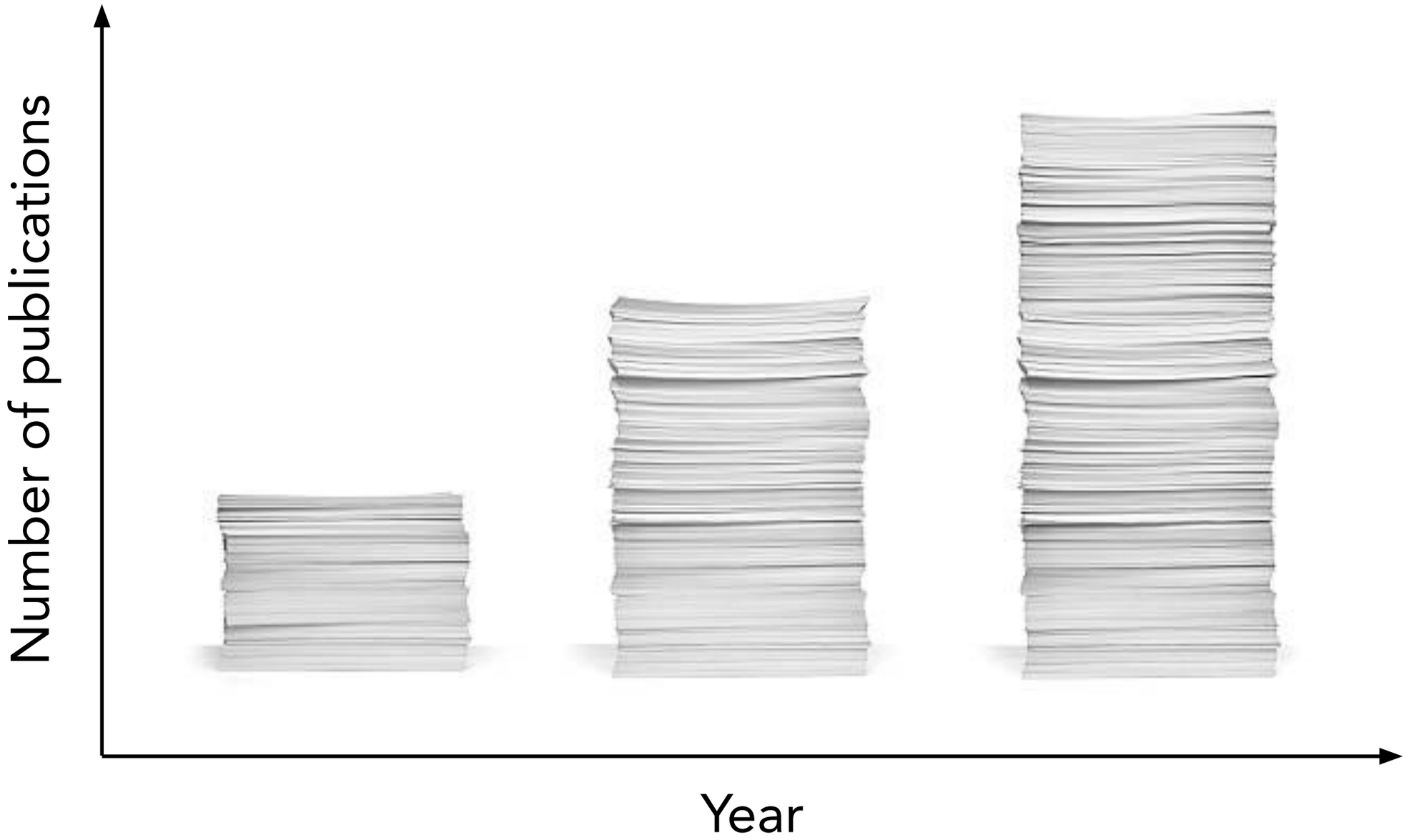
Learning Materials Science from Text

Vahe Tshitoyan
Machine Learning @ Goolge ATAP*

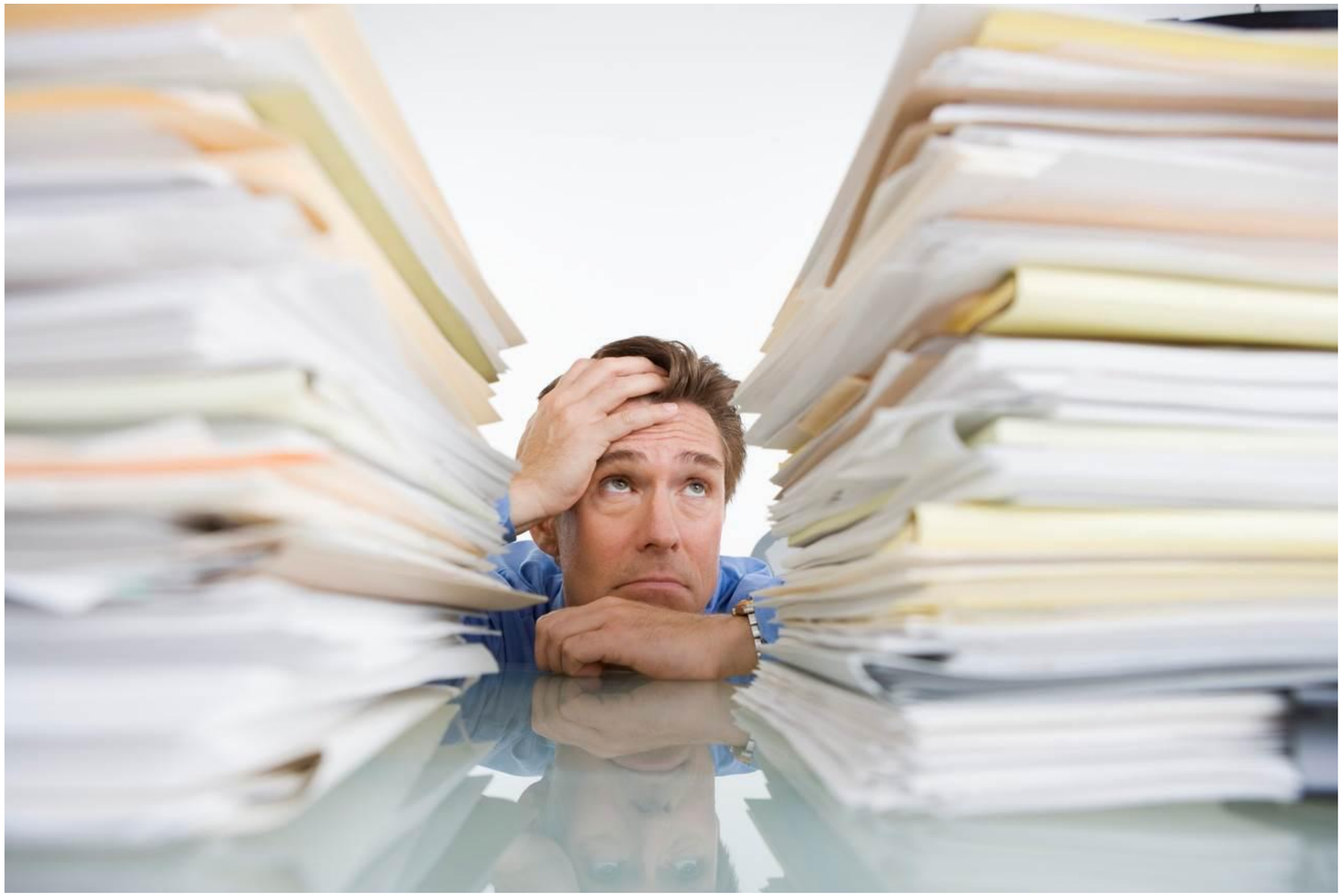
December 14, 2019

* The presented work was carried out while I was a Postdoc @ Berkeley Lab

Science is in text



Problem



Solution

Computers speak in numbers

Word vectors (embeddings)

Map each word to a vector in space (sequence of numbers)

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

	Femininity	Youth	Royalty
Man	0	0	0
Woman	1	0	0
Boy	0	1	0
Girl	1	1	0
Prince	0	1	1
Princess	1	1	1
Queen	1	0	1
King	0	0	1
Monarch	0.5	0.5	1

Each word gets a 1x3 vector

Similar words… similar vectors

Vector algebra with embeddings

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch

	Femininity	Youth	Royalty
Man	0	0	0
Woman	1	0	0
Boy	0	1	0
Girl	1	1	0
Prince	0	1	1
Princess	1	1	1
Queen	1	0	1
King	0	0	1
Monarch	0.5	0.5	1

Each word gets a 1x3 vector

Similar words… similar vectors

- Words with similar meanings have similar vectors

Also…

- queen - king = woman - man = girl - boy
- (prince + princess + queen + king)/4 = monarch

How do we "machine learn" word vectors?

Word2Vec: skip-gram

Sentence: Electrochemical properties of LiCoO2 thin film.

Task: train a neural network to predict the context (target) word given the centre word

Training example 1:

of →? Electrochemical

of →? properties

of →? LiCoO2

of →? thin

Word2Vec: skip-gram

Sentence: Electrochemical properties of LiCoO2 thin film.

Task: train a neural network to predict the context (target) word given the centre word

Training example 2:

