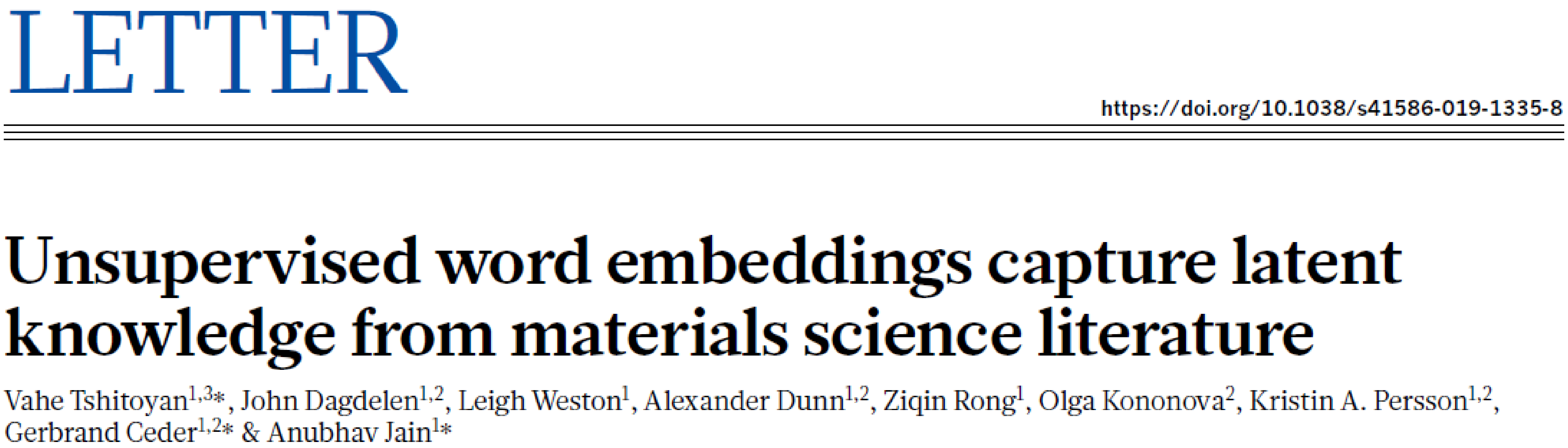


LETTER

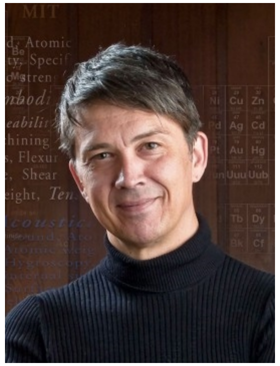
https://doi.org/10.1038/s41586-019-1335-8

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan[1,3]*, John Dagdelen[1,2], Leigh Weston[1], Alexander Dunn[1,2], Ziqin Rong[1], Olga Kononova[2], Kristin A. Persson[1,2], Gerbrand Ceder[1,2]* & Anubhav Jain[1]*

