



UNSW  
SYDNEY



NCI  
AUSTRALIA



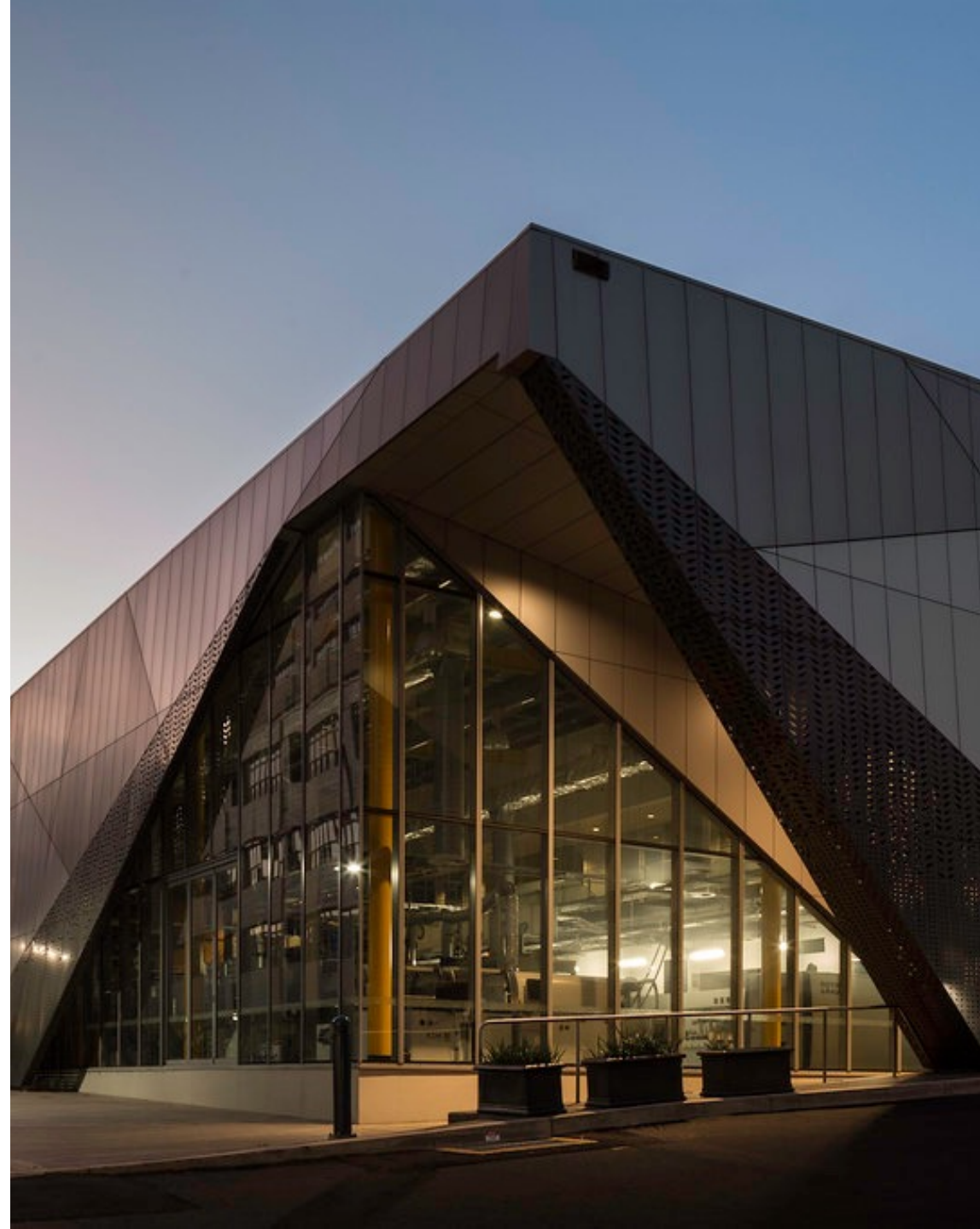
MASTER.AI  
UNSW AI4Science

# Large Language Model as a master key: Unlock the potential of material science

University of New South Wales  
School of Photovoltaics and Renewable Energy

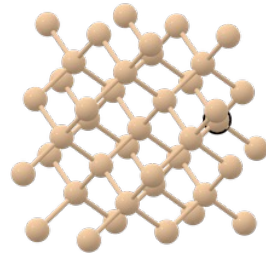
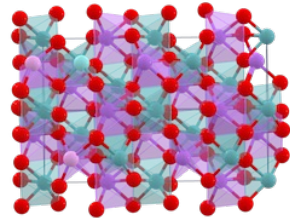
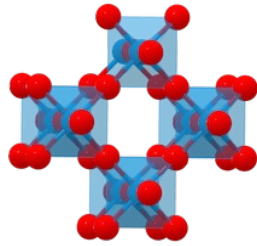
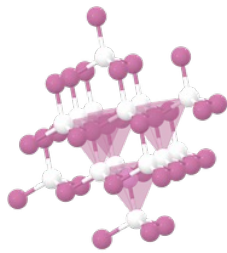
Tong Xie

2023.06



# Problem : Material Design

## Massive candidates



Periodic Table of the Elements

Number		Symbol		Name		Mass				
1	H	Hydrogen	1.008	2	He	Helium	4.003			
3	Li	Lithium	6.941	4	Be	Beryllium	9.012			
11	Na	Sodium	22.990	12	Mg	Magnesium	24.305			
19	K	Potassium	39.098	20	Ca	Calcium	40.078			
21	Sc	Scandium	44.956	22	Ti	Titanium	47.867			
23	V	Vanadium	50.942	24	Cr	Chromium	51.996			
25	Mn	Manganese	54.938	26	Fe	Iron	55.845			
27	Co	Cobalt	58.933	28	Ni	Nickel	58.693			
29	Cu	Copper	63.546	30	Zn	Zinc	65.38			
31	Ga	Gallium	69.723	32	Ge	Germanium	72.631			
33	As	Arsenic	74.922	34	Se	Selenium	78.971			
35	Br	Bromine	79.904	36	Kr	Krypton	83.798			
37	Rb	Rubidium	85.468	38	Sr	Strontium	87.62			
39	Y	Yttrium	88.906	40	Zr	Zirconium	91.224			
41	Nb	Niobium	92.906	42	Mo	Molybdenum	95.95			
43	Tc	Technetium	98.907	44	Ru	Ruthenium	101.07			
45	Rh	Rhodium	102.906	46	Pd	Palladium	106.42			
47	Ag	Silver	107.868	48	Cd	Cadmium	112.414			
49	In	Indium	114.818	50	Sn	Tin	118.710			
51	Sb	Antimony	121.760	52	Te	Tellurium	127.6			
53	I	Iodine	126.904	54	Xe	Xenon	131.293			
55	Cs	Cesium	132.905	56	Ba	Barium	137.328			
