Karsten W. Jacobsen
Department of Physics
Technical University of Denmark

Predicting material properties without knowing where the atoms are.

Computational materials screening

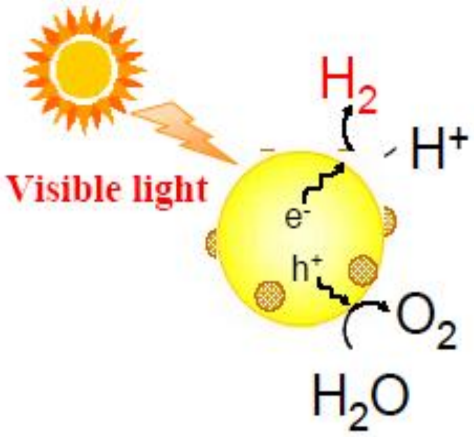
- Finding new and better materials for
  - solar cells
  - solar to fuel conversion
  - catalysis
  - structural materials
  - electronics

Some questions:

- Which materials properties should be computed?
  - “descriptors”
- How should the descriptors be computed (DFT, many-body perturbation theory…)?
- How should one search in materials space?

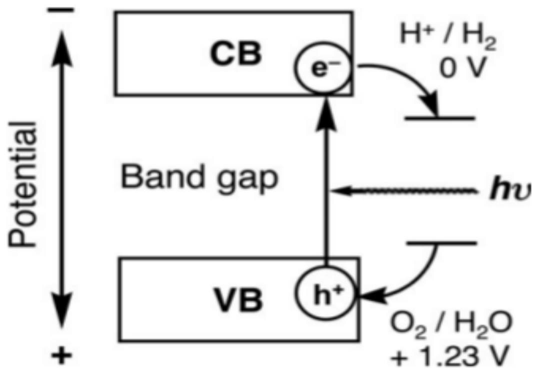
- Can machine learning be used to speed up computational screening?

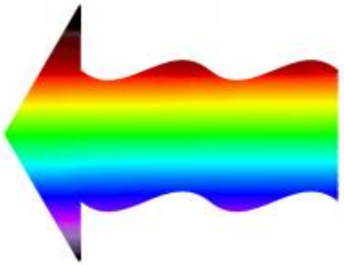
An example:
light-induced water splitting



Descriptors:

- **Stability of material**
  - Heat of formation
- **Good light absorption**
  - Bandgap in the visible range
- **Photogenerated charges at right potentials**
  - Band edges straddle the water redox potentials



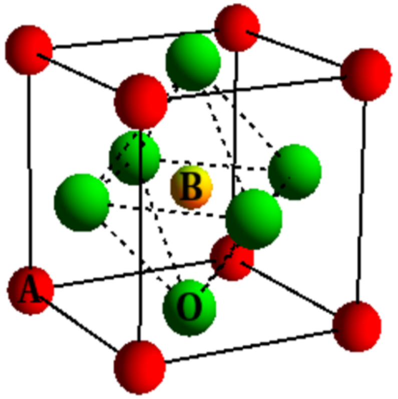


Principle of water splitting using semiconductor photocatalysts.

(Fujishima and Honda, Nature 1972)

Cubic perovskites ABX_3

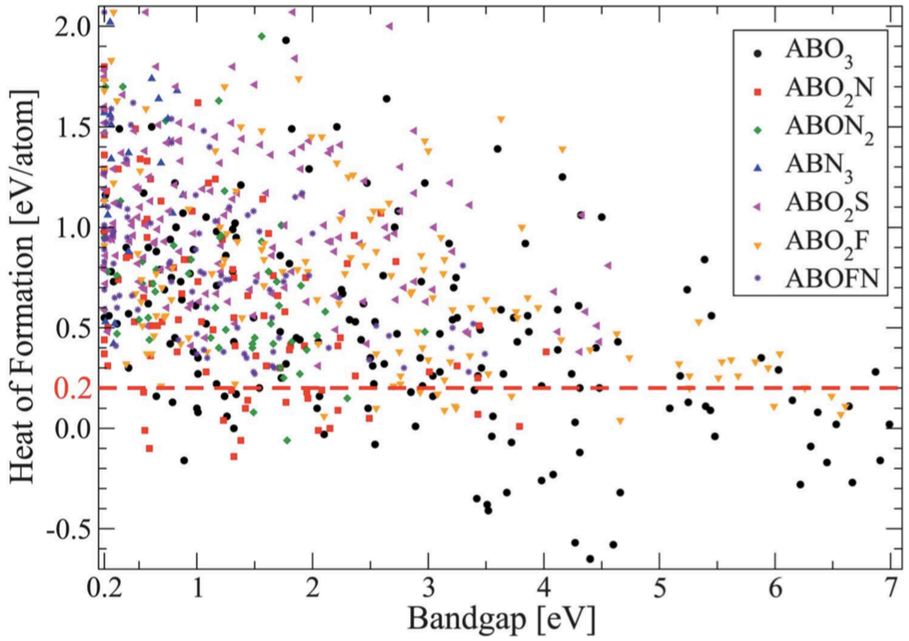
(X_3= O_3, O_2N, ON_2, N_3, O_2S, O_2F, OFN)



Screening criteria:

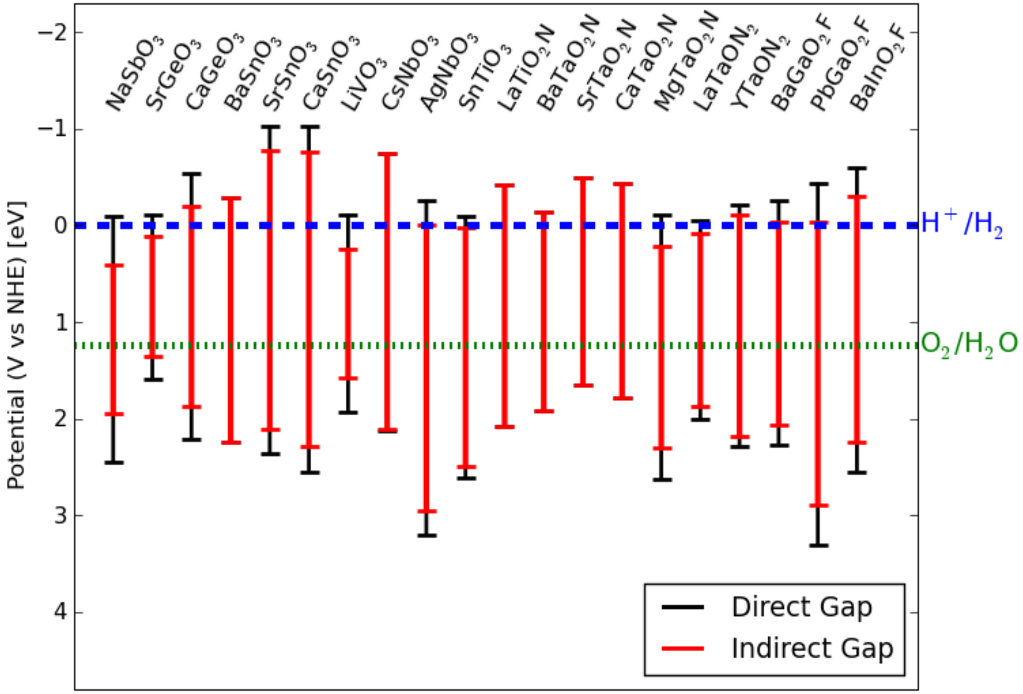
- Stability (heat of formation)
- Band gap
- Band alignment

~19000 materials



→

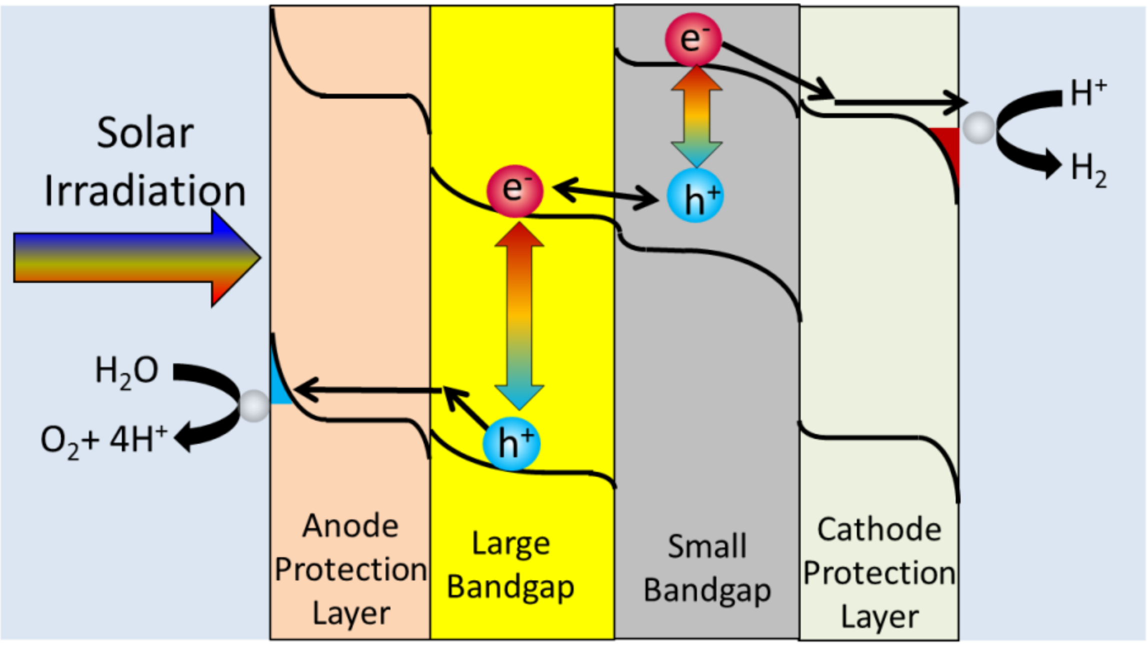
20 candidate materials
About half are known



