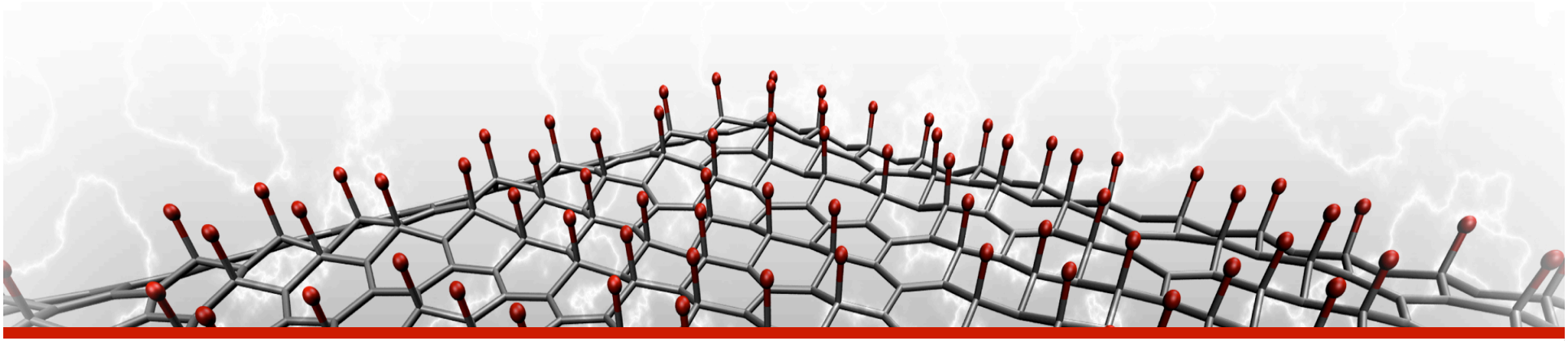
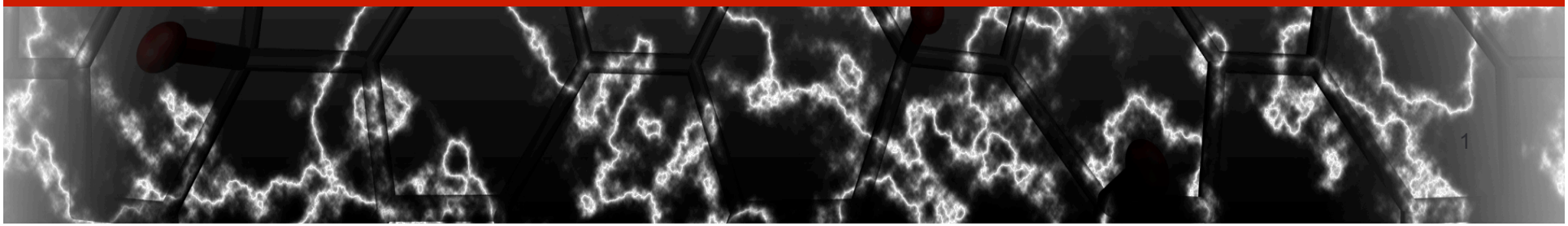


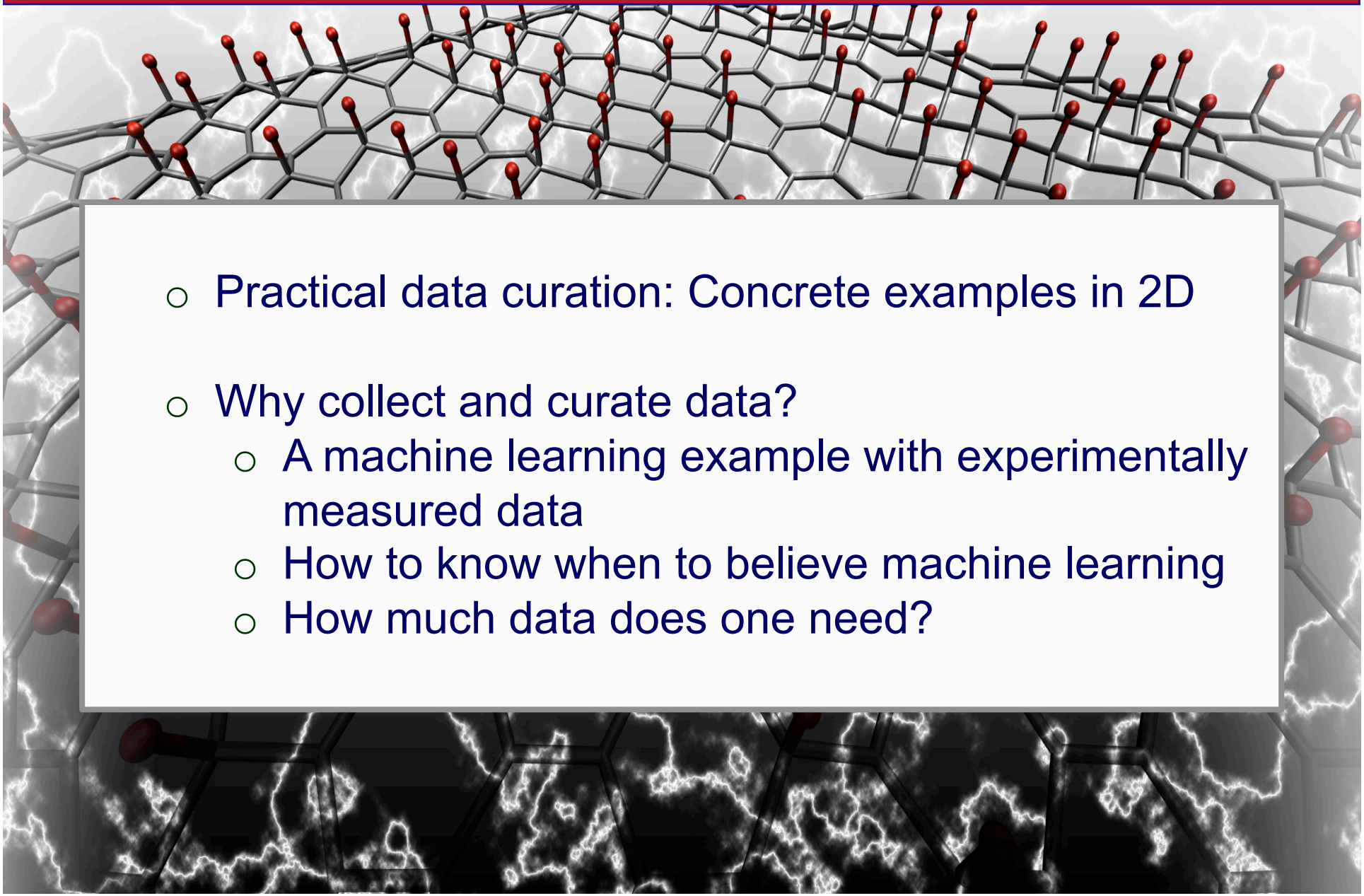
How data curation has potential to help you do science

Austin Sendek, E. Dogus Cubuk, Gowoon Cheon, Qian Yang, Yi Cui, Evan Reed*

**Department of Materials Science and Engineering*
Stanford University



THE PLAN

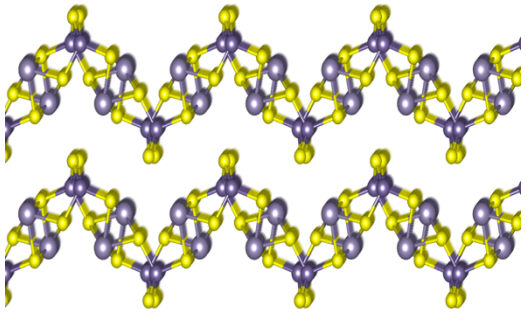


- Practical data curation: Concrete examples in 2D
- Why collect and curate data?
  - A machine learning example with experimentally measured data
  - How to know when to believe machine learning
  - How much data does one need?