Vahe Tshitoyan

Anubhav Jain

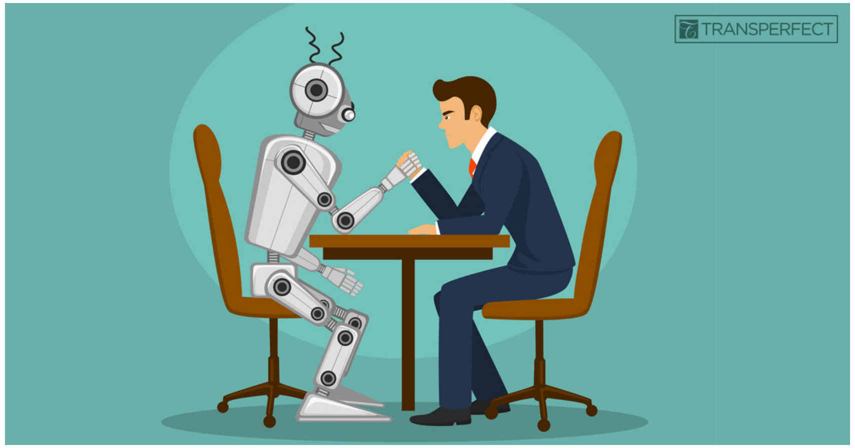


Gerbrand Ceder

Cancan Huang
09/11/2019

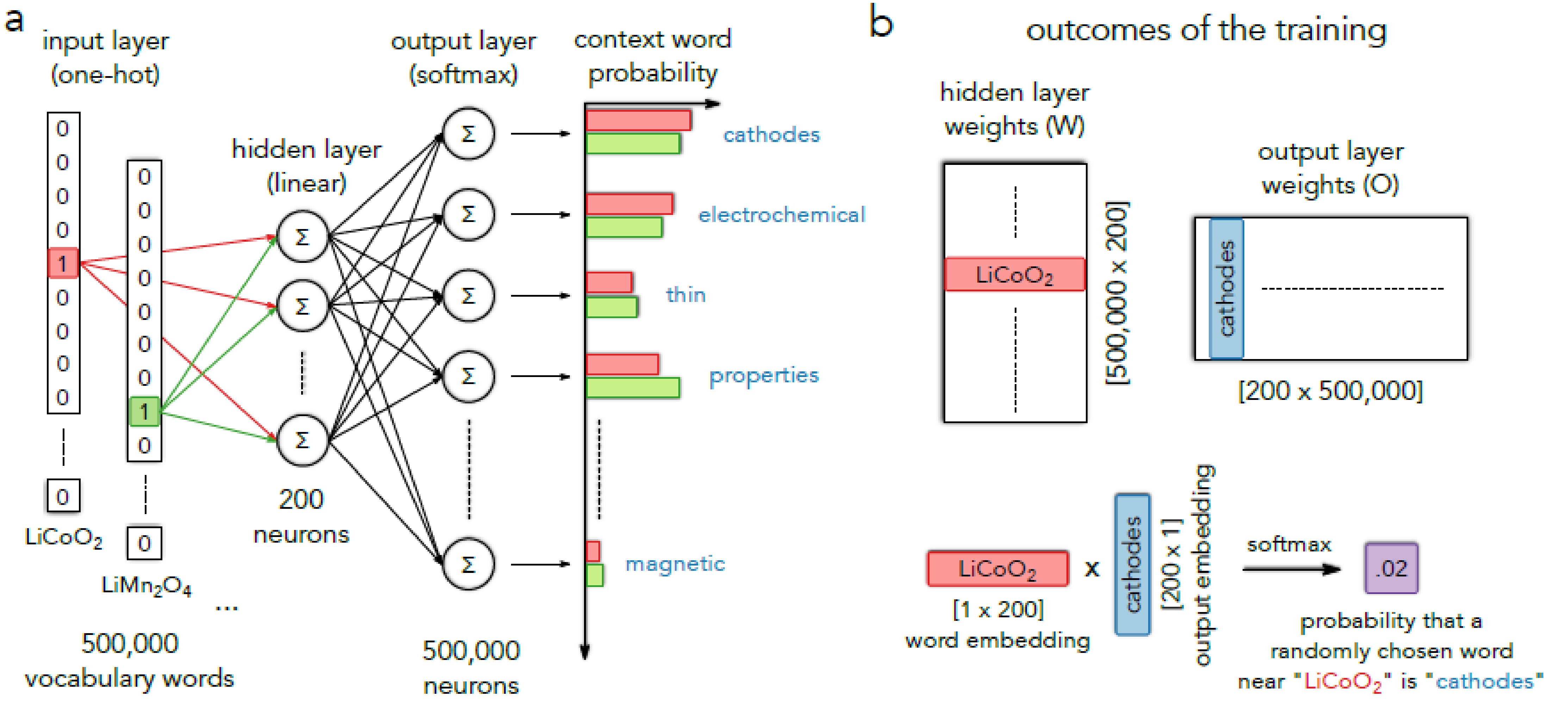
Background

- **Researches are published in the form of text**
- **Current researches based on structured property databases**
  - Only cover a small fraction of knowledge in literature
- **Natural language processing helps extract information in text**
- **Supervised machine learning requires large hand-labelled datasets**