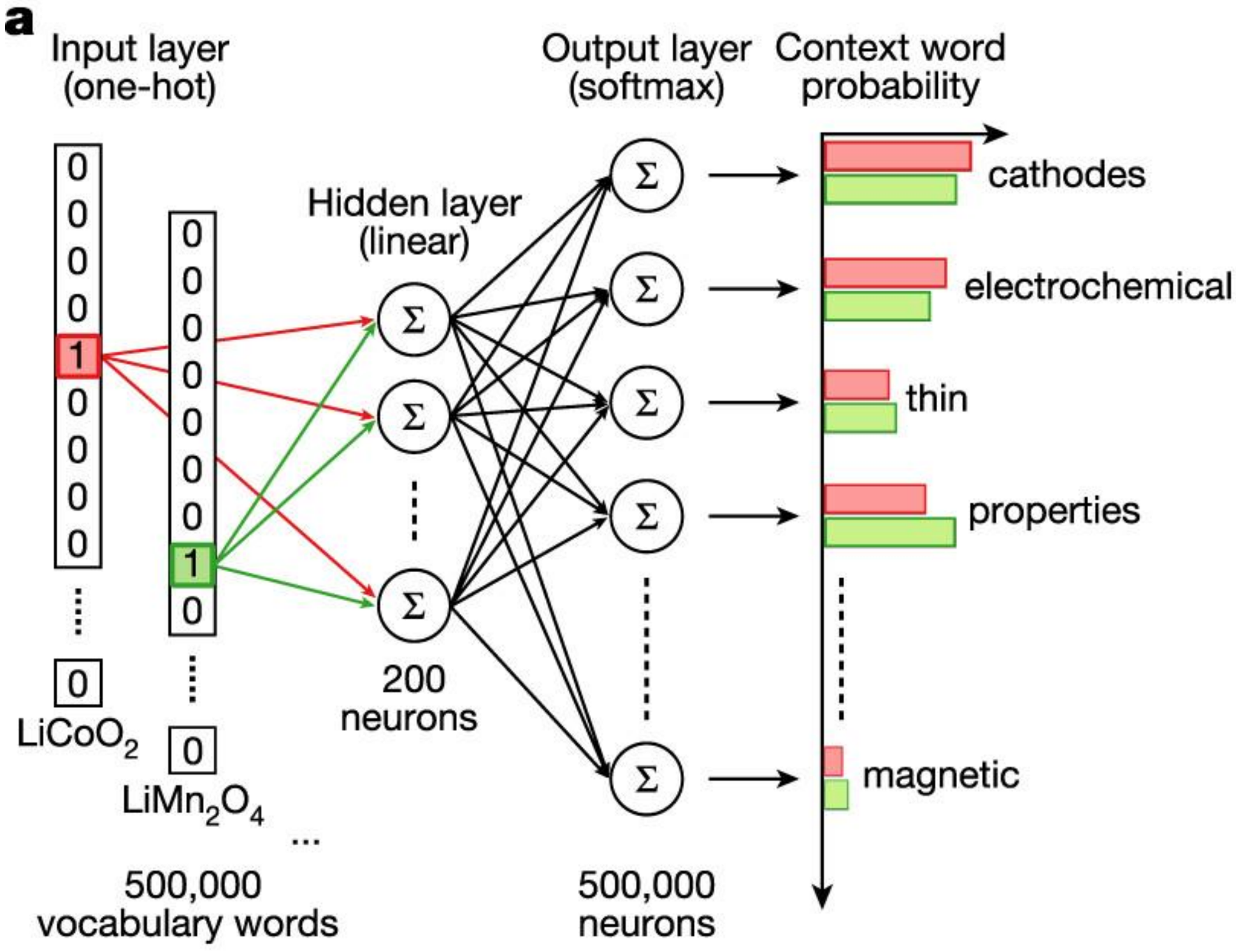
LiCoO2 →? properties

LiCoO2 →? of

LiCoO2 →? thin

LiCoO2 →? film

The Neural Network



Word2Vec: skip-gram

Sentence: Electrochemical properties of LiCoO2 thin film.

Task: train a neural network to predict the context (target) word given the centre word

Training example 2:

LiCoO2 →? properties

LiCoO2 →? of

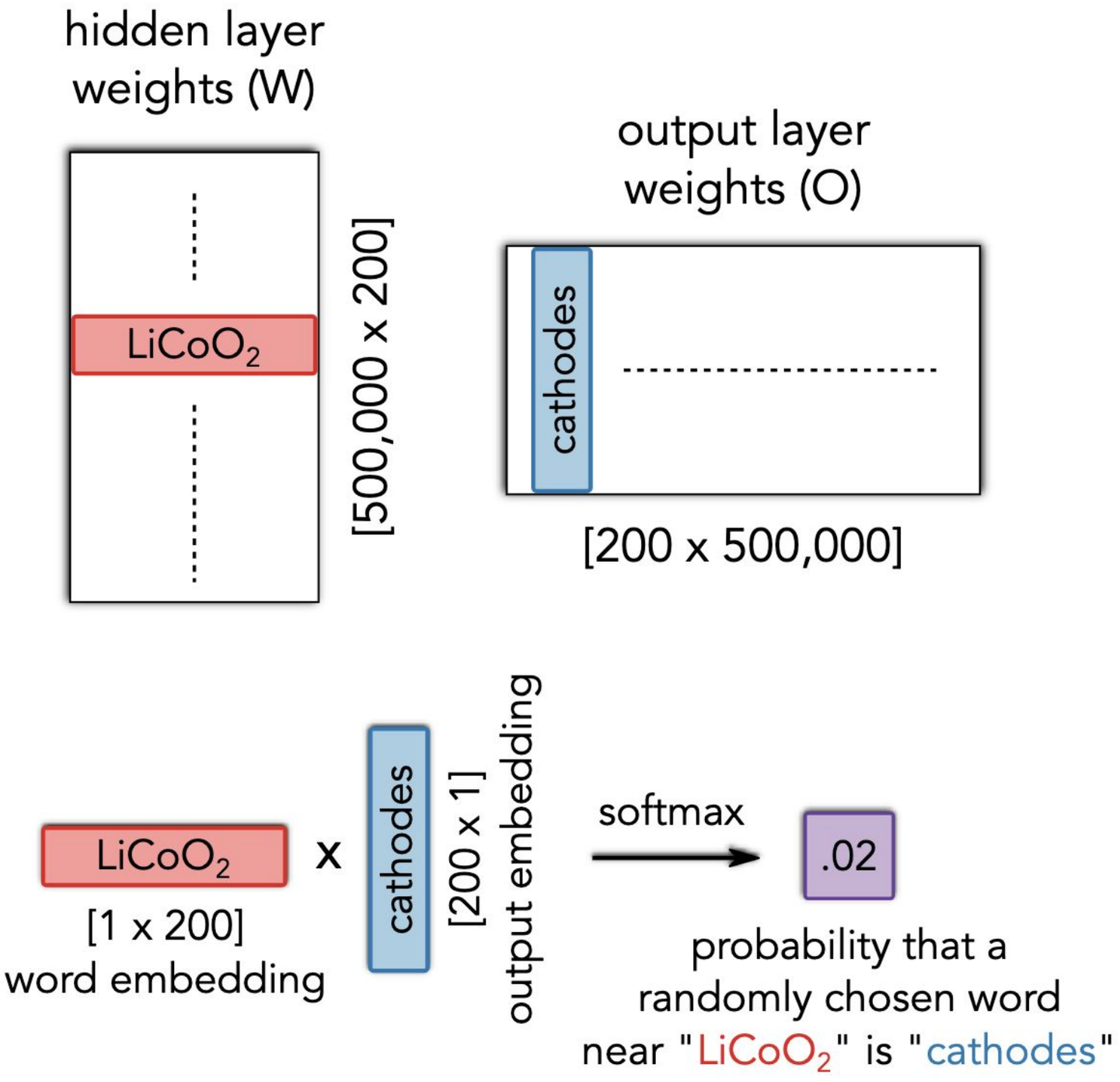
LiCoO2 →? thin

LiCoO2 →? film

Objective: learn vector representations for words

If two different words have similar contexts, then the neural network needs to output similar results for these two words. The network will output similar context predictions for these two words if the word vectors are similar.

Outcomes of the training



Training Data

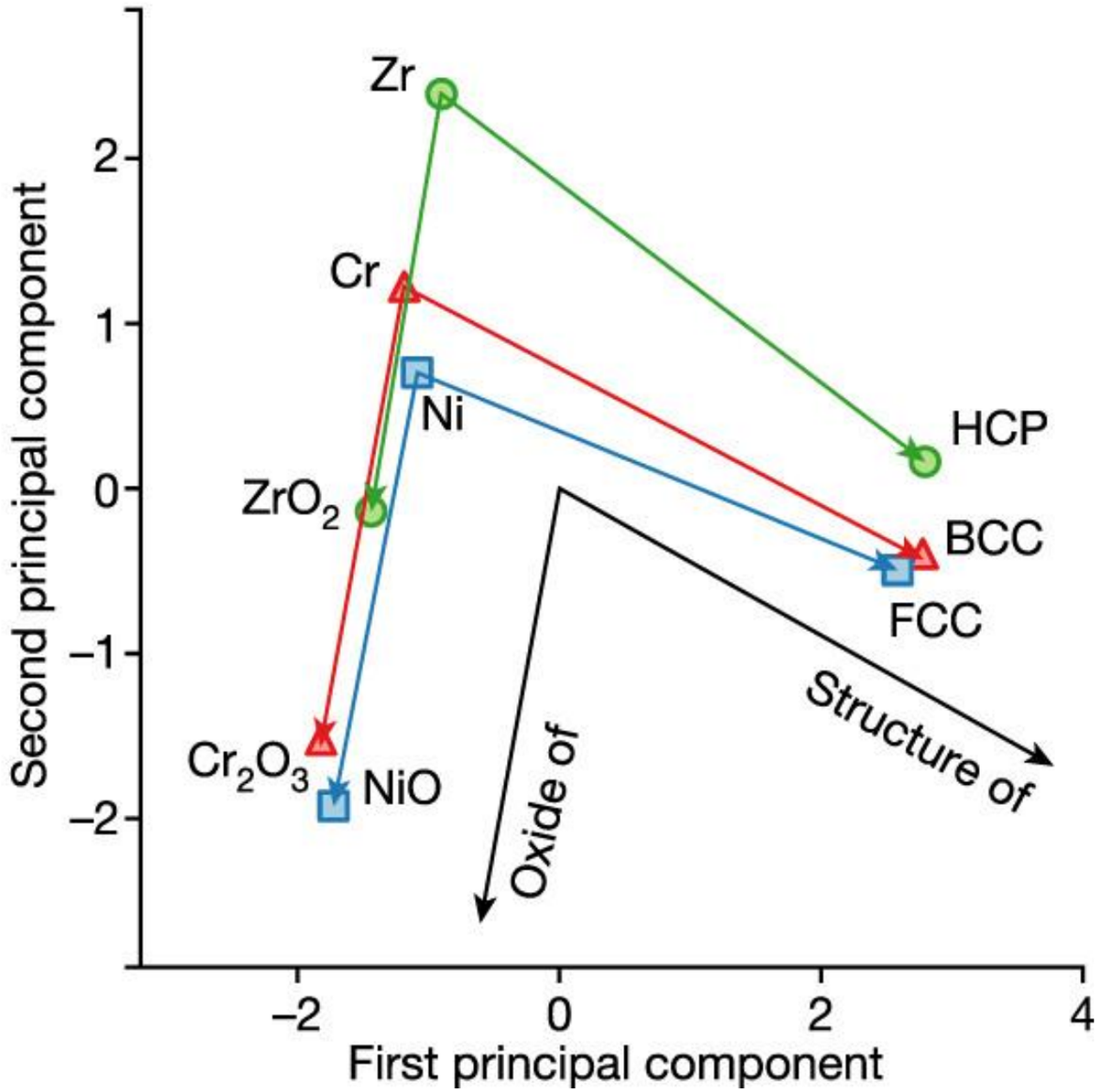
- A total of 3.3 million abstracts downloaded
- 1.5 million were classified as inorganic materials science
  - ~300 million words across the abstracts
  - ~500k unique vocabulary words and phrases
  - ~100k unique material formulae

Word similarity

Words most similar to ___ are?

- **LiCoO2**: LiMn2O4, LiNi0.5Mn1.5O4, LiNi0.8Co0.2O2, LiNi0.8Co0.15Al0.05O2
- **ferromagnetic**: ferrimagnetic, antiferromagnetic, anti-ferromagnetic, paramagnetic

Analogies



CoO - Co + Al = Al2O3

Li - lithium + helium = He

temperature - K + Pa = pressure

LiCoO2 - cathode + anode = graphite

A famous vector representation in chemistry?