Castelli, Olsen, Datta, Landis, Dahl, Thygesen, Jacobsen, *Energy & Environmental Science*, *5*(2), 5814 (2012).
Castelli, Landis, Thygesen, Dahl, Chorkendorff, Jaramillo, Jacobsen, Energy Environ Sci **5**, 9034 (2012)

Photoelectrochemical water splitting – a more realistic device

Required materials:
Light absorbing materials – small/large band gaps
Protection layers
Catalysts
p-n junctions



● = catalyst

Small bandgap semiconductor: ~1.1 eV Silicon

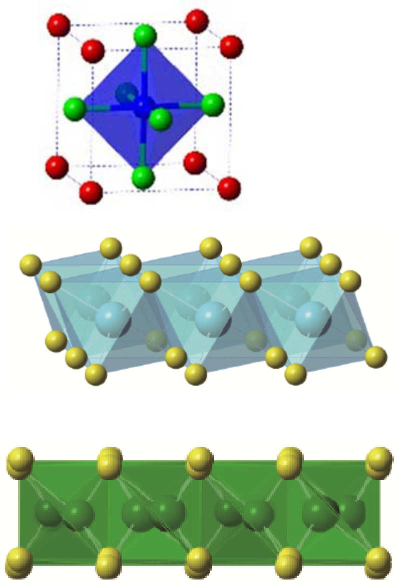
Large bandgap semiconductor: ~1.8 eV ??

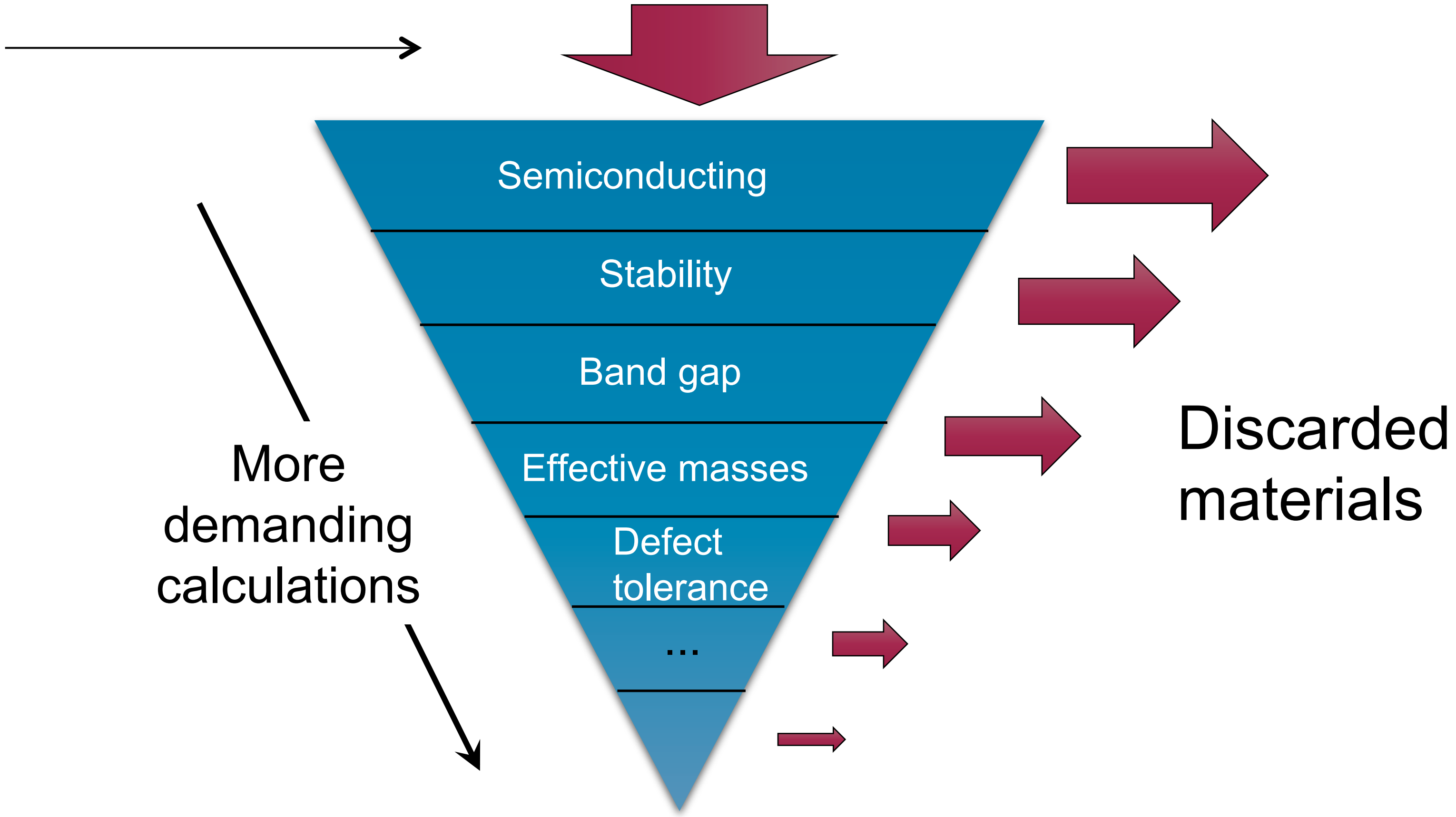
Sulfide ABS_3 – screening funnel

Initial structures and compositions

53*54/2 = 1431 different chemical compositions

8 crystal structures





Semiconducting

Stability

Band gap

Effective masses

Defect tolerance

…

More demanding calculations

Discarded materials

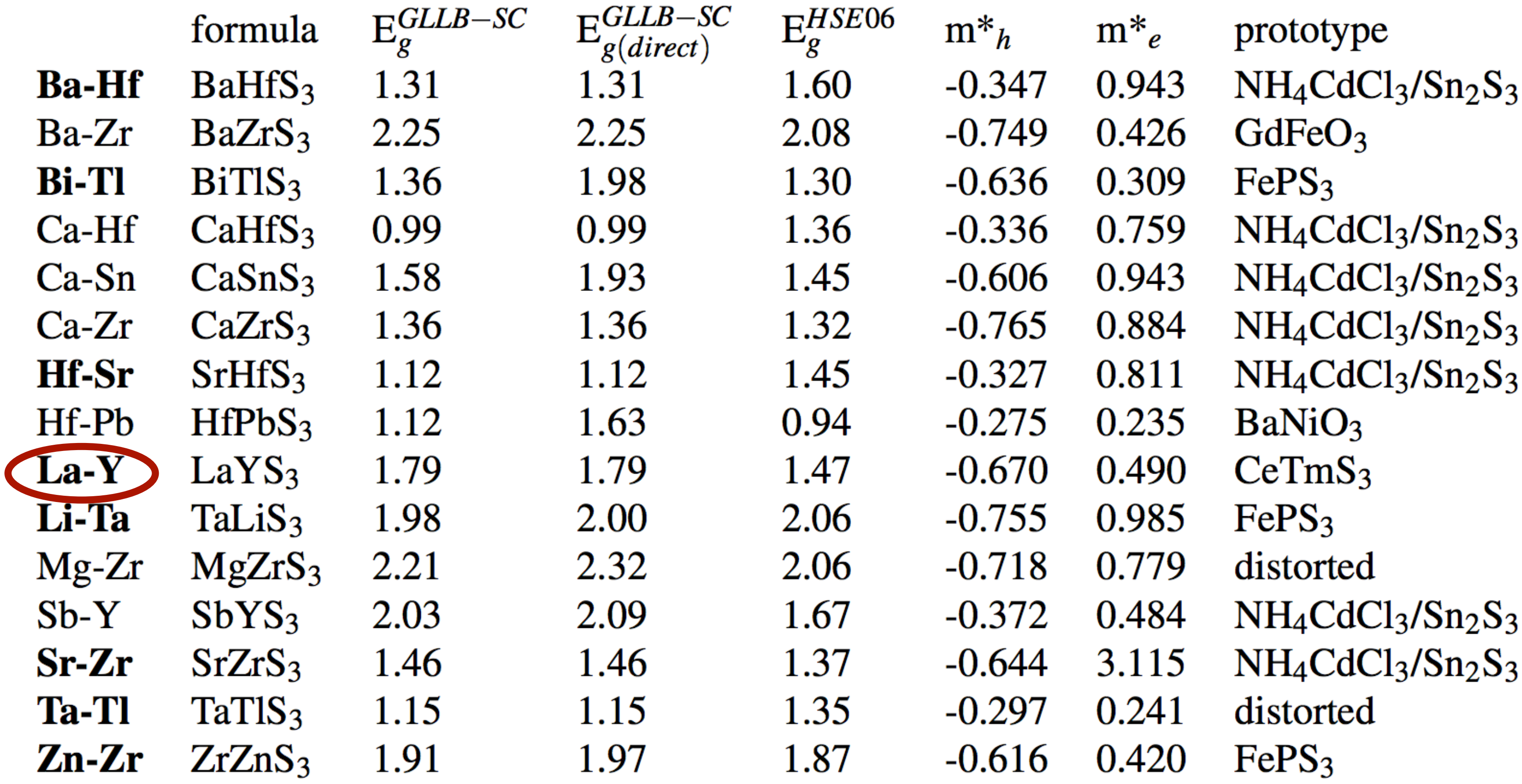
Candidates for experimental investigation

