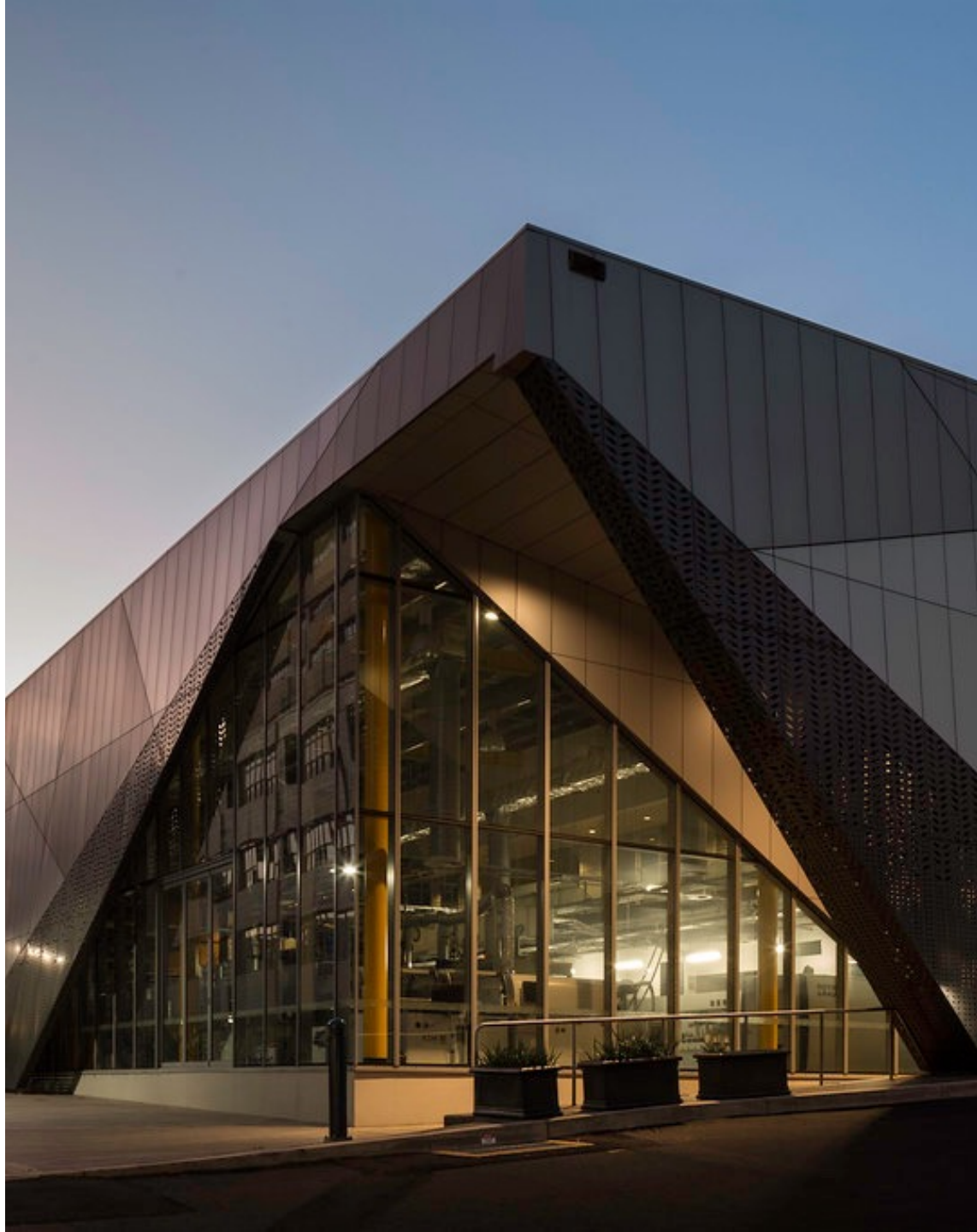
Large Language Model as a master key: Unlock the potential of material science