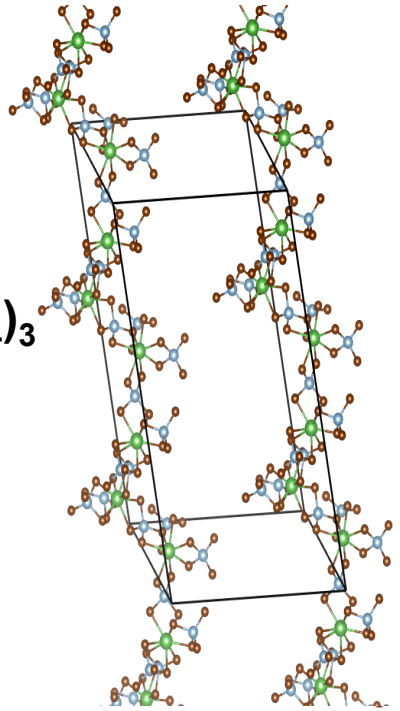
We discover new 2D and 1D materials

- Data mining of public databases leads to the discovery of:
  - **1173** 2D layered materials
  - **325** materials with piezoelectric monolayers
  - **98** bulk vertical lattice-commensurate heterostructures
  - **487** 1D molecular wires

$SnGeS_3$



$La(AlBr_4)_3$



G.Cheon et al., Data mining for new two- and one-dimensional weakly bonded solids and lattice-commensurate heterostructures, *Nano Letters* (2017)

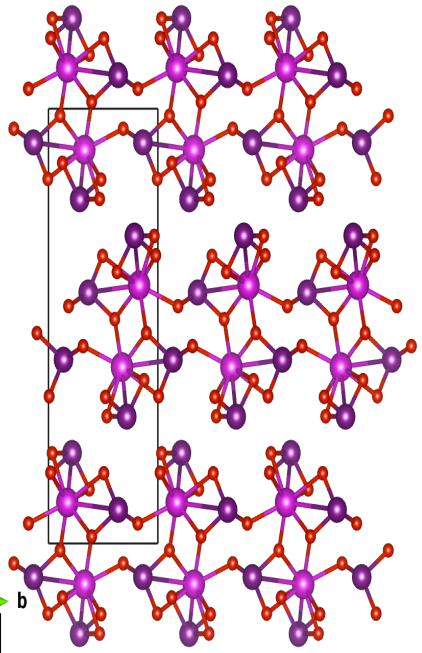
We compile a genome of 1173 2D materials

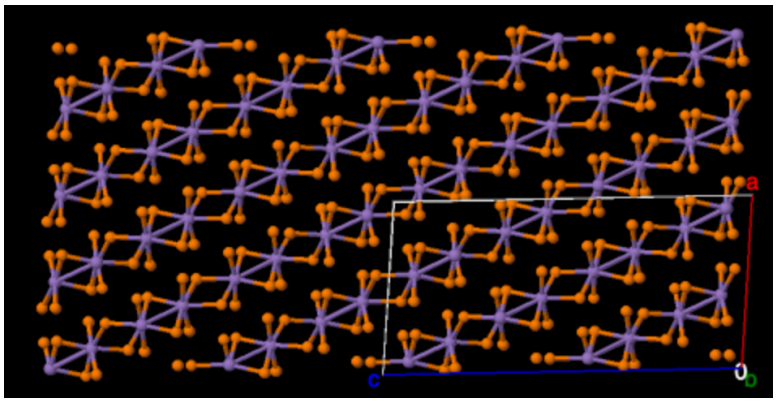
- Diverse spectrum of layered materials
- Materials Project IDs of all layered materials available in Supporting Information

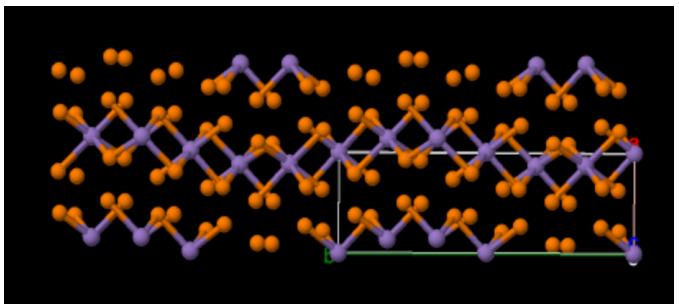
$V_4S_9Br_4$

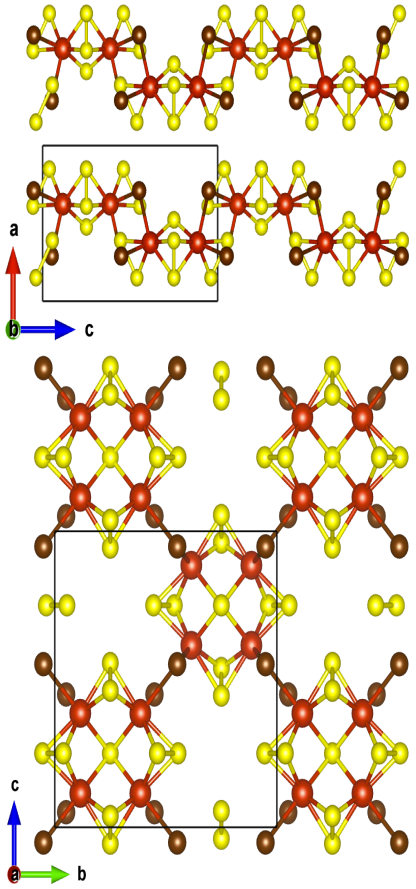
MnP_4

$Cd(IO_3)_2$





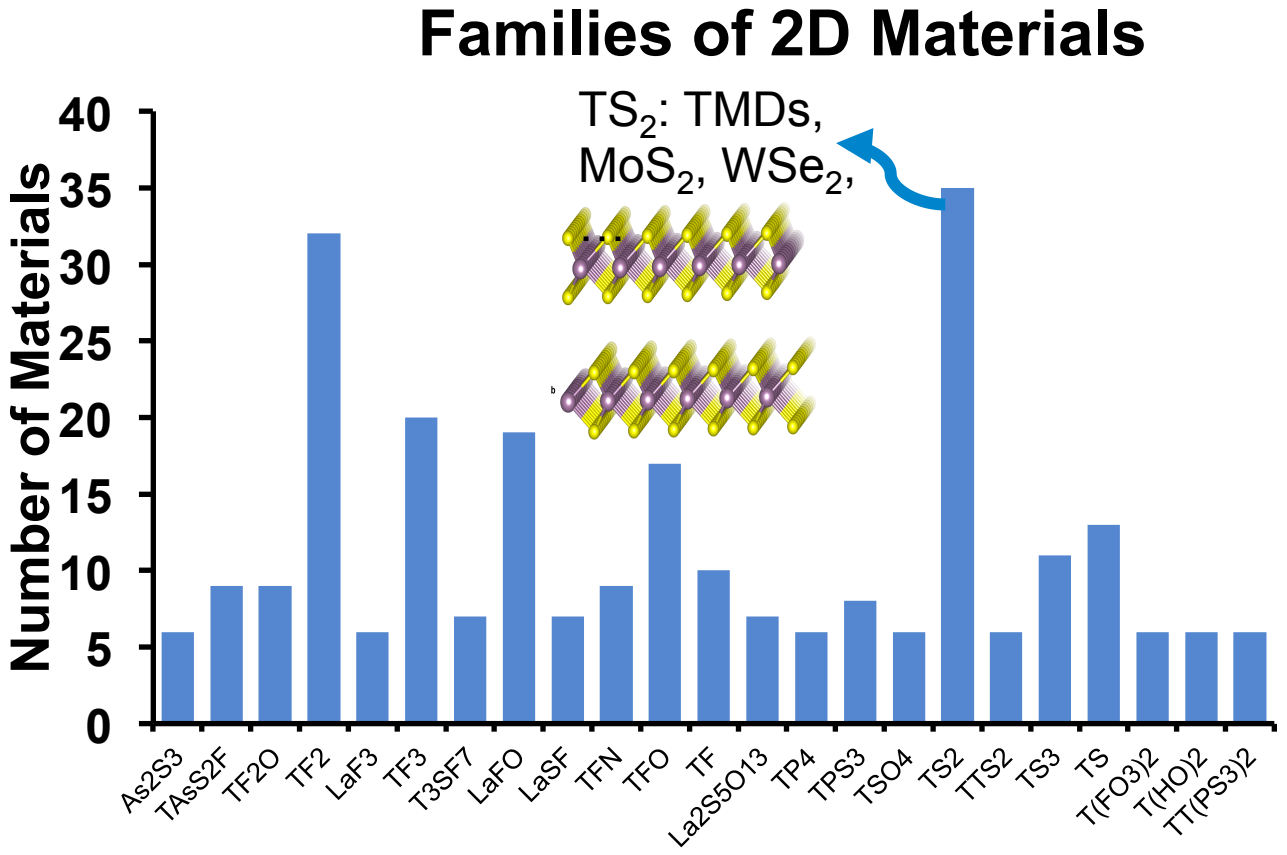




G.Cheon et al., *Nano Letters* (2017)

