intervention. (b) The unsupervised machine learning model uncovers the patterns of elemental materials, their oxides and their applications. (c) The machine learning model successfully clusters different materials into their respective regions such as those for solar cells, photocatalysts and capacitance-related applications. (d) The relationships between the materials (such as $PtSe_2$) and their applications (such as solar cells) can be established via the consideration of the bridging words (such as photocatalysts and battery).

and materials topics from 1960 to 2020. The text preprocessing is performed to delete the punctuation and meaningless labels and achieve a more structured database. In order to validate the NLP process for the solar cell materials, the abstracts from 1981 to 1990 are selected to assess the predictability of the word2vec method for the DSSC materials, and the abstracts from 1999 to 2008 are selected to predict the perovskite solar cell materials to check the validity of the NLP process. The chemical formulas and materials names are screened from the output file using ChemDataExtractor [30], and the new solar cell materials candidates are generated according to the word vector relationships between the chemical formula and the 'solar cell' output. The word2vec model employs the following hyperparameters: the minimum word frequency is 2 (min_count = 2); the vector dimension is 100 (vector_size = 100); the word window size is set to 5 (window = 5); the training algorithm is skip gram; the configuration threshold of the random downsampling of the high-frequency words is 1×10^{-4} (sample = 1×10^{-4}). The first principles calculations are performed in CASTEP [31], using PBE functional and 440 eV cutoff energy. The dispersion correction is considered via the Tkatchenko–Scheffler (TS) scheme [32]. A 10 Å vacuum layer is inserted in the unit cell to avoid unnecessary interlayer interactions for the two-dimensional (2D) material. The convergence criteria for the geometrical optimization is 1.0×10^6 eV for the energy, 0.03 eV Å^{-1} for the force and 0.002 Å for the displacement. The *k*-point for the properties calculations of the predicted material is $8 \times 8 \times 2$. The first-principles molecular dynamics calculations adopt the NVT assemble at the temperature of 298 K and a time step of 1 fs. The overall NLP process including the database preparation, preprocessing, word2vec model construction and results generation is depicted in figure 1.

In order to better illustrate the ideal of the unsupervised machine learning for materials science, figure 2 is presented to demonstrate the true significant of the unsupervised machine

learning process, and the scientific concepts in materials science are extracted in an unsupervised way (without deliberate annotation by human). With the literature abstracts as the only inputs, the unsupervised machine learning model successfully construct the hidden knowledge in materials and chemical science. For example, the periodic table is automatically constructed (figure 2(a)) via the NLP process without any human intervention. In addition, the unsupervised machine learning model uncovers the patterns of elemental materials, their oxides and their applications (figure 2(b)). Also, the model cluster different materials for solar cells, photocatalysts and capacitance-related applications into respective regions (figure 2(c)). Last but not least, the relationships between the materials (such as $PtSe_2$) and their applications (such as solar cell) can be established via the consideration of the bridging words (such as photocatalysts and battery) (figure 2(d)) in the unsupervised machine learning step. Consequently, the model is 'intelligent' enough to automatically formulate materials science concepts without human intervention.

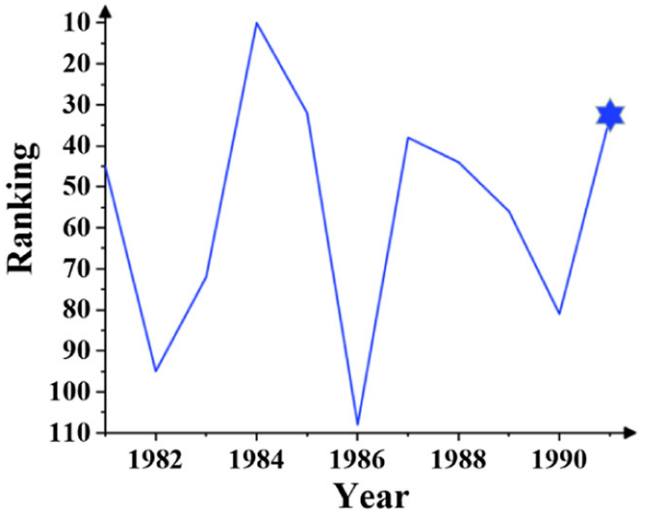

Figure 3. Visualization of the successful prediction of dye solar cell using NLP based on papers published in each year from 1981 to 1990. The star corresponds to the benchmark report of the dye-sensitized solar cell in 1991.