University of New South Wales
School of Photovoltaics and Renewable Energy

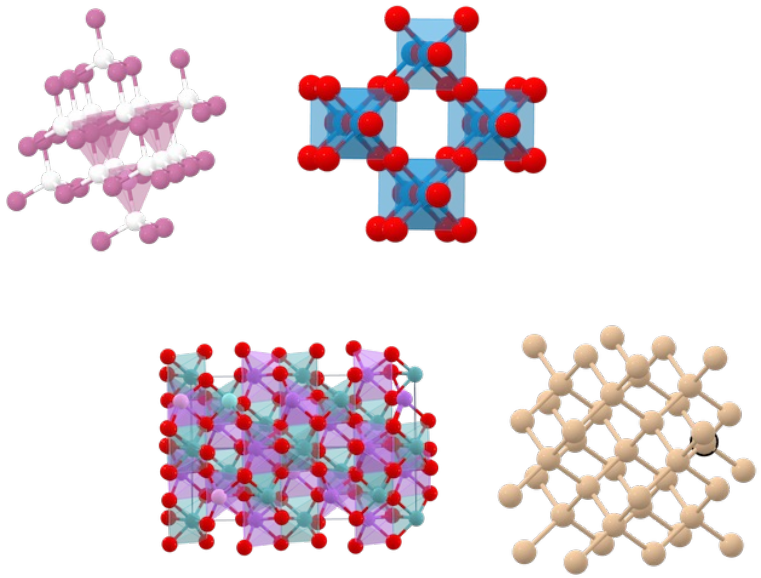
Tong Xie

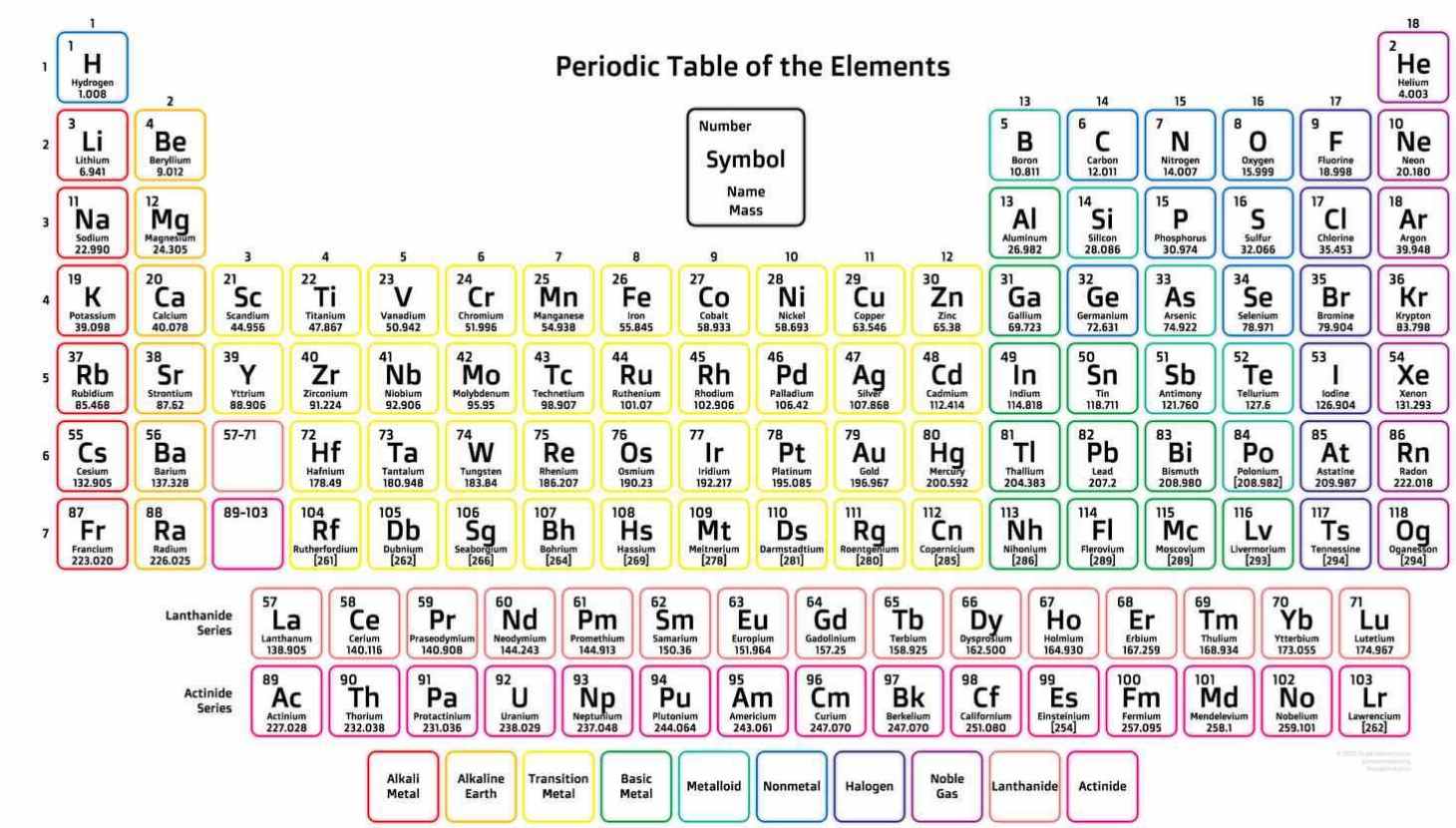
2023.06



Problem： Material Design

Massive candidates





Machine learning is suitable for processing data in tabular format

Embedding material structure or Design special model architecture

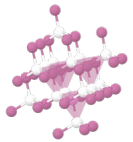
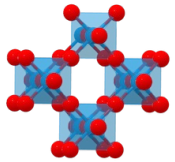
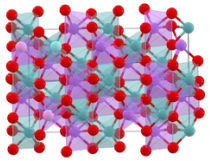
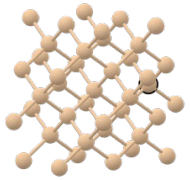
Need plenty of Specialists