Kuhar, Crovetto, Pandey, Thygesen, Seger, Vesborg, Hansen, Chorkendorff, Jacobsen, *Energy Environ. Sci.*, **10**, 2579 (2017)

Final list of ABS_3 sulfides for photoabsorption

Initially
~ 1400 compositions
x 8 structures

Bold: all low-energy phases have relevant band gaps



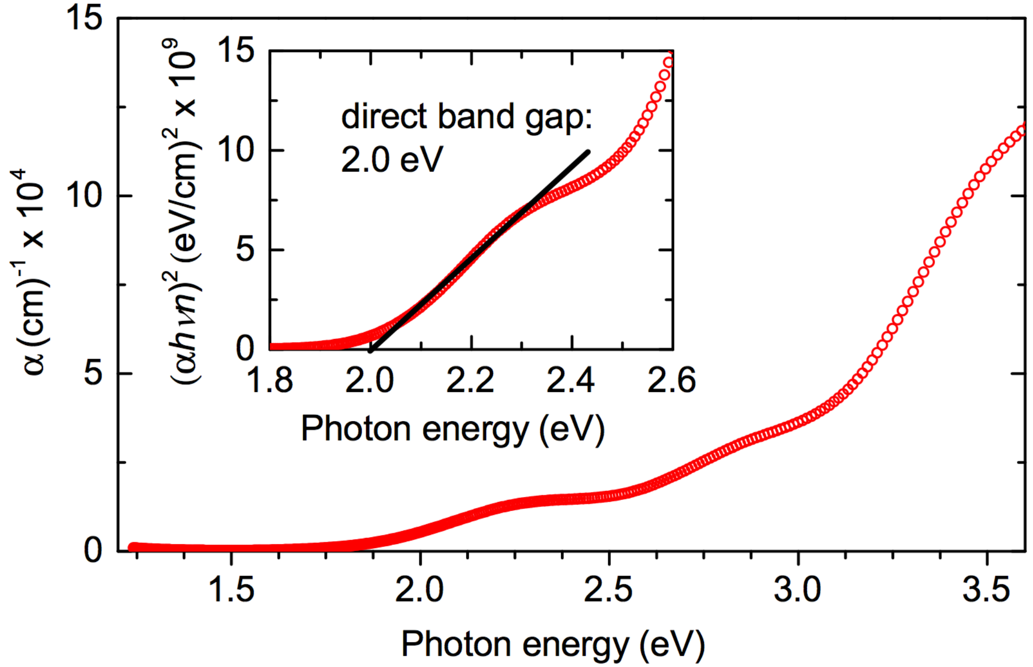
	formula	$E_g^{GLLB-SC}$	$E_{g(direct)}^{GLLB-SC}$	E_g^{HSE06}	m^*_h	m^*_e	prototype
Ba-Hf	$BaHfS_3$	1.31	1.31	1.60	-0.347	0.943	NH_4CdCl_3/Sn_2S_3
Ba-Zr	$BaZrS_3$	2.25	2.25	2.08	-0.749	0.426	$GdFeO_3$
Bi-Tl	$BiTlS_3$	1.36	1.98	1.30	-0.636	0.309	$FePS_3$
Ca-Hf	$CaHfS_3$	0.99	0.99	1.36	-0.336	0.759	NH_4CdCl_3/Sn_2S_3
Ca-Sn	$CaSnS_3$	1.58	1.93	1.45	-0.606	0.943	NH_4CdCl_3/Sn_2S_3
Ca-Zr	$CaZrS_3$	1.36	1.36	1.32	-0.765	0.884	NH_4CdCl_3/Sn_2S_3
Hf-Sr	$SrHfS_3$	1.12	1.12	1.45	-0.327	0.811	NH_4CdCl_3/Sn_2S_3
Hf-Pb	$HfPbS_3$	1.12	1.63	0.94	-0.275	0.235	$BaNiO_3$
La-Y	$LaYS_3$	1.79	1.79	1.47	-0.670	0.490	$CeTmS_3$
Li-Ta	$TaLiS_3$	1.98	2.00	2.06	-0.755	0.985	$FePS_3$
Mg-Zr	$MgZrS_3$	2.21	2.32	2.06	-0.718	0.779	distorted
Sb-Y	$SbYS_3$	2.03	2.09	1.67	-0.372	0.484	NH_4CdCl_3/Sn_2S_3
Sr-Zr	$SrZrS_3$	1.46	1.46	1.37	-0.644	3.115	NH_4CdCl_3/Sn_2S_3
Ta-Tl	$TaTlS_3$	1.15	1.15	1.35	-0.297	0.241	distorted
Zn-Zr	$ZrZnS_3$	1.91	1.97	1.87	-0.616	0.420	$FePS_3$

Kuhar, Crovetto, Pandey, Thygesen, Seger, Vesborg, Hansen, Chorkendorff, Jacobsen, Energy and Environmental Science, **10**, 2579 (2017).

$BaZrS_3$: W. Meng et al., *Chem Mater*, **28**, 821 (2016)

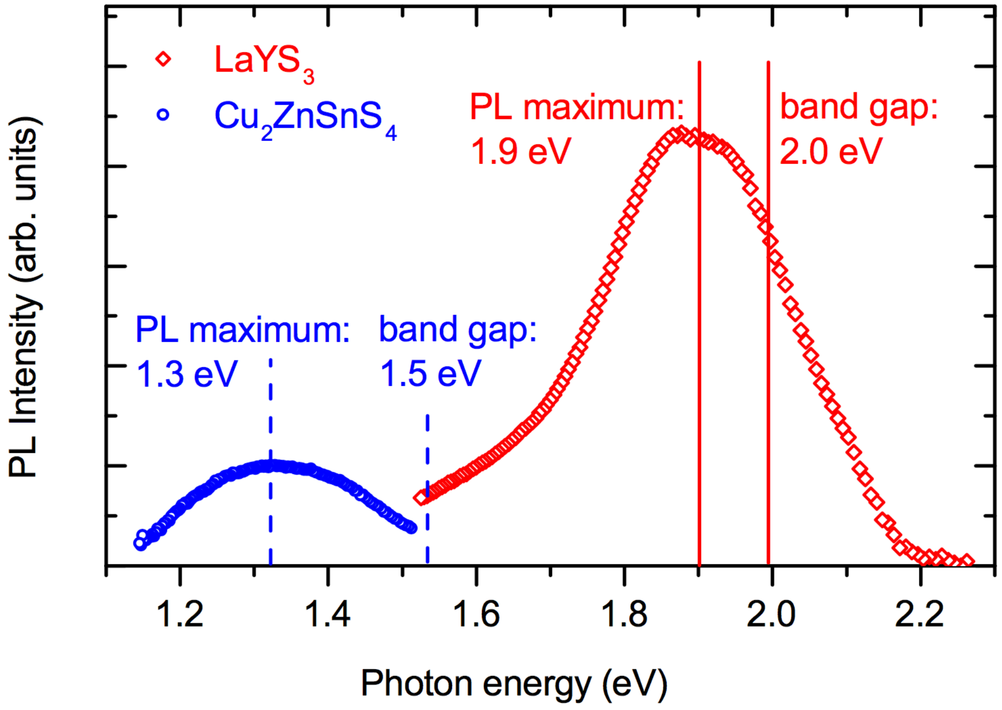
$LaYS_3$ – experiments

Spectroscopic ellipsometry – light absorption coefficient



Direct band gap determined from absorption coefficient and refractive index

Photoluminescence



Kuhar, Crovetto, Pandey, Thygesen, Seger, Vesborg, Hansen, Chorkendorff, Jacobsen, Energy and Environmental Science, **10**, 2579 (2017).