The periodic table of elements

(0, 0)

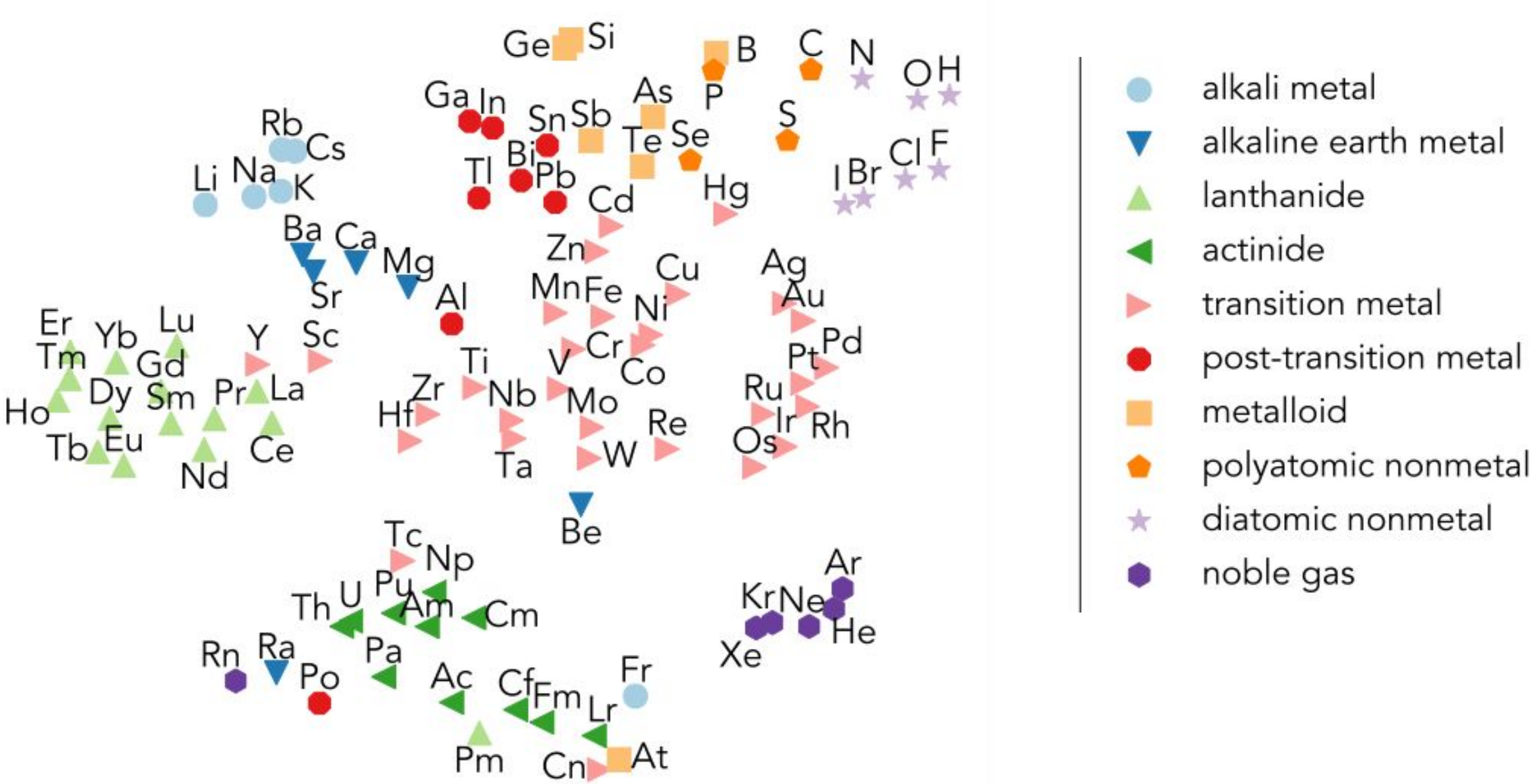
2D representation :

H - (1, 1)
Li - (1, 2)
…

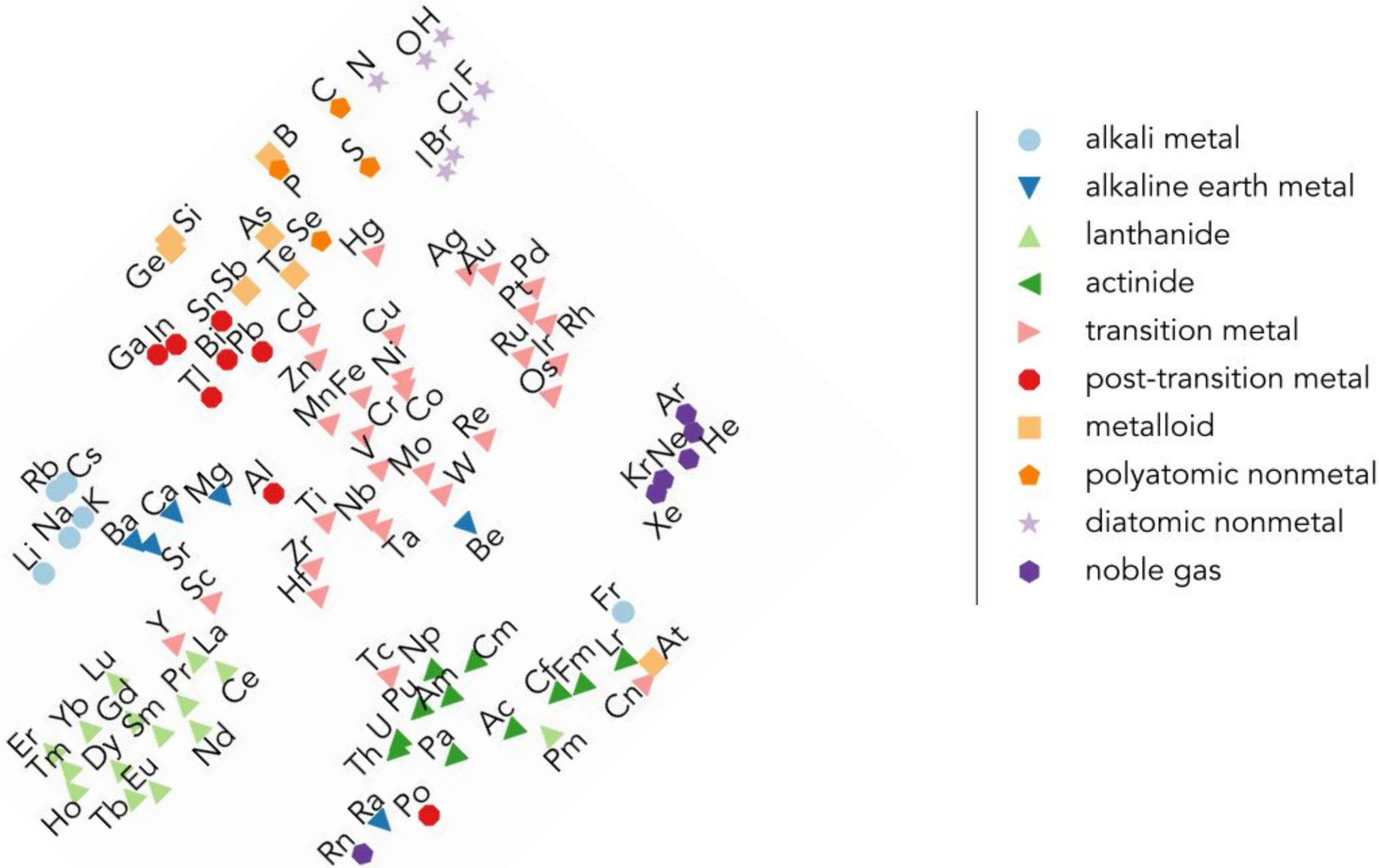
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba		Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra		Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	Lv	Ts	Og

