Application of the Question Answering method to extract information from materials science literature

MLM4MS

13.05.2024-17.05.2024 Ljubljana

M. Sipilä[1], F. Mehryary[2], F. Ginter[2], S. Pyysalo[2] & M. Todorović[1]

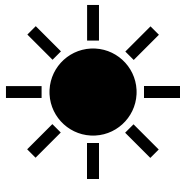
1. Department of Mechanical & Materials Engineering, University of Turku, Finland

2. Department of Computing, University of Turku, Finland

UNIVERSITY OF TURKU

Challenges in materials design

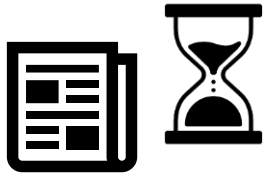
Perovskites are promising candidates for solar cell materials.



Compositional engineering is used to tune the functional properties.

It is difficult to know which composition would lead to best properties.

There is too much information in articles to be parsed manually.



Could we use text as data in machine learning?

UNIVERSITY OF TURKU

Current methods for information extraction

Rule based

The **[property]** of **[material]** can vary from **[number] [unit]** to **[number] [unit].**

The **[material]** has a **[property]** of **[number] [unit].**

Supervised machine learning based

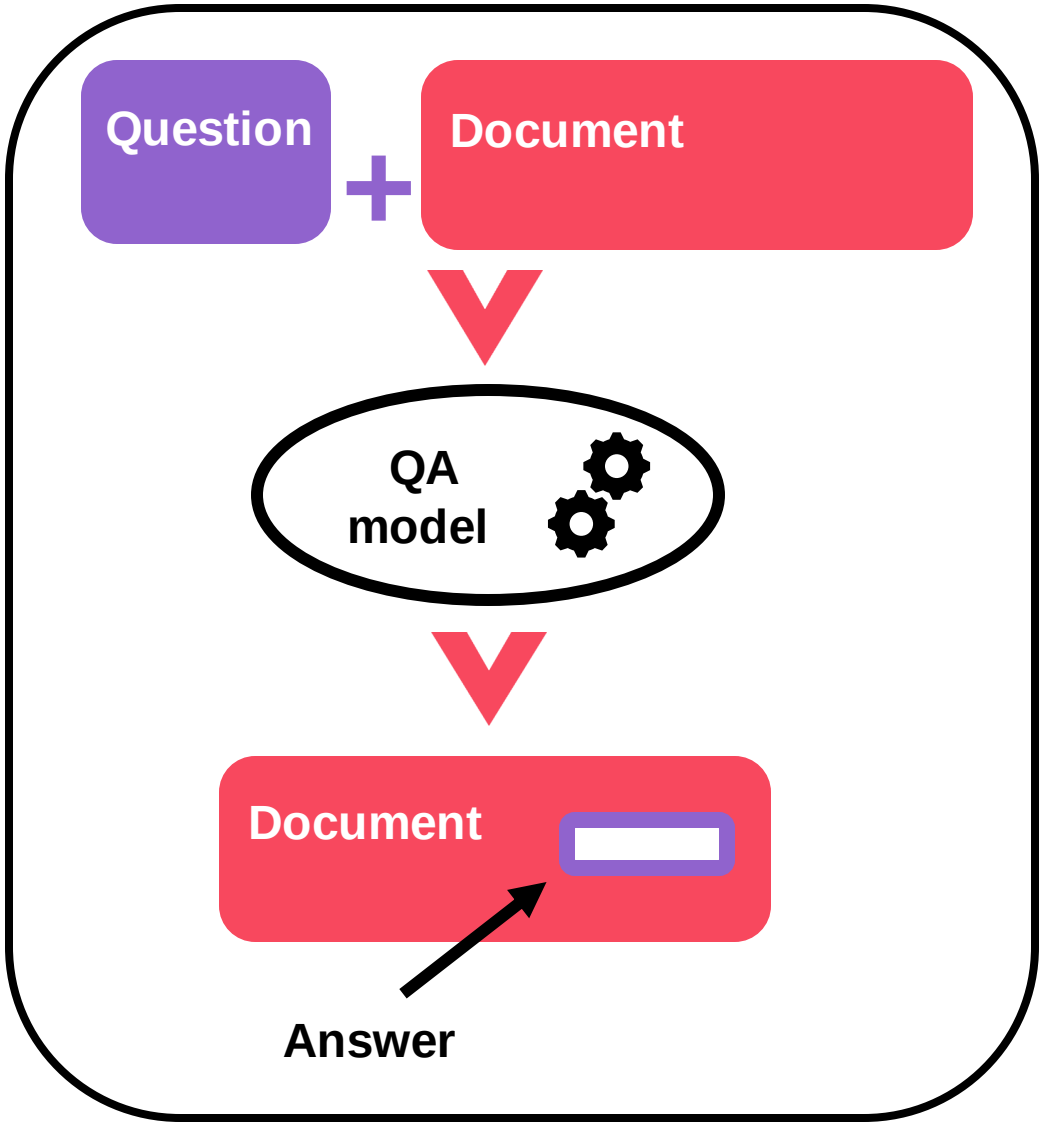
MAPB has a bandgap of 1.3 eV.