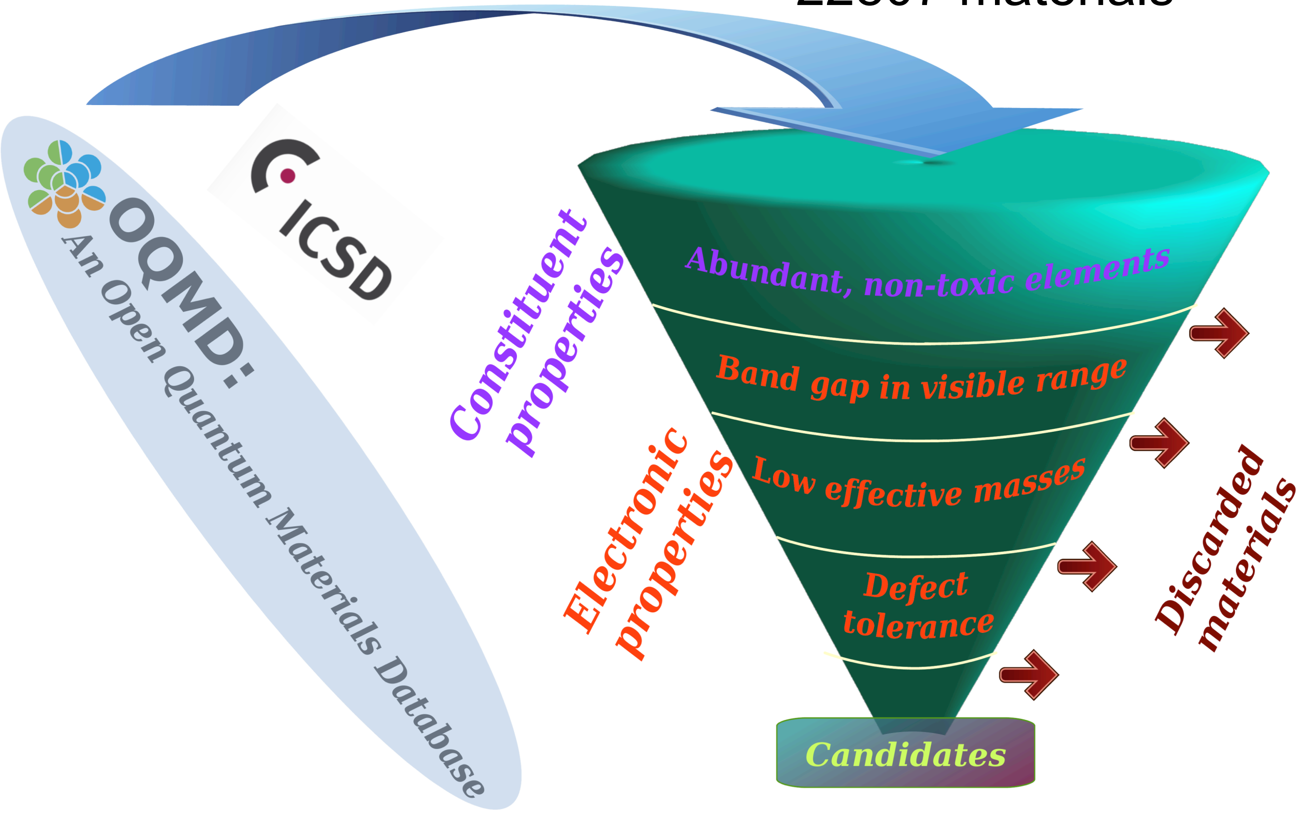
Screening of *known* materials for photovoltaics or water splitting

Advantages:

Materials known to
be stable or metastable

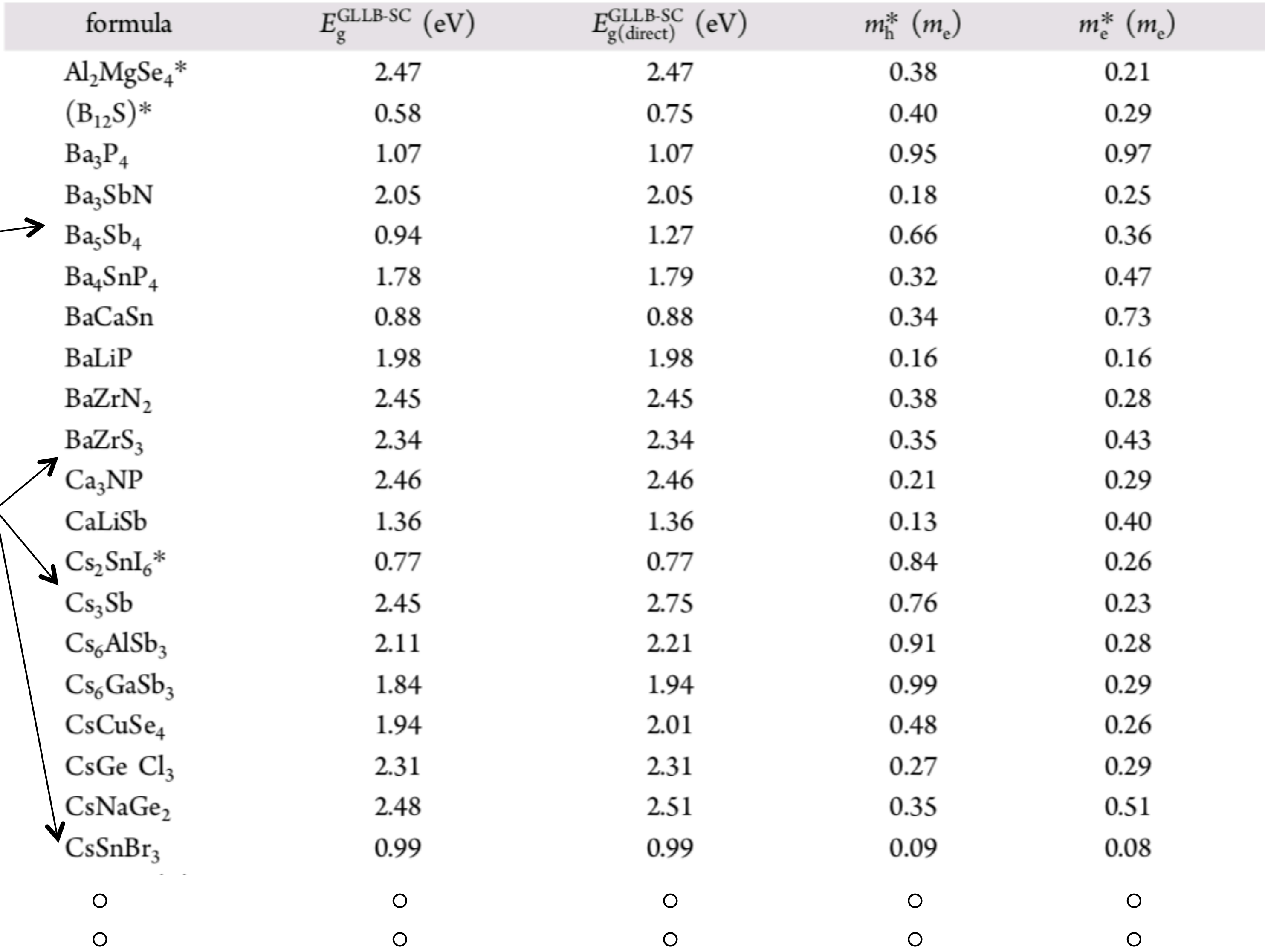
Known synthesis procedures

Current limitations:
Binary or ternary compounds
Non-magnetic compounds



K. Kuhar, M. Pandey, K. S. Thygesen, and K. W. Jacobsen, ACS Energy Lett., 3, 436 (2018)

Screening results (74 materials)



formula	$E_g^{GLLB\text{-}SC}$ (eV)	$E_{g(direct)}^{GLLB\text{-}SC}$ (eV)	m_h^* (m_e)	m_e^* (m_e)
Al_2MgSe_4*	2.47	2.47	0.38	0.21
($B_{12}S$)*	0.58	0.75	0.40	0.29
Ba_3P_4	1.07	1.07	0.95	0.97
Ba_3SbN	2.05	2.05	0.18	0.25
Ba_5Sb_4	0.94	1.27	0.66	0.36
Ba_4SnP_4	1.78	1.79	0.32	0.47
BaCaSn	0.88	0.88	0.34	0.73
BaLiP	1.98	1.98	0.16	0.16
$BaZrN_2$	2.45	2.45	0.38	0.28
$BaZrS_3$	2.34	2.34	0.35	0.43
Ca_3NP	2.46	2.46	0.21	0.29
CaLiSb	1.36	1.36	0.13	0.40
Cs_2SnI_6*	0.77	0.77	0.84	0.26
Cs_3Sb	2.45	2.75	0.76	0.23
Cs_6AlSb_3	2.11	2.21	0.91	0.28
Cs_6GaSb_3	1.84	1.94	0.99	0.29
$CsCuSe_4$	1.94	2.01	0.48	0.26
$CsGe\ Cl_3$	2.31	2.31	0.27	0.29
$CsNaGe_2$	2.48	2.51	0.35	0.51
$CsSnBr_3$	0.99	0.99	0.09	0.08
○	○	○	○	○
○	○	○	○	○

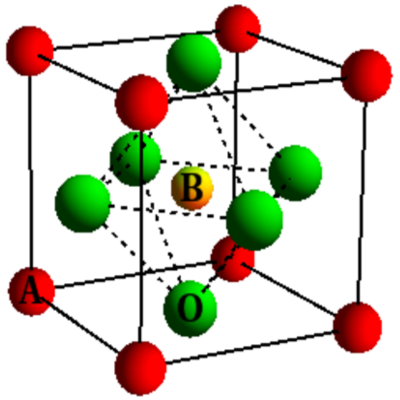
Antiperovskite

Known perovskites

Database available on-line at https://cmr.fysik.dtu.dk

K. Kuhar, M. Pandey, K. S. Thygesen, and K. W. Jacobsen, ACS Energy Lett., 3, 436 (2018)

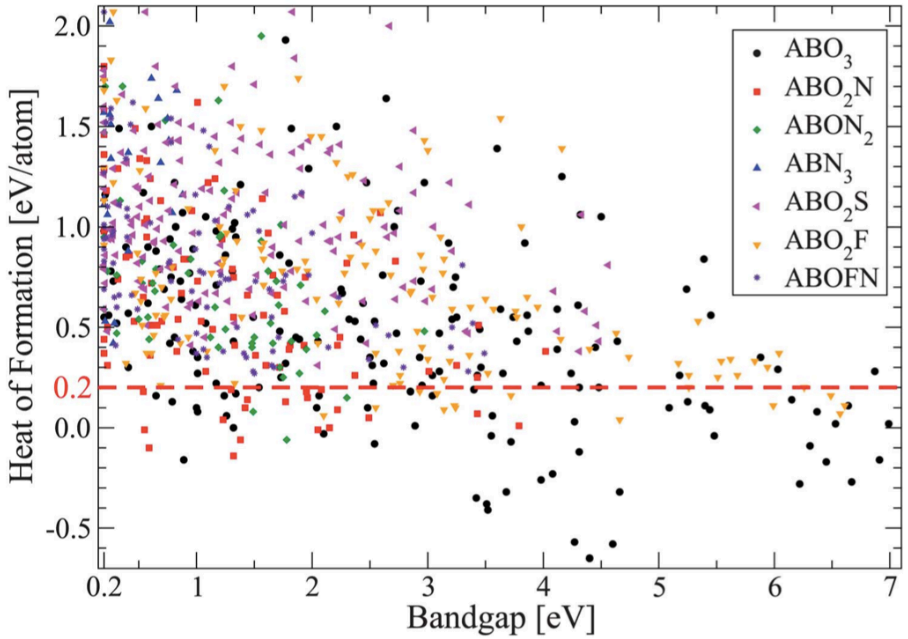
Back to the cubic perovskites ABX_3



Screening criteria:

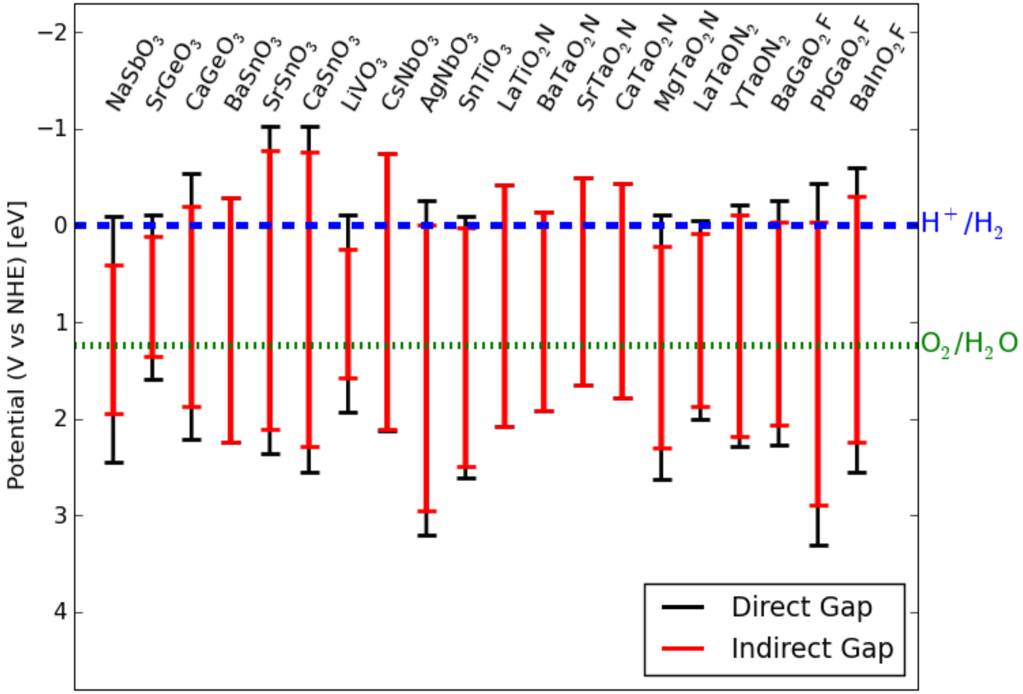
- Stability (heat of formation)
- Band gap
- Band alignment

~19000 materials



→

20 candidate materials
About half are known



Castelli, Olsen, Datta, Landis, Dahl, Thygesen, Jacobsen, *Energy & Environmental Science*, *5*(2), 5814 (2012).
Castelli, Landis, Thygesen, Dahl, Chorkendorff, Jaramillo, Jacobsen, Energy Environ Sci **5**, 9034 (2012)

Machine learning: Kernel regression

Fitting a function *f(x)* based on data points $y_i = f(x_i)$

Drop a Gaussian on each data point:

$$k(x, x_i) = \exp(-|x - x_i|^2/2\rho^2)$$

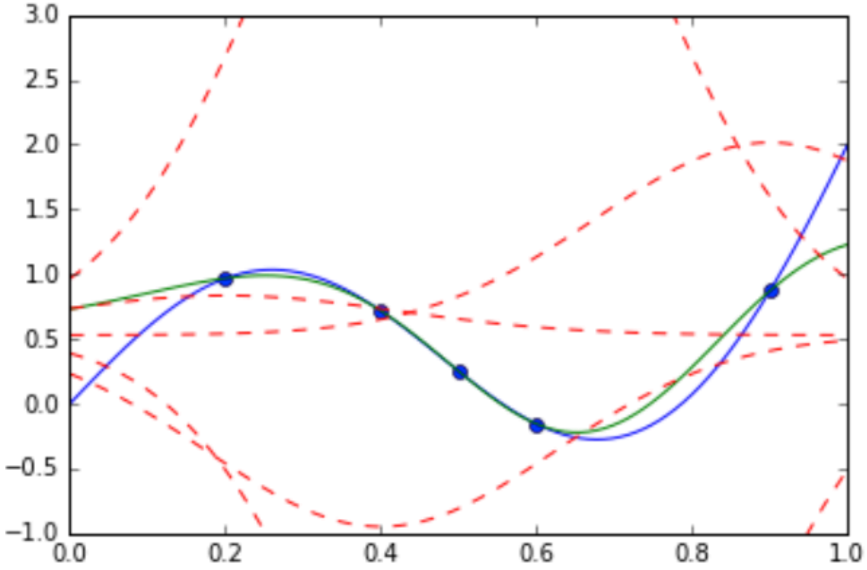
Interpolation: $y(x) = \sum_i k(x, x_i)\alpha_i$

Coefficients determined by data points:

$$y_j = \sum_i k(x_j, x_i)\alpha_i = \sum_i K_{ji}\alpha_i \rightarrow \mathbf{y} = \mathbf{K}\boldsymbol{\alpha} \rightarrow \boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$$

Interpolation: $y(x) = \mathbf{k}^T\mathbf{K}^{-1}\mathbf{y}$

with $k_i = k(x, x_i)$



Green: f(x)
Blue: fit
Red: Gaussians

Kernel regression with uncertainties: Gaussian process

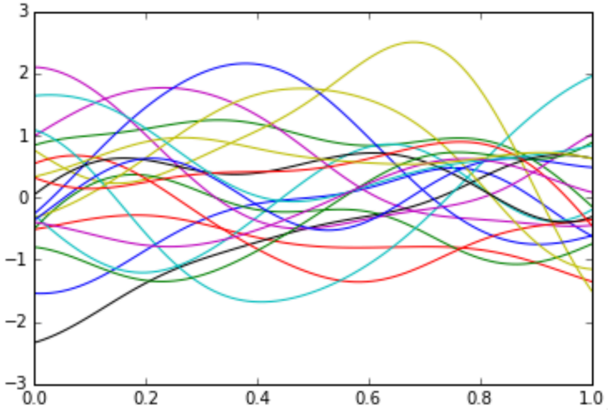
Based on Bayes theorem: $P(\mathrm{Model}|\mathrm{Data}) = \frac{1}{P(\mathrm{Data})} P(\mathrm{Data}|\mathrm{Model}) P_0(\mathrm{Model})$

“Reinterpretaion” of kernel function as correlation: $K_{ij} = \langle y(x_i) y(x_j) \rangle = k(x_i, x_j) = \exp(-|x_i - x_j|^2 / 2\rho^2)$

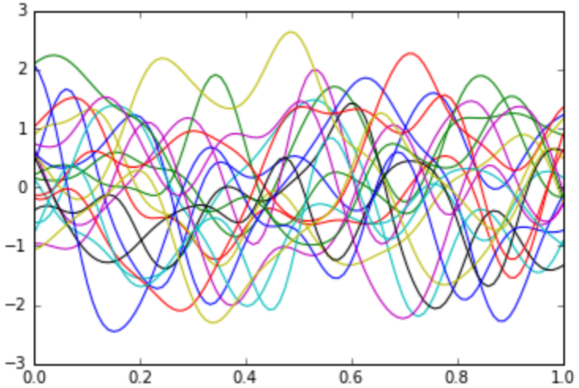
Prior probability distribution (i.e. *without* data points):

$$P_0(\mathbf{y}) = \frac{1}{\sqrt{2\pi det(\mathbf{K})}} \exp\left(-\frac{1}{2}\mathbf{y^T K^{-1} y}\right), \ \mathbf{y}^T = (y(x_1), y(x_2), \ldots, y(x_N))$$

$\rho^2 = 0.1$



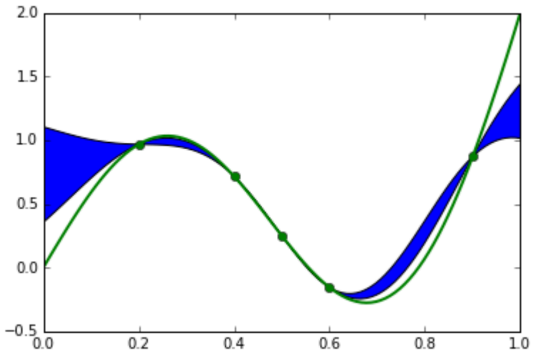
$\rho^2 = 0.01$



Gaussian process

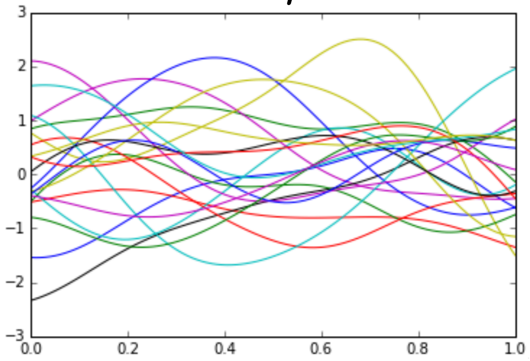
Fitting a function f(x) based on data points y_i=$f(x_i)$