3. Results and discussion

3.1. Prediction of dye solar cell

The NLP method successfully extracts the dye concept and predict the existence of the dye solar cells based on the papers published before 1991 (figure 3). The dye-sensitized solar cell is reported in 1991 by O'Brian and Grätzel [33]; since then, DSSCs receive significant attentions and rapidly develop over the years, reaching efficiencies beyond 10% using organometallic dyes. In a DSSC device, the molecular dyes are responsive for the light absorption and the charge injection and play the central role in the electricity generation. The NLP-based machine learning technique will be demonstrated to outperform the human since the NLP process ingeniously captures the dye solar cell concept and unsupervisedly predicts the existence of the dye solar cell many years prior to the year 1991. For example, using only the papers published in 1981 (a decade before the discovery of DSSCs) as the inputs, the NLP method successfully predicts the existence of the dye solar cell, and exhibits the high ranking of the dye as the solar cell material (45th) according to their correlations. Over the years, the ranking of the dye for solar cell fluctuates, such as the dramatic droppings in 1982 (95th) and 1986 (108th); nevertheless, the long-lasting presence of the dye solar cell in the top 120 ranking list are indicative of the dye solar cells. In particular, the ranking of the dye for the solar cells reaches an all-time high (10th place) in 1984, signifying the importance of dyes for the solar cell applications. Apart from capturing the dye solar cell concept, the NLP-based data-driven technique can predict the specific dyes. For example, the N3 dye, which is a widely accepted DSSC-active ruthenium bipyridine dye with the carboxylic acid anchor, ranks 72nd in 1983 among all the suitable solar cell material candidates after the unsupervised learning. Apart from the dyes, the materials with higher rankings are common solar cells materials such as TiO_2, Si, Ge, ZnO and CdS, as well as the common elements and ingredients that constitute the solar cell materials (figure 4 and table S1 (https://stacks.iop.org/JPCM/34/095902/mmedia)). For example, silicon ranks 1st in 1984 and TiO_2 ranks 1st in 1985, and thus it is natural to combine the N3 dye and TiO_2 to form the new type of solar cell because of the high rankings of the two materials. To sum up, the NLP process successfully proposes the existence of dyes solar cells a decade before 1991, demonstrating the NLP-enabled competitive intelligence of the machine to the human.

3.2. Prediction of perovskite solar cell

The perovskite solar cell is first reported by Kojima and Miyasaka in 2009, and has been the research hotspot in the past decade with the expectation to potentially disrupt the silicon industry. The NLP method is also attempted on the perovskite solar cells to verify the applicability for solar cell materials. Same as the DSSC case, the NLP process successfully predicts the existence of perovskite solar cells using literatures published before 2009. For example, based on the papers published in 1999 as the inputs, the perovskite material exhibits the highest ranking (9th). Over the years, its ranking fluctuates over the years with a lower ranking (70th) in 2002 (i.e., papers published in 2002 are used as the only inputs) and another high ranking (21st) is achieved in 2004 (i.e., papers published in 2004 are used as the only inputs) (figure 5). Apart from the perovskites, the materials with the high rankings in the years from 1999 to 2008 are the mainstream solar cell materials (figure 6); for example, silicon ranks 1st in most years including 1999, 2002 and 2004. Other high-ranking materials correspond to the metal oxides and the constituent elemental species; for example, TiO_2 exhibit high rankings in 2002 (10th), 2004 (10th) and 2006 (8th), while ZnO ranks high in 2000 (8th) and 2006 (1st). Apart from that, GaN, CuO and

1981	1982	1983	1984	1985
1. Co	1. S	1. Si	1. K^+	1. TiO_2
2. Cr	2. $CaSO_4$	2. K	2. Si	2. Si
3. Ga_2Se_3	3. Cu/I	3. S	3. Oxide	3. Ni^{2+}
4. CuBr	4. Gd_2O_3	4. 2H-TaS_2	4. Na^+	4. $MgMoO_4$
5. ZnO	5. S-Sn	5. CuO	5. Cu-Cl	5. KCl
...	...	...	...	...
45. dye	95. dye	72. N3	10. dye	32. dyes

1986	1987	1988	1989	1990
1. CdS	1. Ga	1. Na^+	1. Nd	1. Ni
2. Ge	2. Mo	2. H_2O	2. Fe_2O_3	2. PbO_2
3. Ga	3. ZnO	3. GaAs/GaAlAs	3. SnO_2	3. Si
4. La_2CuO_4	4. SnO_2	4. S	4. Pb	4. Cu
5. MgO	5. $HClO_4$	5. Co	5. InSe	5. KCN
...	...	...	...	...
108.dye	38. dye	44. dye	56. dye	81. dye

Figure 4. Detailed materials outputs of the NLP process trained by the literatures published in each year from 1981 to 1990.