Solution: using unsupervised word embeddings to capture latent knowledge from materials science literature

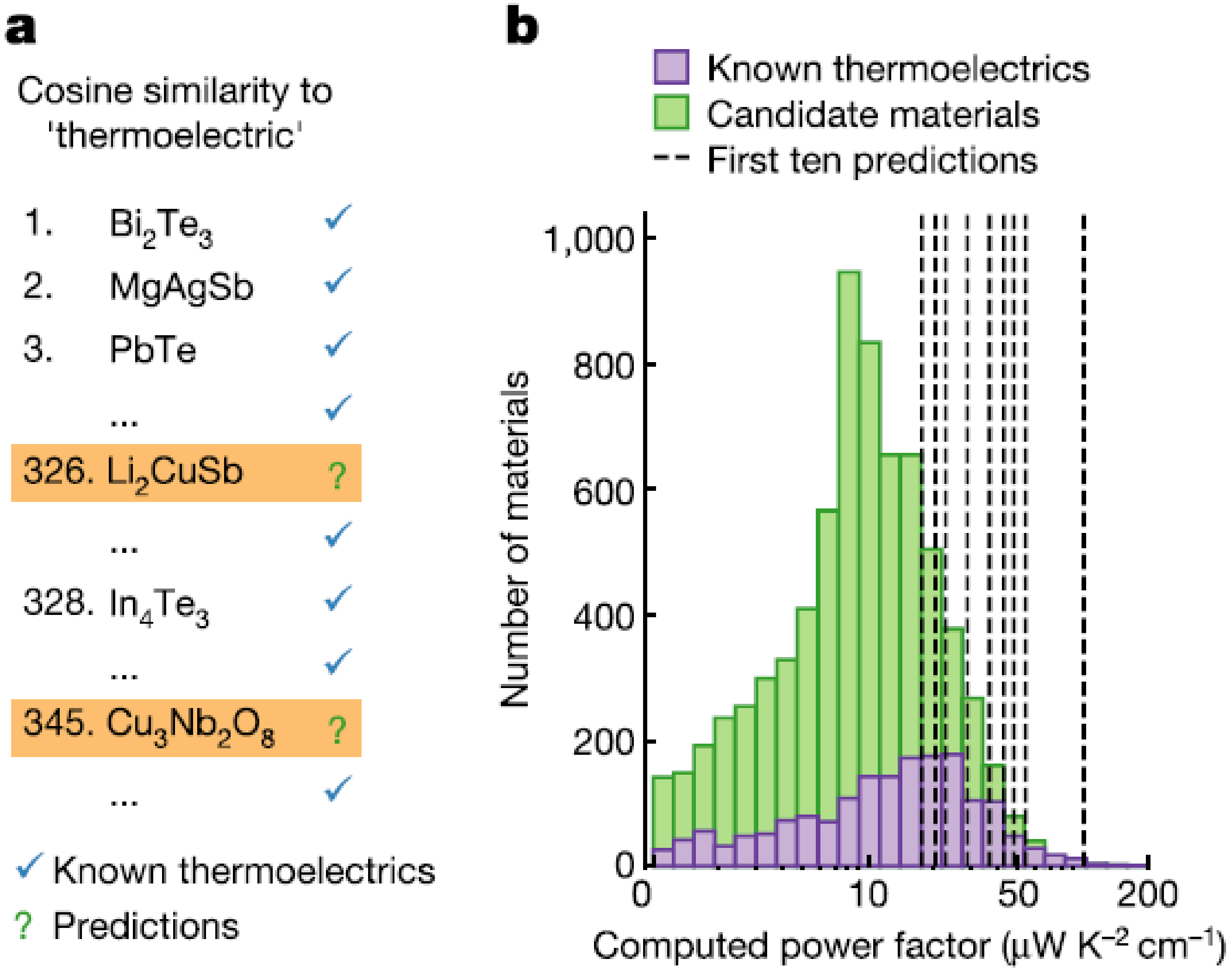
Word2Vec skip-gram



- **Words with similar meanings often appear in similar contexts, the corresponding embeddings will also be similar.**