La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

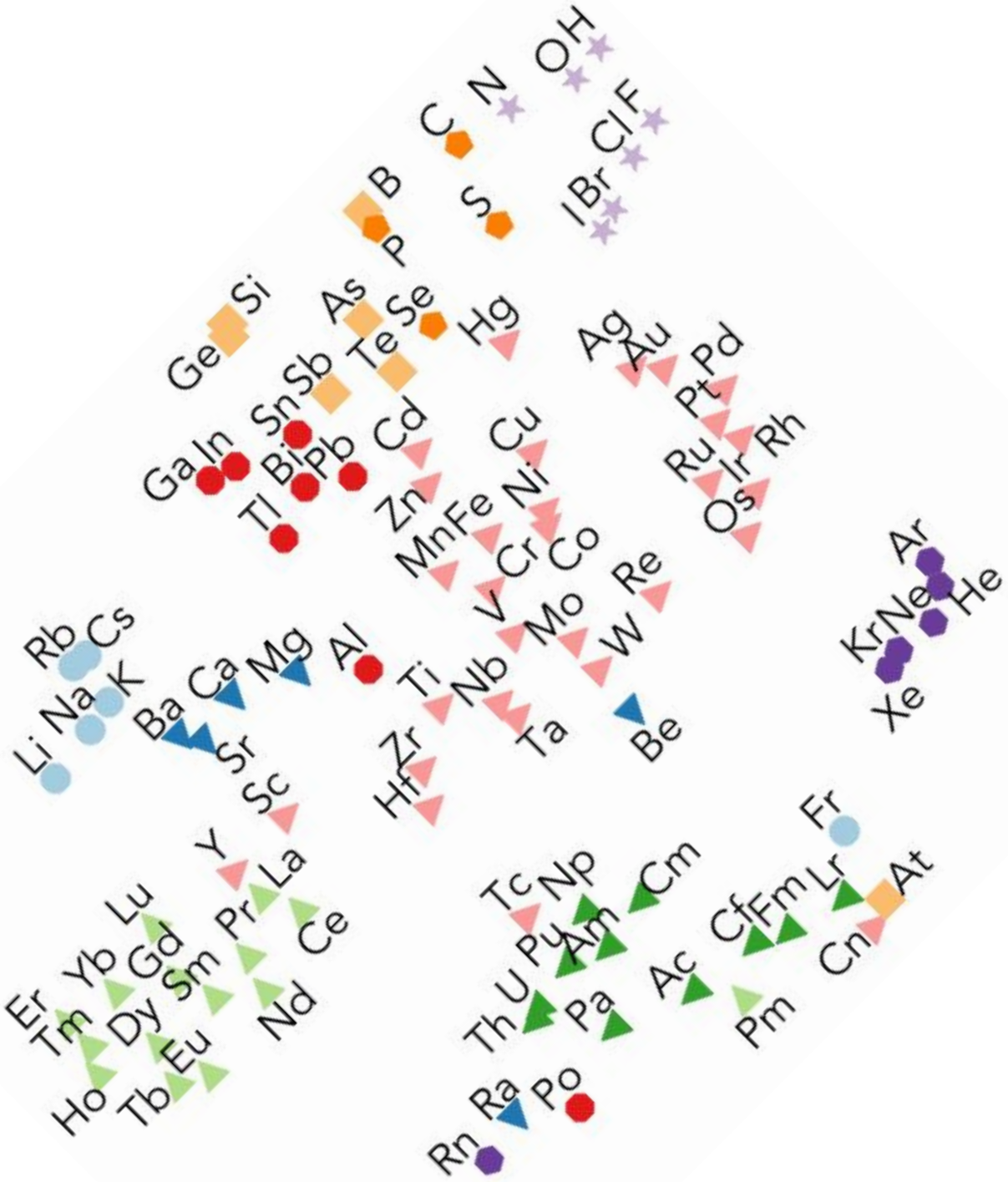
Word vectors of elements in 2D

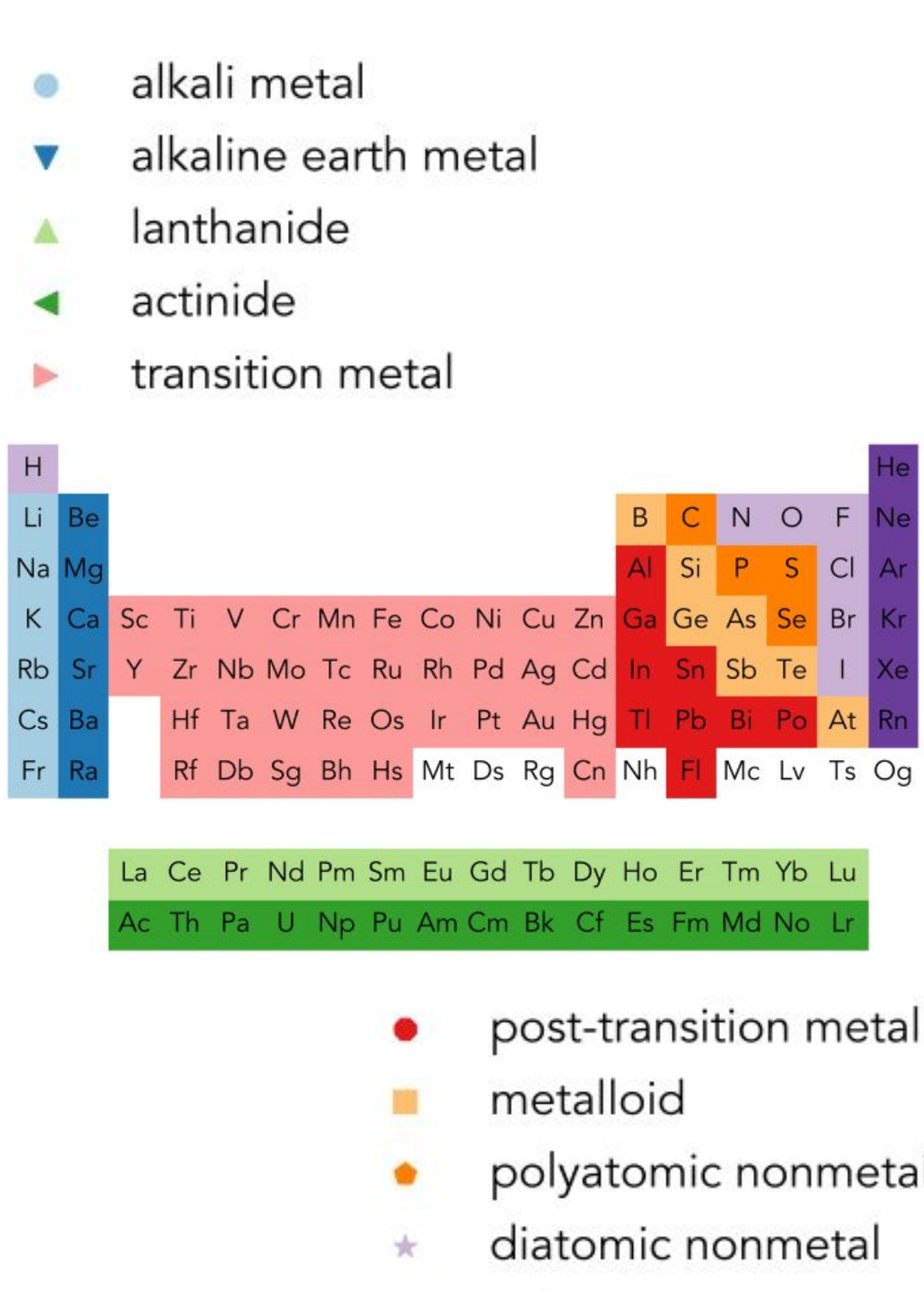


Word vectors of elements in 2D

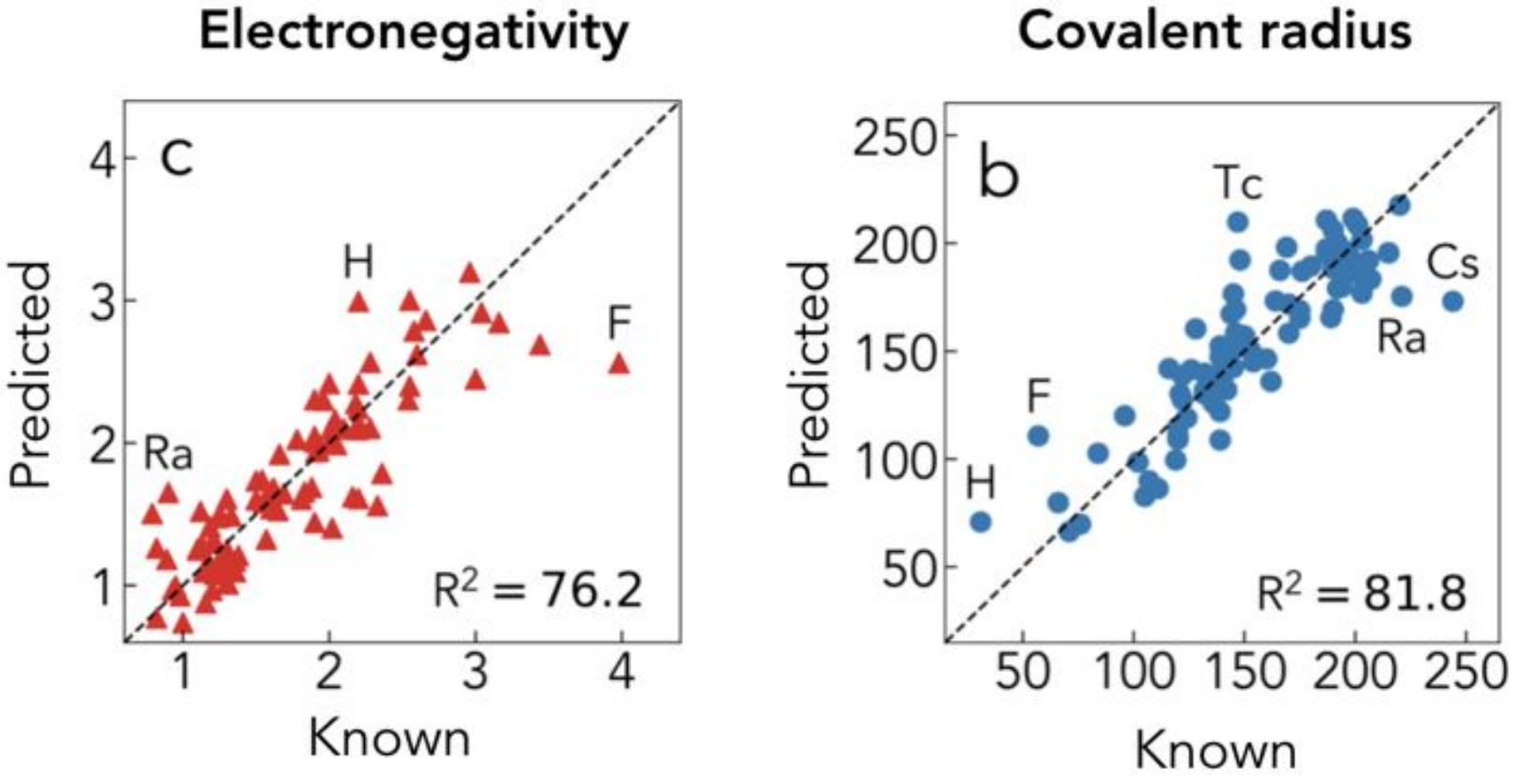


Word vectors of elements in 2D





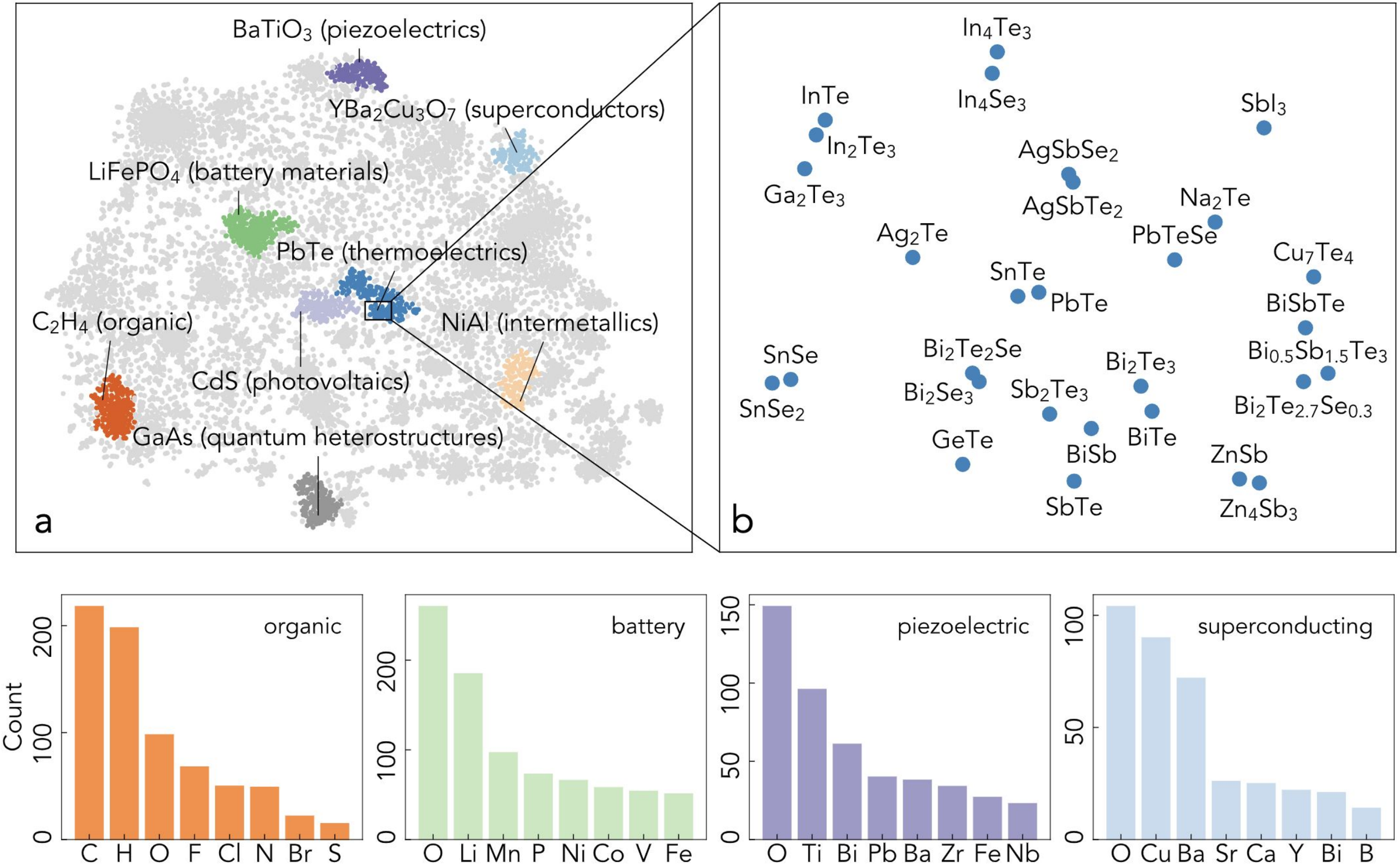
Directions in latent latent space correspond to elemental properties



Property	R^2
Mendeleev number	74.4
Atomic weight	72.4
Melting T	73.9
Covalent radius	81.8
Electronegativity	76.2
Row	80.7
Column	69.2