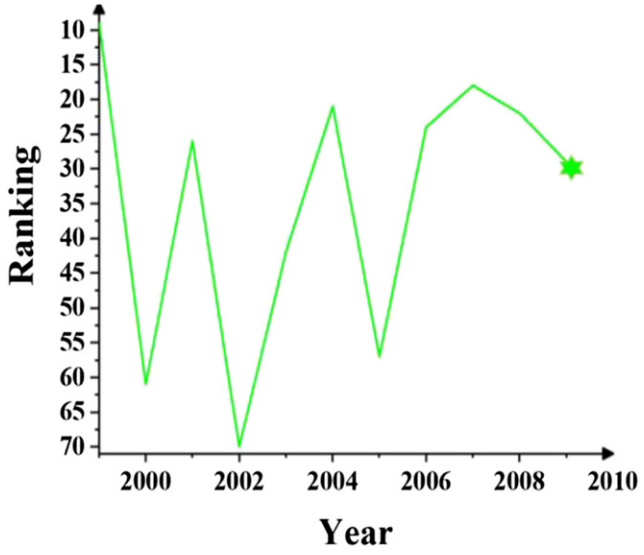

Figure 5. The evolutions of the rankings of the perovskite materials for the solar cell application via NLP from 1999 to 2009. The star represents the first report of the perovskite solar cell in 2009.

Al_2O_3 appears multiple times in the top 10 list from 1999 to 2008 (table S2).

3.3. Prediction of future solar cell material

Encouraged by the successful prediction of the DSSC and perovskite solar cell via the NLP process, we further prepare a database consisting of paper abstracts published before 2021 (1960–2020) to identify future solar cell materials. The top 5 solar cell materials predicted using the dataset include CdS, kesterite (CZTS), dye-sensitized, c-Si and ZnO, which are common solar cell materials that have been widely reported in the literature. In fact, all the top 100 materials based on the large dataset correspond to the existing common solar cell materials or common ingredients inside the device (table S3). Nevertheless, an uncommon new material, $PtSe_2$, ranking 168th (figure 7), is suggested as a possible new solar cell material that has not been received sufficient attention. As a result, the optoelectronic properties of $PtSe_2$ are explored via the first-principles calculations.

3.4. First-principles calculations on predicted solar cell material

The NLP-predicted sample is simulated via the first-principles calculations to evaluate the optoelectronic properties. The 2D system is focused to examine the dimensional tailoring effects on the layered structures that are commonly adopted for the solar cell materials. The band structure demonstrates an indirect band gap of 1.36 eV for the 2D $PtSe_2$ (figure 8 and table 1). For the projected density of states (PDOS) spectra, the valence band is mainly contributed by the Se-p orbitals while the conduction band is contributed by both Pt-s and Se-p orbitals. The UV-vis absorption spectra of $PtSe_2$ exhibit decent absorptions in both UV and visible regions; the absorption peak wavelength is 350 nm while the large absorption