Test 1: comparison with available computational data

- **9,483** compounds overlap in total (fig. b)
  - mentioned more than 3 times in **text corpus**
  - Thermoelectric power factors reported in **dataset**
  - **7,663** never mentioned with thermoelectric keywords **acting as prediction**
- 7,663 Ranked by the **dot product of their normalized output embedding with the word 'thermoelectric'** (fig. a)
  - Interpreted as the likelihood that that material will co-occur with the word 'thermoelectric' in a scientific abstract
- **Conclusion: Top 10 predictions have greater thermoelectric power factor than means!**



a

Cosine similarity to 'thermoelectric'

1. Bi_2Te_3 ✓
2. MgAgSb ✓
3. PbTe ✓

... ✓

326. Li_2CuSb ?

... ✓

328. In_4Te_3 ✓

... ✓

345. $Cu_3Nb_2O_8$?

... ✓

✓ Known thermoelectrics

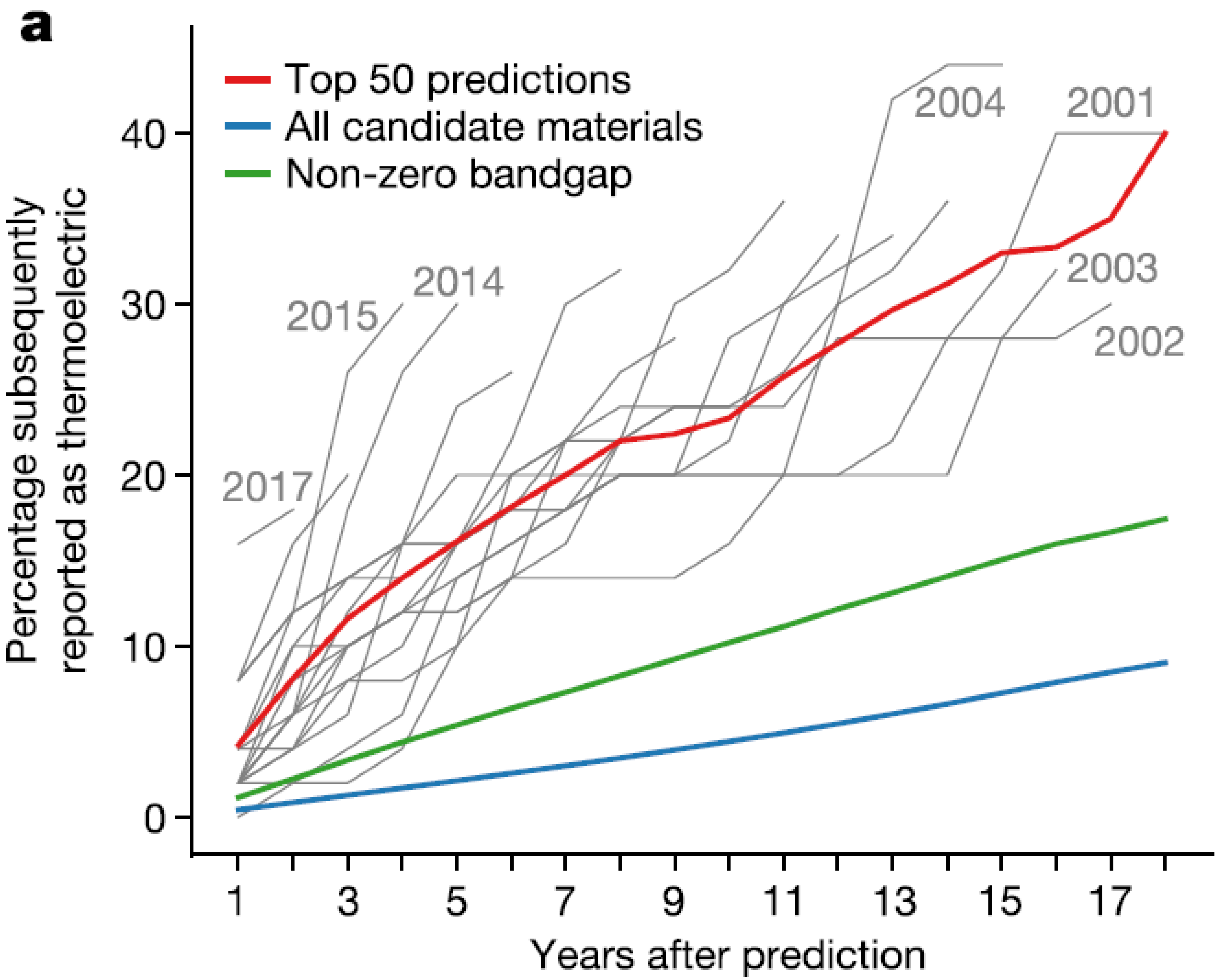
? Predictions

Test 2: Predict thermoelectric materials

- **Dataset:** 18 different text corpora before cutoff years between 2001 and 2018

- **Goal:** predict the top 50 thermoelectric materials that were likely to be reported in the future years

- **Conclusion**
  - 8 times than randomly chosen from all
  - 3 times than random material with a non-zero DFT bandgap
  - More recent data improve performance indicated by steeper slope.

