Statistical Language Models

Week 12.0

From Cookbook data (week 9)

- library(wordVectors)
- model = train_word2vec("cookbooks.txt","cookbook_vectors.bin",
- vectors=200,threads=4,
- cbow=1,window=12,
- iter=500,**negative_samples**=0) #<— **negative samples can be used to update only a percentage of model weights rather than all of them for each word. This updating focuses on a subset of the words where the model should be outputting a zero, i.e. negative words.**
- Iter is number of times it passes through the corpus. With many words smaller iters (fewer passes) is not so bad, for smaller corpus expect to need .
- Wod2Vec can take many hours to train (iters=5 takes ~1 minute on 4 cores)

From Cookbook data (week 9)

- #Find the fish words:
- model %>% closest_to("fish")
- #Expand out to search for more fish words:
- model %>% closest_to(model[[c("fish","salmon","trout","shad","flounder","carp","roe", "eels")]],50)
- #Can be used to find all recipes that relate to fish things, find ways of cooking fish,…

Goal

- Can chemical interpretation and scientific knowledge be captured through positions of words in scientific abstracts?
- Can these insights be used to discover new chemicals with desired properties?

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]*

The overwhelming majority of scientific knowledge is published as text, which is difficult to analyse by either traditional statistical analysis or modern machine learning methods. By contrast, the main source of machine-interpretable data for the materials research community has come from structured property databases[1,2], which encompass only a small fraction of the knowledge present in the research literature. Beyond property values, publications contain valuable knowledge regarding the connections and relationships between data items as interpreted by the authors. To improve

behave consistently with chemical intuition when they are combined using various vector operations (projection, addition, subtraction). For example, many words in our corpus represent chemical compositions of materials, and the five materials most similar to $LiCoO_2$ (a well-known lithium-ion cathode compound) can be determined through a dot product (projection) of normalized word embeddings. According to our model, the compositions with the highest similarity to $LiCoO_2$ are $LiMn_2O_4$, $LiNi_{0.5}Mn_{1.5}O_4$, $LiNi_{0.8}Co_{0.2}O_2$, $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$ and $LiNiO_2$—all of which are also lithium-ion cathode materials.

Similar to the observation made in the original Word2vec paper[11]

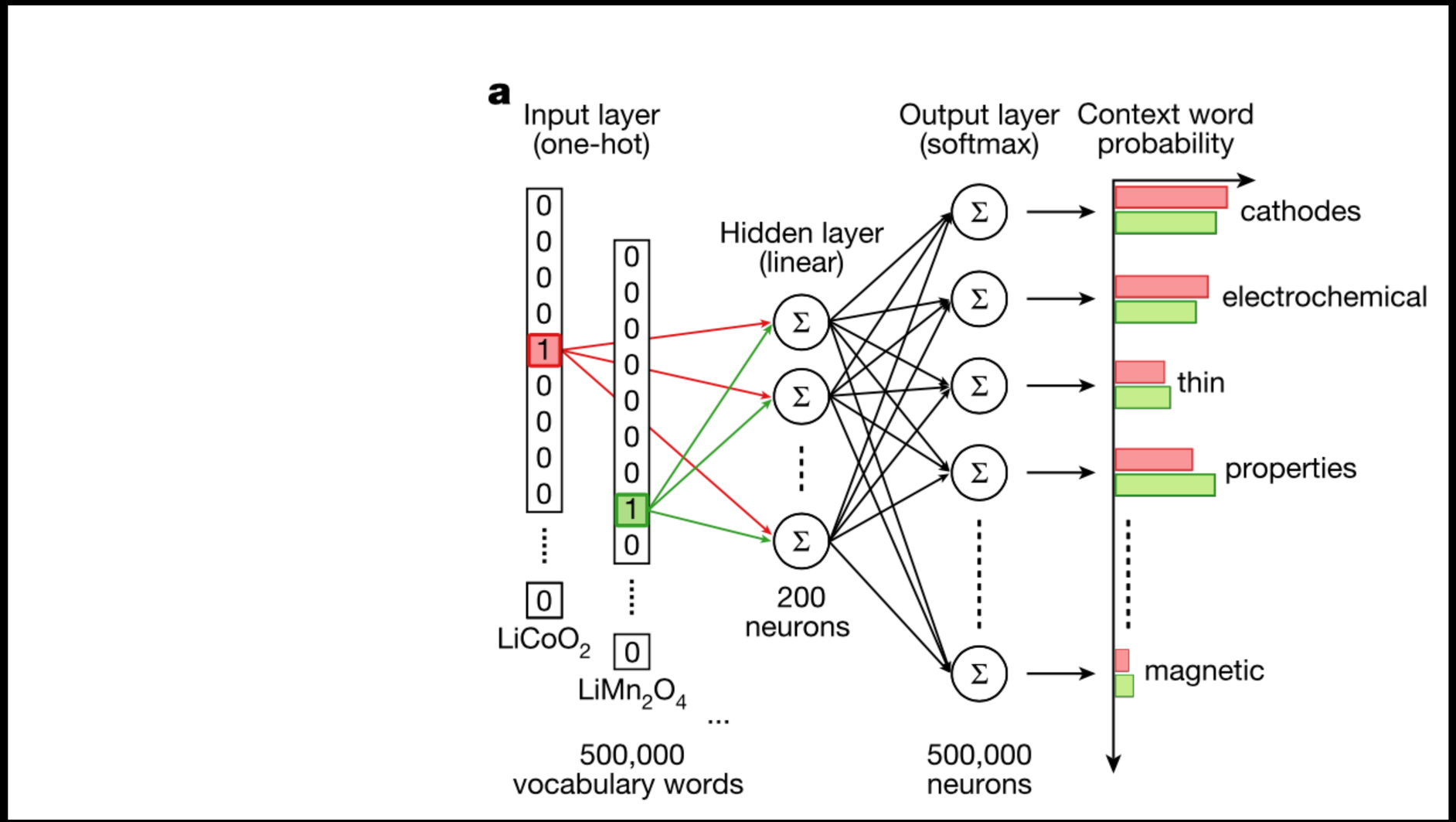
Word Embeddings

- "… when trained on a suitable body of text, such methods should produce a vector representing the word ‘iron’ that is closer by cosine distance to the vector for ‘steel’ than to the vector for 'organic’."
- "To train the embeddings, we collected and processed approximately 3.3 million scientific abstracts published between 1922 and 2018 in more than 1,000 journals deemed likely to contain materials-related research, resulting in a vocabulary of approximately 500,000 words."