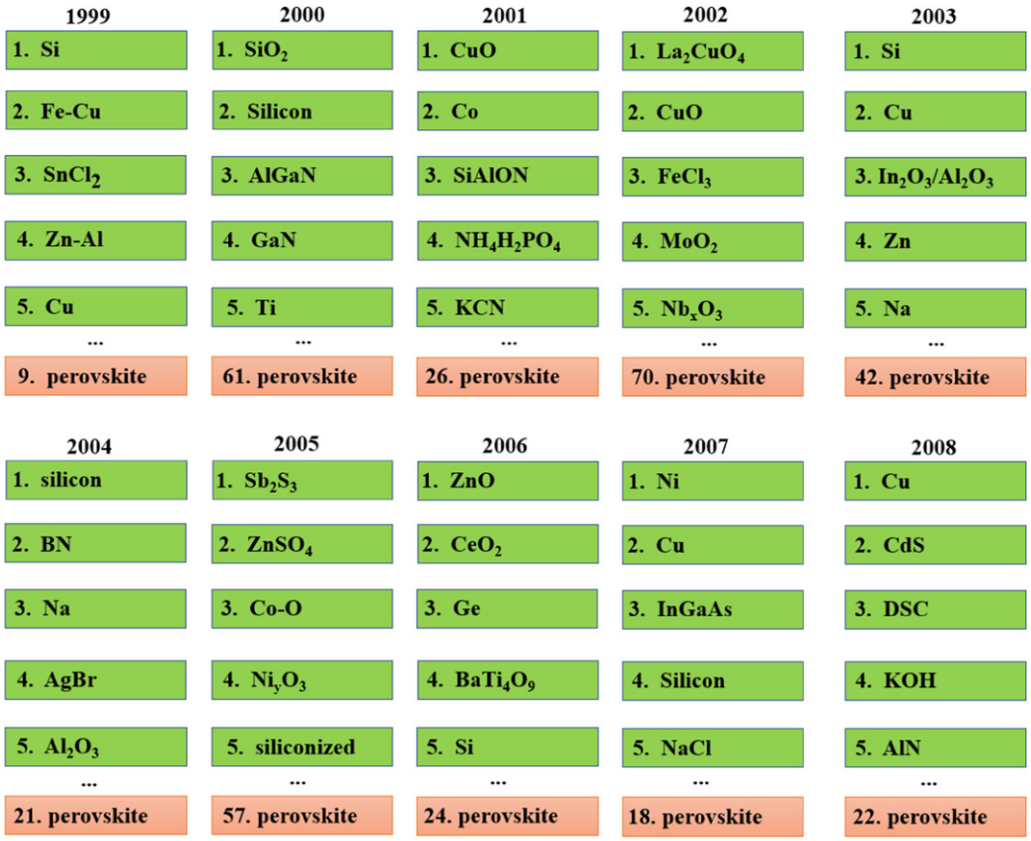

Figure 6. Detailed NLP outputs of the materials for solar cells trained by the literatures published in each year from 1999 to 2009.

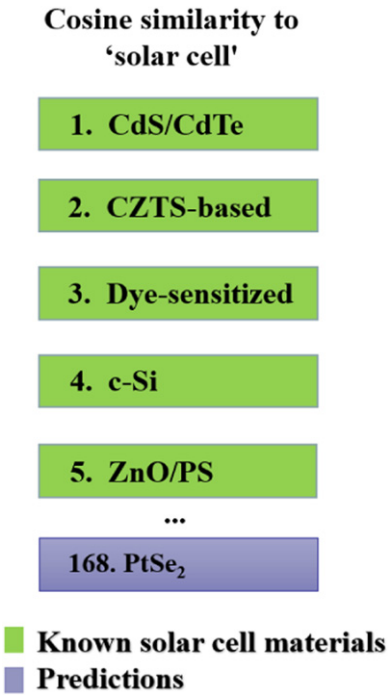

Figure 7. List of top ranking materials most related to solar cell predicted by the word2vec model.

intensity is present in the visible region (for example, the intensity is 75 000 cm^{-1} at 500 nm). The solid $PtSe_2$ demonstrates a smaller band gap of 0.86 eV and stronger light absorption in the visible region (figure S1). Nevertheless, such small band gap that may cause severe charge recombination, and we suggest that the 2D $PtSe_2$ is more valid as the solar cell materials compared with the bulk counterpart. Nevertheless, the simulated properties of the bulk three dimensional (3D) $PtSe_2$ (enjoying more valid structures but suffering from slightly lower band gap) and the 2D $PtSe_2$ (suffering from thin layer but enjoying the more appropriate band gap value) suggest its suitability as the solar cell materials, while more detailed structural optimizations such as the employment of the 2D/3D configuration that combines the merits of the individual constituents can be helpful to further improve the solar cell materials performance [34]. The first-principles molecular dynamic calculations using the NVT ensemble demonstrate the stable atomic structures at the ambient condition and suggest the excellent thermal stability of the selected material in both bulk and low-dimensional forms (figure 9). The first-principles calculations suggest the 2D $PtSe_2$ as a viable stable and high-performance solar cell material and verifies the NLP prediction process.