We compile a genome of 1173 2D materials

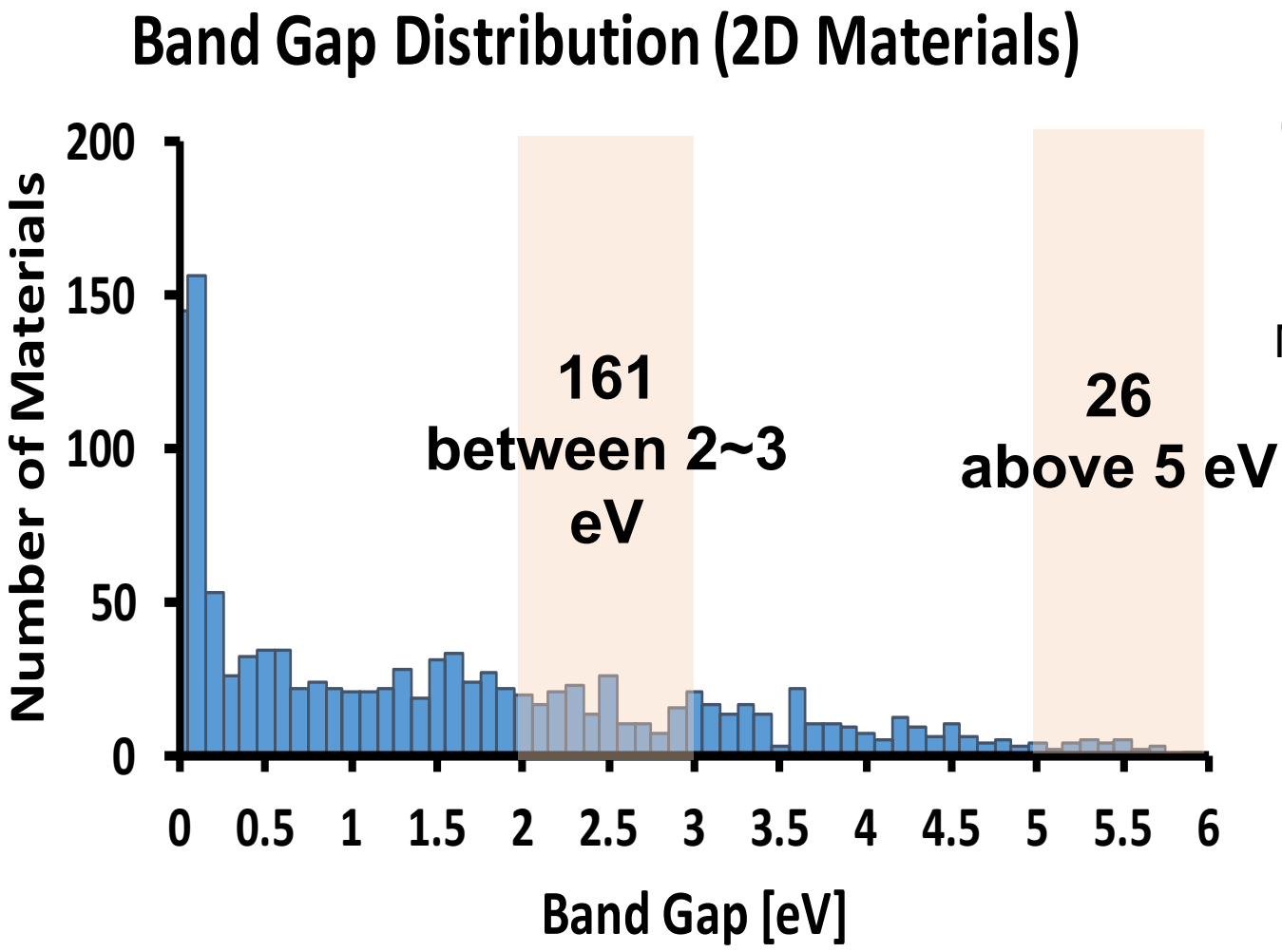
- **1173 weakly bonded layered materials** identified, lots of new candidates!
- **23 families** of similar chemical compositions (>5 materials), but >80% don't belong to a family



Ac: actinides
As: large pnictogens (As, Sb, Bi)
F: halogens
S: chalcogens excluding O
La: lanthanides
T: transition metals

G.Cheon et al., *Nano Letters* (2017)

We find a wide spectrum of 2D material band gaps

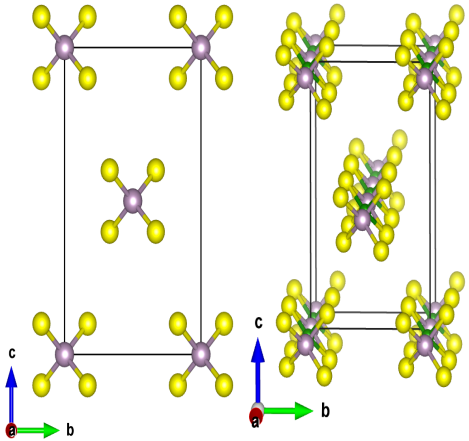


(Semi-local DFT band gaps for bulk materials from the Materials Project database)

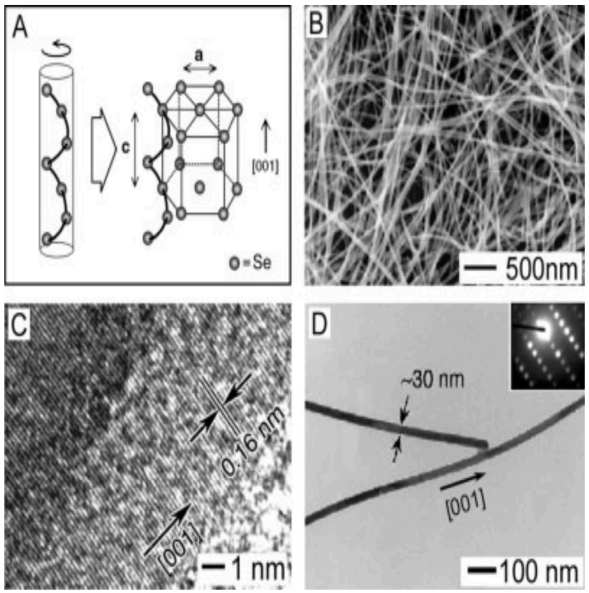
G.Cheon et al., *Nano Letters* (2017)

We discover 487 1D materials

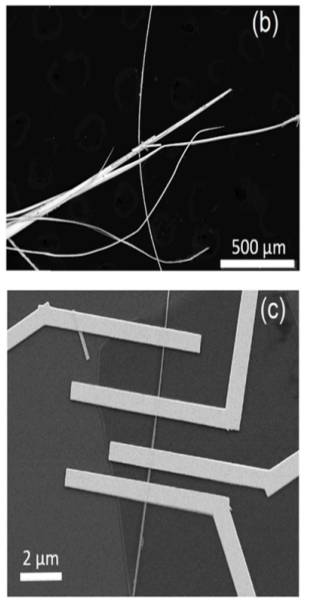
- Our algorithm can find 1D dimensional subunits



BPS_4, a material found in this work



Chain-like structure of t-Se that grow into nanowires.
Xia et al., *Adv. Materials* (2003)

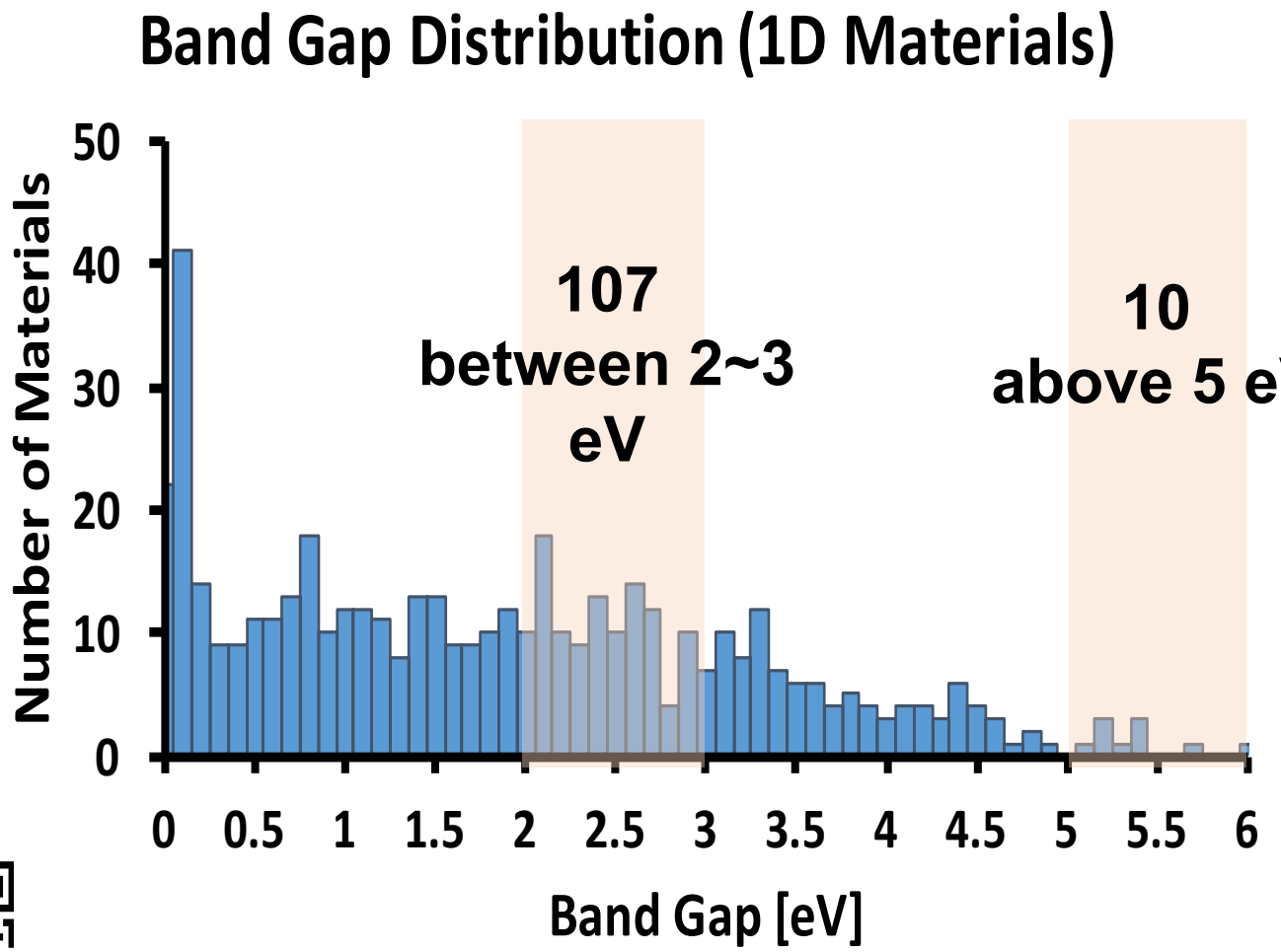


Quasi-1D TaSe3 low-noise nanowire devices
Liu et al., *Nano Lett.* (2017)