Successful cases

materials virtual·lab

The Materials Project

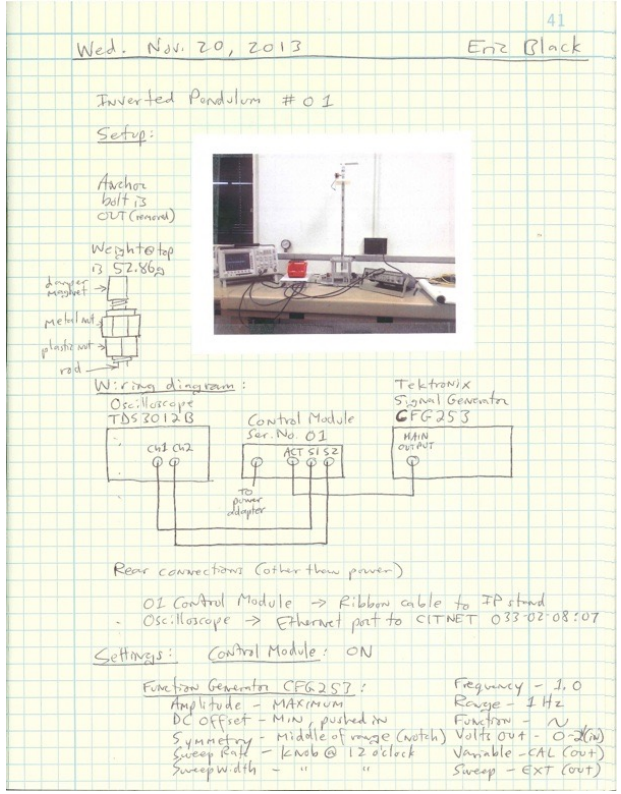
matminer

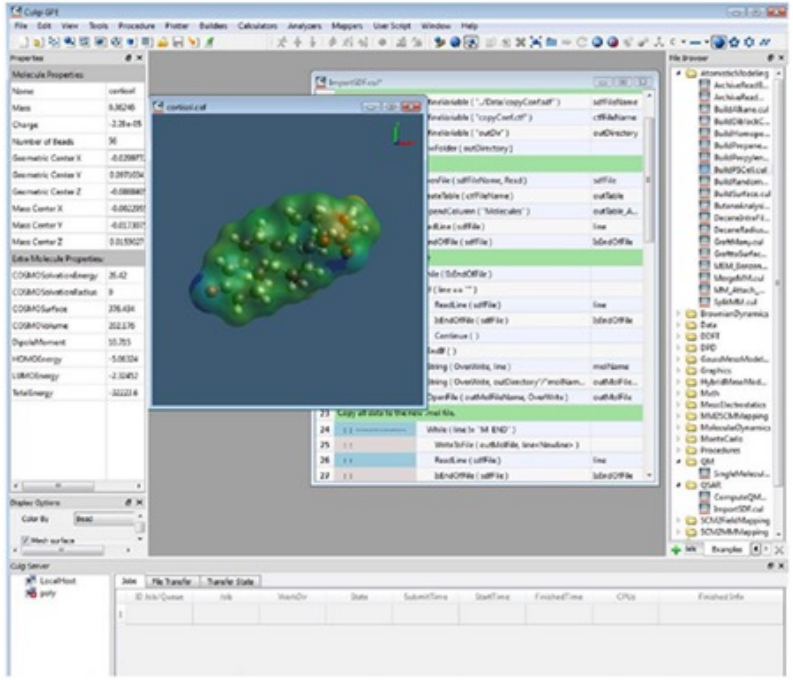
	Material descriptor								Label
	1	22	11.7	52	78	0	1	2	126
	2	31	15	28.7	0	3	5	6	11
	36.7	115	62	0	32	5	9	2	3
	86	115	0	0	32	5	4	29	37.8

Why we need Large Language Model (LLM) ?