The band-edge of $CsPbI_3$ is between 1.6 eV and 1.7 eV.

UNIVERSITY OF TURKU

Question Answering

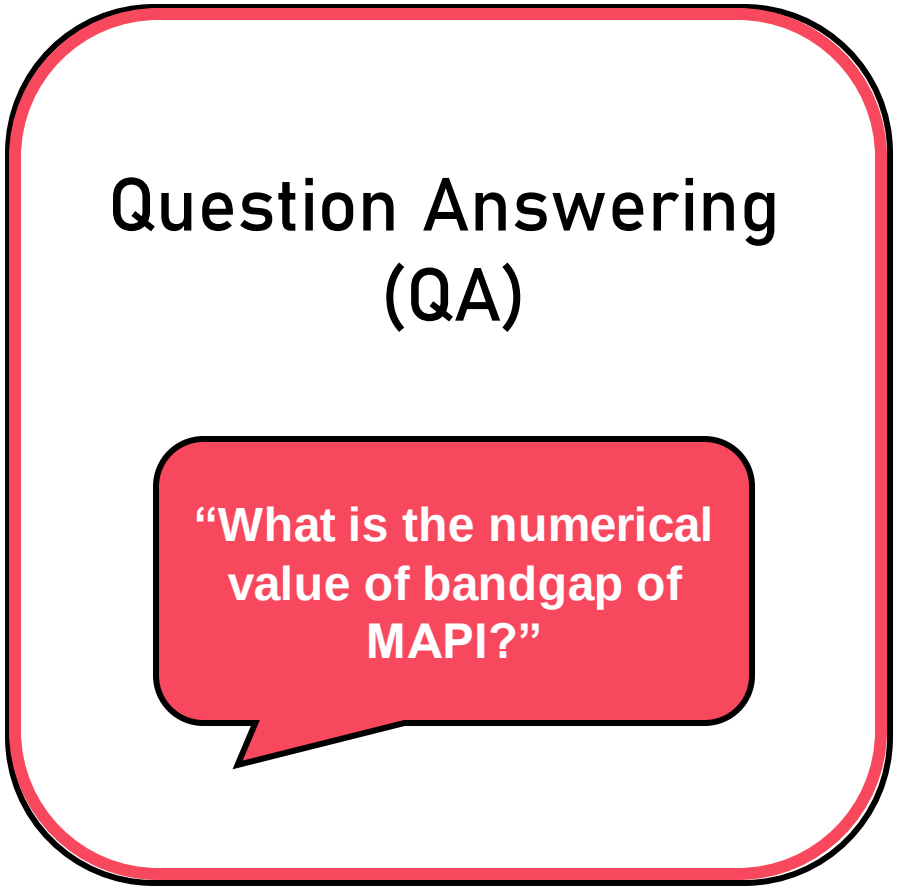
- QA model is a large language model trained to extract information from documents based on the provided questions.
- Does not require re-training for different materials or properties.
- From an end-user perspective easy to use.