- Some inorganic 'molecular wires' have been predicted to possess structural stability and versatile material properties, **but only ~20 known**

G.Cheon et al., *Nano Letters* (2017)

We find a wide spectrum of 1D material band gaps



(Semi-local DFT band gaps for bulk materials from the Materials Project database)

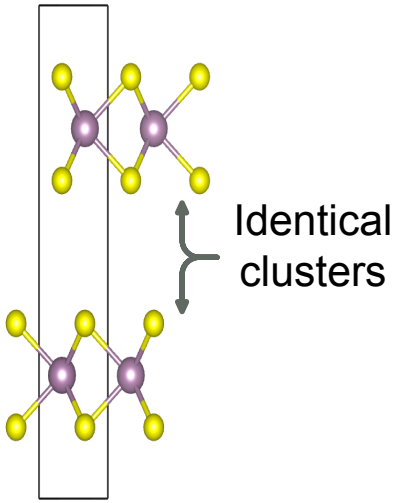
G.Cheon et al., *Nano Letters* (2017)

We discover lattice-commensurate vertical heterostructures

- We discover **intrinsic, lattice-commensurate heterostructures** that preclude the need for artificial stacking:

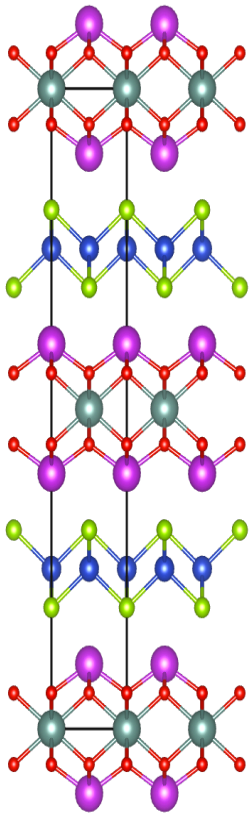
Not a heterostructure

MoS_2

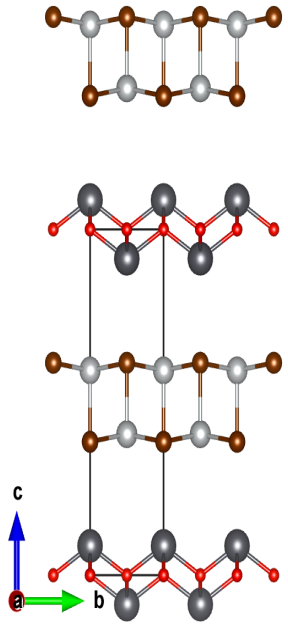


Heterostructures:
different
chemical compositions
or
number of atoms
in each layer

$YCu_2Bi_2(SeO_2)_2$



AgPbBrO

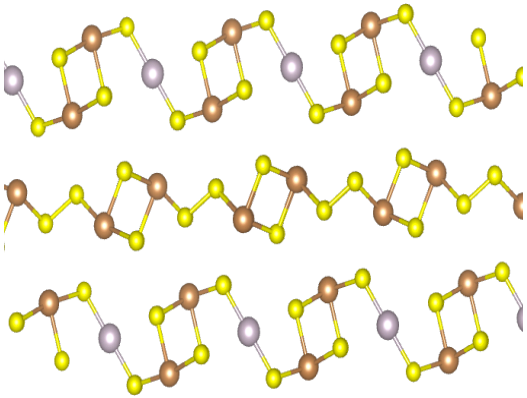


G.Cheon et al., *Nano Letters* (2017)

We discover lattice-commensurate heterostructures

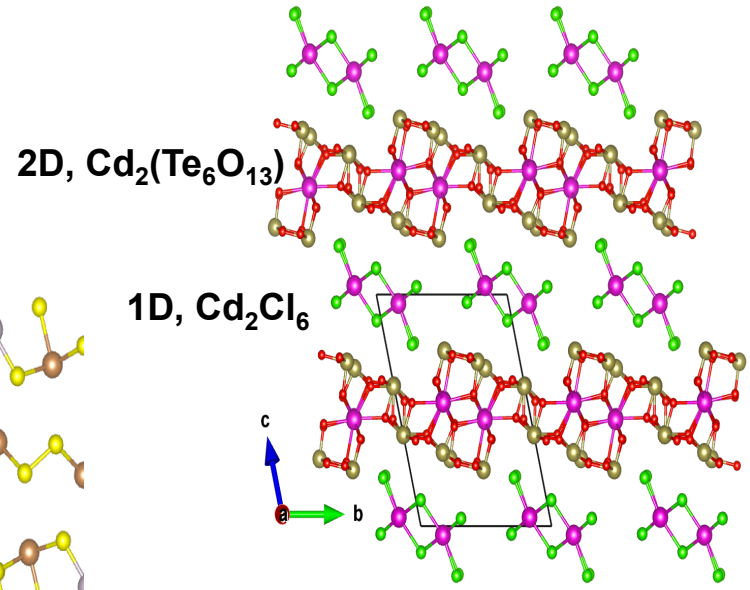
- We identify 98 lattice-commensurate heterostructures:
- Experimentally reported in bulk crystals

$HgSb_4S_8$(Livingstonite)



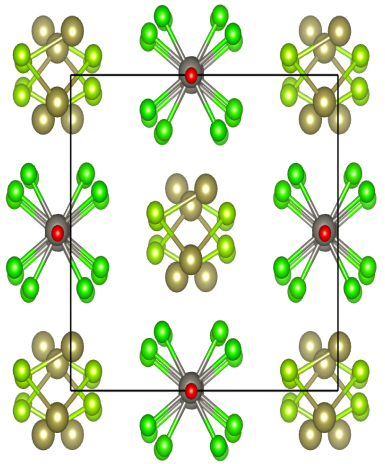
2D

$Cd_4Te_6Cl_6O_{13}$



2D+1D

$Te_3W_2Se_4(Cl_4O)_2$



1D

G.Cheon et al., *Nano Letters* (2017)

SOME PRACTICAL CURATION APPROACHES

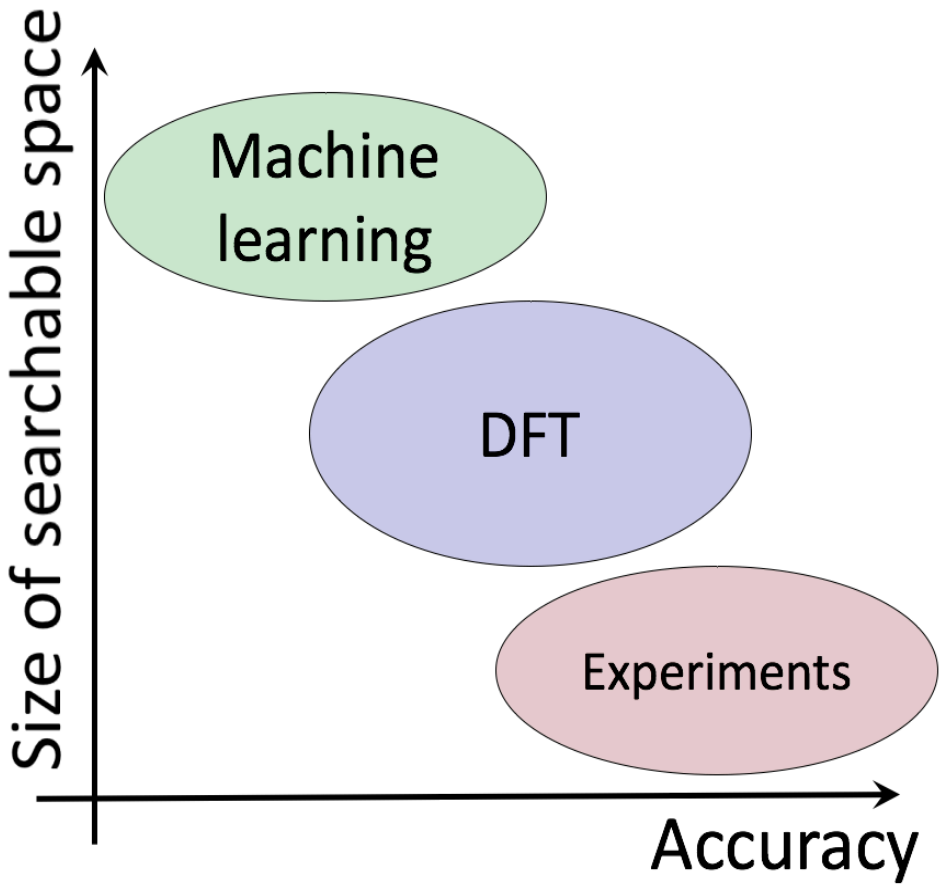
Lower effort:

- Supporting information of publications
- File on your webpage
- File on Materials Data Facility (NIST)

More effort, but broader utility:

- Work with others to fold into existing databases:
  - NSF 2DCC at Penn State (Vin Crespi, Richard Hennig, et al)
  - Jarvis at NIST (Francesca Tavazza, et al)
  - Materialsweb.org (Richard Hennig)
  - DOE's Materials Project (Kristin Persson et al)

MACHINE LEARNING HAS THE POTENTIAL TO FILL A GAP, ENABLING SEARCHES OF LARGE SPACES OF MATERIALS



WE COLLECT 39 EXPERIMENTAL MEASUREMENTS OF LI ION CONDUCTIVITY FOR SOLIDS

Composition	RT bulk ionic conductivity (S cm^{-1})
$LiLa(TiO_3)_2$	1×10^{-3}
$Li_{9.81}Sn_{0.81}P_{2.19}S_{12}$	5.5×10^{-3}
$Li_{10}Ge(PS_6)_2$	1.4×10^{-2}
$Li_{10.35}Si_{1.35}P_{1.65}S_{12}$	6.5×10^{-3}
$Li_{14}ZnGe_4O_{16}(2)$	1.0×10^{-6}
$Li_2Ca(NH)_2$	6.4×10^{-6}
$Li_2Ge_7O_{15}$	5.0×10^{-6}
Li_2NH	2.5×10^{-4}
Li_2S	1.0×10^{-13}
$Li_{13.6}Si_{2.8}S_{1.2}O_{16}$	6.0×10^{-7}
$Li_{14}Ge_2V_2O_{16}$	7.0×10^{-5}
$Li_{15}Ge_3V_2O_4$	6.03×10^{-6}
$Li_{14.8}Ge_{3.4}W_{0.6}O_4$	4.0×10^{-5}
$Li_3Fe_2P_3O_{12}$	1.0×10^{-7}
Li_3N	5.75×10^{-4}
Li_3P	1.0×10^{-3}
γ-Li_3PS_4	3.0×10^{-7}
$Li_3Sc_2P_3O_{12}$	1.0×10^{-10}
β_{II}-Li_3VO_4	4.4×10^{-8}
$Li_4B_7O_{12}Cl$	1.0×10^{-7}
$Li_4BN_3H_{10}$	2.0×10^{-4}
γ-Li_4GeO_4	3.1×10^{-12}
Li_4SiO_4	2.4×10^{-10}
$Li_5La_3Bi_2O_{12}$	2.0×10^{-5}
$Li_5La_3Nb_2O_{12}$	8.0×10^{-6}
$Li_5La_3Ta_2O_{12}$	1.5×10^{-6}
Li_5NI_2	1.5×10^{-7}
$Li_6BaLa_2Ta_2O_{12}$	4.0×10^{-5}
Li_6FeCl_8	1.0×10^{-4}
Li_6NBr_3	1.5×10^{-7}
$Li_6SrLa_2Ta_2O_{12}$	7.0×10^{-6}
$Li_7La_3Zr_2O_{12}$	3.5×10^{-4}
$Li_7P_3S_{11}$	4.1×10^{-3}
$LiAlH_4$	2.0×10^{-9}
$LiAlSiO_4$	1.4×10^{-5}
$LiBH_4$	2.0×10^{-8}
LiI	1.0×10^{-6}
$LiNH_2$	4.0×10^{-10}
α'-$LiZr_2P_3O_{12}$	5.0×10^{-8}

- We adopt a binary classification strategy with a 10^{-4} S/cm boundary, motivated by engineering requirements
- Training set includes 8 “good” conductors, 31 “bad” conductors

A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

WE DRAW ON WISDOM/PROPOSALS IN THE LITERATURE FOR READILY COMPUTABLE FEATURES (NO DFT!)

	Feature	Pearson correlation coefficient
1	Volume per atom[a]	0.20
2	Standard deviation in Li neighbour count	0.22
3	Standard deviation in Li bond ionicity	−0.04
4	Li bond ionicity[a]	−0.18
5	Li neighbour count[a]	−0.19
6	*Li–Li bonds per Li*[a]	*0.06*
7	*Bond ionicity of sublattice*[a]	*−0.28*
8	Sublattice neighbour count[a]	−0.13
9	*Anion framework coordination*[a]	*−0.06*
10	Minimum anion–anion separation distance[a] (Å)	0.09
11	Volume per anion (Å^3)	−0.01
12	*Minimum Li–anion separation distance*[a] *(Å)*	*0.20*
13	*Minimum Li–Li separation distance*[a] *(Å)*	*−0.10*
14	Electronegativity of sublattice[a]	−0.16
15	Packing fraction of full crystal	0.16
16	Packing fraction of sublattice	0.19
17	Straight-line path width[a] (Å)	0.07
18	Straight-line path electronegativity[a]	−0.29
19	Ratio of features (4) and (7)	−0.03
20	Ratio of features (5) and (8)	−0.18
	Constant term	—

No single feature has strong correlation with ionic conductivity across the broad spectrum of 39 materials

A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

WE DRAW ON WISDOM/PROPOSALS IN THE LITERATURE FOR READILY COMPUTABLE FEATURES (NO DFT!)

	Feature
1	Volume per atom[a]
2	Standard deviation in Li neighbour count
3	Standard deviation in Li bond ionicity
4	Li bond ionicity[a]
5	Li neighbour count[a]
6	*Li–Li bonds per Li*[a]
7	*Bond ionicity of sublattice*[a]
8	Sublattice neighbour count[a]
9	*Anion framework coordination*[a]
10	Minimum anion–anion separation distance[a] (Å)
11	Volume per anion (Å^3)
12	*Minimum Li–anion separation distance*[a] *(Å)*
13	*Minimum Li–Li separation distance*[a] *(Å)*
14	Electronegativity of sublattice[a]
15	Packing fraction of full crystal
16	Packing fraction of sublattice
17	Straight-line path width[a] (Å)
18	Straight-line path electronegativity[a]
19	Ratio of features (4) and (7)
20	Ratio of features (5) and (8)
	Constant term

A physics-based model for a single crystal, with implicit assumptions:

$$\sigma(T) = \frac{Dnq^2}{k_B T} = \frac{\left(fa^2 \nu_0 p_{occ} zj\right) nq^2}{k_B T} e^{-E_a/k_B T}$$

A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

WE DRAW ON WISDOM/PROPOSALS IN THE LITERATURE FOR READILY COMPUTABLE FEATURES (NO DFT!)

	Feature
1	Volume per atom[a]
2	Standard deviation in Li neighbour count
3	Standard deviation in Li bond ionicity
4	Li bond ionicity[a]
5	Li neighbour count[a]
6	*Li–Li bonds per Li*[a]
7	*Bond ionicity of sublattice*[a]
8	Sublattice neighbour count[a]
9	*Anion framework coordination*[a]
10	Minimum anion–anion separation distance[a] (Å)
11	Volume per anion (Å^3)
12	*Minimum Li–anion separation distance*[a] *(Å)*
13	*Minimum Li–Li separation distance*[a] *(Å)*
14	Electronegativity of sublattice[a]
15	Packing fraction of full crystal
16	Packing fraction of sublattice
17	Straight-line path width[a] (Å)
18	Straight-line path electronegativity[a]
19	Ratio of features (4) and (7)
20	Ratio of features (5) and (8)
	Constant term

$$\sigma(T) = \frac{Dnq^2}{k_B T} = \frac{\left(fa^2\nu_0 p_{occ} zj\right)nq^2}{k_B T} e^{-E_a/k_B T}$$

A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

WE EMPLOY LOGISTIC REGRESSION (TWO-CLASS CLASSIFIER)

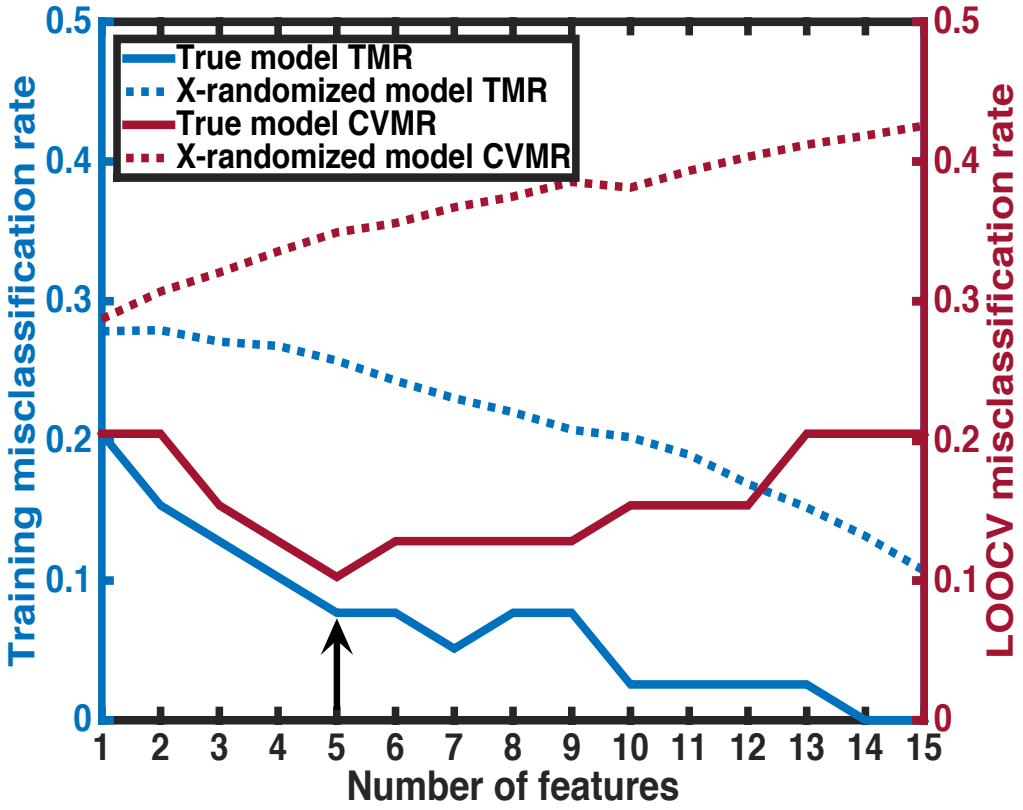
Assuming a logistic form, we search for the maximally predictive set of features