- We then applied the skip-gram variation of Word2vec, , which is trained to predict context words that appear in the proximity of the target word as a means to learn the **200-dimensional embedding** of that target word, to our text corpus"

- **200-dimensional embedding**
- Window = 8
- Minimum word count = 5
- Negative sampling = 15
- Iterations = 30

Figure 1a



Word2Vec capacity

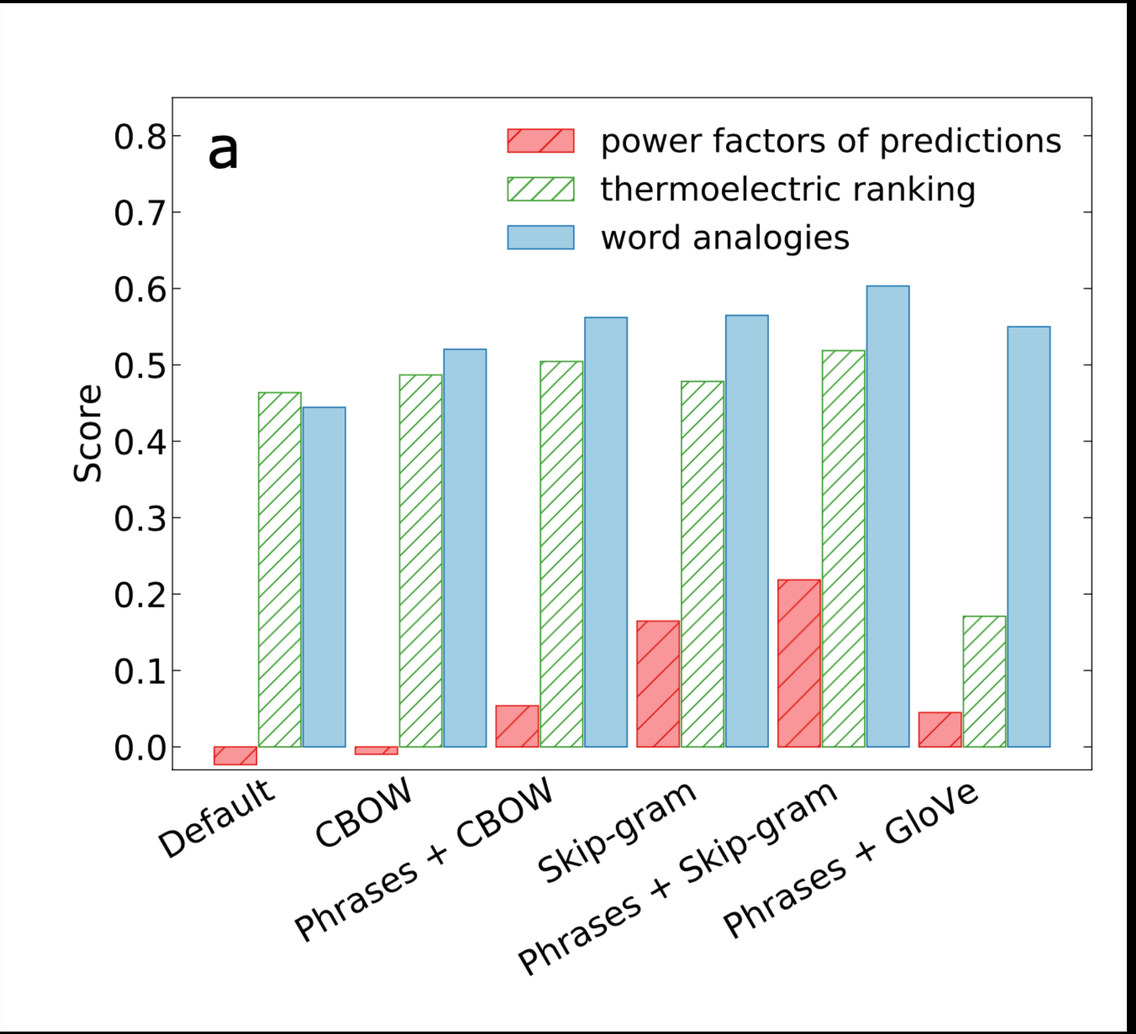
- Find materials with the smallest cosine distance from $LiCoO_2$
- Find Analogies: ‘NiFe’ is to ‘ferromagnetic’ as ‘IrMn’ is to 'antiferromagnetic’
    - ferromagnetic–NiFe + IrMn ≈ antiferromagnetic

But does it work?

- Chemistry context sanity checks: Zr-ZrO_2 ≈ Cr-Cr_2O_3 ≈ Ni - NiO
- Chemical Structure sanity checks: (Zr–HCP≈Cr–BCC≈ Ni – FCC).
- "The positions of the embeddings in space encode materials science knowledge such as the fact that zirconium has a hexagonal close packed (HCP) crystal structure under standard conditions and that its principal oxide is ZrO_2."

But does it work?