UNIVERSITY OF TURKU

Objectives

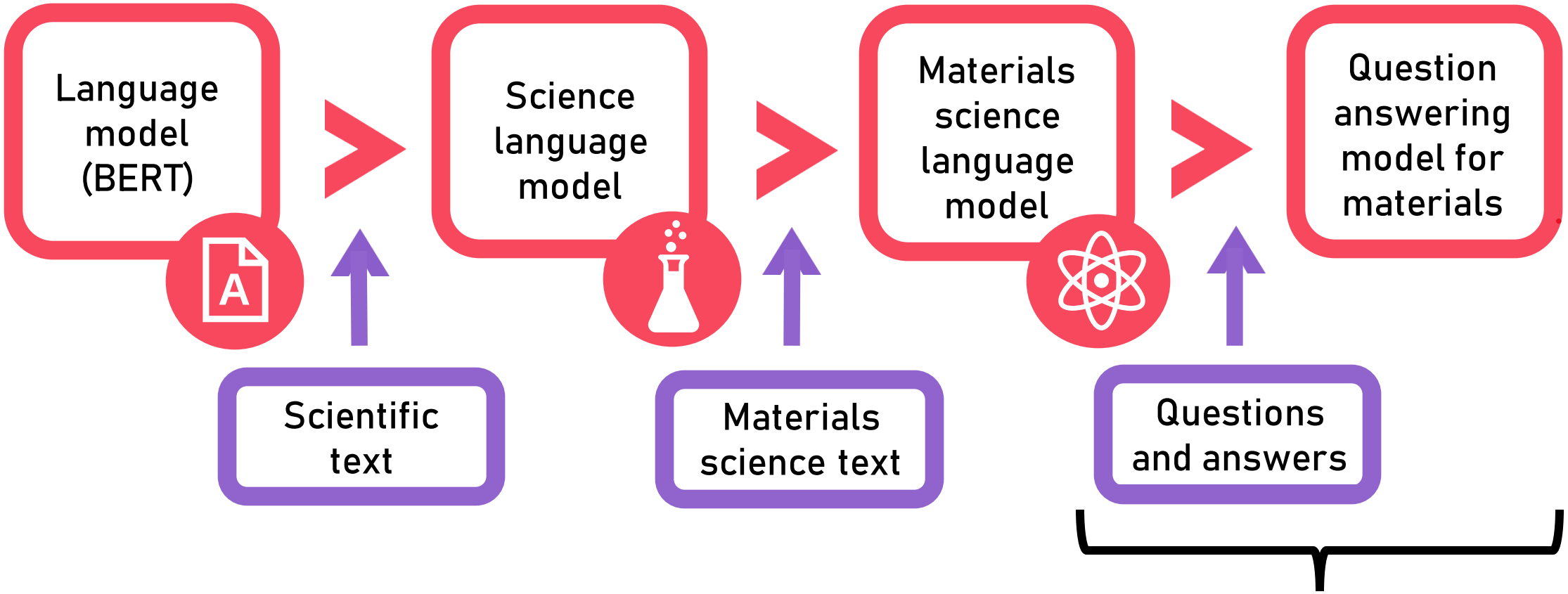
- Develop a QA method for information extraction.
- Evaluate the method performance and different model choices.
- Apply the method to large scale information extraction.



UNIVERSITY OF TURKU

Methodology

Question answering language model



Based on transformer neural networks and transfer learning.

UNIVERSITY OF TURKU

Workflow

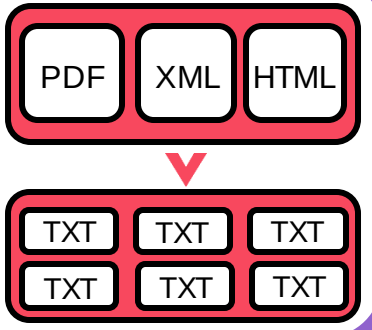
1. Data acquisition

- Web scraping
- API querying

ELSEVIER CORE Springer arXiv ROYAL SOCIETY OF CHEMISTRY

2. Data processing

- Converting to text
- Cleaning and normalising
- Segment formatting



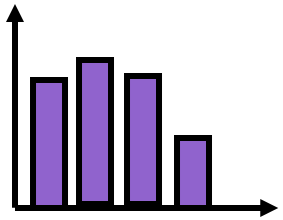
3. Machine learning

- Selecting the QA model
- Applying model to the segments

"What is the numerical value of bandgap of MAPI?"