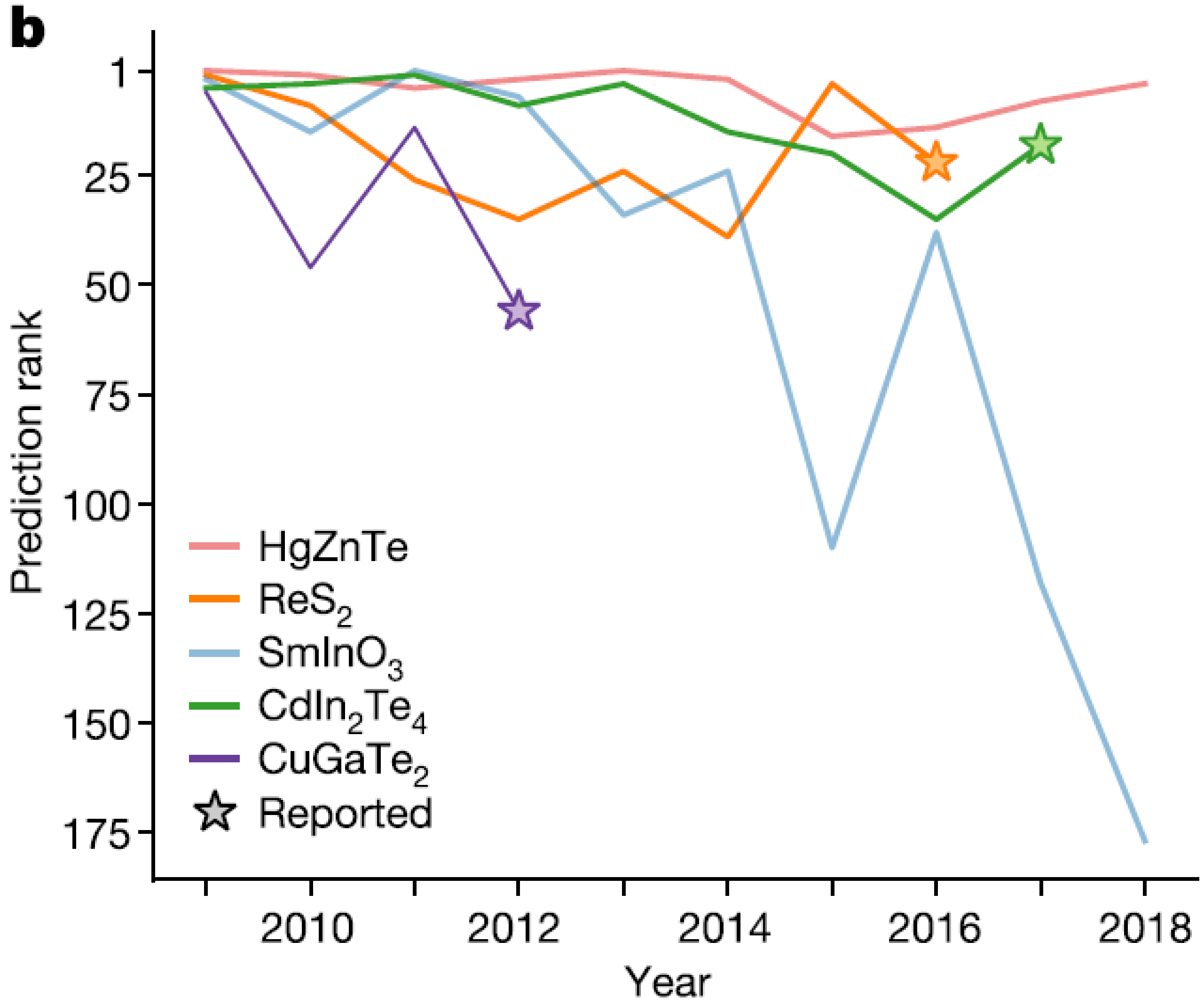


Test 2: Predict thermoelectric materials

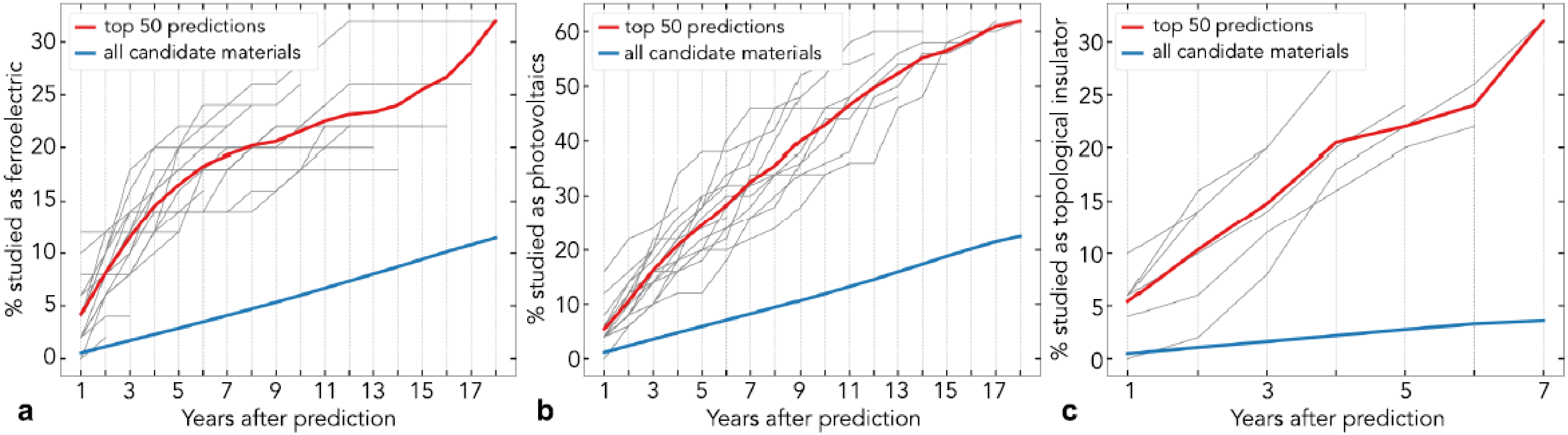
- **Top 5 predictions using data before 2009**

- **Marker:** the year of first published report as a thermoelectric

- **ReS_2 & $CdIn_2Te_4$**: 8-9 years
- **$CuGaTe_2$**: 4 years
- **$SmInO_3$**: expensive
- **HgZnTe:** toxic



Test 2: Predict thermoelectric materials



Ferroelectric

photovoltaics

topological insulator

Conclusions

- Without any explicit insertion of chemical knowledge, embeddings capture complex materials science concepts.

- An unsupervised method can recommend materials for functional applications several years before their discovery.

- This can enable a new paradigm of machine-assisted scientific breakthroughs.

Questions?