3.5. Rationalization and suggestions on NLP for materials prediction

The present study constructs the unsupervised machine learning model and predicts the potentiality of $PtSe_2$ as the suitable solar cell material. A drawback of this study is the lack of experimental validation of this potential candidate that causes

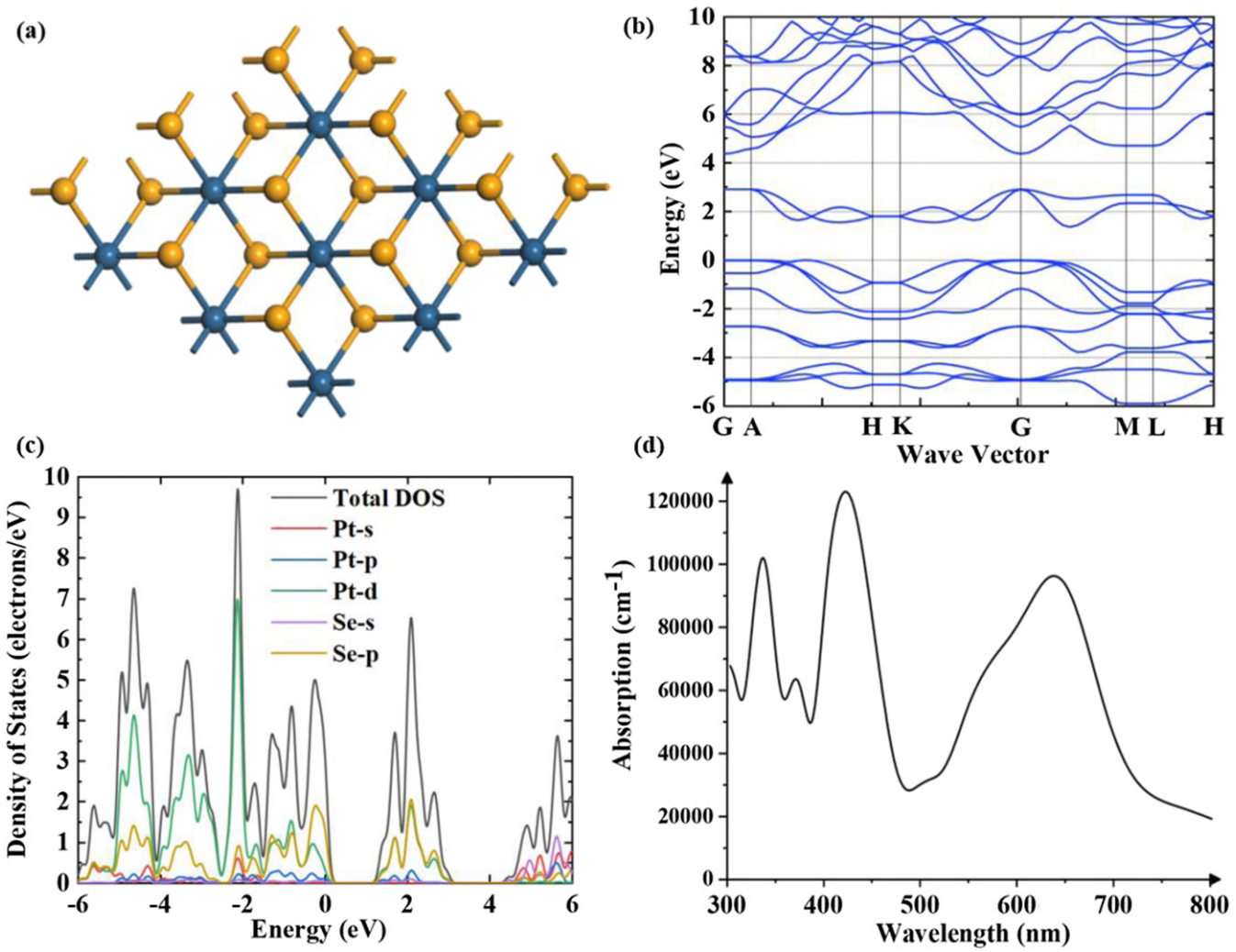

Figure 8. Atomic structure (a), band structure (b), PDOS spectra (c) and UV-vis absorption spectra (d) of the 2D single-layer $PtSe_2$. A band gap of 1.36 eV is obtained.

Table 1. Lattice constant, band gap and optical properties of the 2D $PtSe_2$ and the solid $PtSe_2$ counterpart.