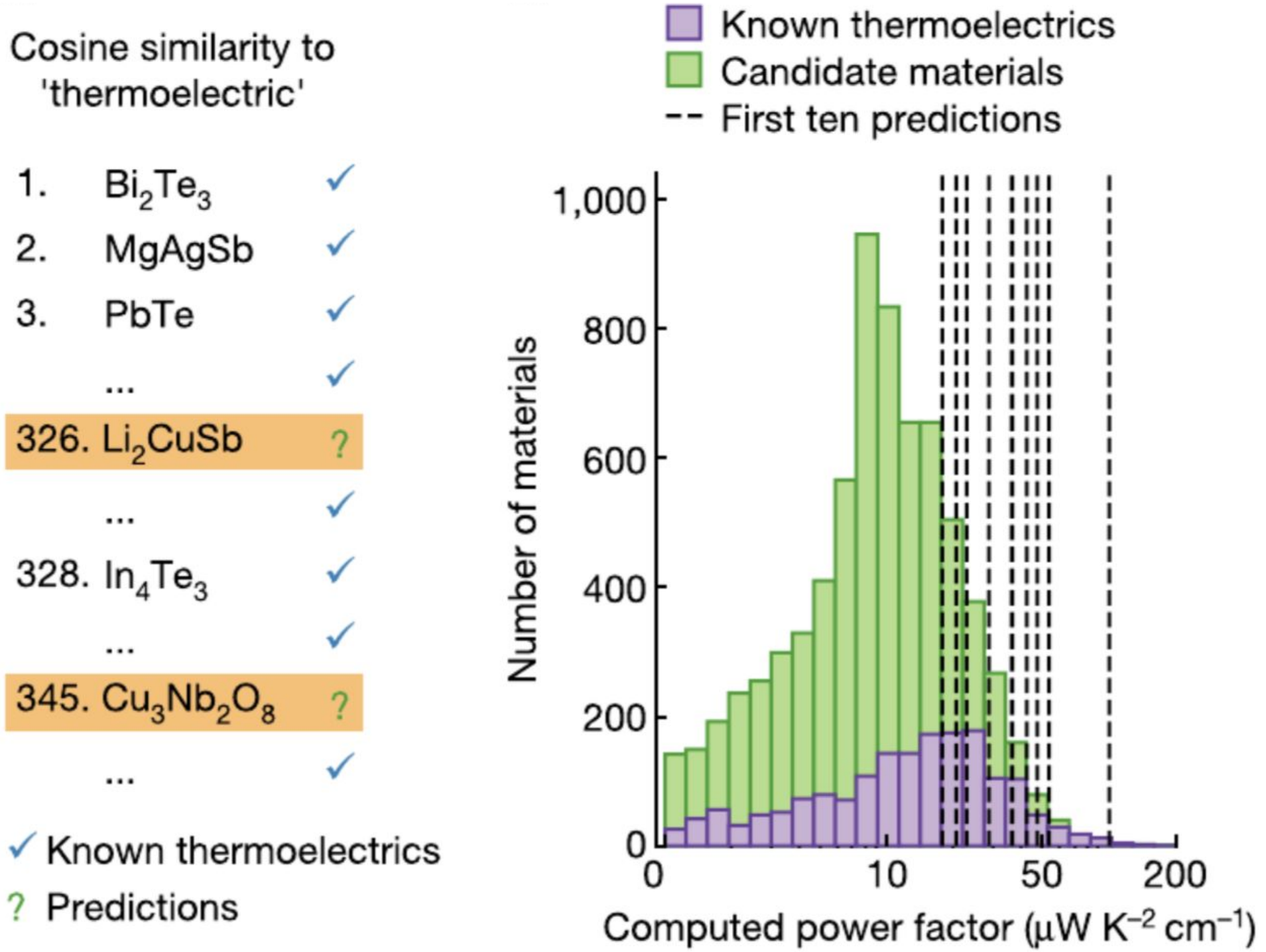
Basic linear regression models fit using only the unsupervised, text-derived embeddings as feature vectors.

Similar materials cluster together



...what is the most similar material to the word "thermoelectric"?

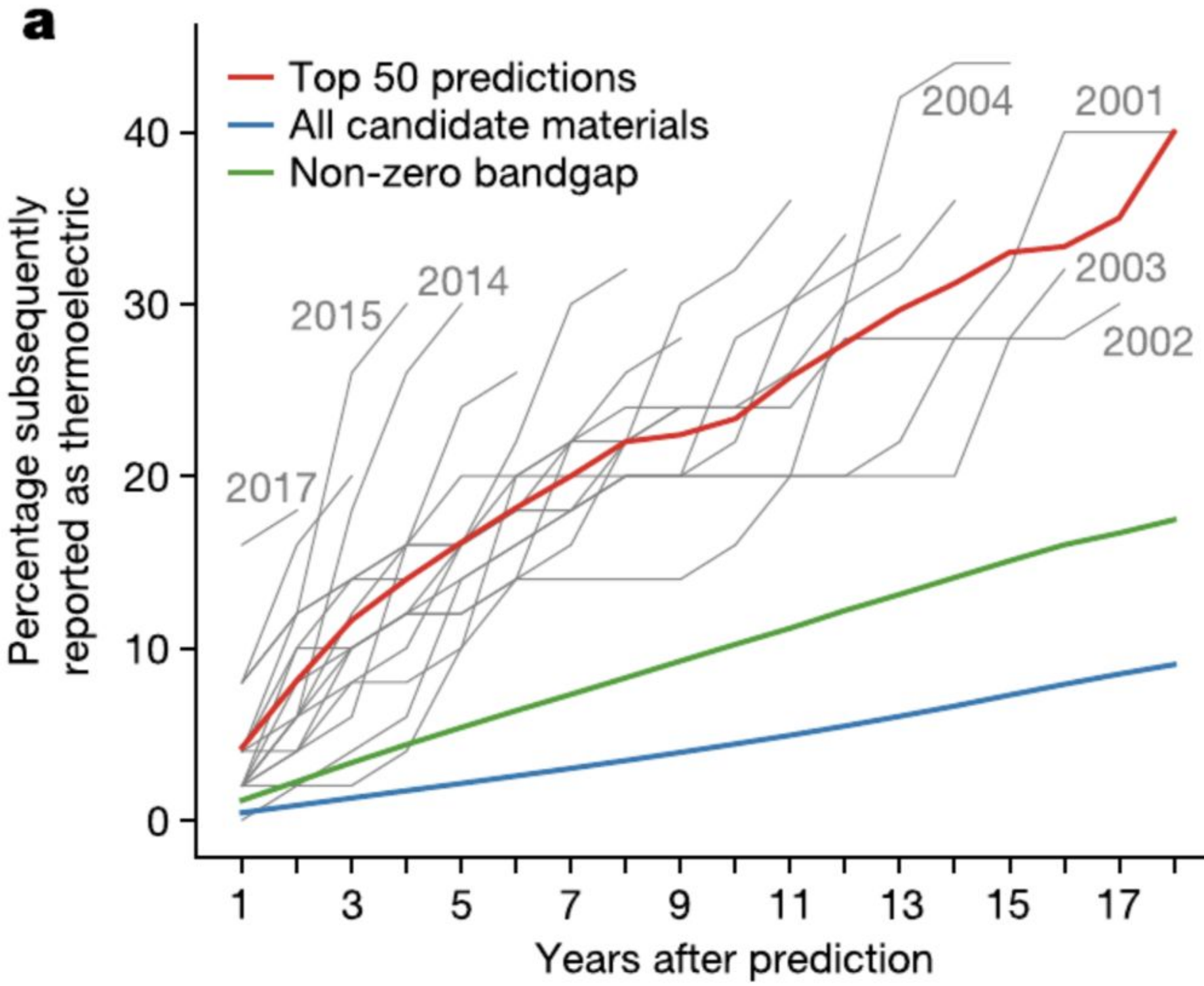
Predictions of new Thermoelectrics



We can examine the plausibility of these predictions by performing experiments “in the past”.

(i.e. remove all abstracts published after a given year, re-train the model, and rank candidates.)