4. Analysis

- Obtaining a list of values extracted
- Evaluating the model performance
- Visualising results



UNIVERSITY OF TURKU

Text segment (snippet) formatting

Snippet

Table D presents the photovoltaic characteristics of this structure in which Jsc, Voc, FF and PCE were calculated 18.11 mA/cm2, 0.96 V, 0.82 and 14.37 %. The results exhibit a good agreement with the earlier results. Then, a layer of CIGS with 100 nm was incorporated between CH3NH3PbI3 and CuSCN. Owing to the lower **bandgap** of CIGS (1.2 **eV**) compared to **MAPI** (1.56 **eV**), it is places beneath the CH3NH3PbI3 for absorbing the incoming light. While the CIGS has quite a strong absorption in the wavelength range of 800 to 1100 nm. In other words, the light absorption spectrums of CH3NH3PbI3 and CIGS effectively complement each other thus, inserting the CIGS layer beside the CH3NH3PbI3 layer enables the cell to harvest the wider range of the solar spectrum.

- Using the dataset of full articles would be inefficient.
- Using 7-sentence snippets enables more context than single sentences.
- Snippet has to contain name of the **property** ('bandgap'), name of the **material** ('MAPI') and the **unit** ('eV'). We used regular expressions for this.

UNIVERSITY OF TURKU

Evaluation dataset

- Manually annotated dataset of 600 snippets.
- Snippets are evenly balanced between five different perovskites.
- Annotators were materials science experts and every snippet was annotated by two people.

Question

"What is the numerical value of bandgap of MAPI?"

Table D presents the photovoltaic characteristics of this structure in which Jsc, Voc, FF and PCE were calculated 18.11 mA/cm2, 0.96 V, 0.82 and 14.37 %. The results exhibit a good agreement with the earlier results. Then, a layer of CIGS with 100 nm was incorporated between CH3NH3PbI3 and CuSCN. Owing to the lower bandgap of CIGS (1.2 eV) compared to MAPI (1.56 eV), it is places beneath the CH3NH3PbI3 for absorbing the incoming light. While the CIGS has quite a strong absorption in the wavelength range of 800 to 1100 nm. In other words, the light absorption spectrums of CH3NH3PbI3 and CIGS effectively complement each other thus, inserting the CIGS layer beside the CH3NH3PbI3 layer enables the cell to harvest the wider range of the solar spectrum.

Answer

1.56 eV

Results

Evaluating language models

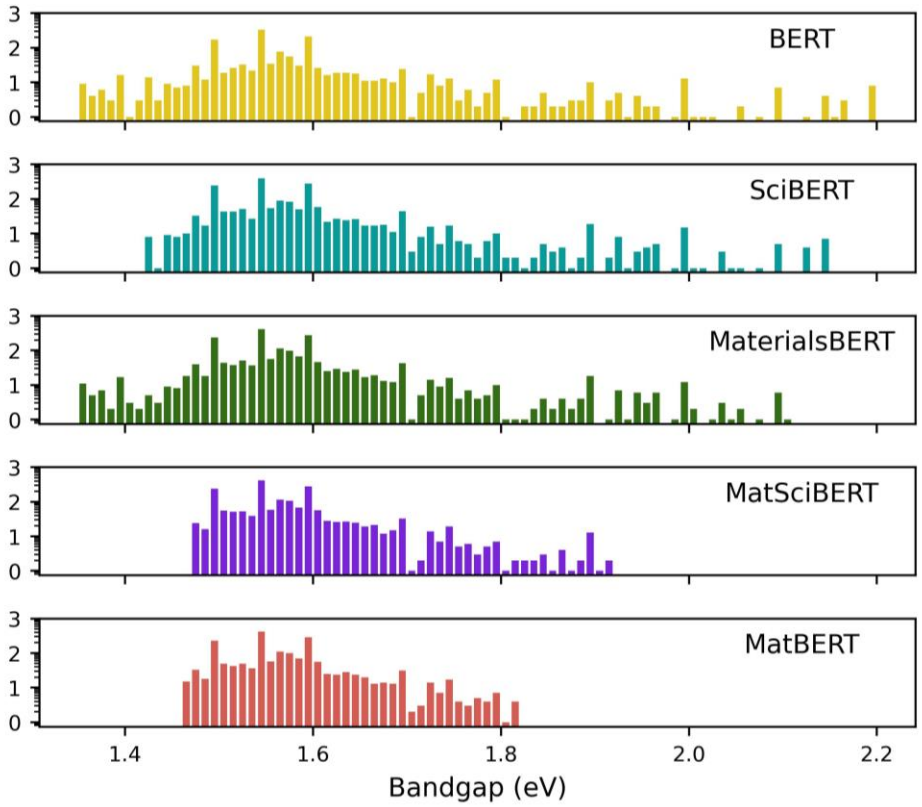
- Evaluation data: 600 manually annotated snippets.
- We tested base BERT and 4 other BERTs which were trained with scientific or materials science text from base BERT.