- (From supplement): "Skip-gram performs approximately 4% better than CBOW"
- Model Performance was measured based on (1) word prediction, (2) correlation between predictions and experimentally measured thermoelectric figures of metic from 80 materials, and (3) analogy scores for material science and grammatical analogies

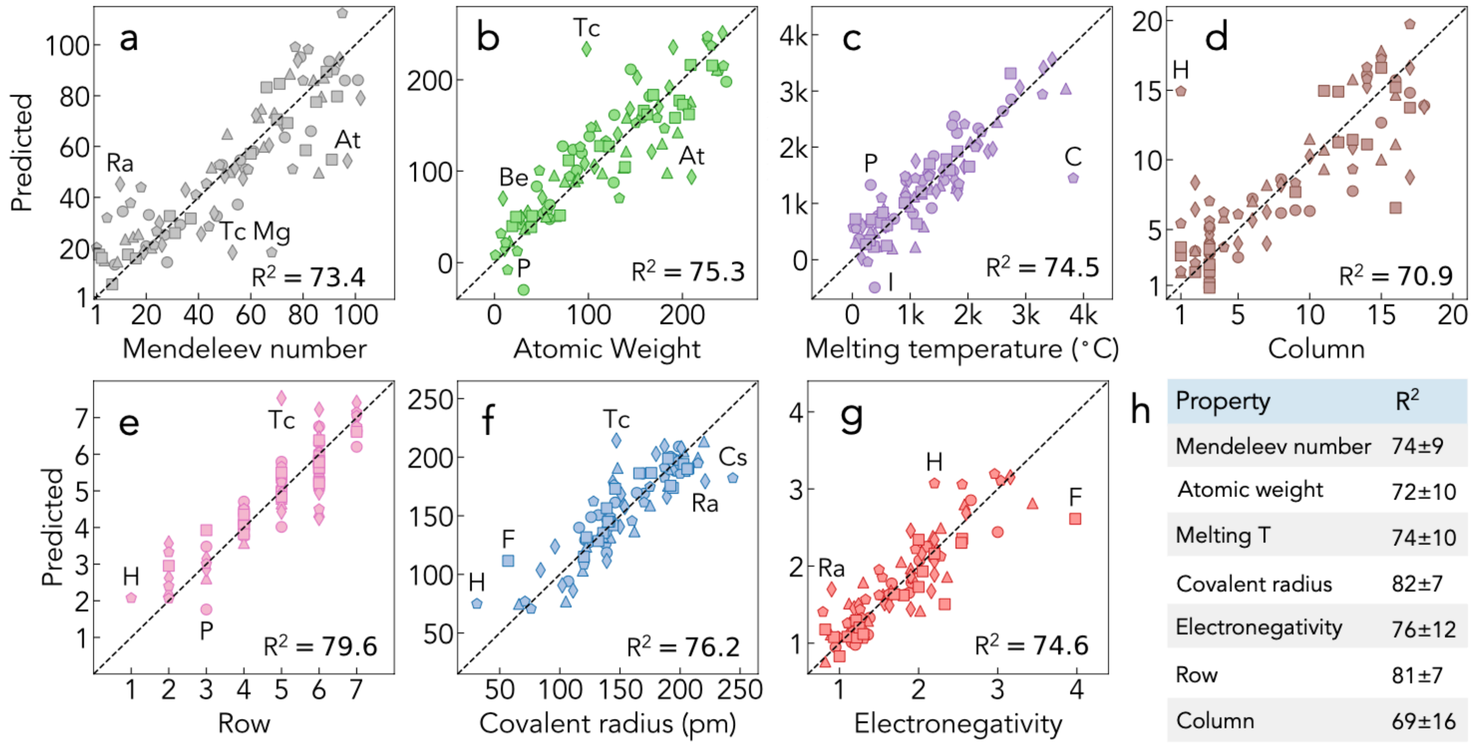


But does it work?

- 5-fold cross-validated predictions of 7 elemental properties using linear regression. The first 15 principal components of word embeddings of element names (e.g. “hydrogen”) were used as features. The 5 different shapes indicate the exact cross-validation split- ting, such that each shape (e.g. square) represents a set of validation elements predicted using the training elements represented by the 4 other shapes (e.g. triangles, diamonds, circles, pentagons). The splitting was determined randomly. h. Means and standard deviations of validation R2 scores (in percent) from 20 random 80% (training) / 20% (validation) splits

- (from supplement)