The majority of material data is in text

Including journals, paper, lab reports and etc.





Use unsupervised learning to find carrier selective material

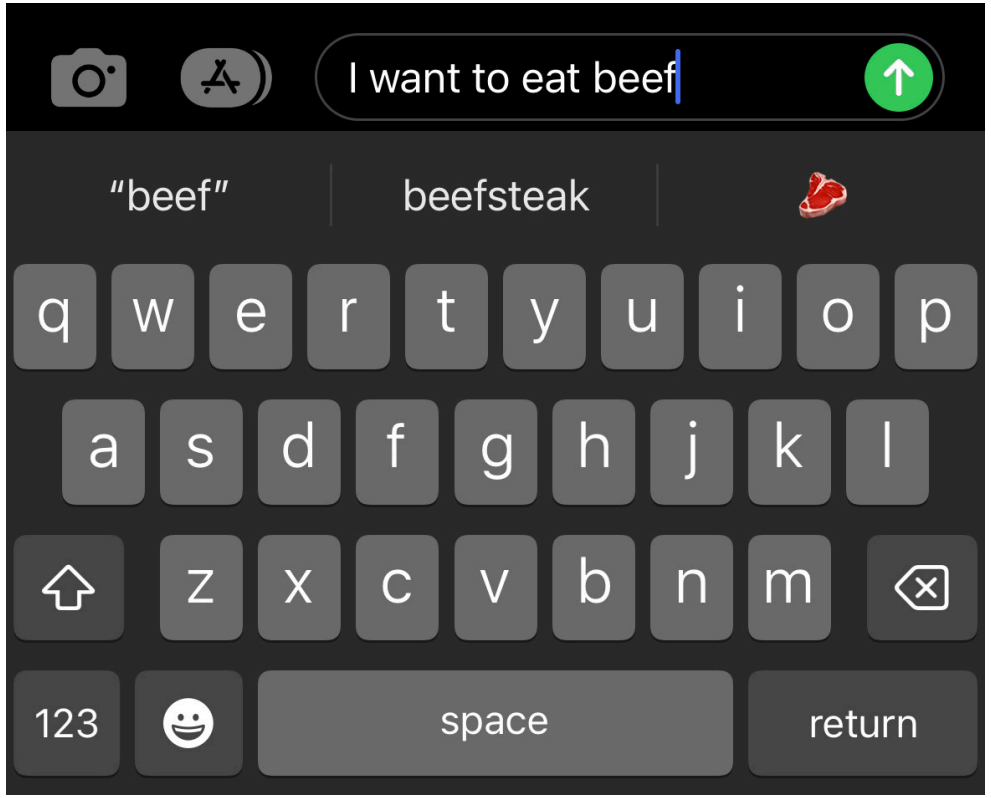
What is the most possible next token?
By learning similar texts

Apply it on material reports
What is the most potential related material for Carrier Selective ?