Properties	2D $PtSe_2$	Solid $PtSe_2$
Lattice constant a (Å)	3.761	3.761
Lattice constant b (Å)	3.761	3.761
Lattice constant c (Å)	12.613	6.118
Lattice constant α (°)	90	90
Lattice constant β (°)	90	90
Lattice constant γ (°)	120	120
Band gap (eV)	1.362	0.681
Type of band gap	Indirect	Indirect
Peak intensity (cm^{-1})	125 000	138 000

problem for the absolute understanding of the model accuracies. However, the accuracies of the unsupervised machine learning model can be illustrated via figures 3 and 5. The concept of the DSSCs are first reported in 1991 by Grätzel *et al*, and data mining the literature published before 1991 using the traditional methods cannot develop the dye solar cell concept; however, the present unsupervised machine learning model successfully predicts the presence of dye solar cell based on the papers published before 1991. For example, using the papers published in 1981, which is ten years before the first dye solar cell report, the unsupervised machine learning model intelligently ranks the dye solar cell in the top 50 solar cell candidates list (figure 3). As a result, the present unsupervised machine learning can predict the solar cell materials accurately. The same situation happens for perovskite solar cells: the first perovskite solar cell report is published in 2009 in a serendipitous way, and it is thus expected that the normal machine learning method cannot extract the concept of perovskite solar cells using papers published before 2009. Interestingly, the unsupervised machine learning model can predict the existence of the perovskite solar cell using papers published long before 2009. To sum up, the present machine learning model quite smartly predicts the potentials of dye sensitized solar cells and perovskite solar cells only using papers published ten years before their first reports (figure 5).

Specifically, this study focuses on the potential light absorbing materials for solar cells, while the unsupervised machine learning model cannot predict the more detailed device

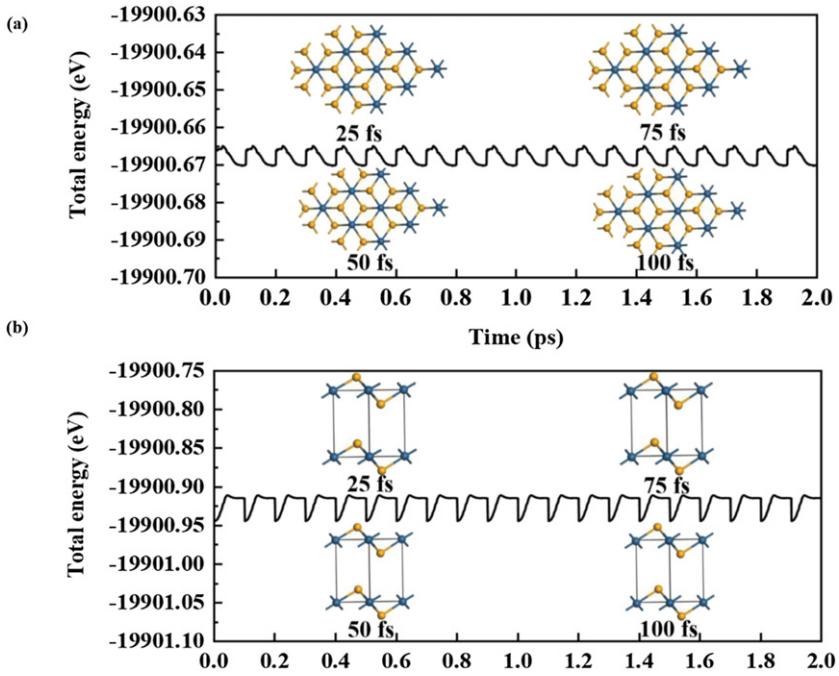

Figure 9. (a) Thermal stability of the 2D $PtSe_2$ material. (b) Thermal stability of the solid $PtSe_2$ material. The dynamic calculation is performed at 298 K.

configurations and neighboring materials at the moment, because these data requires human intervention to prepare the detailed configurations and their efficiency, which detriments the true nature of the unsupervised artificial intelligence concept. While the supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately.