$$P_{\text{superionic}}(\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})}$$

$$\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x} = ?$$

A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

WE EMPLOY LEAVE-ONE OUT CROSS VALIDATION TO DETERMINE OPTIMAL FEATURES



$$\mathrm{TMR} = \frac{1}{M}\sum_{i=1}^{M} 1\left\{\tilde{\sigma}^{(i)} \neq \hat{\sigma}^{(i)}\right\}$$

$$\mathrm{CVMR} = \frac{1}{M}\sum_{i=1}^{M} 1\left\{\tilde{\sigma}^{(i)} \neq \hat{\sigma}^{(i)}_{\mathrm{LOO}}\right\}$$

- We search over all possible combinations of 20 features (>10^6 models)
- Optimal leave-one-out cross-validated misclassification rate = 10%
- Optimal model performance against random guessing: 3-4x improvement

A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

WE DISCOVER 5 FEATURES THAT BEST CLASSIFY ION CONDUCTORS

	Feature
1	Volume per atom[a]
2	Standard deviation in Li neighbour count
3	Standard deviation in Li bond ionicity
4	Li bond ionicity[a]
5	Li neighbour count[a]
6	*Li–Li bonds per Li*[a]
7	*Bond ionicity of sublattice*[a]
8	Sublattice neighbour count[a]
9	*Anion framework coordination*[a]
10	Minimum anion–anion separation distance[a] (Å)
11	Volume per anion (Å^3)
12	*Minimum Li–anion separation distance*[a] *(Å)*
13	*Minimum Li–Li separation distance*[a] *(Å)*
14	Electronegativity of sublattice[a]
15	Packing fraction of full crystal
16	Packing fraction of sublattice
17	Straight-line path width[a] (Å)
18	Straight-line path electronegativity[a]
19	Ratio of features (4) and (7)
20	Ratio of features (5) and (8)
	Constant term

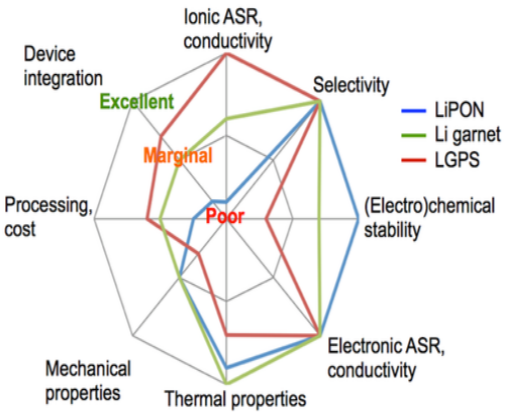
$$\sigma(T) = \frac{Dnq^2}{k_B T} = \frac{\left(fa^2\nu_0 p_{occ} zj\right)nq^2}{k_B T} e^{-E_a/k_B T}$$

A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

WE PERFORM THE FIRST HOLISTIC STRUCTURE SCREENING OF ALL >12,000 LI-CONTAINING SOLIDS IN THE MATERIALS PROJECT DATABASE

Ionic conductivity is not all that matters! We also screen for:

- High stability against oxidation ← Gibbs free energy of atomization
- High stability against reduction ← Presence of transition metals
- Low electronic conductivity ← Band gap
- High phase stability ← Convex hull
- Low cost ← Cost of raw elements involved
- High earth abundance ← Abundance of elements in Earth's crust



All Li-containing materials: >12,000

Phase stable: 1,472

All requirements met except ionic conductivity: 317

All requirements: 21

Ionic conductivity alone: 1,408

A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

WE PROPOSE 21 NEW PROMISING SOLID ELECTROLYTE CANDIDATES BY SCREENING >12,000

MPID	Chemical formula	P_{LR}	d	ε	A	E_{gap}	$\tilde{V}_{ox}$	USD/m^2 (10 μm thick)	I_A	Related study
mp-554076	$BaLiBS_3$	0.589	1.049	0.048	1	2.153	9.697	23	38	
mp-532413	$Li_5B_7S_{13}$	0.897	1.228	0.024	1	3.553	5.454	42	38	95
mp-569782[a]	$Sr_2LiCBr_3N_2$	1.000	6.852	0.000	0	3.973	13.968	16	45	
mp-558219	$SrLi(BS_2)_3$	0.518	1.556	0.114	1	2.91	13.964	38	38	
mp-15797	$LiErSe_2$	0.543	1.505	0.056	1	1.615	6.778	170	67	
mp-29410	$Li_2B_2S_5$	0.994	1.855	0.003	1	2.538	4.895	29	38	95
mp-676361	Li_3ErCl_6	0.655	0.974	0.042	1	5.211	7.794	70	44	96 and 97
mp-643069[a]	Li_2HIO	0.652	2.081	0.079	0	4.319	4.054	2.40	60	
mp-19896	Li_2GePbS_4	0.604	1.063	0.090	1	2.265	4.591	13	54	90
mp-7744[a]	$LiSO_3F$	1.000	4.097	0.000	0	5.792	13.446	10	34	
mp-22905[b]	$LiCl$	0.837	1.381	0.031	1	6.25	4.214	0.94	34	98
mp-34477	$LiSmS_2$	0.89	1.33	0.028	1	1.921	8.536	6.50	40	
mp-676109	Li_3InCl_6	0.656	1.013	0.058	1	3.373	6.215	5.50	63	96 and 97
mp-559238	$CsLi_2BS_3$	0.812	1.642	0.055	1	3.094	4.798	160	49	
mp-866665[a]	$LiMgB_3(H_9N)_2$	1.000	5.149	0.000	0	6.511	11.222	30	38	
mp-8751	$RbLiS$	0.775	1.279	0.051	1	2.745	4.22	240	34	
mp-15789	$LiDyS_2$	0.901	1.339	0.025	1	1.935	8.736	9.20	39	
mp-15790	$LiHoS_2$	0.899	1.327	0.025	1	1.965	8.749	300	55	
mp-15791	$LiErS_2$	0.899	1.319	0.025	1	2.008	8.761	190	44	
mp-561095[a]	$LiHo_3Ge_2(O_4F)_2$	0.984	3.247	0.009	0	4.163	53.18	370	55	
mp-8430	$KLiS$	0.76	1.243	0.052	1	3.057	4.348	14	34	

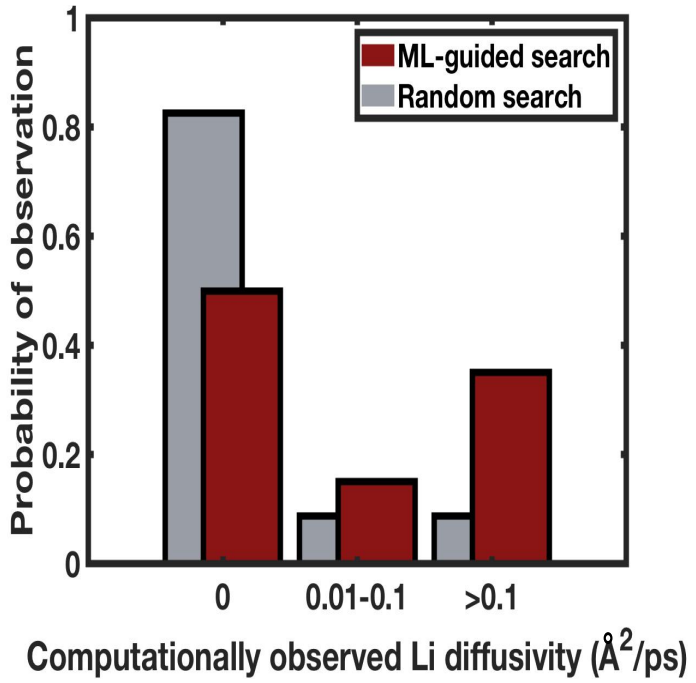
A. D. Sendek, E. J. Reed, et al, Energy and Environmental Science (2017).

How well does the model do?