S5 Linear regression for elemental properties

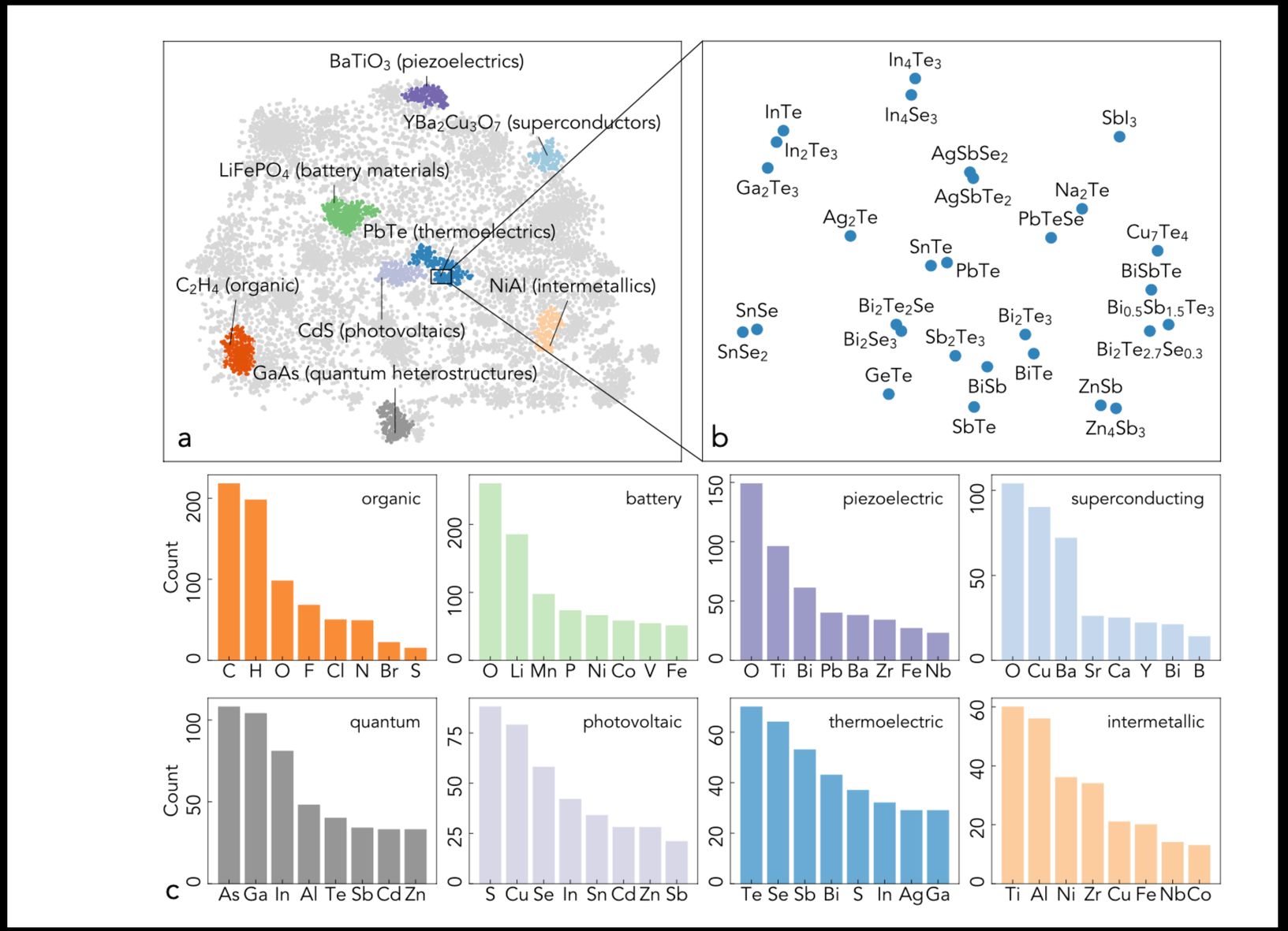


h

Property	R^2
Mendeleev number	74±9
Atomic weight	72±10
Melting T	74±10
Covalent radius	82±7
Electronegativity	76±12
Row	81±7
Column	69±16

But does it work?

- First perform 2D t-SNE on the 12,340 materials mentioned more than 10 times in the corpus
- Clustered using DBSCAN.
- Used graph to consider co-occurence graphs of material. "Most connected" materials are labelled. Barcharts: most common elements in each cluster



But does it work?

- First perform 2D t-SNE on the 12,340 materials mentioned more than 10 times in the corpus

- "We can create dynamic visualizations of material / keywords similarities by coloring each material on a 2D map according to its cosine similarity to that keyword - be that an application word (e.g. “thermoelectric”), a class of materials (e.g. “alloys”) or a crystal structure (e.g. “perovskite”). "