57-71	Lanthanide Series						72	Hf	Hafnium	178.45
73	Ta	Tantalum	180.948	74	W	Tungsten	183.84			
75	Re	Rhenium	186.207	76	Os	Osmium	190.23			
77	Ir	Iridium	192.227	78	Pt	Platinum	195.085			
79	Au	Gold	196.967	80	Hg	Mercury	200.592			
81	Tl	Thallium	204.383	82	Pb	Lead	207.2			
83	Bi	Bismuth	208.980	84	Po	Polonium	[209]			
85	At	Astatine	[209]	86	Rn	Radon	222.018			
87	Fr	Francium	223.020	88	Ra	Radium	226.025			
89-103	Actinide Series						104	Rf	Rutherfordium	[261]
105	Db	Dubnium	[262]	106	Sg	Seaborgium	[266]			
107	Bh	Bhassium	[264]	108	Hs	Hassium	[269]			
109	Mt	Mitlerium	[278]	110	Ds	Darmstadtium	[281]			
111	Rg	Röntgenium	[280]	112	Cn	Copernicium	[285]			
113	Nh	Nihonium	[286]	114	Fl	Flerovium	[289]			
115	Mc	Moscovium	[289]	116	Lv	Livermorium	[293]			
117	Ts	Tennesine	[294]	118	Og	Oganesson	[294]			
57	La	Lanthanum	138.905	58	Ce	Cerium	140.116			
59	Pr	Praseodymium	140.908	60	Nd	Neodymium	144.243			
61	Pm	Promethium	144.913	62	Sm	Samarium	150.36			
63	Eu	Europium	151.964	64	Gd	Gadolinium	157.25			
65	Tb	Terbium	158.925	66	Dy	Dysprosium	162.500			
67	Ho	Holmium	164.930	68	Er	Erbium	167.259			
69	Tm	Thulium	168.934	70	Yb	Ytterbium	173.055			
71	Lu	Lutetium	174.967	89	Ac	Actinium	227.028			
90	Th	Thorium	232.038	91	Pa	Protactinium	231.036			
92	U	Uranium	238.029	93	Np	Neptunium	237.048			
94	Pu	Plutonium	244.064	95	Am	Americium	243.061			
96	Cm	Curium	247.070	97	Bk	Berkelium	247.070			
98	Cf	Californium	251.080	99	Es	Einsteinium	[254]			
100	Fm	Fermium	257.095	101	Md	Mendelevium	258.1			
102	No	Nobelium	259.101	103	Lr	Lawrencium	[262]			