We discover ten new solids that are superionic conductors, doubling known superionic conductors

Model success rate: 39% (melting cases excluded)

Random success rate: 14% (melting cases excluded)



Learning from only 40 data points = 3x improvement over guesswork

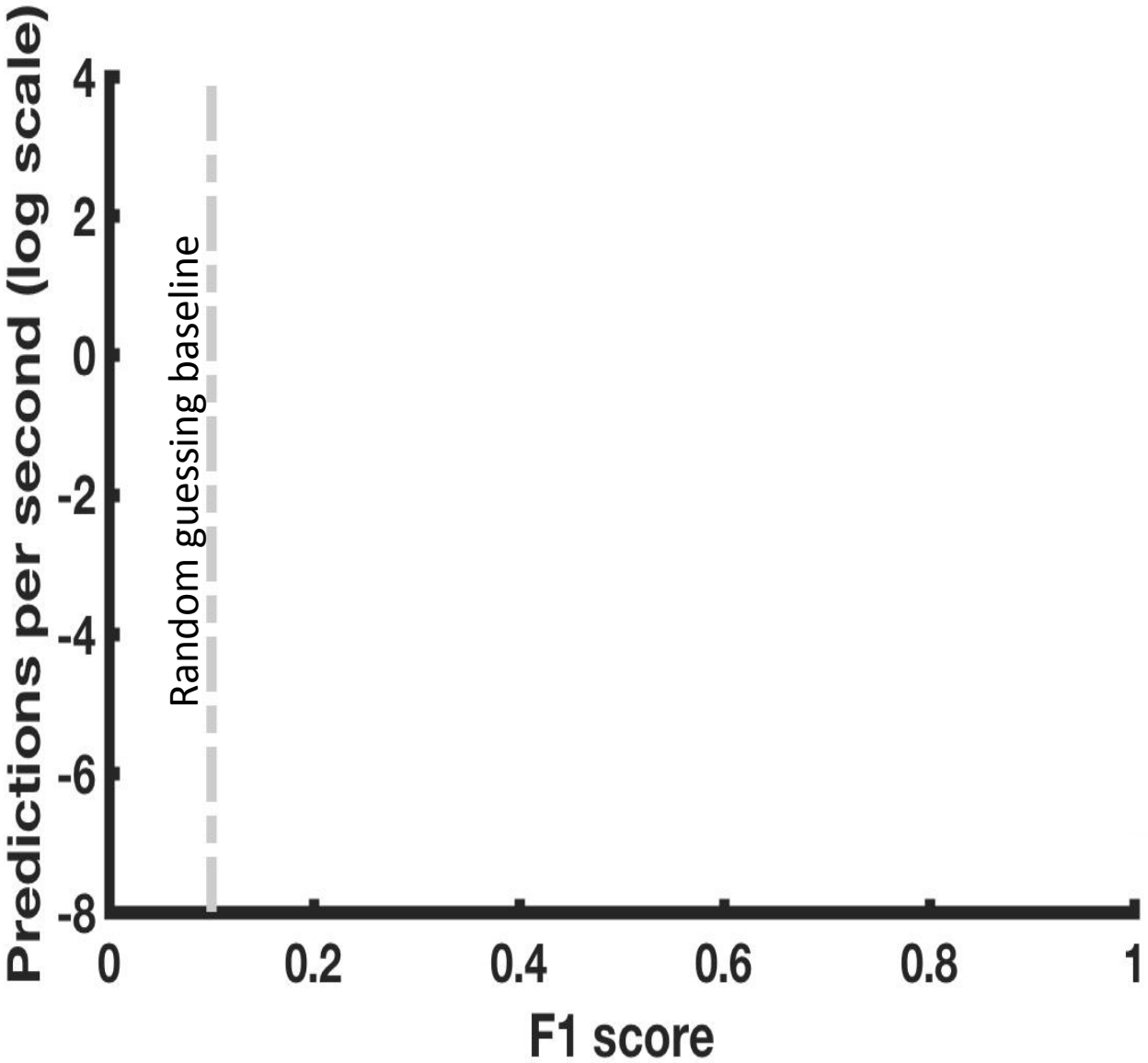
AD Sendek, ED Cubuk, G Cheon, ER Antoniuk, Y Cui, E. J. Reed, Chemistry of Materials (2018).

TWO METRICS FOR MODEL PERFORMANCE

- How much better than random guessing is our model?
  - Approximately three times better
  - Quantifies the state of humanity's knowledge
- Are the false positives acceptable in number?
  - 61% are false positives
  - Need to synthesize two materials to get one that works
  - This is probably good enough in practice, but could be better

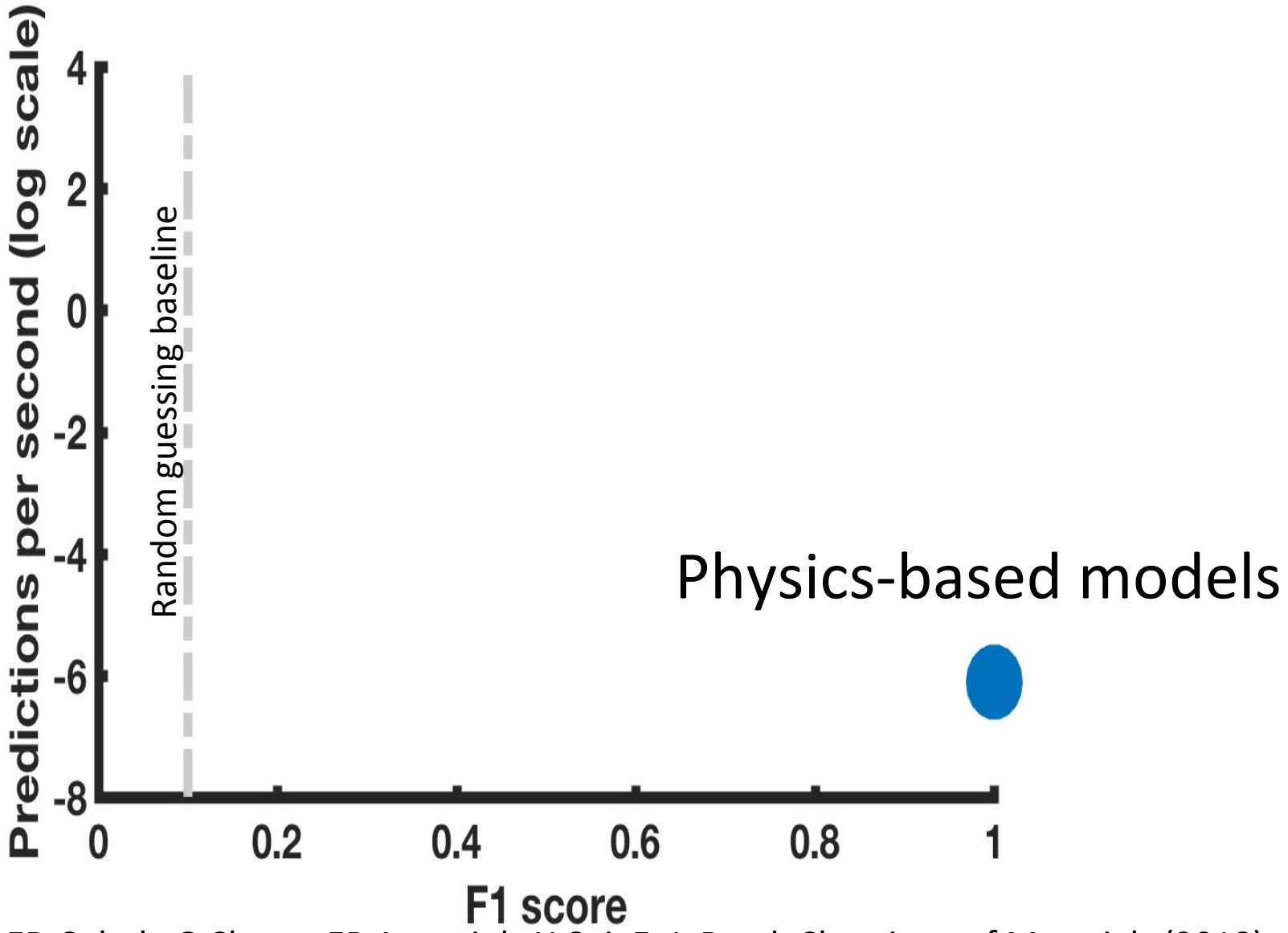
Scientists aren't truly guessing at random – but can they beat the machine?