Kernel: $k(x_i, x_j) = \exp(-|x_i - x_j|^2/2\rho^2)$

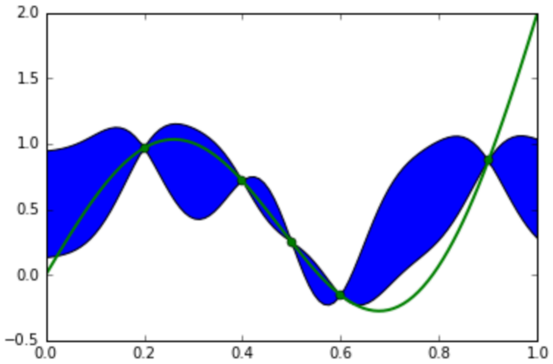


Update of model because of data

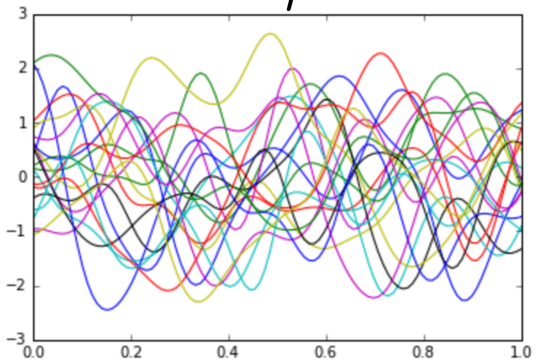
$\rho^2 = 0.1$



The value of ρ can be addressed by so-called cross validation

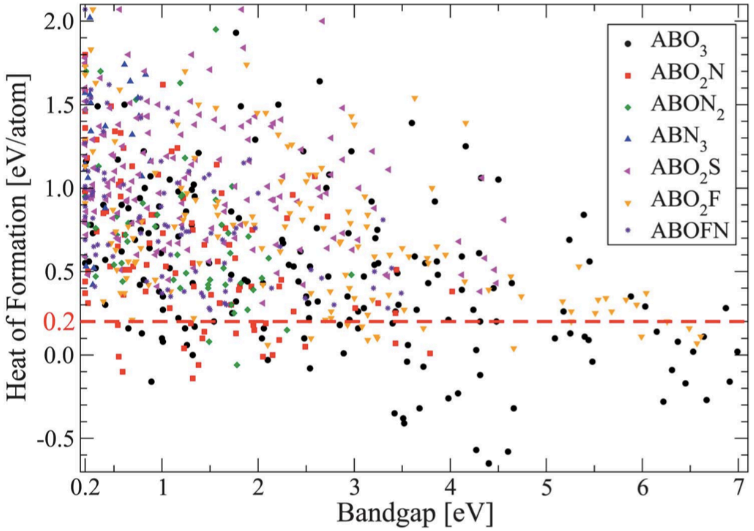


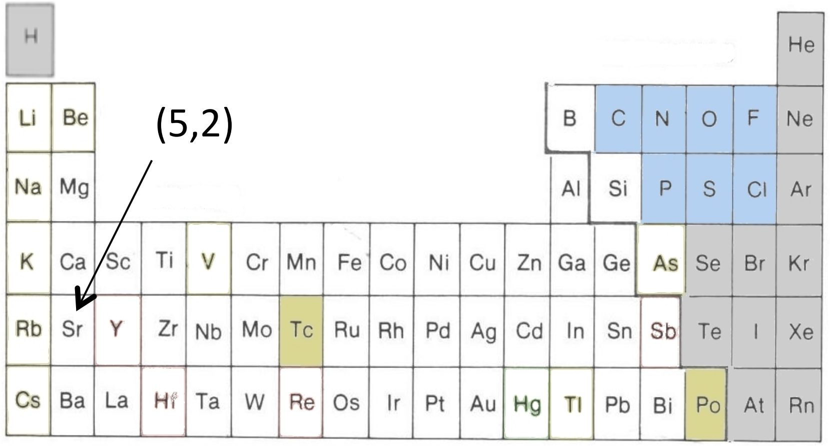
$\rho^2 = 0.01$



Back to water splitting with machine learning

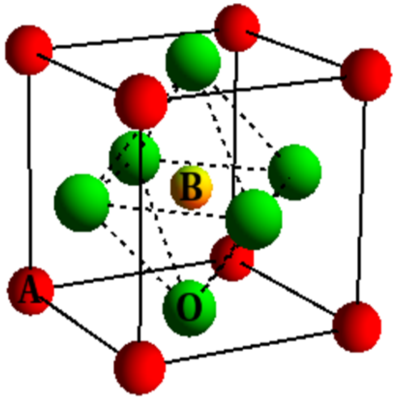
About 19000 cubic perovskites oxides, oxynitrides, oxysulfides, oxyfluorides, oxyfluornitrides





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	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn

(5,2)



ABO$_3$, ABON$_2$, ..

Fingerprint (x-vector):

$$x(\mathrm{SrTaO_2N}) = (5, 2, 6, 5, 2, 1, 0, 0)$$

O, N, S, F

Sr “coordinates”

Kernel function:

$$k(x_i, x_j) = \exp(-|x_i - x_j|^2/2\rho^2)$$

Water splitting with Gaussian process

Training on 500 perovskites (~2.6 % of the total dataset).

Prediction: $y(x) = \mathbf{k}^T \mathbf{K}^{-1} \mathbf{y}$ with

Only determined by “metric” (not by data)