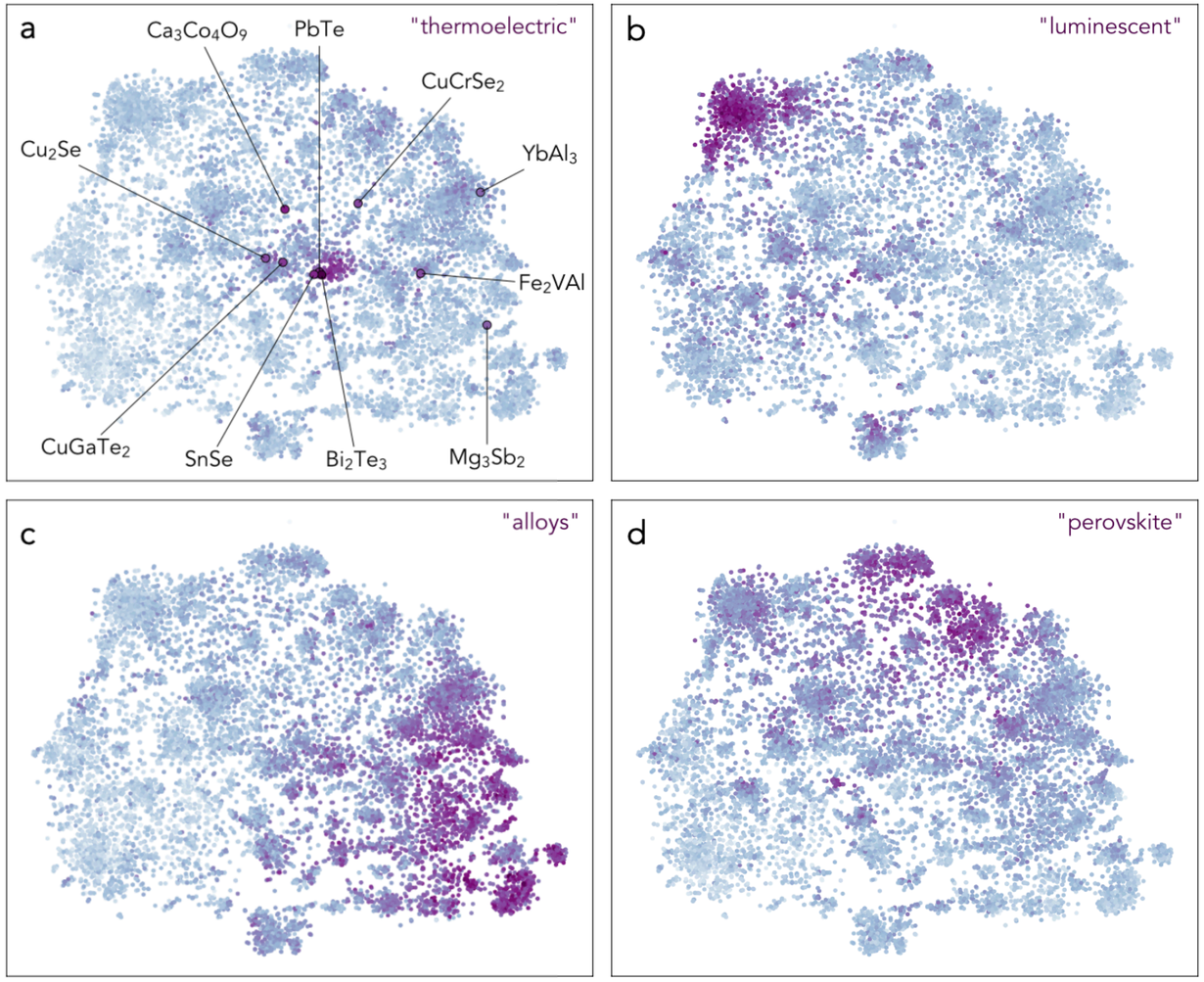
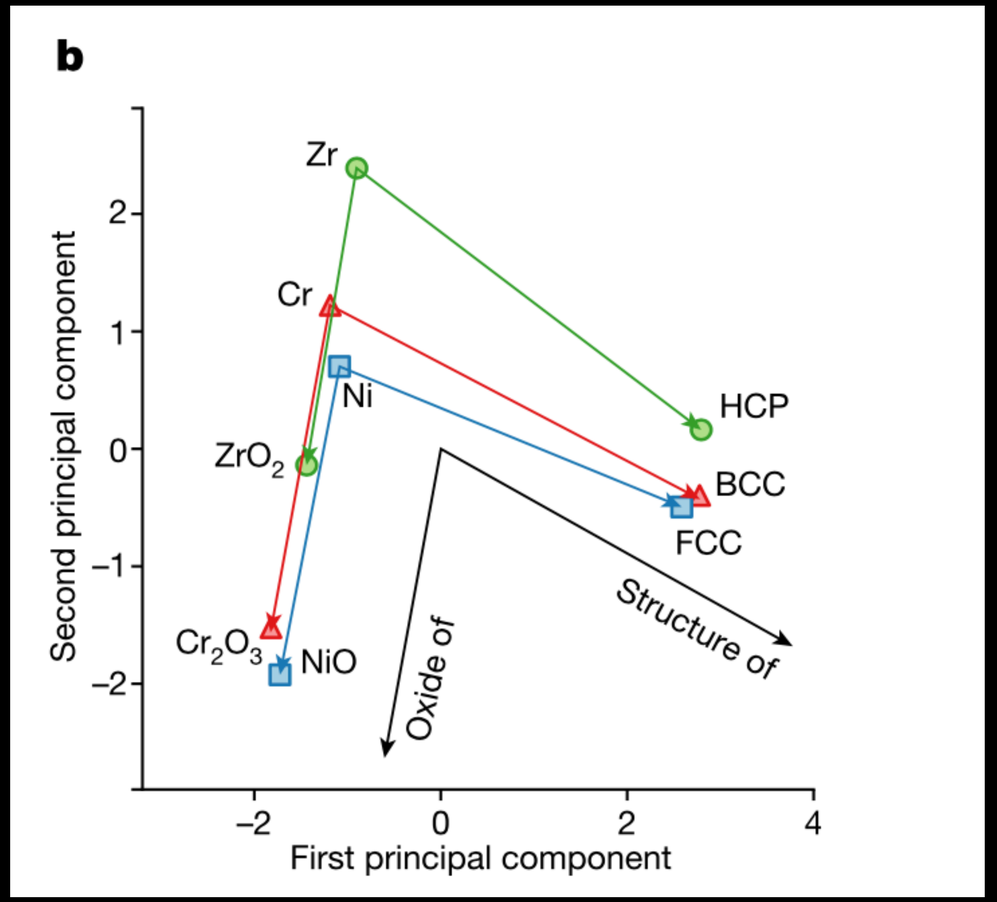


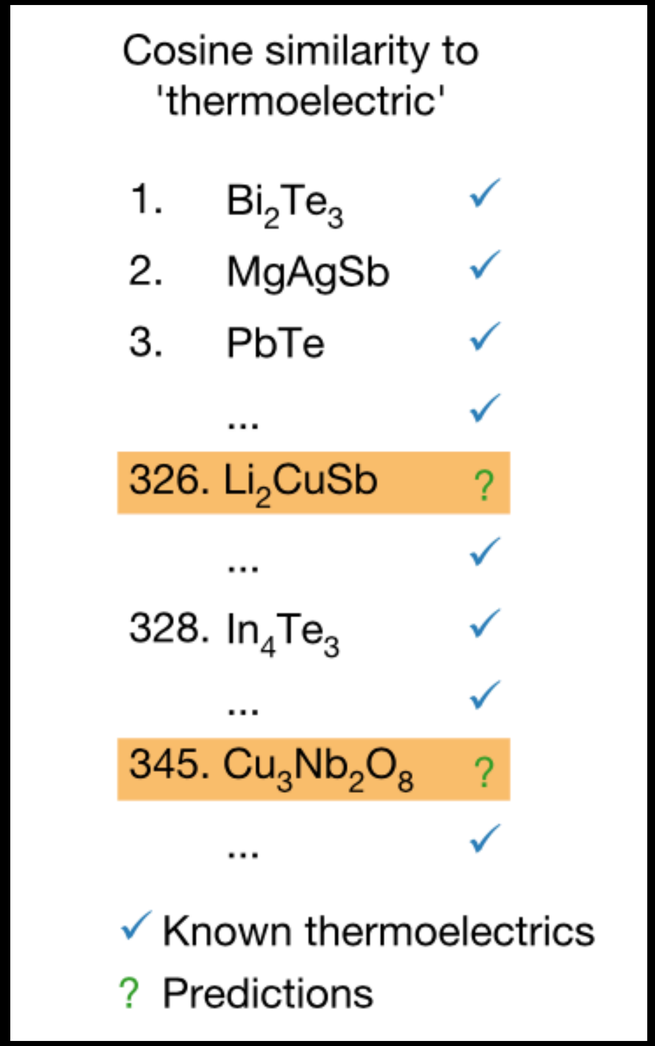
Figure S8: Dynamic material maps. Material maps highlighted according to various keywords. Darker colors correspond to more similarity.