Alkali Metal   Alkaline Earth   Transition Metal   Basic Metal   Metalloid   Nonmetal   Halogen   Noble Gas   Lanthanide   Actinide

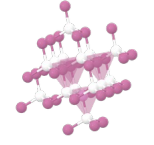
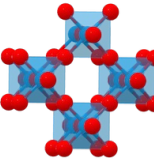
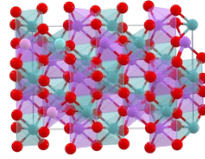
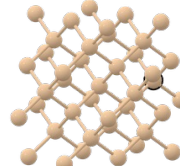
# Machine learning is suitable for processing data in tabular format

Embedding material structure or  
Design special model architecture

Need plenty of  
Specialists

Successful cases



	Material descriptor	Label								
	<table border="1"><tr><td>1</td><td>22</td><td>11.7</td><td>52</td><td>78</td><td>0</td><td>1</td><td>2</td></tr></table>	1	22	11.7	52	78	0	1	2	<u>126</u>
1	22	11.7	52	78	0	1	2			
	<table border="1"><tr><td>2</td><td>31</td><td>15</td><td>28.7</td><td>0</td><td>3</td><td>5</td><td>6</td></tr></table>	2	31	15	28.7	0	3	5	6	<u>11</u>
2	31	15	28.7	0	3	5	6			
	<table border="1"><tr><td>36.7</td><td>115</td><td>62</td><td>0</td><td>32</td><td>5</td><td>9</td><td>2</td></tr></table>	36.7	115	62	0	32	5	9	2	<u>3</u>
36.7	115	62	0	32	5	9	2			
	<table border="1"><tr><td>86</td><td>115</td><td>0</td><td>0</td><td>32</td><td>5</td><td>4</td><td>29</td></tr></table>	86	115	0	0	32	5	4	29	<u>37.8</u>
86	115	0	0	32	5	4	29			

# Why we need Large Language Model (LLM) ?

The majority of material data is in text  
Including journals, paper, lab reports and etc.

**Machine learning for material science**  
comment  
REVIEW ARTICLE  
Unsupervised distributed word embeddings enhanced with natural language processing for material science data analysis  
Jonathan Schmidt, Minjie Ye, Shihong Wang, John Dunning, Levent Erkoc, Tyler A. Pritchard, Alexander Dementyev, Ziqun Dong, Olga Kovaleva, Kristin A. Persing, Gabriela Ceder, Wolfgang E. Kerzendorf, Ferdinand Patat, Dominic Boreland, Glenn van de Ven, et al.

One of the most powerful tools that have been developed in the last few years is machine learning. In the past few years, machine learning has become a dominant force in many scientific fields, including materials science. This is due to the fact that machine learning algorithms can automatically learn from large amounts of data and make predictions about new data. In this paper, we describe a new machine learning approach for material science data analysis. This approach uses a combination of unsupervised learning and natural language processing to extract information from text-based data. The results show that this approach is able to identify patterns in the data that were not previously known. This work is a step towards the development of a more powerful machine learning tool for material science data analysis.