Method	QA BERT	QA SciBERT	QA MatSciBERT	QA MatBERT	QA Materials-BERT	ChemData-Extractor2
Precision	62.4	58.1	72.4	61.8	64.4	81.1
Recall	36.9	63.2	59.3	68.0	56.1	34.9
F1-score	45.8	60.5	64.4	64.6	59.4	48.8

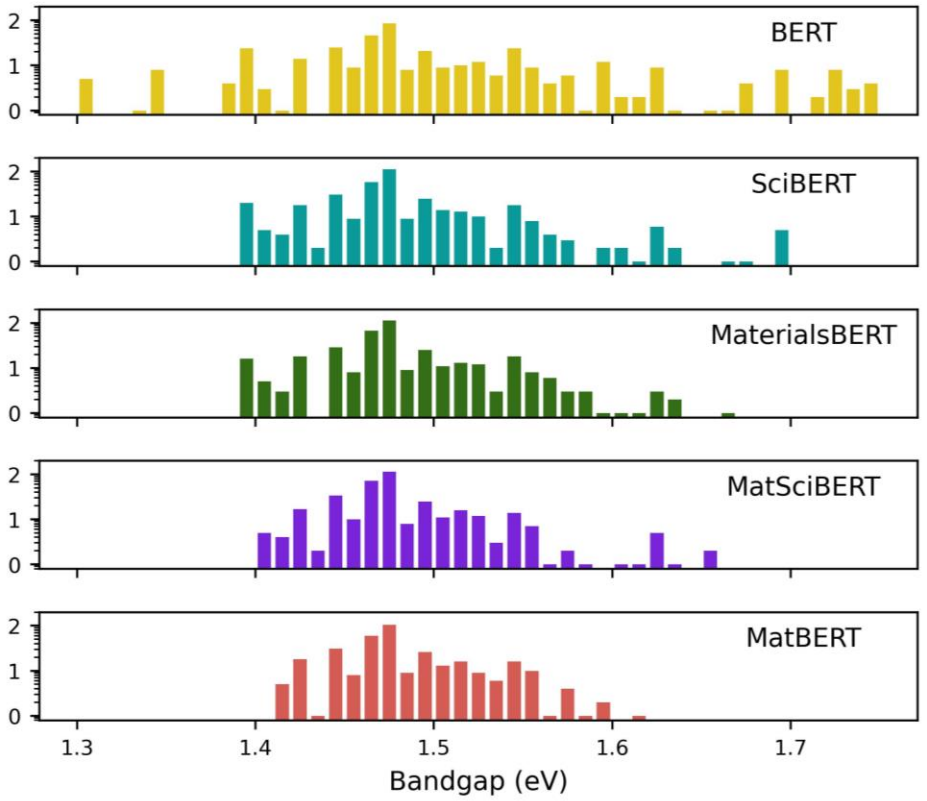
UNIVERSITY OF TURKU

Information extraction with different BERTs

MAPI (7,283 snippets)