Find the ground-breaking
Eletron transport layer or
hole transport layer material

~ 60K paper , 200 dimension

Synthesis



[1] T. Xie et al. Deep learning accelerated material disocery for Si-based solar cells passivating contact layers material disocery & synthesis, PVSEC-33, Japan, 2022

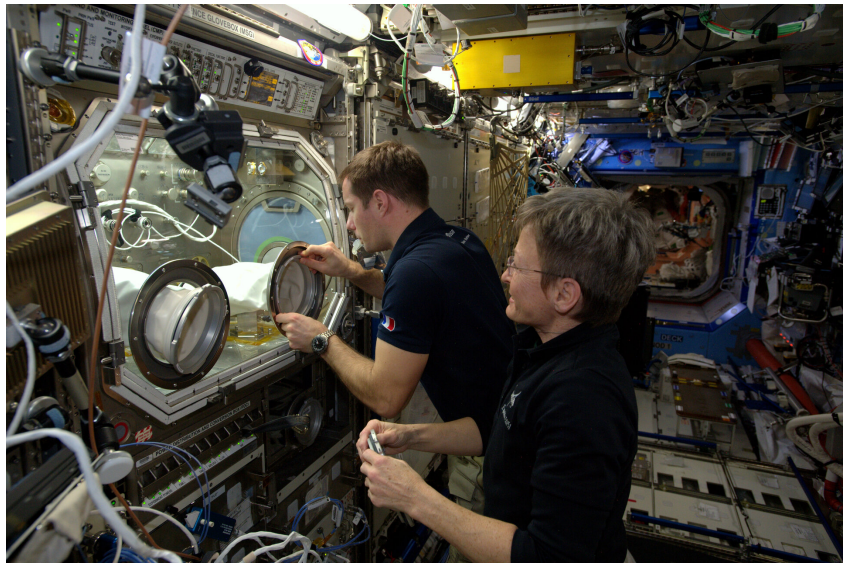
MASTER.AI
UNSW AI4Science
UNSW SYDNEY

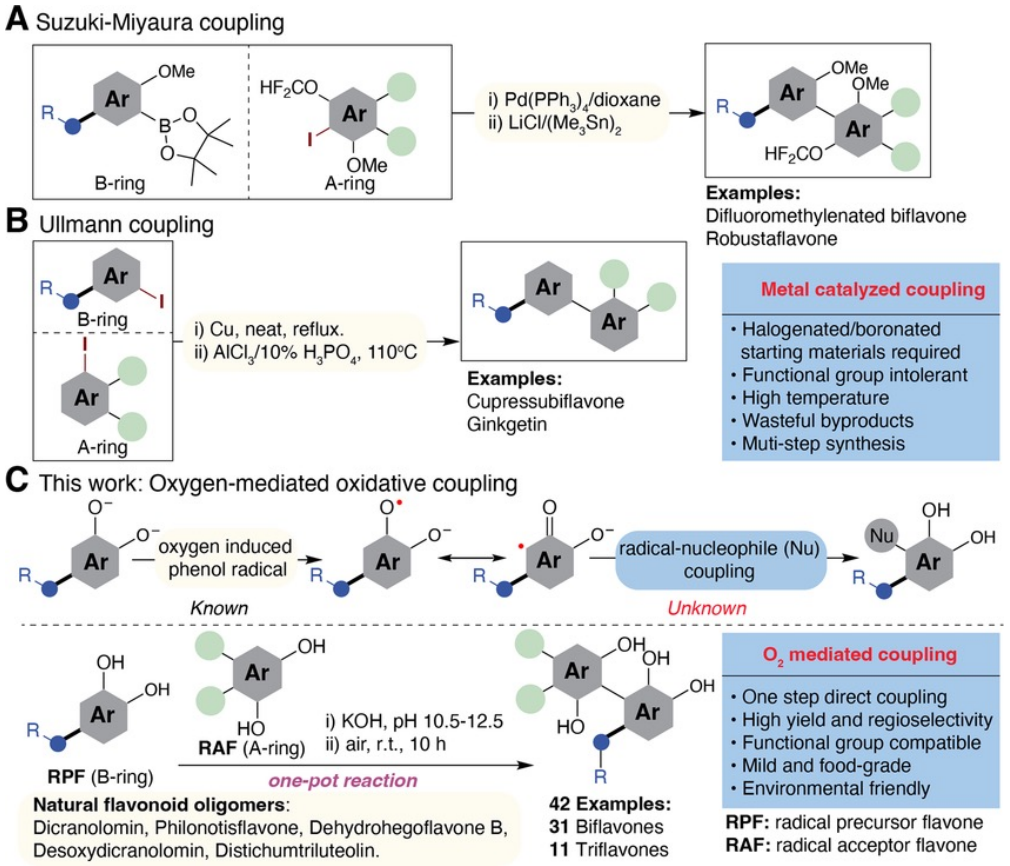
Why we need Large Language Model ?