- Figure 1b: considering a 2D PCA projection of the 200D embedding and vectors thereof



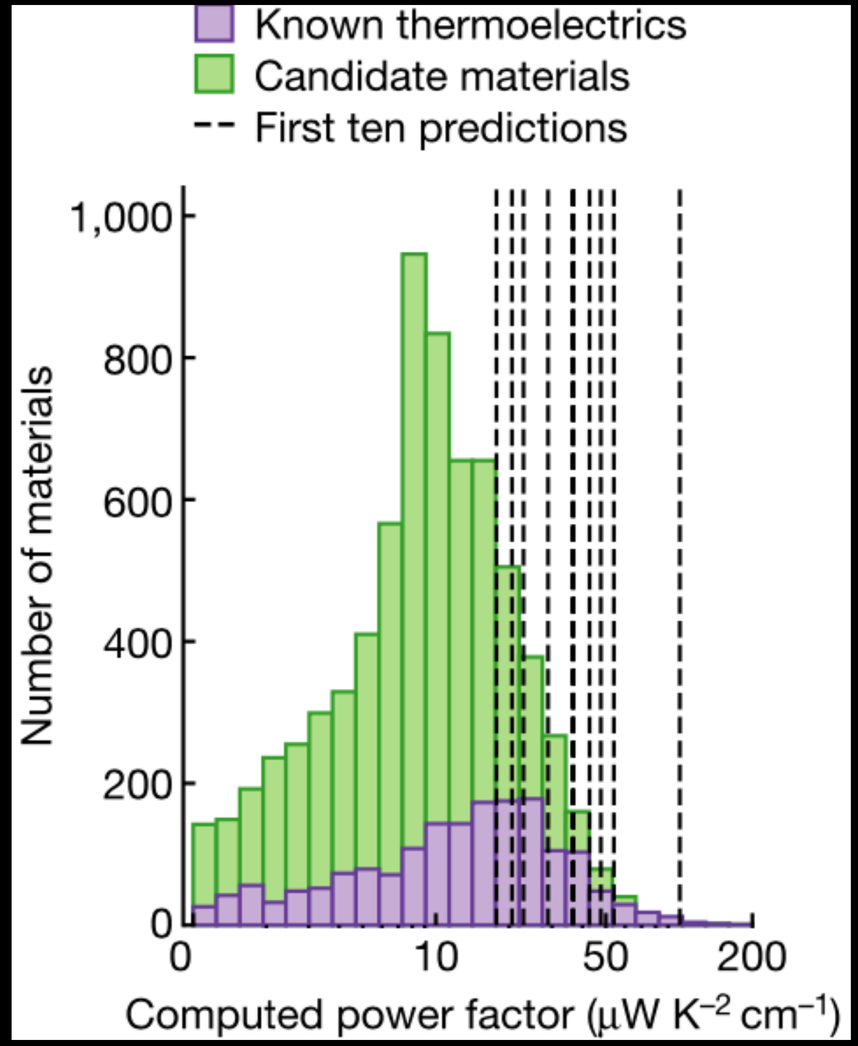
New materials

- "However, we found that a number of materials that have relatively high cosine similarities to the word ‘thermoelectric’ never appeared explicitly in the same abstract with this word, or any other words that unequivocally identify materials as thermoelectric… Rather than dismissing these instances as spurious, we investigated whether such cases could be usefully interpreted as predictions of novel thermoelectric materials. "



New materials

- "We find that the top ten predictions all exhibit computed power factors significantly greater than the average of candidate materials (green), and even slightly higher than the average of known thermoelectrics (purple) … Moreover, the three highest power factors from the top ten predictions are at the 99.6th, 96.5th and 95.3rd percentiles of known thermoelectrics. "