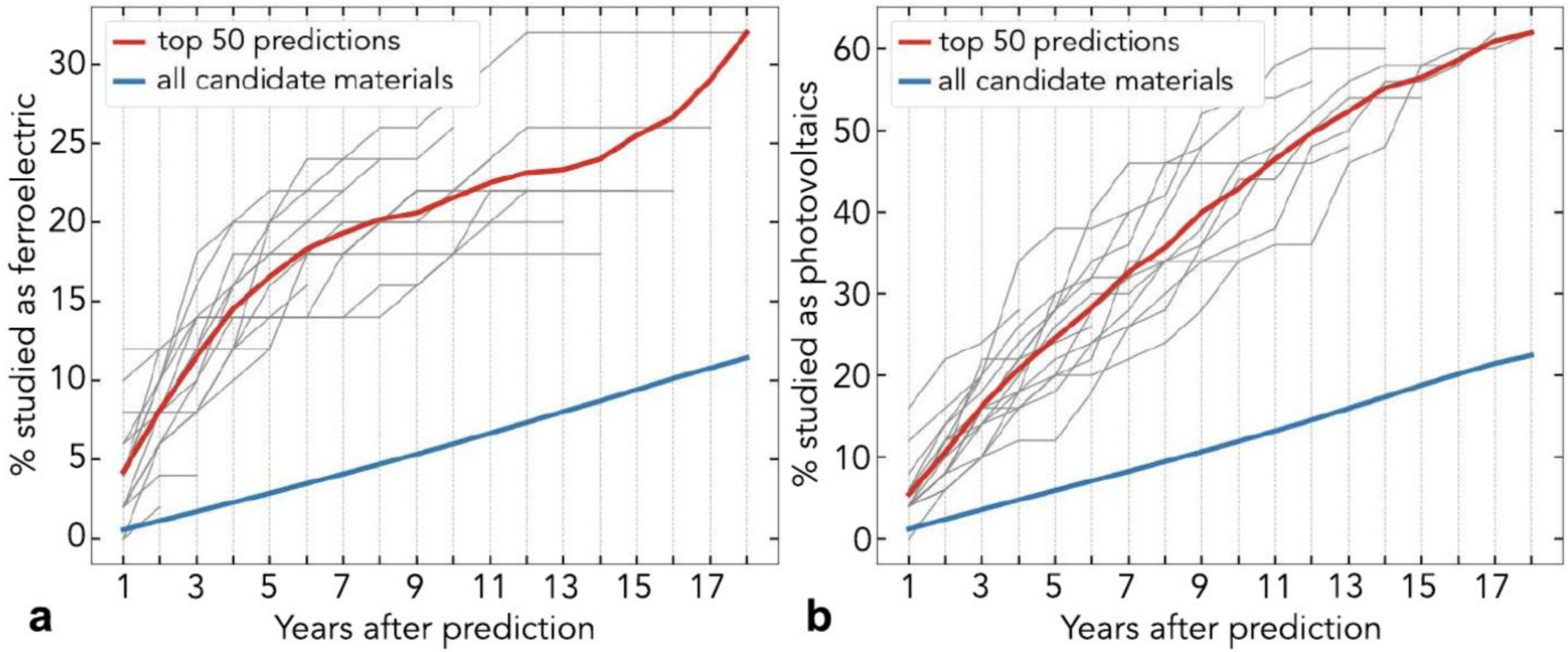
Historical validation



Changing the application is just a matter of changing the query words

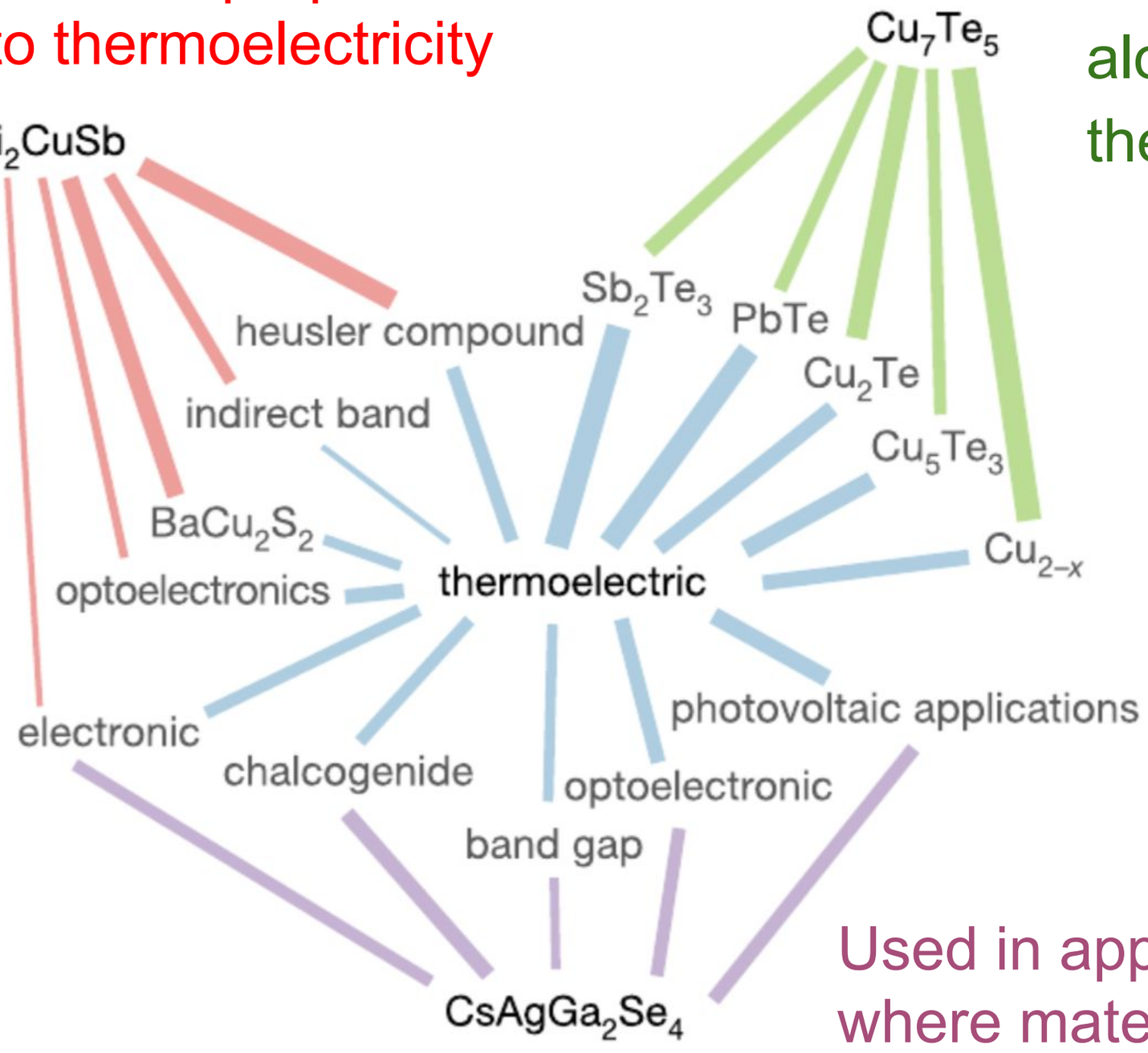
"ferroelectric"

"photovoltaics"



Explaining the predictions

Associated with properties related to thermoelectricity



Commonly referenced alongside known thermoelectrics

Used in applications where materials are often also thermoelectric

MATSCHOLAR

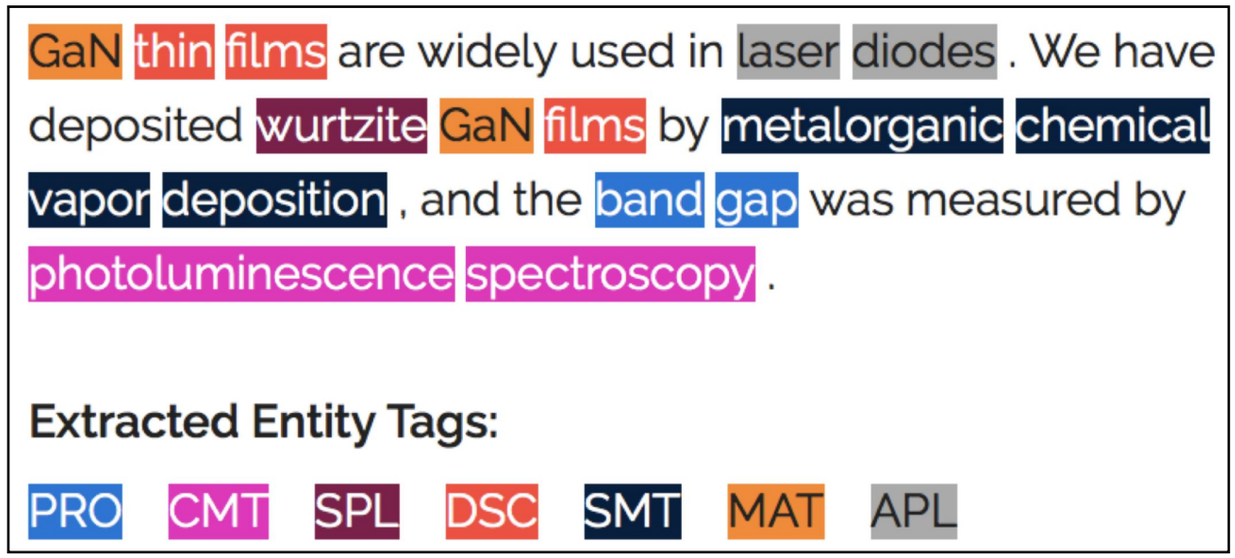
BETA

https://www.matscholar.com

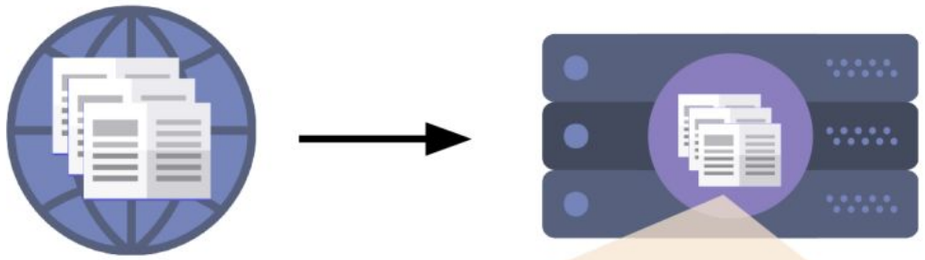
Database of all published results

Information retrieval from literature:

- Material (MAT)
- Material descriptor (DSC)
- Symmetry / Phase (SPL)
- Property (PRO)
- Application (APL)
- Synthesis method (SMT)
- Characterization method (CMT)



1. *Abstracts collected and stored in Matscholar corpus*



"Olivine LiFePO4 is a promising candidate cathode material for lithium-ion batteries..."

2. *Tokenization*

Olivine	LiFePO4	is	a	promising	candidate	cathode
material	for	lithium	ion	batteries		

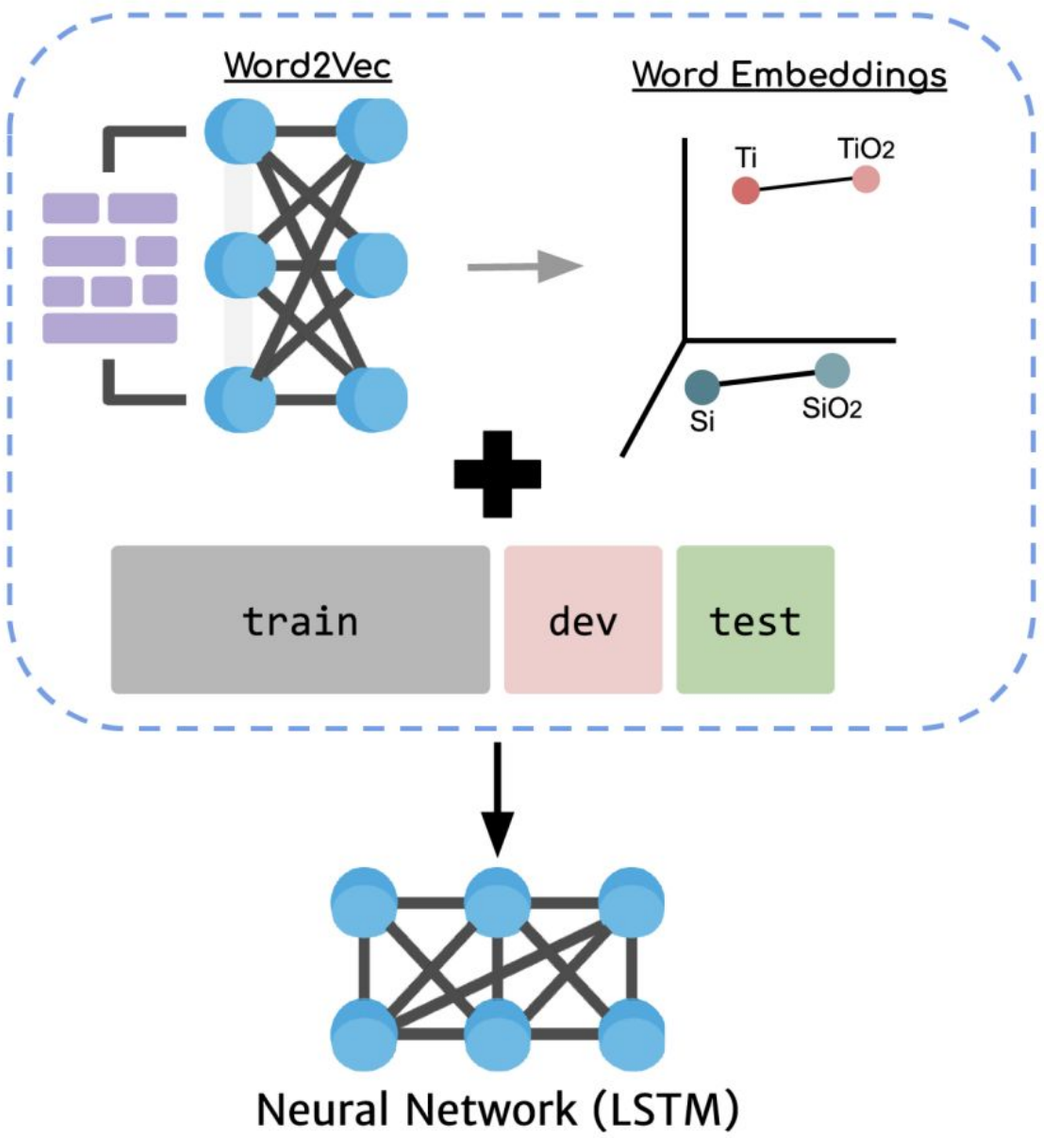
3. *Labeling*

	SPL	MAT					APL
	Olivine	LiFePO4	is	a	promising	candidate	cathode
			APL	APL			
material	for		lithium	ion	batteries		

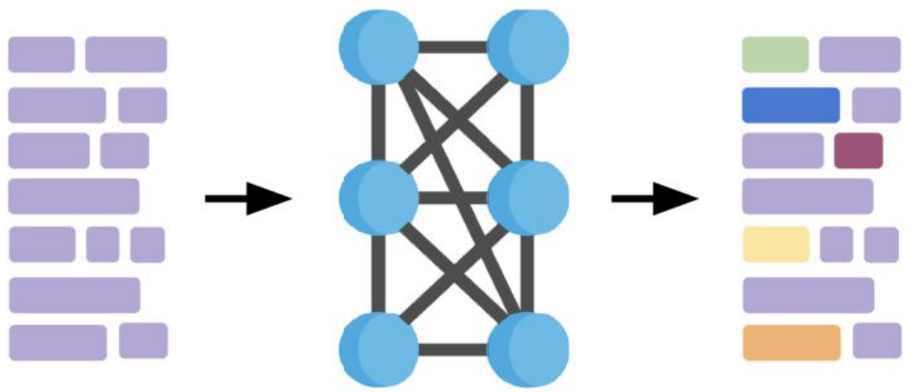
SPL
MAT
APL
...