Suggestions are provided to help predict new solar cell materials in the future via the NLP process. Firstly, the NLP process should be optimized to predict more details of the solar cell materials, including the possibility of defects, dimensions and interfacial engineering routes to further identify the materials more accurately and comprehensively. Secondly, the NLP technique should be employed to predict the fabrication and design routes of the material candidate, since the text-based fabrication details in the paper is exclusively and elaborately described in the paragraphs. Particularly, the NLP process is expected be more powerful to predict the detailed fabrication routes of the solar cell materials such as the perovskites including the material precursors, the ingredients and the detailed actions to obtain the candidate. Thirdly, the combination of materials should be predicted by the NLP process, since the proper function of the solar cells relies on the careful selection of a series of materials rather than a single one. For example, the perovskite solar cell should incorporate the appropriate electron-transporting materials and the hole-transporting materials in addition to the central perovskite layer. It is advised that the sum of the vectors of the materials names should be calculated to include the phrases rather than the single word and help examine the relationships between multiples words and the target word. Fourthly, a series of recently developed BERT-derived word embedding techniques should be employed to improve the accuracy of the NLP process for materials science benefited by the better handling of the contexts. Last but not least, the multi-lingual textual data should be considered to enrich the textual database and improve the accuracy of the materials prediction process. The successful predictions of the dye solar cell and the perovskite solar cell demonstrate the viability of NLP as a reliable tool to predict the promising solar cell materials in the future. We rationalize the effectiveness of NLP for the materials prediction via the applicability of these potential solar cell materials to both optoelectronic and photocatalytic applications before their first solar cell report. For example, the dyes have been well-demonstrated in the water-splitting and photocatalytic papers before 1991 and the perovskites are serving as the photo-catalytic materials before 2009. High similarities between these materials and the solar cells are present, which can be extracted via the NLP process. As a result, the NLP-based machine learning technique outperforms the human due to its capability to predict the existence of the dye solar cell materials many years prior to their adsorption by human.

4. Conclusions

The word embedding-based NLP technique successfully predicts the existence of dye-sensitized solar cell and perovskite solar cell based on the training database consisting of papers published prior to their first reports. Encouraged by this, we apply the NLP method to analyze the 3.2 million literatures published before 2021 and identify possible solar cell materials in the future. A possible solar cell material $PtSe_2$ is predicted via the NLP process, and the atomic structures, electronic and optical properties of $PtSe_2$ are closely examined

via the first-principles calculations, suggesting the excellent stability and optoelectronic properties of the 2D $PtSe_2$ material and the efficacy of the NLP technique for the solar cell material prediction. The present study calls for extensive employment of NLP to predict new functional materials and optoelectronic materials that can be tailored for particular applications.

Acknowledgments

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

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Author contributions

The author has given approval to the final version of the manuscript.