Wed. Nov. 20, 2013 Eric Black

Inverted Pendulum # 01

Setup:

Anchor  
buff 13  
OUT (normal)

Weight top  
13 52.86g

Wiring diagram:

- Oscilloscope TDS 3012B
- Control Module Ser. No 01
- Tektronix Signal Generator CFG 253

Rear connections (other than power)

- 01 Control Module → Ribbon cable to IP stand
- Oscilloscope → Ethernet port to CITNET 033-02-08:07

Settings: Control Module: ON

Function Generator CFG 253:

- Amplitude - MAXIMUM
- DC offset - MIN, pushed in
- Symmetry - Middle of range (notch)
- Sweep Rate - knob @ 12 o'clock
- Sweep width - " "

Frequency - 1.0  
Range - 1 Hz  
Function - V  
Volts out - 0-2(V)  
Variable - CAL (out)  
Sweep - EXT (out)

3D visualization of a material structure (possibly a protein or crystal lattice).

Property	Value
Name	control.vcf
Mass	3.8246
Charge	-2.38e-05
Number of Beads	36
Geometric Center X	-0.039971
Geometric Center Y	0.070524
Geometric Center Z	-0.088847
Mass Center X	-0.002290
Mass Center Y	-0.017807
Mass Center Z	0.015027

# Use unsupervised learning to find carrier selective material

**What is the most possible next token?**  
By learning similar texts

**Apply it on material reports**

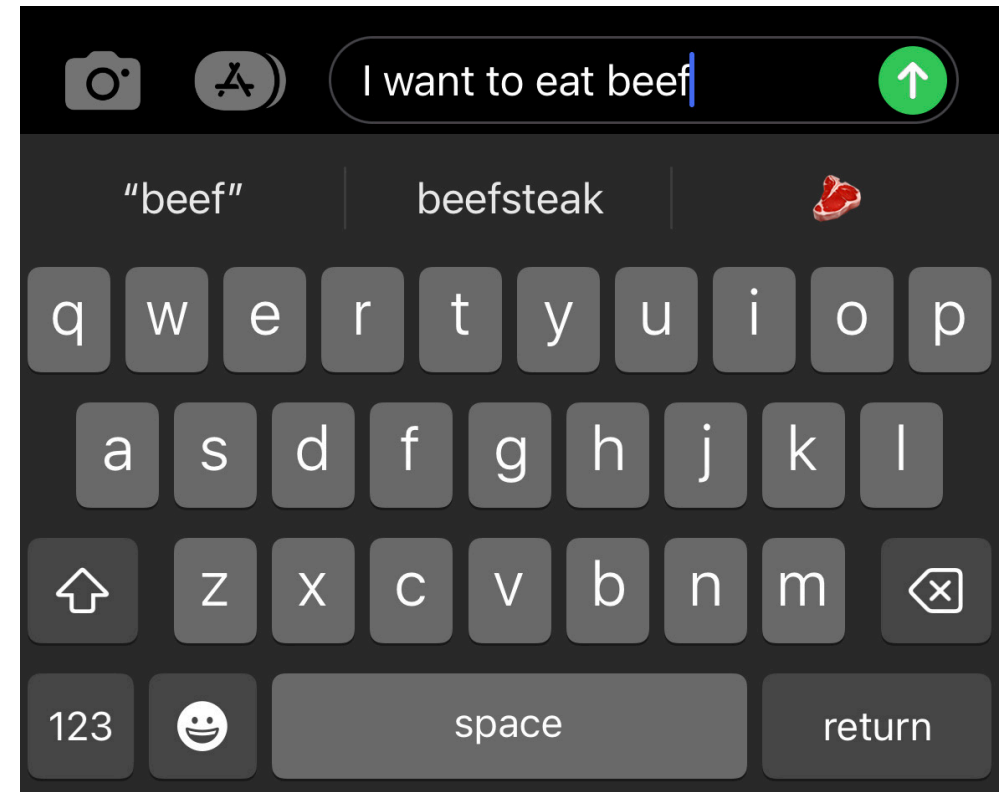
What is the most potential related material for Carrier Selective ?