Training Set	Dev Set	Test Set
1011011110110111011 1001100110001110 01 ...	1111011 0110001 ...	0011010 0111000 ...

4. Train model



5. *Extract entities with model*

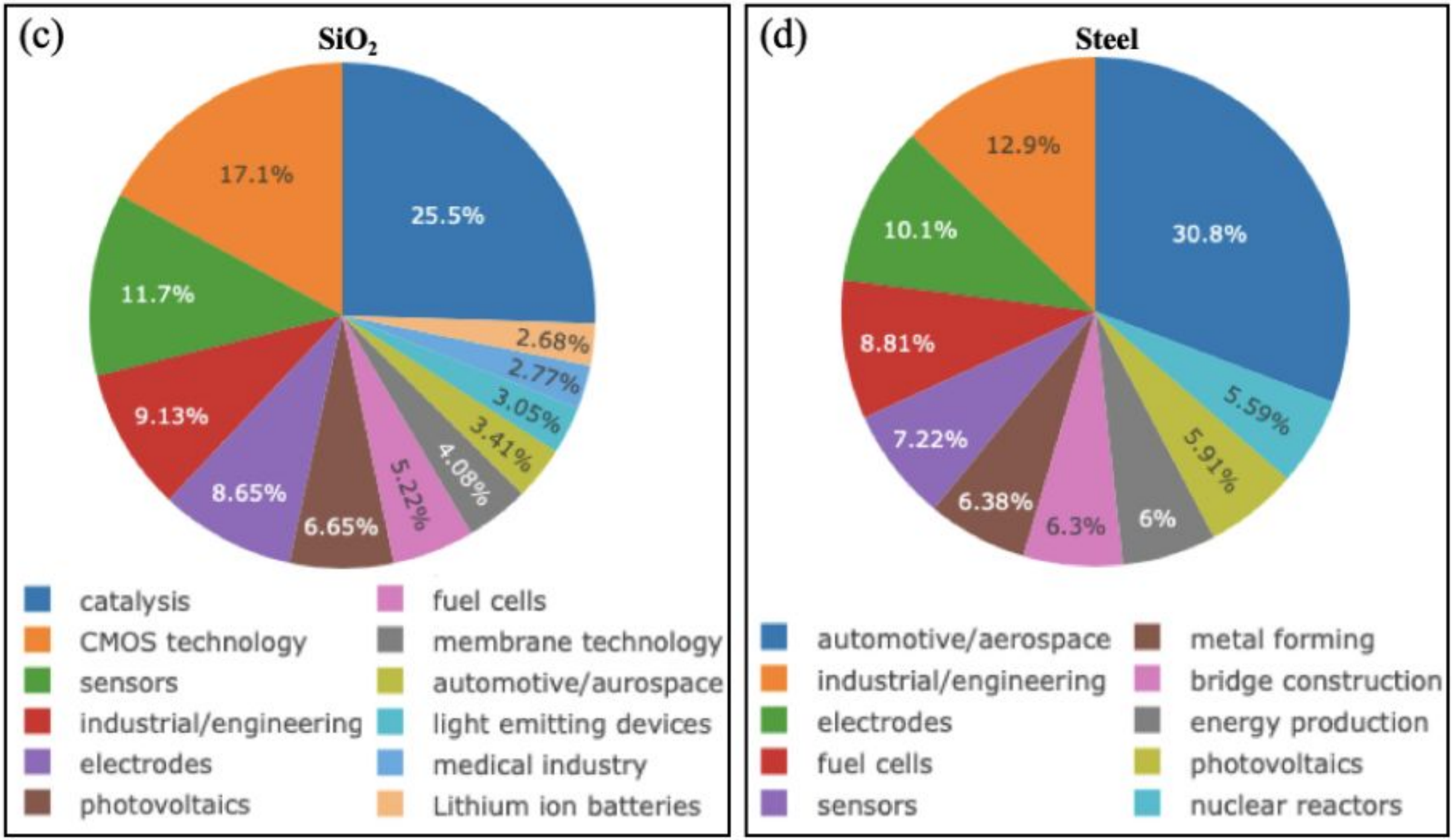


Named Entities in Numbers

> 3.3 million	**31 million**	**19 million**
Abstracts Collected	**Properties**	**Materials Mentions**
8.8 million	**7.5 million**	**5 million**
Characterization Methods	**Applications**	**Synthesis Methods**

Working on full text extraction

What are the most common applications of a material?



References

- Tshitoyan et al. Nature 571, 95–98 (2019).
  - https://github.com/materialsintelligence/mat2vec
- Weston et al. J. Chem. Inf. Model. (2019).
- REST API: https://github.com/materialsintelligence/matscholar
- Website: https://www.matscholar.com