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.

MASTER.AI UNSW AI4Science

UNSW SYDNEY

Introduction

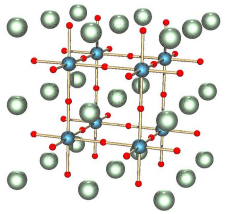
Accelerating materials science discoveries through data-driven and first-principles-based materials design

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

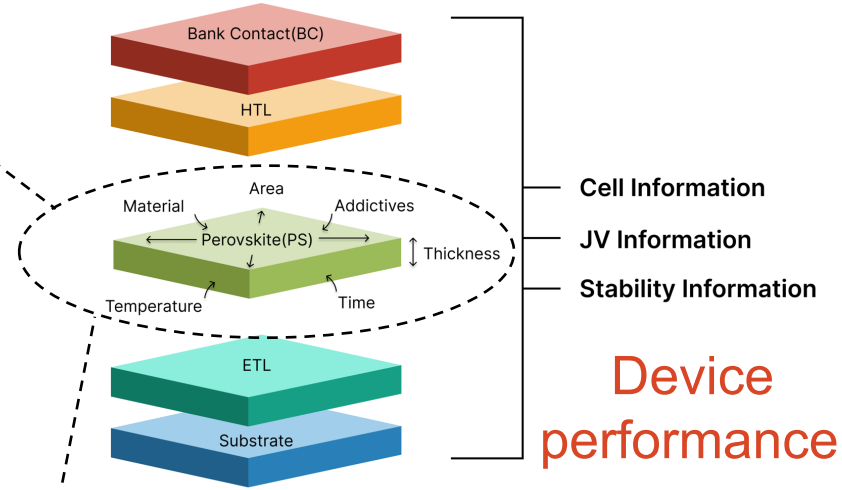
Material properties

Bandgap

K_VRH

Carrier mobility

…



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.

MASTER.AI UNSW AI4Science

UNSW SYDNEY

DARWIN

MASTER.AI
UNSW AI4Science

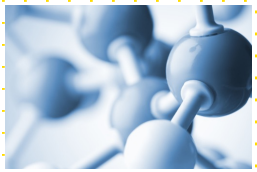
❖ *Customized GPT for natural science*

LLaMA + 6M Paper FAIR Material Database… + UNSW SYDNEY, NCI AUSTRALIA, GreenDynamics

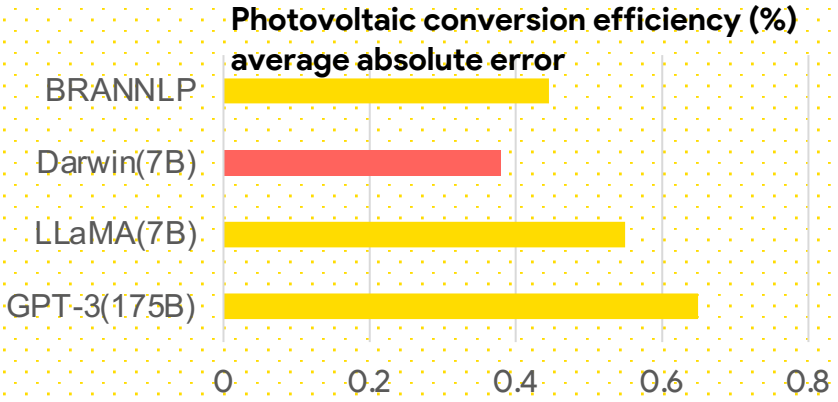
❖ *Powerful materials design capabilities*

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-C3N4) that make it an attractive material for optoelectronic applications?