Find the ground-breaking

**Electron transport layer** or **hole transport layer** material

~ 60K paper , 200 dimension

Synthesis

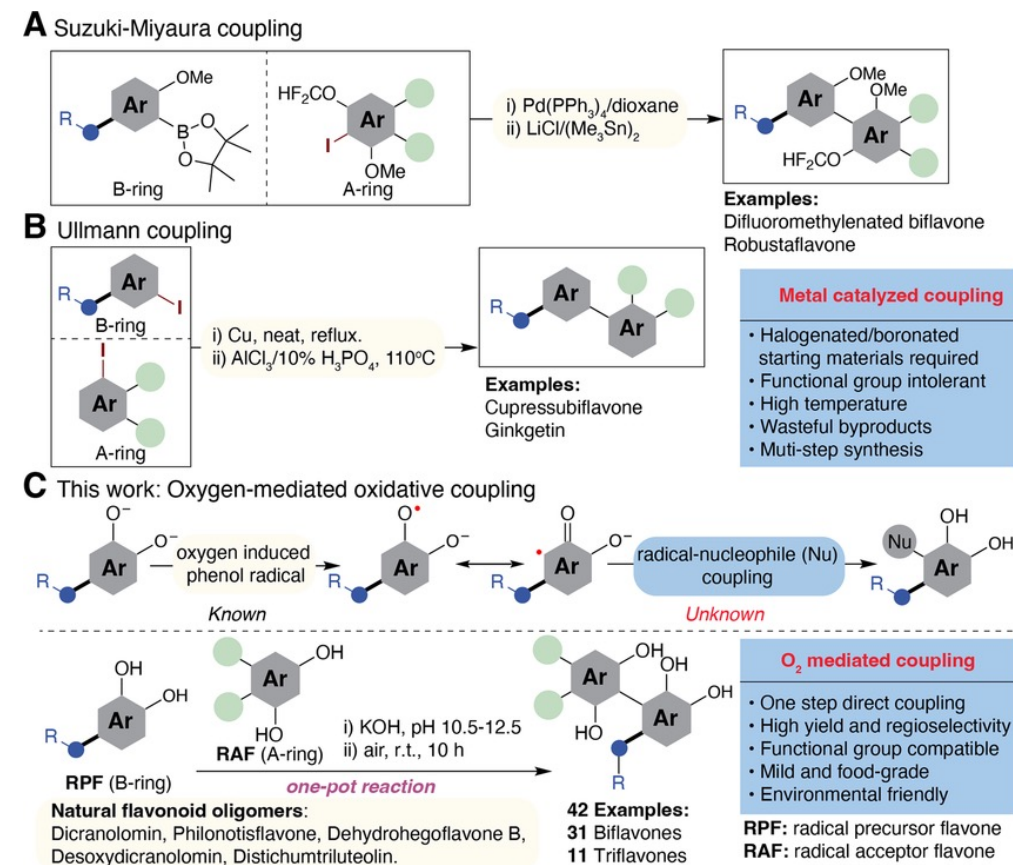


# Why we need Large Language Model ?

Data usually depends on its context

Information such as material manufacture data, synthesis steps and etc ...

It is too complex for data in table format



[1] T. Xie et al. Large Language model as master keys: Unlock the secrets of material science, Patterns, Manuscripts under review.

# Introduction

Accelerating materials science discoveries through data-driven and first-principles-based materials design

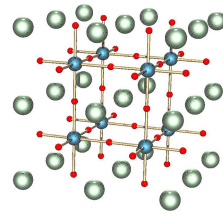
Discovering and mining materials science information based on literature and databases.

Experiment information

Material Science LLM

Understanding the regularity of material structure and properties from unstructured information.

## Material structure



Calculation: DFT, MD ...  
Experimentation: XRD, SEM...

## Material information

Compound:  
Formula, Name, Abbreviation  
Description:  
Processing method, Defects, Modifications, Morphology  
Structure/Phase:  
Crystalline structure, Symmetry, Phase

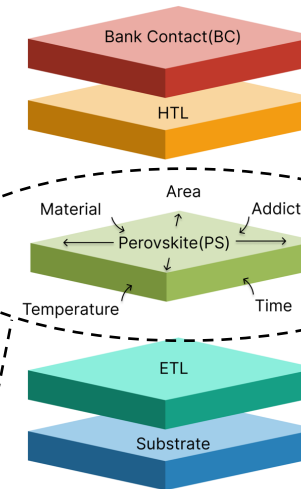
## Material properties

Bandgap

K\_VRH

Carrier mobility

...



Cell Information

JV Information

Stability Information

Device performance

## Device structure

[1] T. Xie et al. Large Language model as master keys: Unlock the secrets of material science, Patterns, Manuscripts under review.