References

[1] Green M A, Ho-Baillie A and Snaith H J 2014 The emergence of perovskite solar cells *Nat. Photon.* **8** 506–14
[2] Qiu L, He S, Ono L K and Qi Y 2020 Progress of surface science studies on ABX 3-based metal halide perovskite solar cells *Adv. Energy Mater.* **10** 1902726
[3] Zhang H, Wu Y, Shen C, Li E, Yan C, Zhang W, Tian H, Han L and Zhu W H 2019 Efficient and stable chemical passivation on perovskite surface via bidentate anchoring *Adv. Energy Mater.* **9** 1803573
[4] Li H *et al* 2020 Intermolecular $\pi-\pi$ conjugation self-assembly to stabilize surface passivation of highly efficient perovskite solar cells *Adv. Mater.* **32** 1907396
[5] Zimmermann I, Gratia P, Martineau D, Grancini G, Audinot J-N, Wirtz T and Nazeeruddin M K 2019 Improved efficiency and reduced hysteresis in ultra-stable fully printable mesoscopic perovskite solar cells through incorporation of CuSCN into the perovskite layer *J. Mater. Chem.* A **7** 8073–7
[6] Liu M, Johnston M B and Snaith H J 2013 Efficient planar heterojunction perovskite solar cells by vapour deposition *Nature* **501** 395–8
[7] Daus A, Roldán-Carmona C, Domanski K, Knobelspies S, Cantarella G, Vogt C, Grätzel M, Nazeeruddin M K and Tröster G 2018 Metal-halide perovskites for gate dielectrics in field-effect transistors and photodetectors enabled by PMMA lift-off process *Adv. Mater.* **30** 1707412
[8] Khan U, Zhinong Y, Khan A A, Zulfiqar A and Khan Q U 2019 Organic-inorganic hybrid perovskites based on methylamine lead halide solar cell *Sol. Energy* **189** 421–5
[9] Khan U, Zhinong Y, Khan A A, Zulfiqar A and Ullah N 2019 High-performance $CsPbI_2Br$ perovskite solar cells with zinc and manganese doping *Nanoscale Res. Lett.* **14** 116
[10] Min H *et al* 2021 Perovskite solar cells with atomically coherent interlayers on SnO_2 electrodes *Nature* **598** 444–50
[11] Agrawal A and Choudhary A 2016 Perspective: materials informatics and big data: realization of the 'fourth paradigm' of science in materials science *APL Mater.* **4** 053208
[12] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 Unsupervised word embeddings capture latent knowledge from materials science literature *Nature* **571** 95–8
[13] Toyao T, Maeno Z, Takakusagi S, Kamachi T, Takigawa I and Shimizu K-i 2020 Machine learning for catalysis informatics: recent applications and prospects *ACS Catal.* **10** 2260–97
[14] Guo J, Ibanez-Lopez A S, Gao H, Quach V, Coley C W, Jensen K F and Barzilay R 2021 Automated chemical reaction extraction from scientific literature *J. Chem. Inf. Model.* **61** 4124
[15] Zhou X, Nurkowski D, Mosbach S, Akroyd J and Kraft M 2021 Question answering system for chemistry *J. Chem. Inf. Model.* **61** 3868–80
[16] Wilbraham L, Mehr S H M and Cronin L 2021 Digitizing chemistry using the chemical processing unit: from synthesis to discovery *Acc. Chem. Res.* **54** 253–62
[17] Gao W and Coley C W 2020 The synthesizability of molecules proposed by generative models *J. Chem. Inf. Model.* **60** 5714–23
[18] Ji B, Li S, Yu J, Ma J, Tang J, Wu Q, Tan Y, Liu H and Ji Y 2020 Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models *J. Biomed. Inf.* **104** 103395
[19] Jaeger S, Fulle S and Turk S 2018 Mol2vec: unsupervised machine learning approach with chemical intuition *J. Chem. Inf. Model.* **58** 27–35
[20] Zheng S, Yan X, Yang Y and Xu J 2019 Identifying structure-property relationships through SMILES syntax analysis with self-attention mechanism *J. Chem. Inf. Model.* **59** 914–23
[21] He T, Sun W, Huo H, Kononova O, Rong Z, Tshitoyan V, Botari T and Ceder G 2020 Similarity of precursors in solid-state synthesis as text-mined from scientific literature *Chem. Mater.* **32** 7861–73
[22] Grancini G *et al* 2017 One-year stable perovskite solar cells by 2D/3D interface engineering *Nat. Commun.* **8** 15684
[23] Boyd P G, Lee Y and Smit B 2017 Computational development of the nanoporous materials genome *Nat. Rev. Mater.* **2** 17037
[24] Filip M R, Eperon G E, Snaith H J and Giustino F 2014 Steric engineering of metal-halide perovskites with tunable optical band gaps *Nat. Commun.* **5** 5757
[25] Schouwink P, Ley M B, Tissot A, Hagemann H, Jensen T R, Smrčok L and Černý R 2014 Structure and properties of complex hydride perovskite materials *Nat. Commun.* **5** 5706
[26] Pham T A, Ping Y and Galli G 2017 Modelling heterogeneous interfaces for solar water splitting *Nat. Mater.* **16** 401–8
[27] Choudhary K, Zhang Q, Reid A C E, Chowdhury S, Van Nguyen N, Trautt Z, Newrock M W, Congo F Y and Tavazza F 2018 Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms *Sci. Data* **5** 180082
[28] Chen X, Hou T, Persson K A and Zhang Q 2019 Combining theory and experiment in lithium-sulfur batteries: current progress and future perspectives *Mater. Today* **22** 142–58
[29] Zhang L, He M and Shao S 2020 Machine learning for halide perovskite materials *Nano Energy* **78** 105380
[30] Swain M C and Cole J M 2016 ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature *J. Chem. Inf. Model.* **56** 1894–904

[31] Segall M D, Lindan P J D, Probert M J, Pickard C J, Hasnip P J, Clark S J and Payne M C 2002 First-principles simulation: ideas, illustrations and the CASTEP code *J. Phys.: Condens. Matter* **14** 2717–44

[32] Tkatchenko A and Scheffler M 2009 Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data *Phys. Rev. Lett.* **102** 073005

[33] O'Regan B and Grätzel M 1991 A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films *Nature* **353** 737–40

[34] Wang Z, Lin Q, Chmiel F P, Sakai N, Herz L M and Snaith H J 2017 Efficient ambient-air-stable solar cells with 2D–3D heterostructured butylammonium-caesium-formamidinium lead halide perovskites *Nat. Energy* **2** 17135