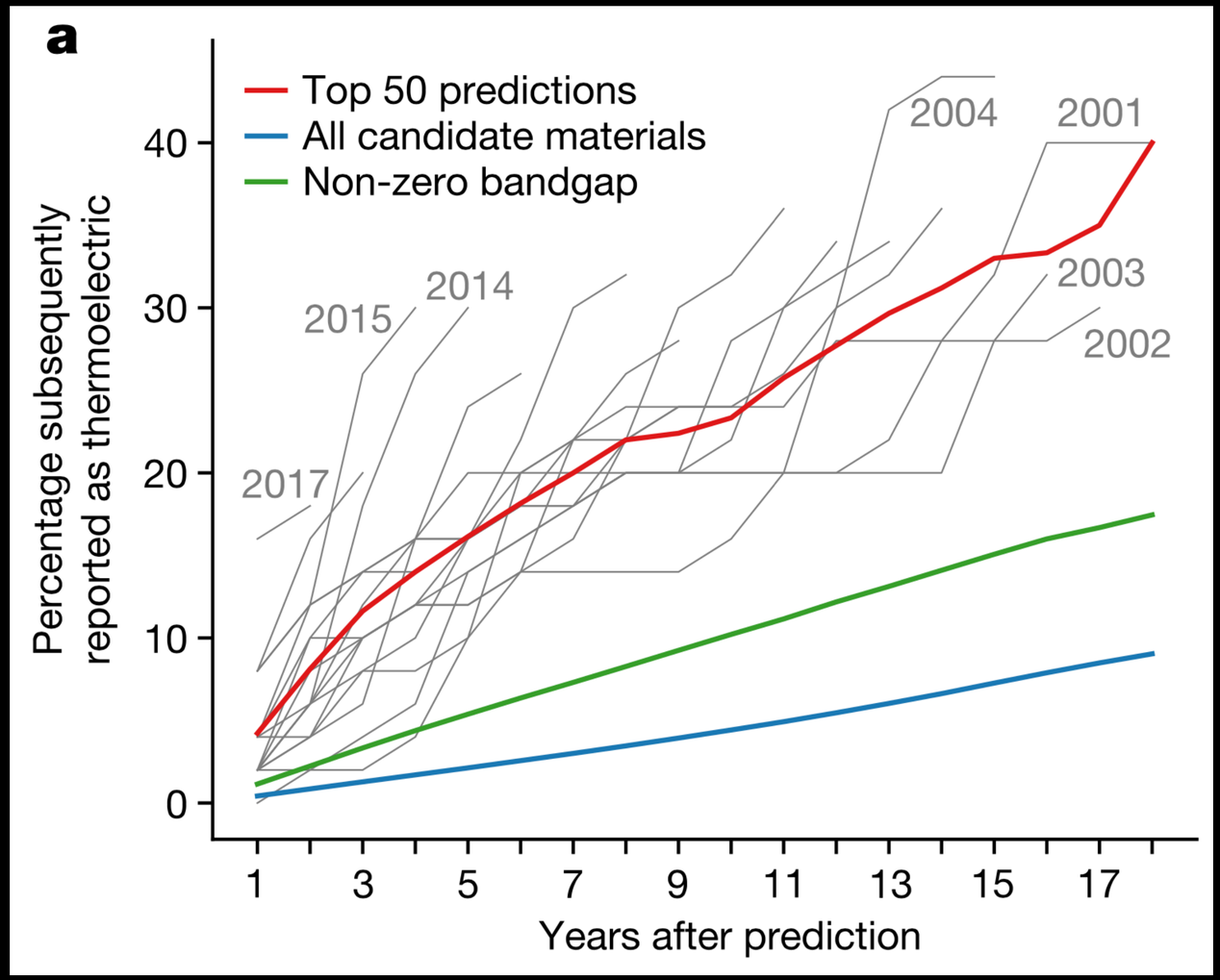


New materials

- "Finally, we tested whether our model —if trained at various points in the past—would have correctly predicted thermoelectric materials reported later in the literature. Specifically, we generated 18 different ‘historical’ text corpora consisting only of abstracts published before cutoff years between 2001 and 2018."

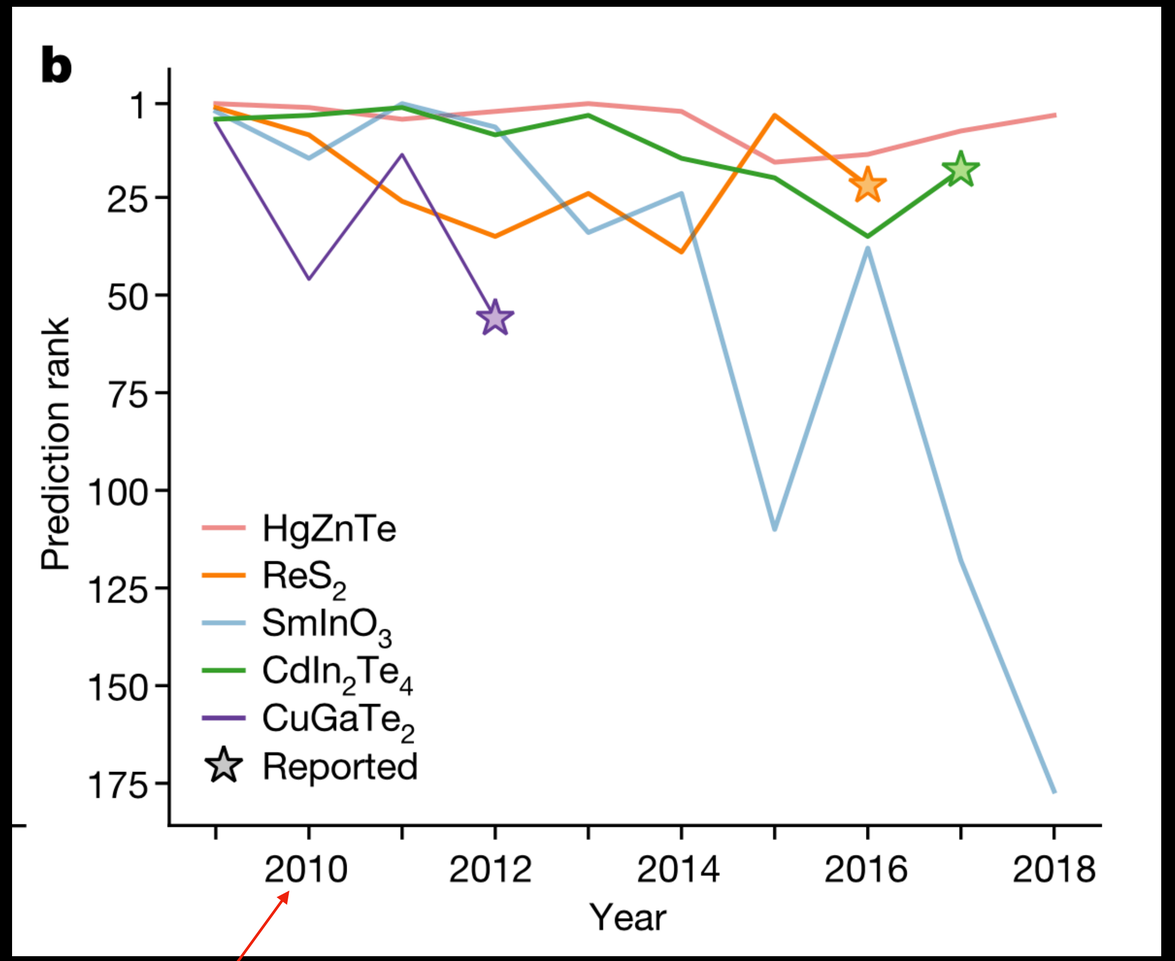
- Each was trained based only on the historical dataset.