## ❖ Customized GPT for natural science



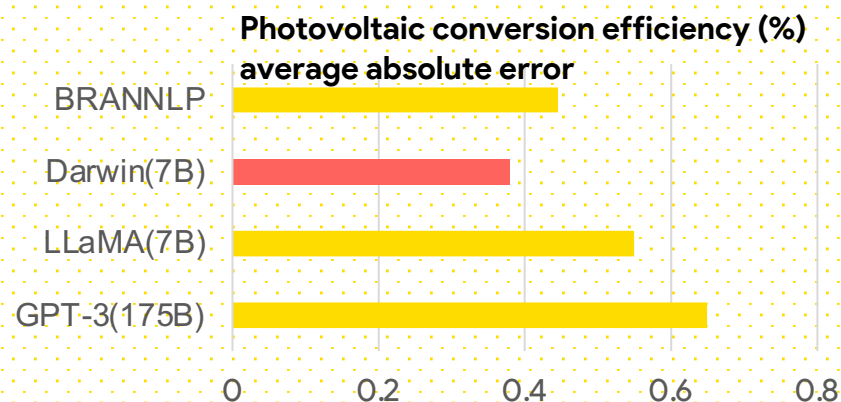
## ❖ Powerful materials design capabilities

**Darwin (7B) > Fine Tuned GPT-3 (175B)**  
**Multi-task training >> Single-task training**  
**One Darwin, All Tasks and Open Source!**

Case:





 **HOPV15**



## ❖ Better scientific answering capabilities than GPT-4


What are the key properties of graphitic carbon nitride (g-C<sub>3</sub>N<sub>4</sub>) that make it an attractive material for optoelectronic applications?

 Graphitic carbon nitride (g-C<sub>3</sub>N<sub>4</sub>) possesses a large bandgap, high thermal and chemical stability, and excellent photocatalytic properties, making it an attractive material for optoelectronic applications such as solar cells, sensors and photocatalyst,

 Graphitic carbon nitride (g-C<sub>3</sub>N<sub>4</sub>) exhibits excellent semiconductor properties, high thermal and chemical stability, and strong optical absorption in the visible light range. Its bandgap can be tuned for specific applications. Moreover, it's non-toxic, abundant, and eco-friendly, making it ideal for sustainable optoelectronic applications.

What's the bandgap of Ag<sub>2</sub>CsSb<sub>2</sub>I<sub>3</sub>?

 1.96

 As of my knowledge cutoff in September 2021, there isn't specific information available on the bandgap of Ag<sub>2</sub>CsSb<sub>2</sub>I<sub>3</sub>. The bandgap of a material is a very specific property, ..., I recommend checking the most recent scientific literature for updates or conducting an experiment if possible.





## ➤ Performance on material prediction task

**Better than State-of-The-Art model on most tasks**

### Berkeley- Matbench (Solid material)

Task	Metric	SOTA	DARWIN
Bandgap (Regression)	MAE	0.3310	<b>0.2790</b>
Metal (Classification)	ROCAUC	0.9209	<b>0.9650</b>

### Harvard - HOPV15 (Organic solar cells)

Task	Metric	SOTA	DARWIN
PCE (Regression)	MAE	0.42	<b>0.38</b>

## ➤ Performance on device design

**Automated formulation of perovskite solar cell product**

```
(base) halona@L-9H75PT3:/mnt/d/tong$ conda activate darwin
(darwin) halona@L-9H75PT3:/mnt/d/tong$ python inference.py darwin/training_2904
loading model, path: darwin/training_2904
Loading checkpoint shards: 100%
| 3/3 [00:40<00:00, 13.62s/it]
User: 
```

[1] T. Xie et al. Large Language model as master keys: Unlock the secrets of material science, Patterns, Manuscripts under review.