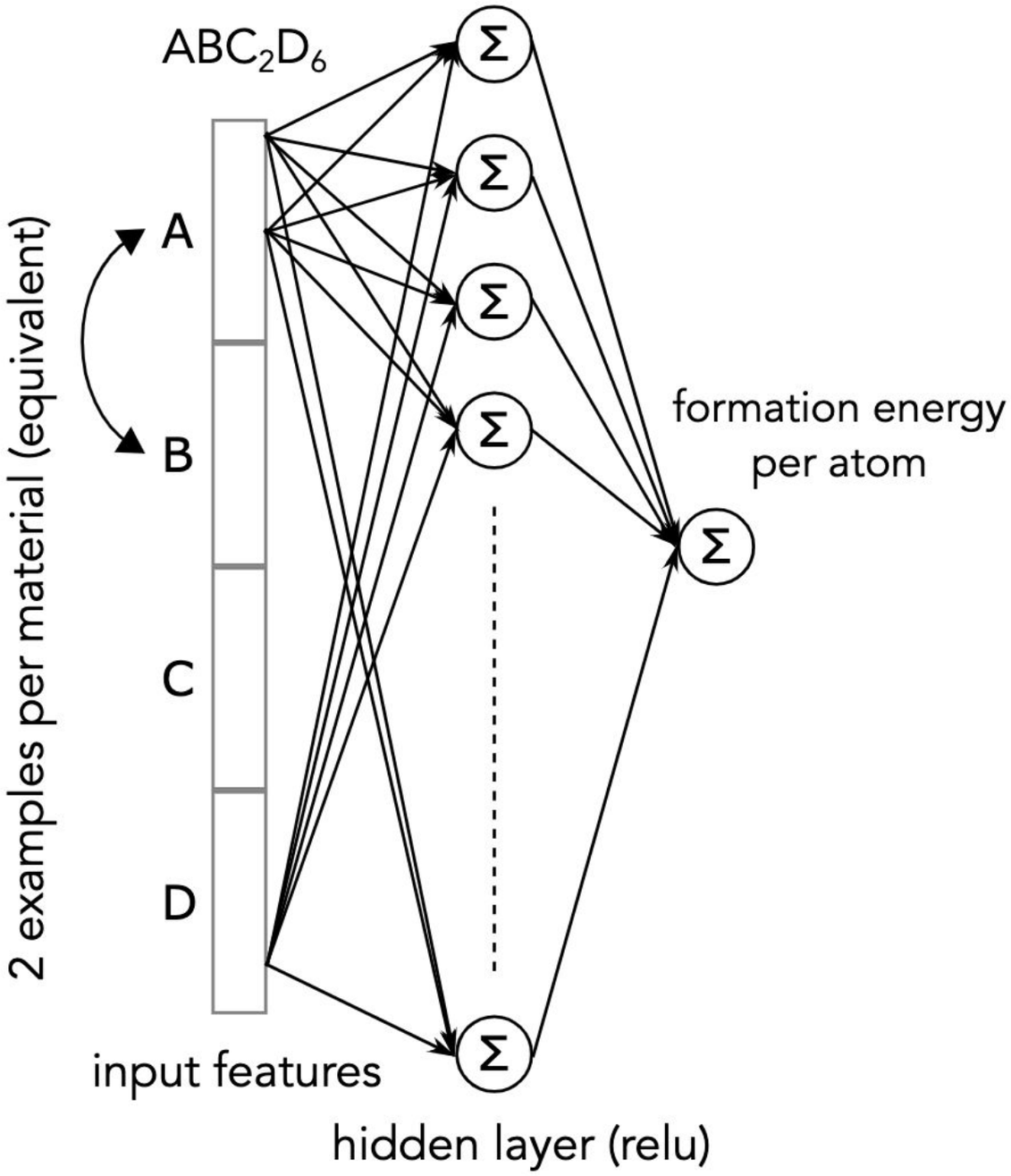
MATSCHOLAR
BETA

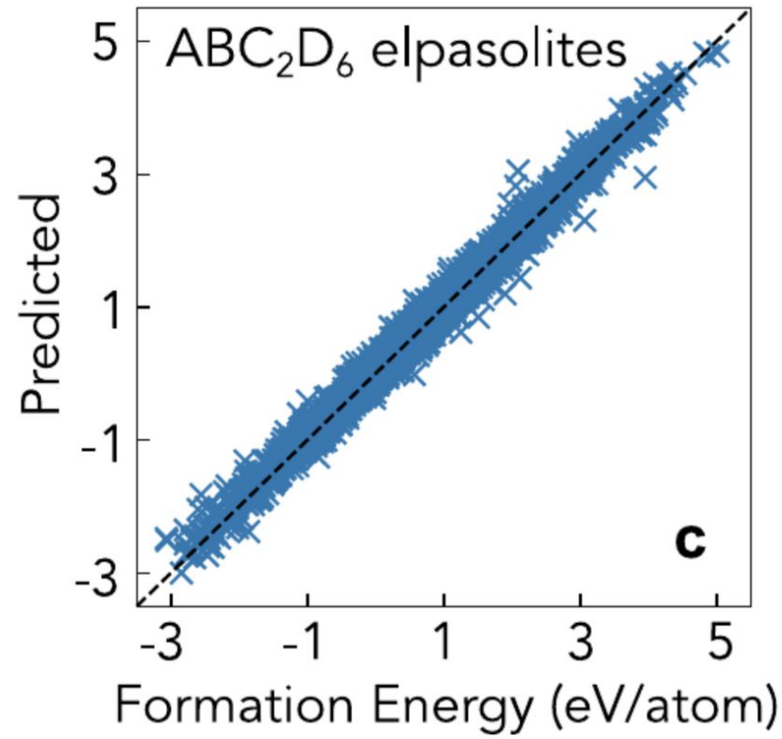
Contact

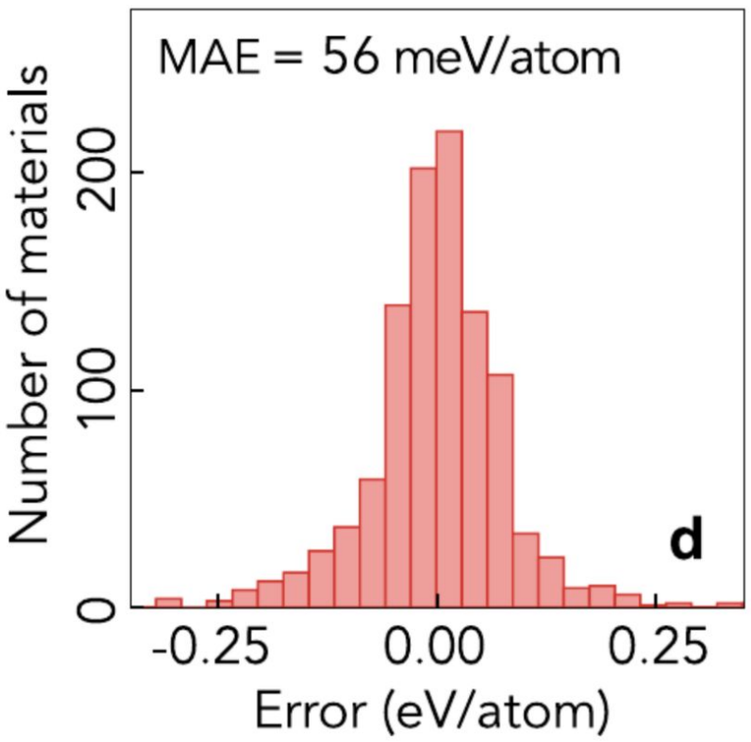
- Email: vahe.tshitoyan@gmail.com
- LinkedIn: /in/vahe-tshitoyan/
- **Project contact**: Anubhav Jain - ajain@lbl.gov

Extra slides

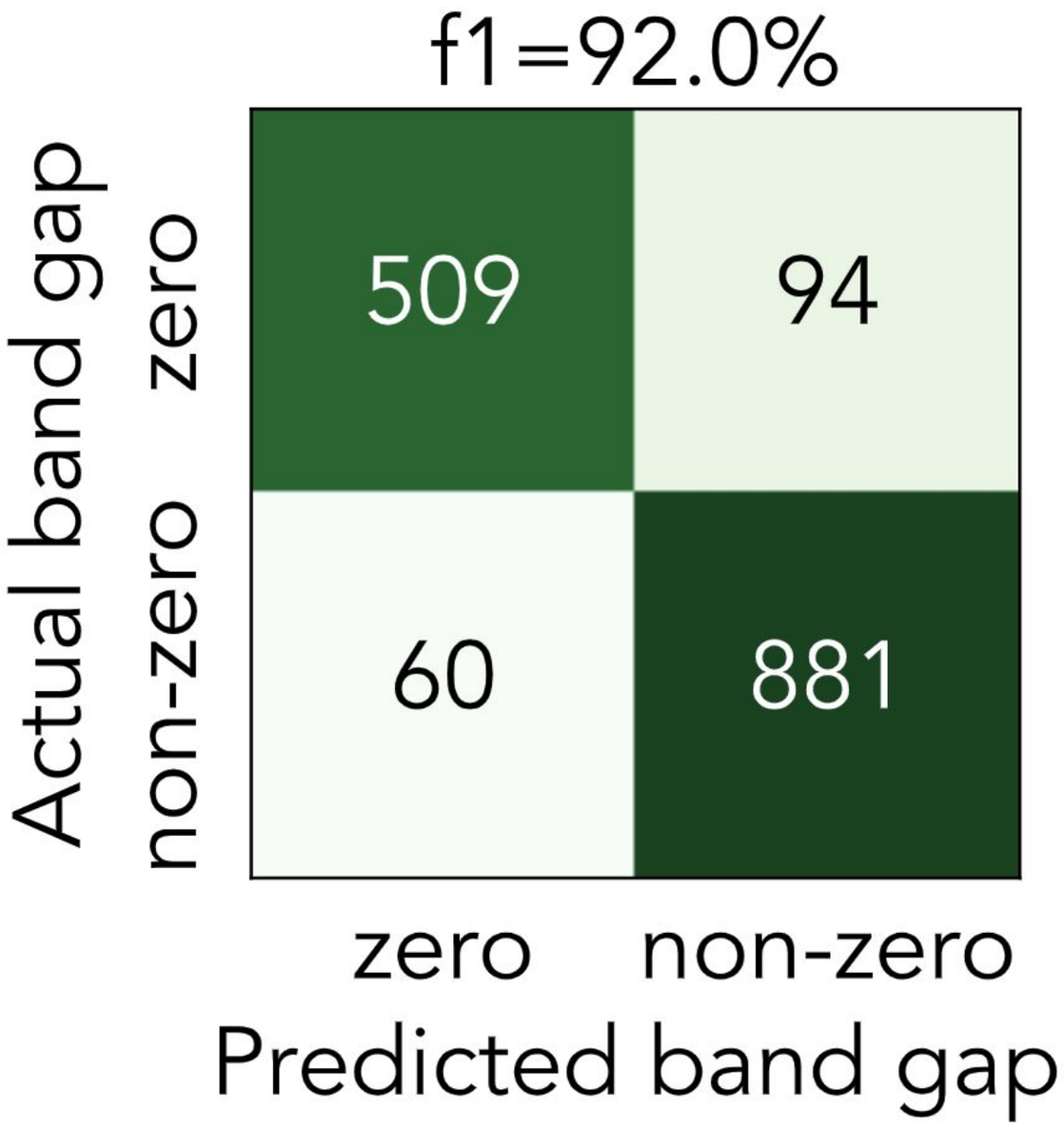
Formation Energy Prediction

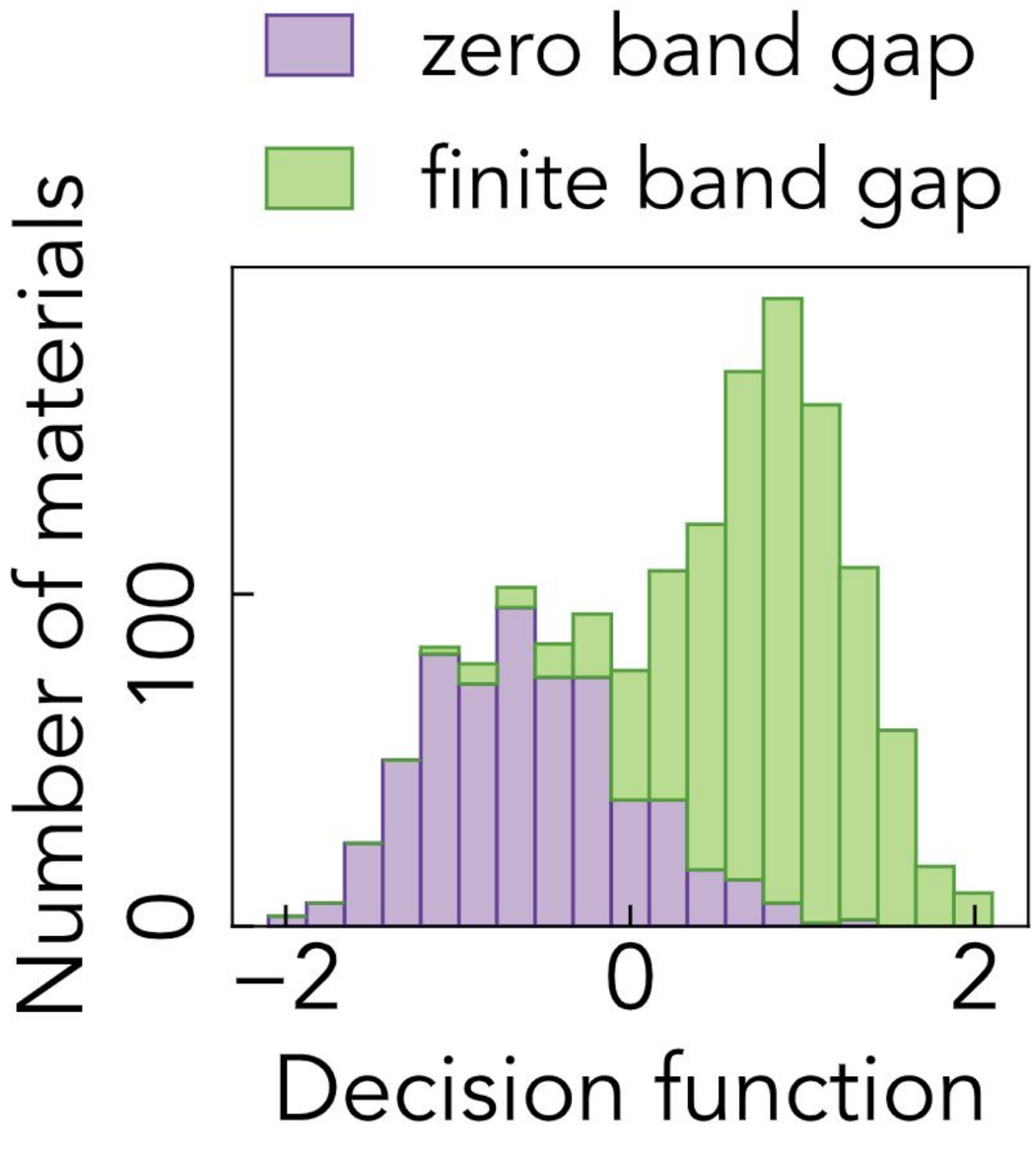






Band gap vs no band gap





The team at Berkeley

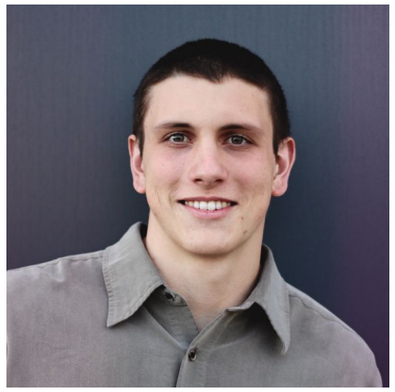
Gerbrand Ceder

Anubhav Jain

Kristin Persson

Leigh Weston

John Dagdelen



Alex Dunn

Olga Kononova

Ziqin Rong

BERKELEY LAB

THE UNIVERSITY OF CALIFORNIA BERKELEY 1868 LET THERE BE LIGHT

TOYOTA RESEARCH INSTITUTE