- Blue: random material from dataset, Green: random material from non-zero density functional theory bandgap



Viewing / Predicting the future from before "My World"

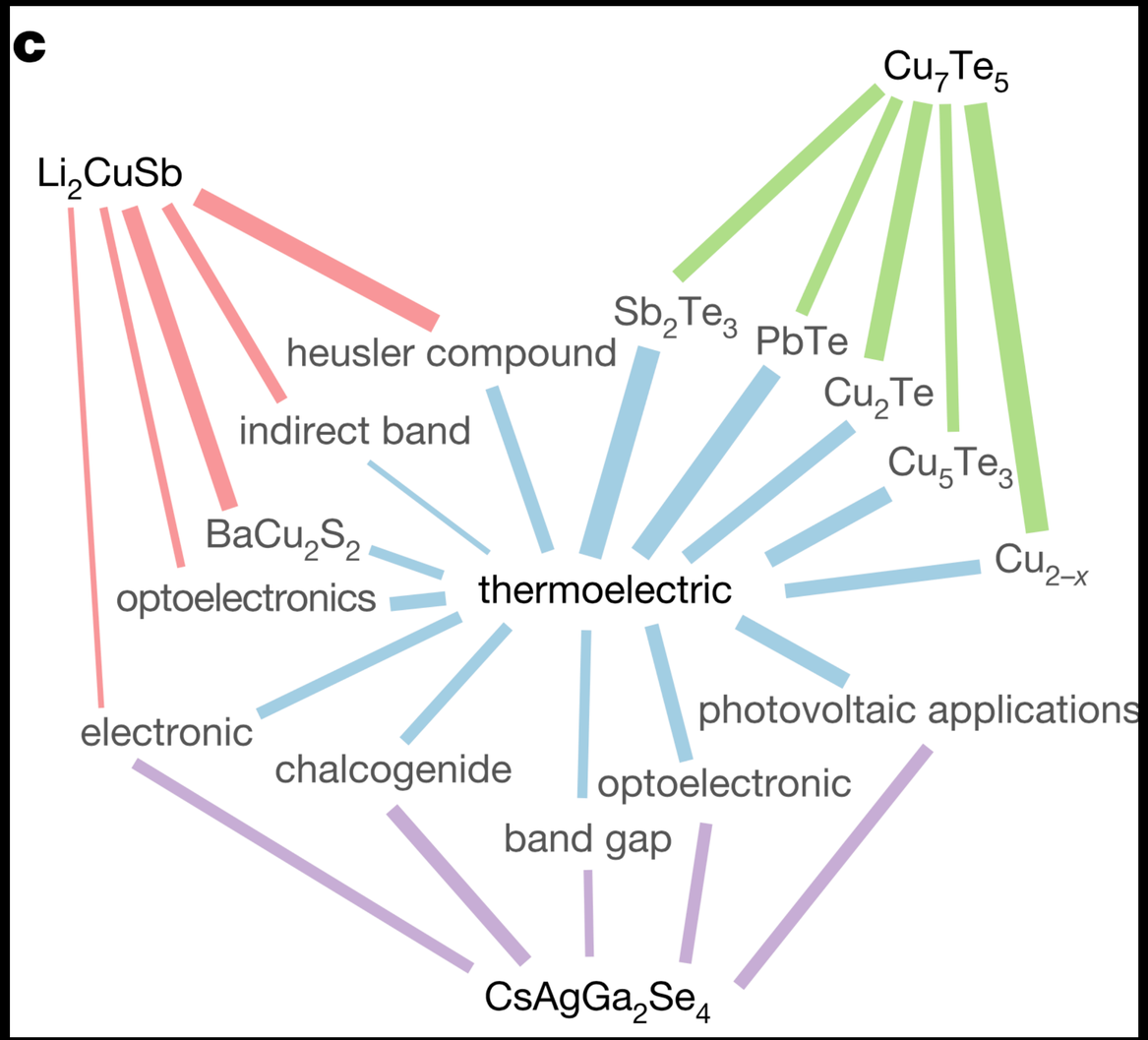
- Select top 5 predicted compounds based on pre-2009 data.
- Track the material prediction rank evolution over time
- CuGaTe2 was published in 2012 "one of the best present-day thermoelectrics"



My World EP released: Nov 17, 2009
My World 2.0 released: Mar 19, 2010

Interpretability of the model

- Li2CuSb doesn’t appear near "thermoelectric" but has a high prediction.
- Line thickness is proportional to cosine similarity between words.



Focused model:

- "The success of our unsupervised approach can partly be attributed to the choice of the training corpus. The main purpose of abstracts is to communicate information in a concise and straightforward manner, avoiding unnecessary words that may increase noise in embeddings during training."

Extended Data Table 4 Importance of the text corpus

From: Unsupervised word embeddings capture latent knowledge from materials science literature

Text corpus	Materials	Grammar	All	Corpus size
Wikipedia	2.6	72.8	51.0	2.81B words
Wikipedia elements	2.7	72.1	41.4	1.08B words
Wikipedia materials	2.2	72.8	41.3	781M words
All abstracts	43.3	58.3	51.0	643M words
Relevant abstracts	48.9	54.9	52.0	290M words
Pre-trained model from Kim et al[39]	10.4	47.1	30.8	640k papers

Next class(es)

- Movie day! Upload your video
- Be here to answer questions. If you can’t be here in person, let me know.