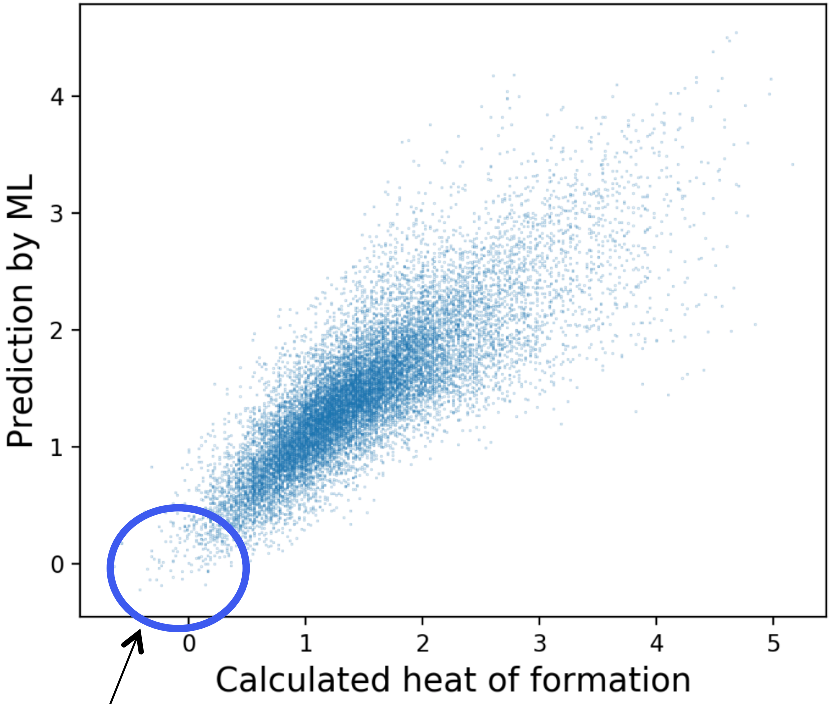
Data

$$k_i = k(x, x_i)$$

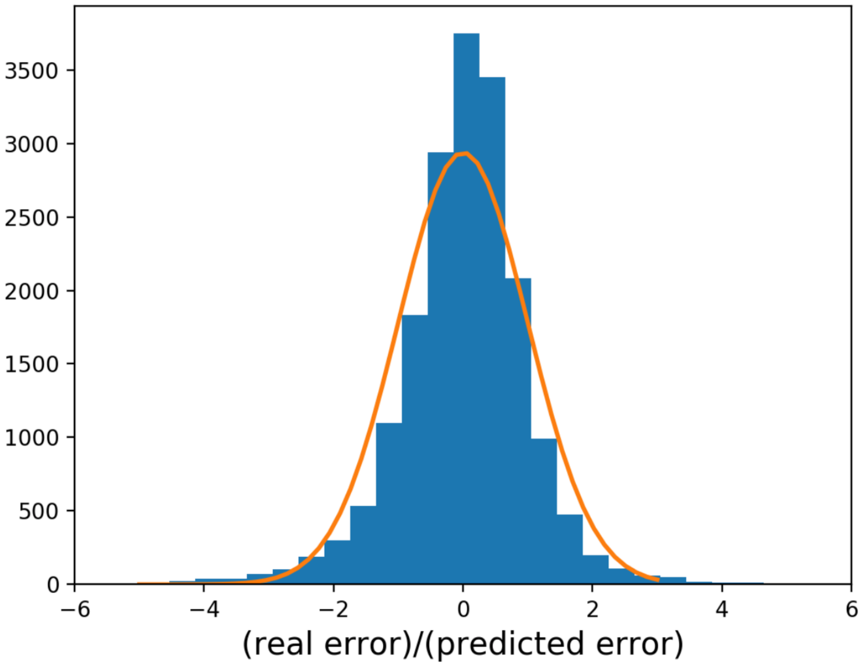
+ error prediction

Example: Heat of formation

Mean Absolute Error: 0.28 eV
Mean Absolute Predicted Error: 0.38 eV

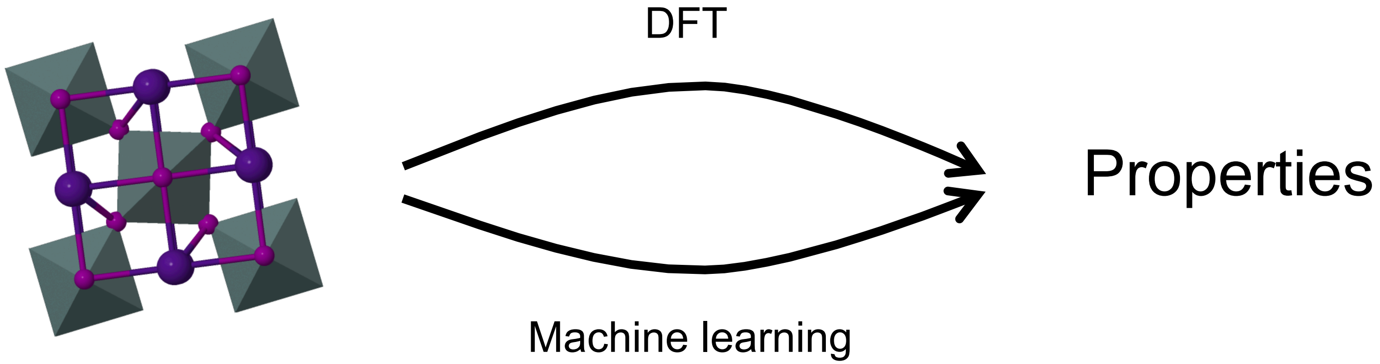


Stable compounds



Large reduction in the number of necessary DFT calculations for stable compounds!

Machine learning accelerated computational screening of *new* materials



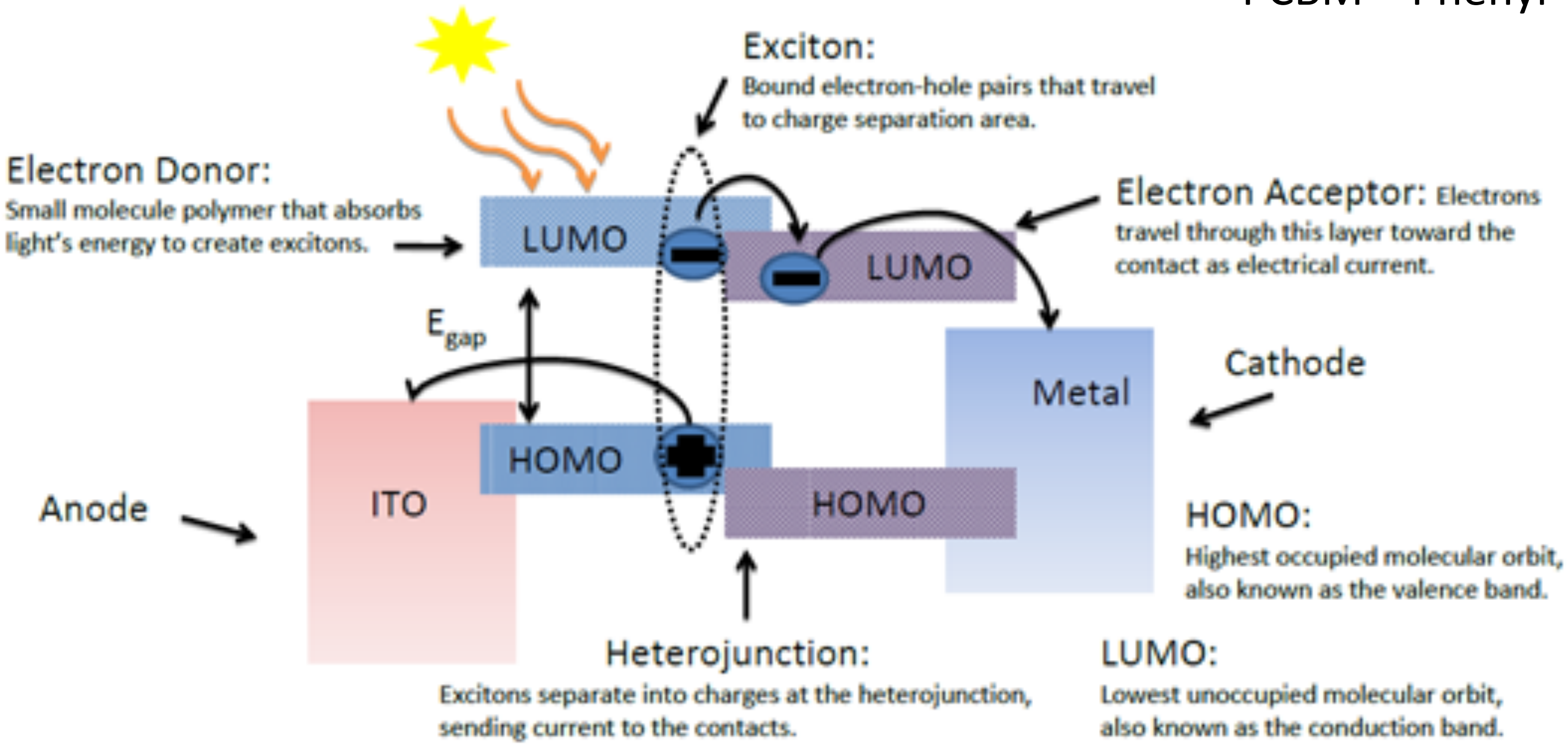
Two challenges:

1) Can we predict material properties for materials in many different structures where the detailed atomic positions are not known?

2) Can we invert the process so we go directly from properties to material? (to avoid evaluation of properties of maybe billions of (irrelevant) materials)

Organic solar cell
(PCBM-based blended polymer solar cell)

PCBM = Phenyl-C’61-Butyric-Acid-Methyl-Ester



Exciton:
Bound electron-hole pairs that travel to charge separation area.
Electron Donor:
Small molecule polymer that absorbs light’s energy to create excitons.
Electron Acceptor: Electrons travel through this layer toward the contact as electrical current.
LUMO
LUMO
E_{gap}
Metal
Cathode
HOMO
HOMO
ITO
Anode
HOMO:
Highest occupied molecular orbit, also known as the valence band.
Heterojunction:
Excitons separate into charges at the heterojunction, sending current to the contacts.
LUMO:
Lowest unoccupied molecular orbit, also known as the conduction band.

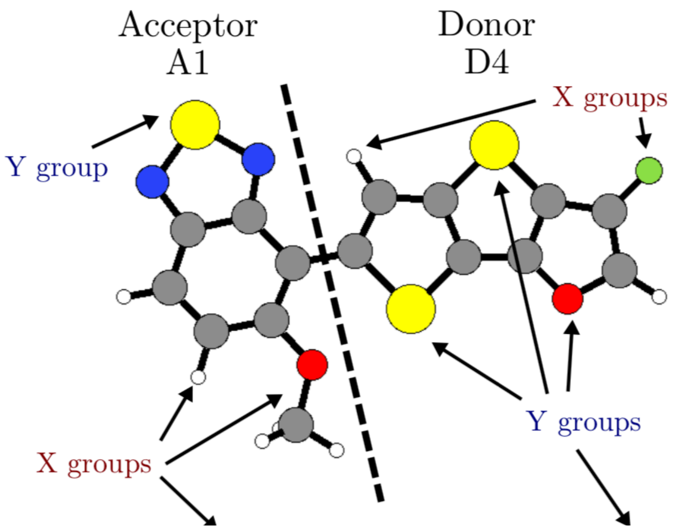
Peter Bjørn Jørgensen, Murat Mesta, Suranjan Shil, Juan Maria García Lastra, Karsten Wedel Jacobsen, Kristian Sommer Thygesen, and Mikkel N. Schmidt
The Journal of Chemical Physics **148**, special issue (2018)

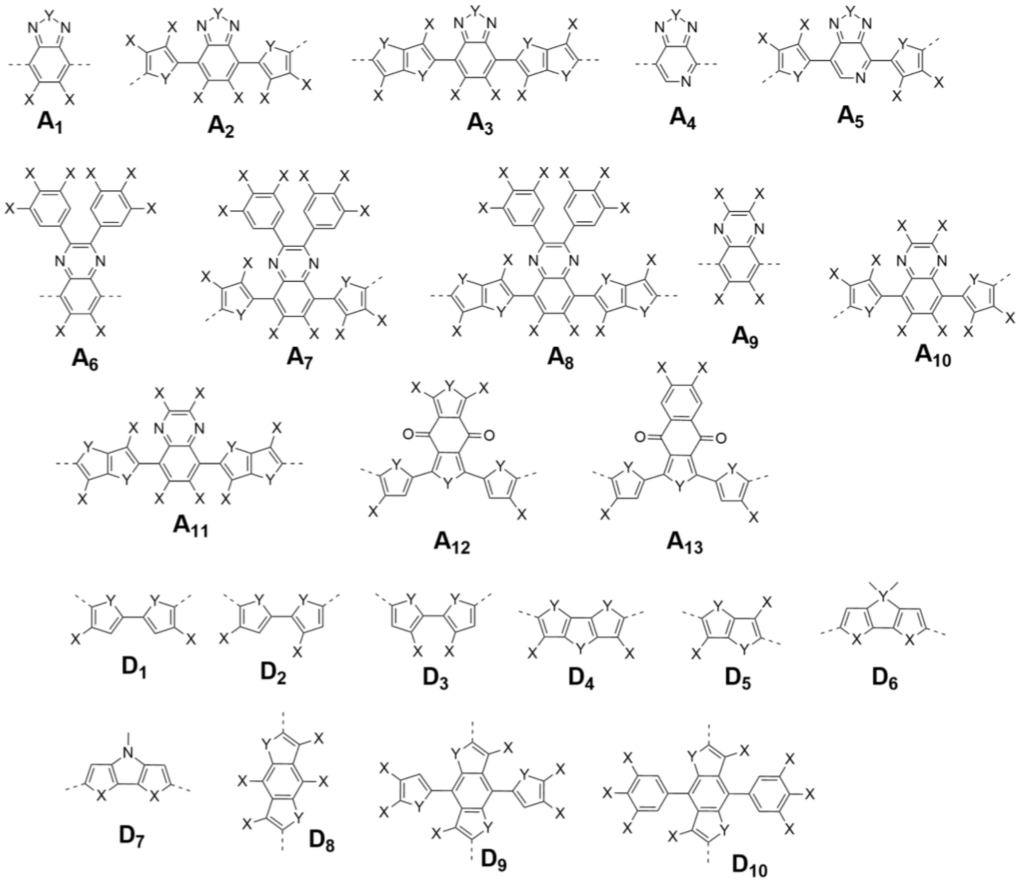
Donor-acceptor molecules (polymer units)

What is the position of the LUMO and the optical gap for these molecules?