Our algorithm outperforms humans in accuracy and speed



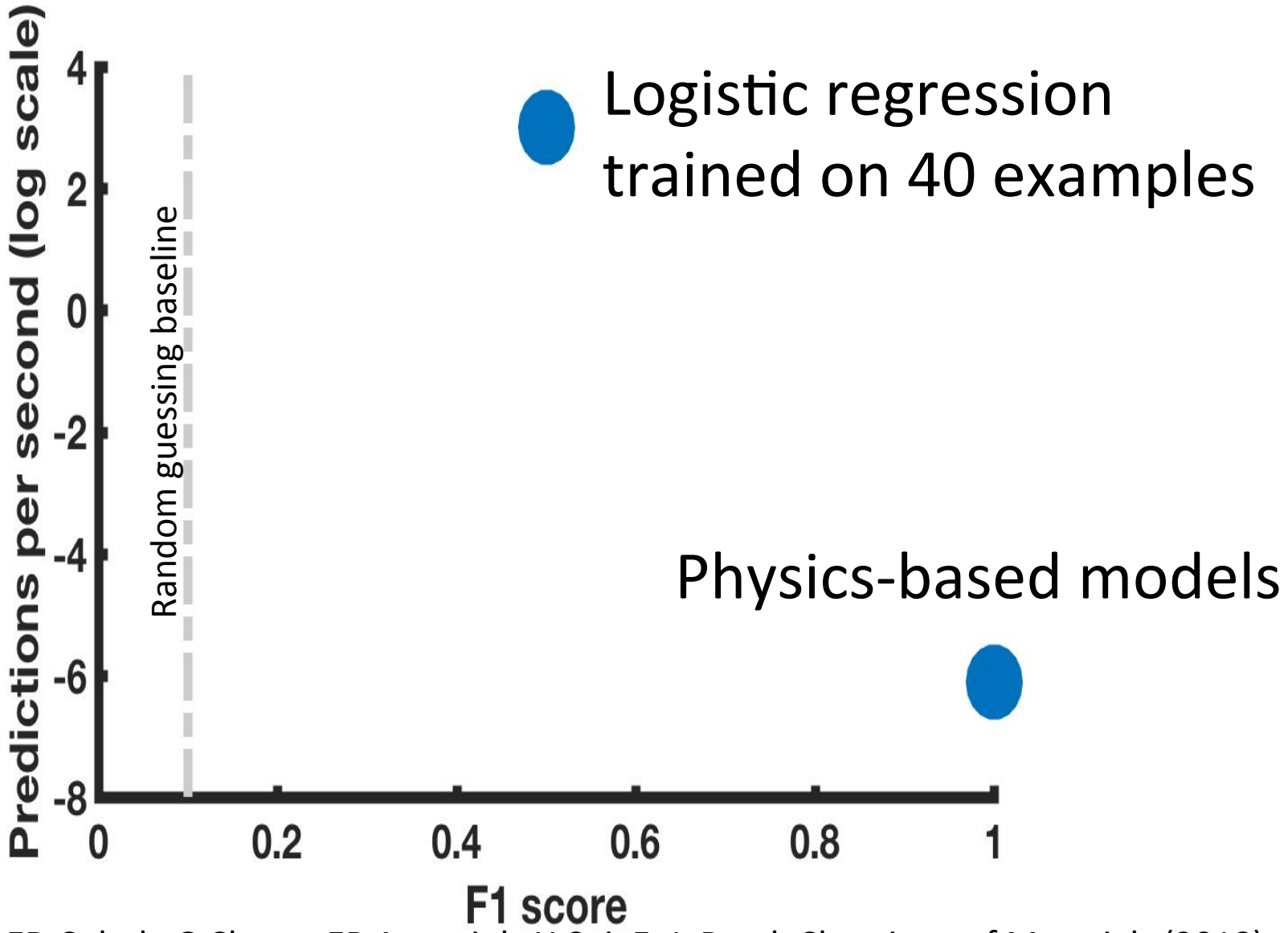
AD Sendek, ED Cubuk, G Cheon, ER Antoniuk, Y Cui, E. J. Reed, Chemistry of Materials (2018).

Our algorithm outperforms humans in accuracy and speed



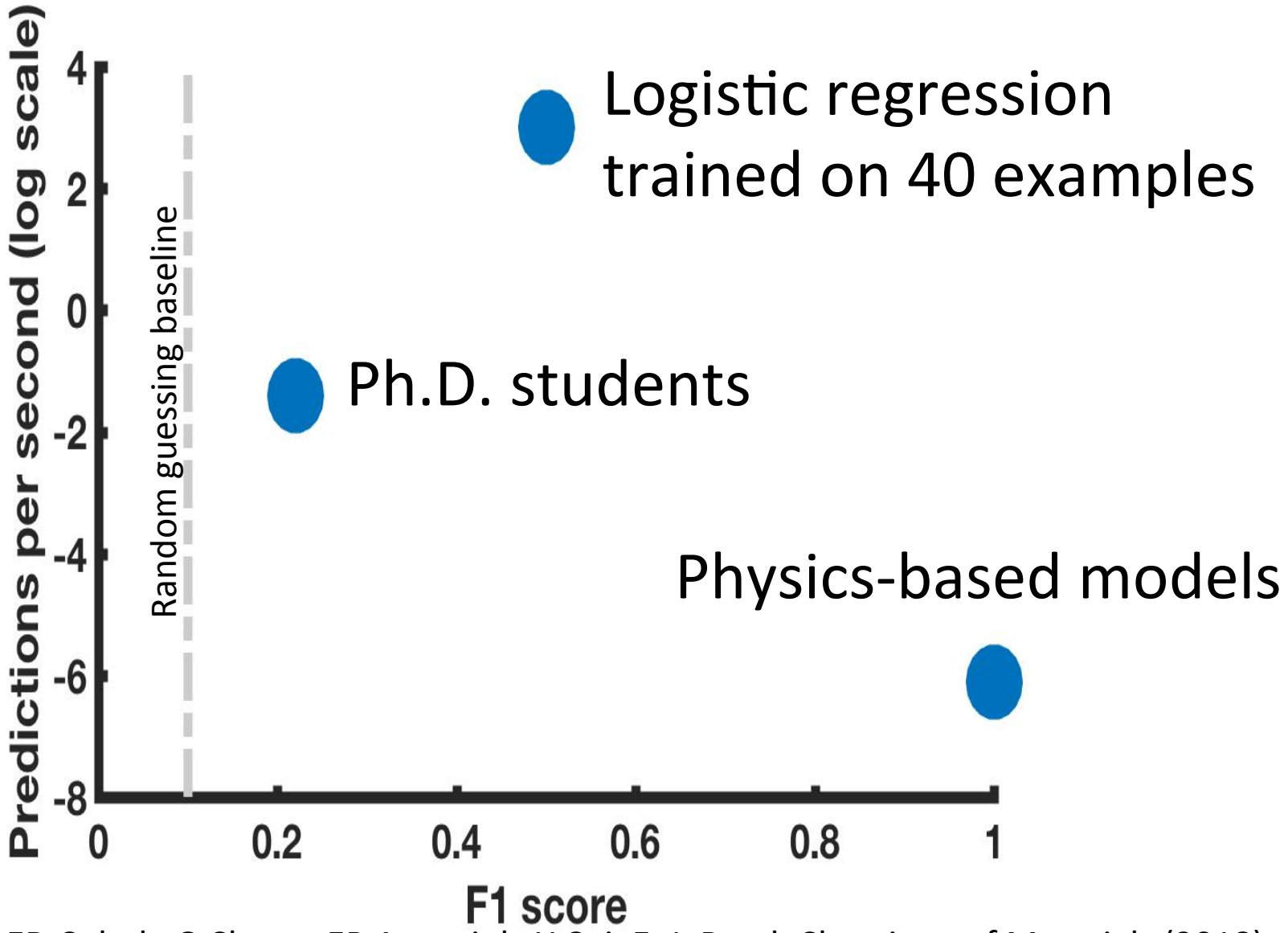
AD Sendek, ED Cubuk, G Cheon, ER Antoniuk, Y Cui, E. J. Reed, Chemistry of Materials (2018).

Our algorithm outperforms humans in accuracy and speed



AD Sendek, ED Cubuk, G Cheon, ER Antoniuk, Y Cui, E. J. Reed, Chemistry of Materials (2018).

Our algorithm outperforms humans in accuracy and speed

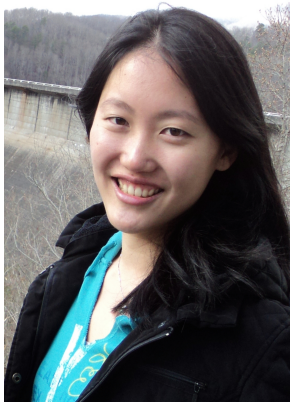


AD Sendek, ED Cubuk, G Cheon, ER Antoniuk, Y Cui, E. J. Reed, Chemistry of Materials (2018).

ACKNOWLEDGEMENTS

Austin Sendek
(now CEO AIONICS)

Ekin Dogus Cubuk
(now at Google Brain)



Qian Yang
(now at U Conn)

Gowoon Cheon

TomKatCenter
FOR SUSTAINABLE ENERGY
STANFORD UNIVERSITY

This work is supported by:

- Stanford TomKat Center for Sustainable Energy Seed grant
- Stanford Graduate Fellowship program

Acknowledgements

This work is supported by:

- Army High Performance Computing Research Center(AHPCRC)
- Army Research Office W911NF-15-1-0570
- NSF EECS-1436626 and DMR-1455050
- Office of Naval Research N00014-15-1-2697

G.Cheon et al., Nano Letters, 2017
Data Mining for New Two- and One-Dimensional Weakly Bonded Solids and Lattice-Commensurate Heterostructures

Gowoon Cheon

Also:

- Austin Sendek
- Karel-Alexander Duerloo
- Chase Porter
- Yuan Chen