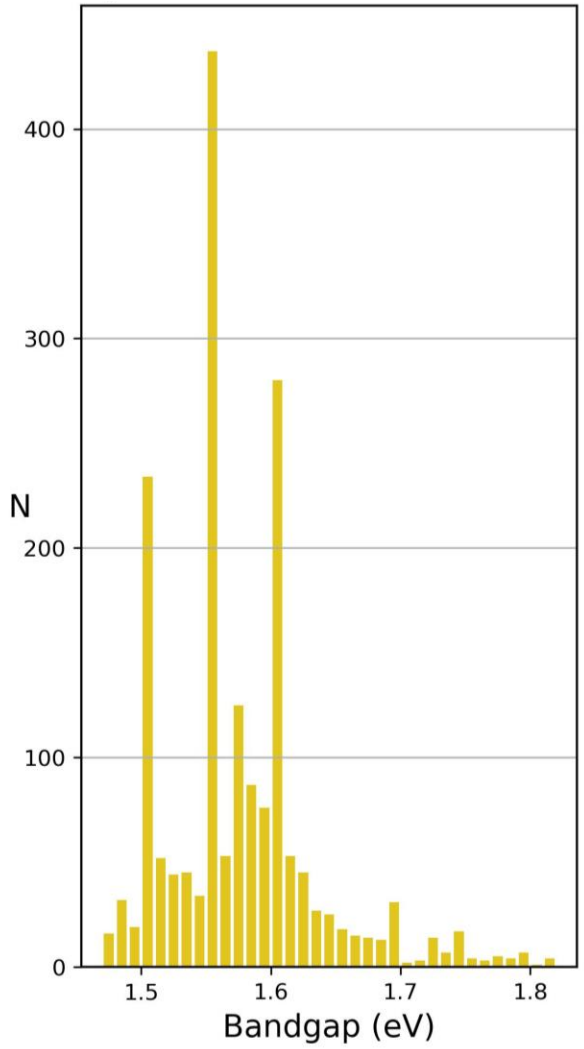
FAPI (1,251 snippets)



UNIVERSITY OF TURKU

Information extraction of MAPI

"What is the numerical value of bandgap of MAPI?"

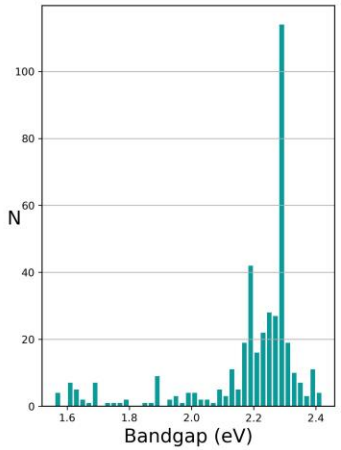


- We extracted 1,845 bandgap values of MAPI using MatBERT.
- The mode 1.55 eV agrees with the accurate literature value.
- Spread of the values is due to different experimental and computational factors.