Graphitic carbon nitride (g-C3N4) 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-C3N4) 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 Ag2CsSb2I3?

1.96

As of my knowledge cutoff in September 2021, there isn't specific information available on the bandgap of Ag2CsSb2I3. 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.

DARWIN

➢Performance on material prediction task

Better thanState-of-The-Art model on most tasks

Berkeley- Matbench (Solid material)

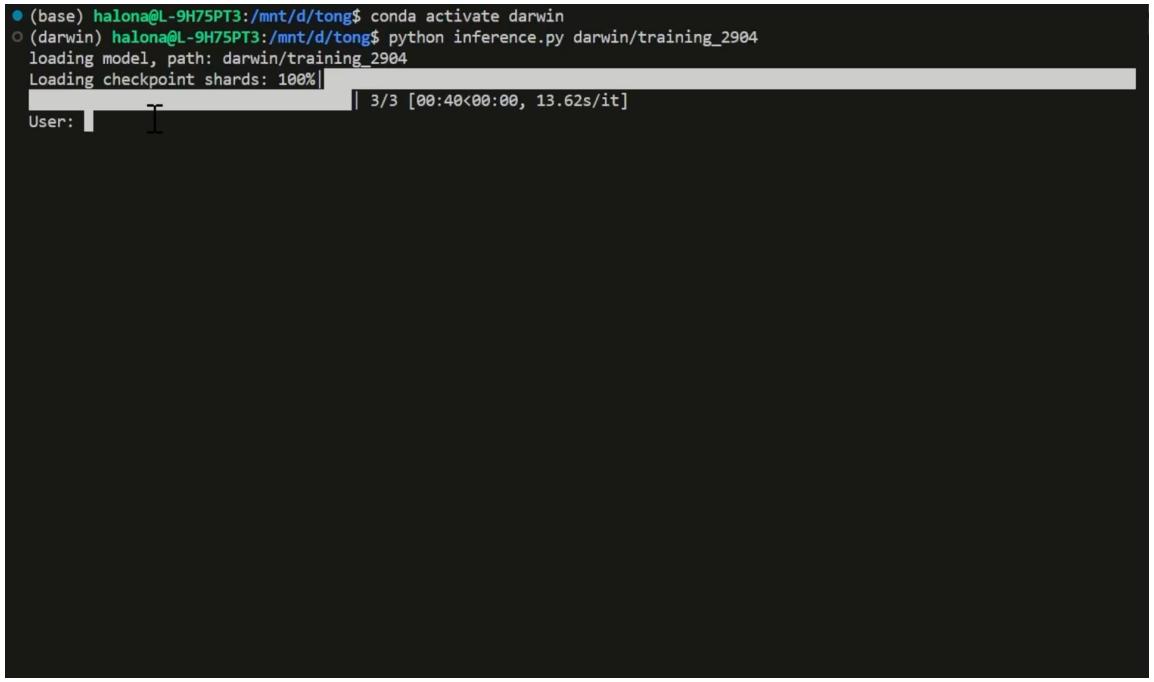
Task	Metric	SOTA	DARWIN
Bandgap (Regression)	MAE	0.3310	0.2790
Metal (Classification)	ROCAUC	0.9209	0.9650

Harvard - HOPV15 (Organic solar cells)

Task	Metric	SOTA	DARWIN
PCE (Regression)	MAE	0.42	0.38

➢Performance on device design

Automated formulation of perovskite solar cell product



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

MASTER.AI
UNSW AI4Science
UNSW SYDNEY

DARWIN

➢Information extraction on paper structure

2-3 years for 50+ scientists

nature energy
RESOURCE
https://doi.org/10.1038/s41560-021-00941-3
Check for updates