## ➤ Information extraction on paper structure

### 2-3 years for 50+ scientists



OPEN

An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles

T. Jesper Jacobsson<sup>1,2,3</sup>, Adam Hultqvist<sup>3</sup>, Alberto García-Fernández<sup>4</sup>, Aman Anand<sup>5,6</sup>, Amran Al-Ashouri<sup>7</sup>, Anders Hagfeldt<sup>8</sup>, Andrea Crovetto<sup>9</sup>, Antonio Abate<sup>10</sup>, Antonio Gaetano Ricciardulli<sup>11</sup>, Anuja Vijayan<sup>2</sup>, Ashish Kulkarni<sup>12</sup>, Assaf Y. Anderson<sup>13</sup>, Barbara Primera Darwich<sup>14</sup>, Bowen Yang<sup>5</sup>, Brendan L. Coles<sup>15</sup>, Carlo A. R. Perini<sup>16</sup>, Carolin Rehermann<sup>1</sup>, Daniel Ramirez<sup>17</sup>, David Fairen-Jimenez<sup>18</sup>, Diego Di Girolamo<sup>19,20</sup>, Donglin Jia<sup>21</sup>, Elena Avila<sup>18</sup>, Emilio J. Juarez-Perez<sup>22</sup>, Fanny Baumann<sup>8,23</sup>, Florian Mathies<sup>3</sup>, G. S. Anaya González<sup>24</sup>, Gerrit Boschloo<sup>2</sup>, Giuseppe Nasti<sup>19</sup>, Gopinath Paramasivam<sup>1,25</sup>, Guillermo Martínez-Denegri<sup>26</sup>, Hampus Näsström<sup>1</sup>, Hannes Michaels<sup>2</sup>, Hans Köbber<sup>10</sup>, Hua Wu<sup>2</sup>, Iacopo Benesperi<sup>2</sup>, M. Ibrahim Dar<sup>27</sup>, Ilknur Bayrak Pehlivan<sup>28</sup>, Isaac E. Gould<sup>29,30</sup>, Jacob N. Vagott<sup>16</sup>, Janardan Dagar<sup>1</sup>, Jeff Kettle<sup>31</sup>, Jie Yang<sup>32</sup>, Jinzhao Li<sup>1</sup>, Joel A. Smith<sup>33,34</sup>, Jorge Pascual<sup>10</sup>, Jose J. Jerónimo-Rendón<sup>35</sup>, Juan Felipe Montoya<sup>17</sup>, Juan-Pablo Correa-Baena<sup>16</sup>, Junming Qiu<sup>21</sup>, Junxin Wang<sup>28,36</sup>, Kári Sveinbjörnsson<sup>7</sup>, Katrin Hirslandt<sup>1</sup>, Krishanu Dey<sup>27</sup>, Kyle Frohna<sup>27</sup>, Lena Mathies<sup>37</sup>, Luigi A. Castriotta<sup>38</sup>, Mahmoud. H. Aldamasy<sup>10,39</sup>, Manuel Vasquez-Montoya<sup>117</sup>, Marco A. Ruiz-Preciado<sup>40,41</sup>, Marion A. Flatken<sup>10</sup>, Mark V. Khenkin<sup>42</sup>, Max Grischek<sup>7,43</sup>, Mayank Kedia<sup>12,35</sup>, Michael Saliba<sup>12,35</sup>, Miguel Anaya<sup>27,44</sup>, Misha Veldhoen<sup>13</sup>, Neha Arora<sup>27</sup>, Oleksandra Shargaieva<sup>1</sup>, Oliver Maus<sup>1</sup>, Onkar S. Game<sup>33</sup>, Ori Yudilevich<sup>13</sup>, Paul Fassl<sup>40,41</sup>, Qisen Zhou<sup>21</sup>, Rafael Betancur<sup>17</sup>, Rahim Munir<sup>1</sup>, Rahul Patidar<sup>15</sup>, Samuel D. Stranks<sup>27,44</sup>, Shahidul Alam<sup>5,6,45</sup>, Shaoni Kar<sup>46</sup>, Thomas Unold<sup>3</sup>, Tobias Abzieher<sup>41</sup>, Tomas Edvinsson<sup>28</sup>, Tudur Wyn David<sup>47</sup>, Ulrich W. Paetzold<sup>40,41</sup>, Waqas Zia<sup>12,35</sup>, Weifei Fu<sup>1</sup>, Weiwei Zuo<sup>35</sup>, Vincent R. F. Schröder<sup>48,49</sup>, Wolfgang Tress<sup>10</sup>, Xiaoliang Zhang<sup>21</sup>, Yu-Hsien Chiang<sup>27</sup>, Zafar Iqbal<sup>10</sup>, Zhiqiang Xie<sup>51</sup> and Eva Unger<sup>1,23,52</sup>