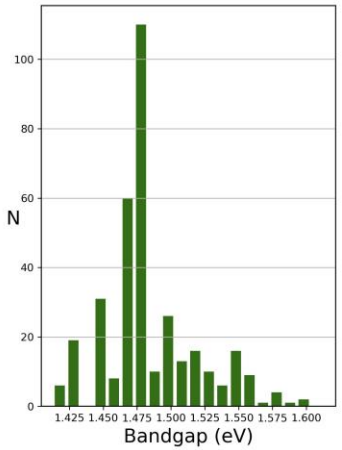
UNIVERSITY OF TURKU

Hybrid and inorganic perovskite results

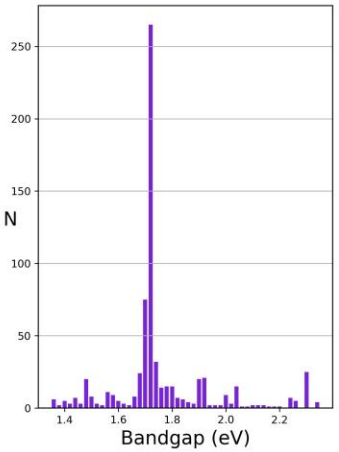
MAPB



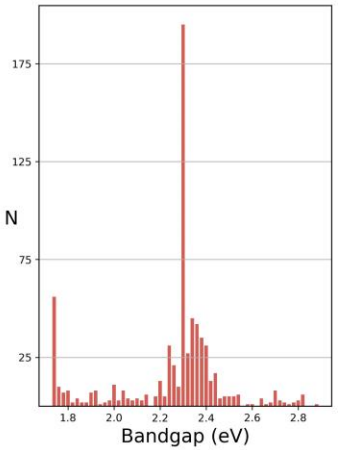
FAPI



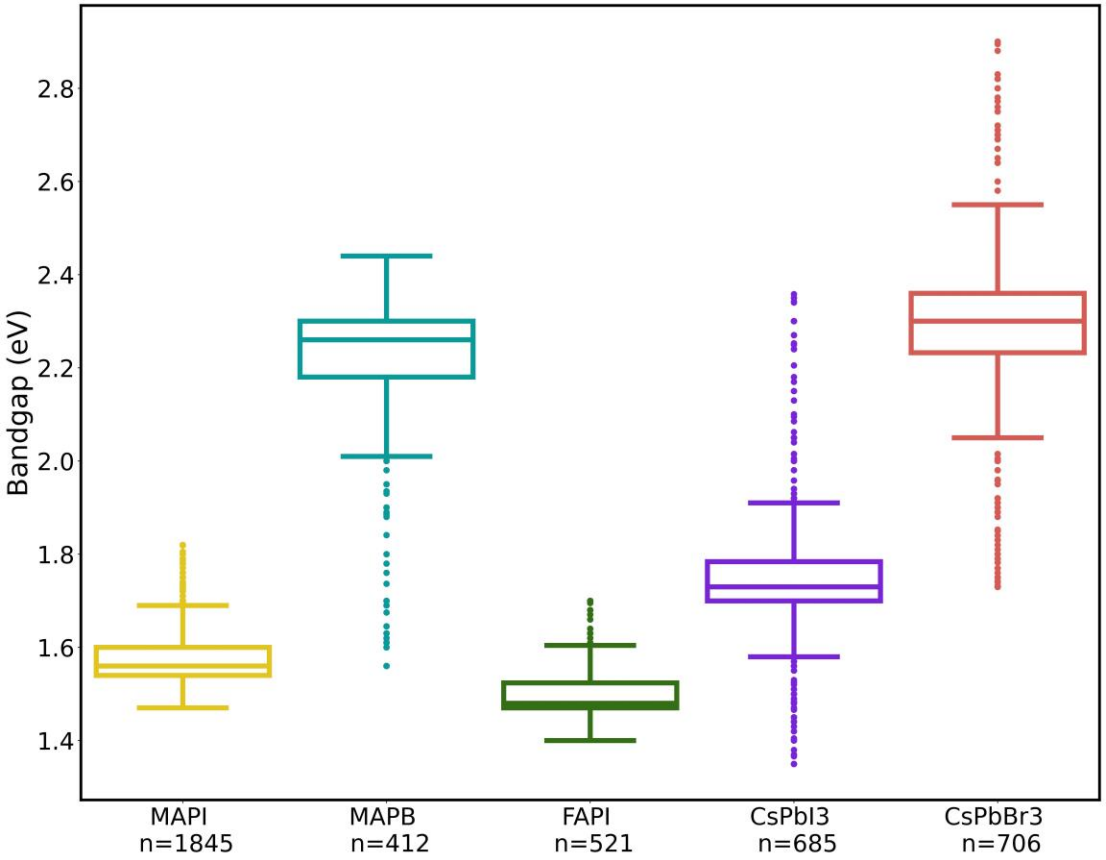
CsPbI3



CsPbBr3



Boxplot of extracted bandgap values



UNIVERSITY OF TURKU

Future work

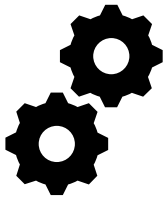
Extract more properties and materials.

- Refractive index
- PCE
- P2
- …

Test the method with alloys (complicated names).

- (CH3NH3)1−x(HC(NH2)2)xPbI3
- FA0.75MA0.25Sn1−xGexI3
- CsPb0.9Cd0.1Br3

Find out whether the extracted value is experimental or computational.



UNIVERSITY OF TURKU

Conclusions

- Given the suitable dataset, material-property relationships can be extracted with QA method.
- Different BERTs have a considerable impact how QA model performs.
- Extracted information can be used to map materials space and to guide materials design.

Acknowledgements:

Suomen Akatemia
Finlands Akademi
Research Council of Finland

DeeperMaterials-project

UNIVERSITY OF TURKU

UNIVERSITY
OF TURKU

Thank you!