OPEN

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

T. Jesper Jacobsson[1,2], Adam Hultqvist[3], Alberto García-Fernández[4], Aman Anand[5,6], Amran Al-Ashouri[7], Anders Hagfeldt[8], Andrea Crovetto[9], Antonio Abate[10], Antonio Gaetano Ricciardulli[11], Anuja Vijayan[2], Ashish Kulkarni[12], Assaf Y. Anderson[13], Barbara Primera Darwich[14], Bowen Yang[8], Brendan L. Coles[15], Carlo A. R. Perini[16], Carolin Rehermann[1], Daniel Ramirez[17], David Fairen-Jimenez[18], Diego Di Girolamo[19,20], Donglin Jia[21], Elena Avila[18], Emilio J. Juarez-Perez[22], Fanny Baumann[8,23], Florian Mathies[1], G. S. Anaya González[24], Gerrit Boschloo[2], Giuseppe Nasti[19], Gopinath Paramasivam[1,25], Guillermo Martínez-Denegri[26], Hampus Näsström[1], Hannes Michaels[2], Hans Köbler[10], Hua Wu[2], Iacopo Benesperi[2], M. Ibrahim Dar[27], Ilknur Bayrak Pehlivan[28], Isaac E. Gould[29,30], Jacob N. Vagott[16], Janardan Dagar[1], Jeff Kettle[31], Jie Yang[32], Jinzhao Li[1], Joel A. Smith[33,34], Jorge Pascual[10], Jose J. Jerónimo-Rendón[35], Juan Felipe Montoya[17], Juan-Pablo Correa-Baena[16], Junming Qiu[21], Junxin Wang[28,36], Kári Sveinbjörnsson[7], Katrin Hirselandt[1], Krishanu Dey[27], Kyle Frohna[27], Lena Mathies[37], Luigi A. Castriotta[38], Mahmoud. H. Aldamasy[10,39], Manuel Vasquez-Montoya[1,17], Marco A. Ruiz-Preciado[40,41], Marion A. Flatken[10], Mark V. Khenkin[42], Max Grischek[7,43], Mayank Kedia[12,35], Michael Saliba[12,35], Miguel Anaya[27,44], Misha Veldhoen[13], Neha Arora[27], Oleksandra Shargaieva[1], Oliver Maus[1], Onkar S. Game[33], Ori Yudilevich[13], Paul Fassl[40,41], Qisen Zhou[21], Rafael Betancur[17], Rahim Munir[1], Rahul Patidar[15], Samuel D. Stranks[27,44], Shahidul Alam[5,6,45], Shaoni Kar[46], Thomas Unold[9], Tobias Abzieher[41], Tomas Edvinsson[28], Tudur Wyn David[47], Ulrich W. Paetzold[40,41], Waqas Zia[12,35], Weifei Fu[11], Weiwei Zuo[35], Vincent R. F. Schröder[48,49], Wolfgang Tress[50], Xiaoliang Zhang[21], Yu-Hsien Chiang[27], Zafar Iqbal[10], Zhiqiang Xie[51] and Eva Unger[1,23]

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 (PbI2 99.99%) was purchased from Tokyo Chemical Industry (TCI, Japan). Organic salts, including methylamium bromide (MABr), methylamium iodide (MAI), and formamidinium iodide (FAI) were purchased from GreatCell Solar (Dyesol, 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'-spirobifluorene (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−1) was modified onto the pre-cleaned FTO glass in the N2-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	TiO2-c
ETL_additives_compounds	Unknown
ETL_deposition_procedure	Spin-coating
Perovskite_composition_long_form	FAPbI3
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.

MASTER.AI
UNSW AI4Science
UNSW SYDNEY

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
- Automatic simulation
- …

Find new material

- Material prediction
- Synthesis methods
- Material design
- …

And more application scenarios…

MASTER.AI UNSW AI4Science

UNSW SYDNEY

Thank you!

E-mail: tong.xie@unsw.edu.au