Large datasets are now ubiquitous as technology enables higher-throughput experiments, but rarely can a research field truly benefit from the research data generated due to inconsistent formatting, undocumented storage or improper dissemination. Here we extract all the meaningful device data from peer-reviewed papers on metal-halide perovskite solar cells published so far and make them available in a database. We collect data from over 42,400 photovoltaic devices with up to 100 parameters per device. We then develop open-source and accessible procedures to analyse the data, providing examples of insights that can be gleaned from the analysis of a large dataset. The database, graphics and analysis tools are made available to the community and will continue to evolve as an open-source initiative. This approach of extensively capturing the progress of an entire field, including sorting, interactive exploration and graphical representation of the data, will be applicable to many fields in materials science, engineering and biosciences.

## DARWIN (1 sec/one paper , Nvidia A5000)

### Perovskite solar cell FAIR dataset

< PREV

Obtain

Next >

All examples are perovskite-related OA papers published after 2021-03  
Records are exactly as our fine-tuned GPT-3 model

Knowledge frame from: [Nature Energy](#) Download GPT-3 Generated Dataset: [Link](#)

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Rui Zhu (iamzhurui@pku.edu.cn). This study did not generate new unique materials. The data that support the findings of this study are available from the corresponding author upon reasonable request. Lead iodide (PbI<sub>2</sub> 99.99%) was purchased from Tokyo Chemical Industry (TCI, Japan). Organic salts, including methylammonium bromide (MABr), methylammonium iodide (MAI), and formamidinium iodide (FAI) were purchased from GreatCell Solar (Dysole, Australia). Methylammonium chloride (MACl) was purchased from Xi'an Polymer Light Technology (China). The li-bis(trifluoromethanesulfonyl)imide (Li-TFSI) was received from Sigma-Aldrich (USA). 2,2',7,7'-Tetrakis(N,N-di-4-methoxyphenylamine)-9,9'-spirofluorene (spiro-OMeTAD) was purchased from Ningbo Borun New Material (China). Poly-4-vinylpyridine (P4VP, M w~50,000) was purchased from TCI. Poly(methyl methacrylate) (PMMA) received from commercial source. 1-butyl-3-methylimidazolium tetrafluoroborate was purchased from Sigma-Aldrich. Methylamine solution (2.0 M in tetrahydrofuran) was purchased from Sigma-Aldrich. Besides, all solvents including N,N-dimethylformamide (DMF, 99.8%), dimethyl sulfoxide (DMSO, 99.7%), methanol, and chlorobenzene (CB, 99.8%) were purchased from commercial sources (Acros) and used without further purification. In addition, acetonitrile (ACN, 99.9%) were obtained from Sigma-Aldrich (USA). Gold (Au) were received from commercial sources with high purity (≥99.99%). First, 1-butyl-3-methylimidazolium tetrafluoroborate dissolved into methanol (5, 10, 20 mg mL<sup>-1</sup>) was modified onto the pre-cleaned FTO glass in the N<sub>2</sub>-filled glove box. After annealing at 70°C for

Paper Link

doi	10.1016/j.joule.2022.04.012.json
-----	----------------------------------

Stack & Synthesis Information

Substrate_stack_sequence	SLG   FTO
ETL_stack_sequence	TiO <sub>2</sub> -c
ETL_additives_compounds	Unknown
ETL_deposition_procedure	Spin-coating
Perovskite_composition_long_form	FAPbI <sub>3</sub>
Perovskite_composition_short_form	FAPbI
Perovskite_additives_compounds	"
Perovskite_deposition_solvents	DMF; DMSO
Perovskite_deposition_procedure	Spin-coating

[1] T. Xie et al. Large Language model as master keys: Unlock the secrets of material science, Patterns, Manuscripts under review.

# Summary

Paper, patents, experimental data



## DARWIN AI Center Hub



Hi, I'm Darwin.  
I could help you with:



Hugging Face



GitHub

### Text-based work

Answer scientific questions

Provide literatures

Data collection

...

Liberate productivity

Data transformation

Device design

...

Find new material

Automatic simulation

Material prediction

Synthesis methods

Material design

...

And more application scenarios...



# Thank you!

E-mail: [tong.xie@unsw.edu.au](mailto:tong.xie@unsw.edu.au)