Training set with 3989 molecules (Gaussian, B3LYP)

In principle 10^{14} molecules!
One prediction 1ms
-> Total 10^{11} sec
~ 3000 years





A(1-13) = Acceptors
D(1-10) = Donors

X = H, F, CH_3, OCH_3, SCH_3
Y (divalent) = O, S, Se, NCH_3
Y (tetravalent) = C, Si, Ge

Data representation

String representation of molecules.

Grammatical production rules.

No specification of atomic coordinates.

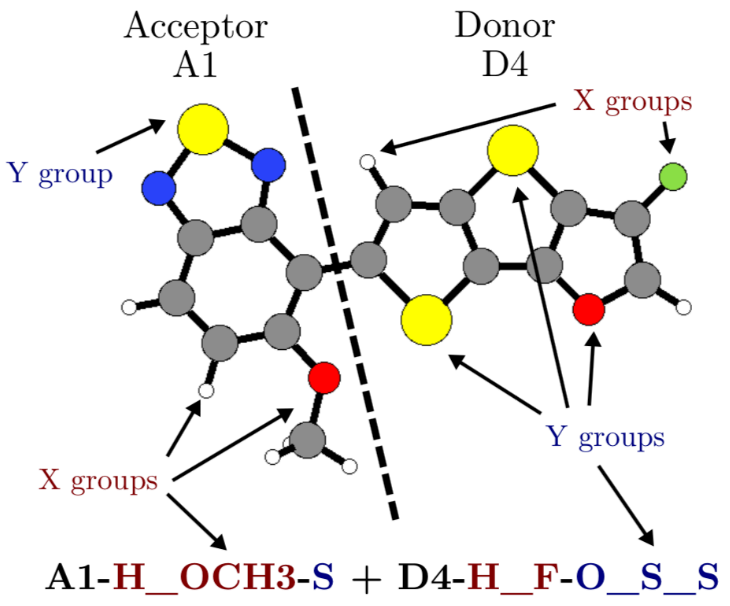


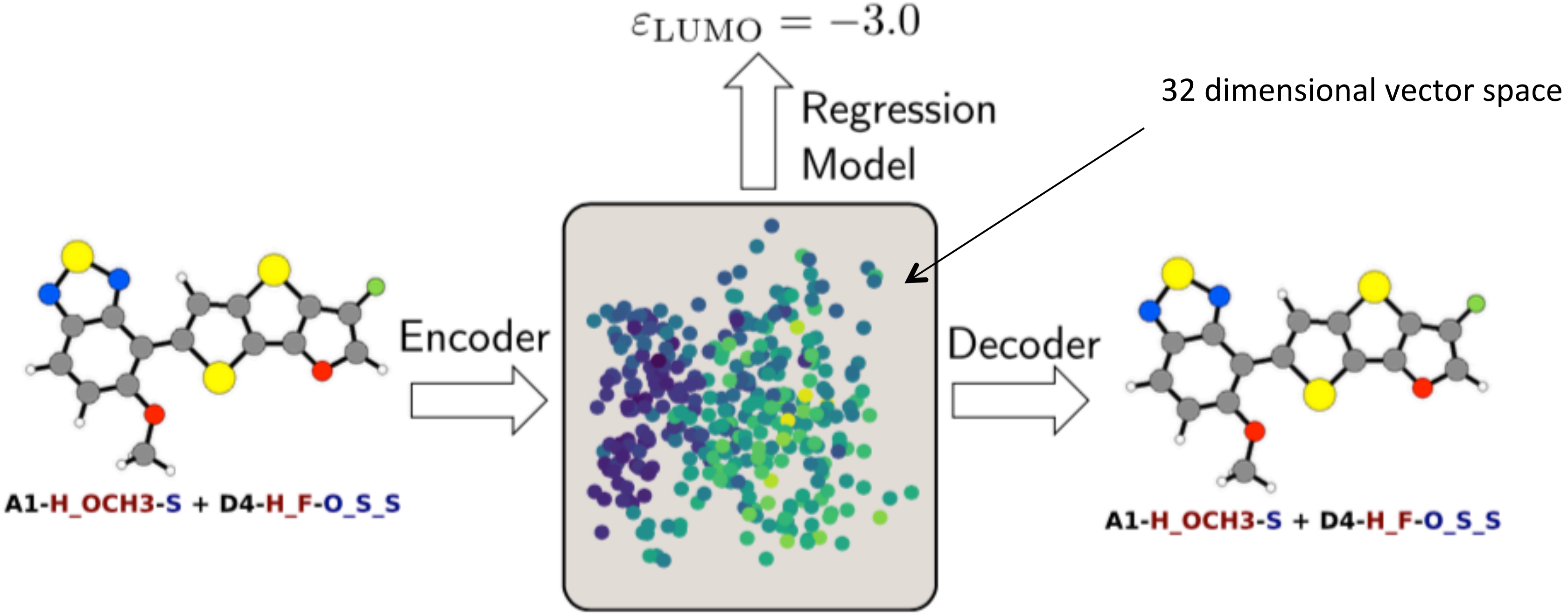
FIG. 2: String representation of one of the molecules of the solar cell dataset: “Acceptor backbone“-“X groups“-“Y groups“+“Donor backbone“-“X groups“-“Y groups“. Whenever no side groups are present “*” character is used instead.

Earlier work uses SMILES to represent molecules:

Gómez-Bombarelli et al. (2016), arXiv:1610.02415 [cs.LG].
Kusner et al. (2017), arXiv:1703.01925 [stat.ML].

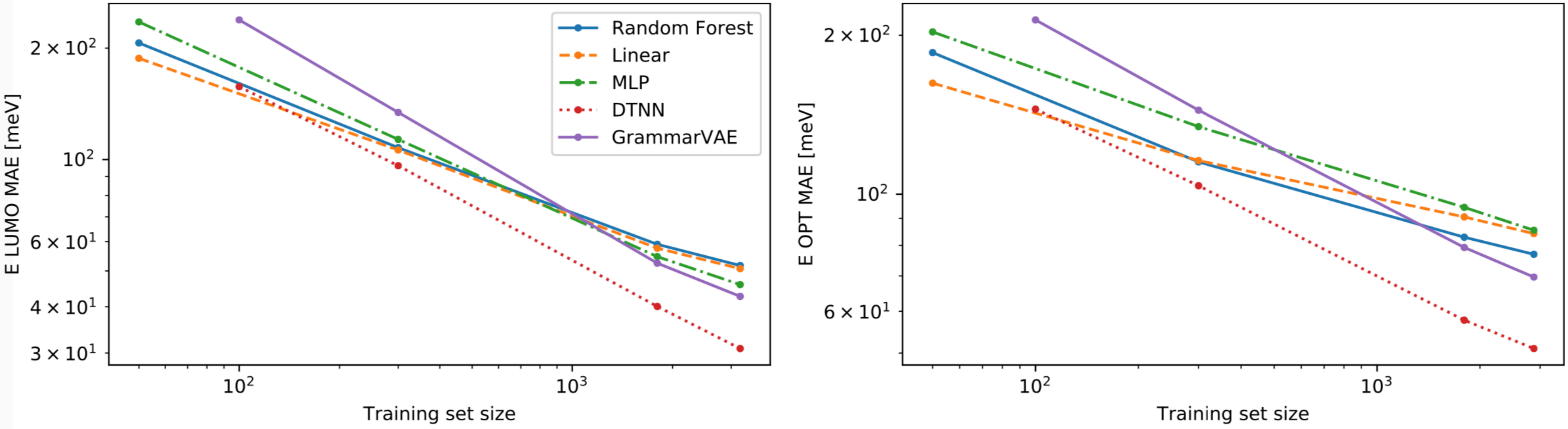
Variational autoencoder

Kingma and Welling [2013], Rezende et al. [2014]


$\varepsilon_{\mathrm{LUMO}} = -3.0$
Regression Model
Encoder
Decoder
A1-H_OCH3-S + D4-H_F-O_S_S
A1-H_OCH3-S + D4-H_F-O_S_S

32 dimensional vector space

Method comparison



Grammar variational autoencoder

Production rules:

m → A - OPTG - OPTG + D - OPTG - OPTG

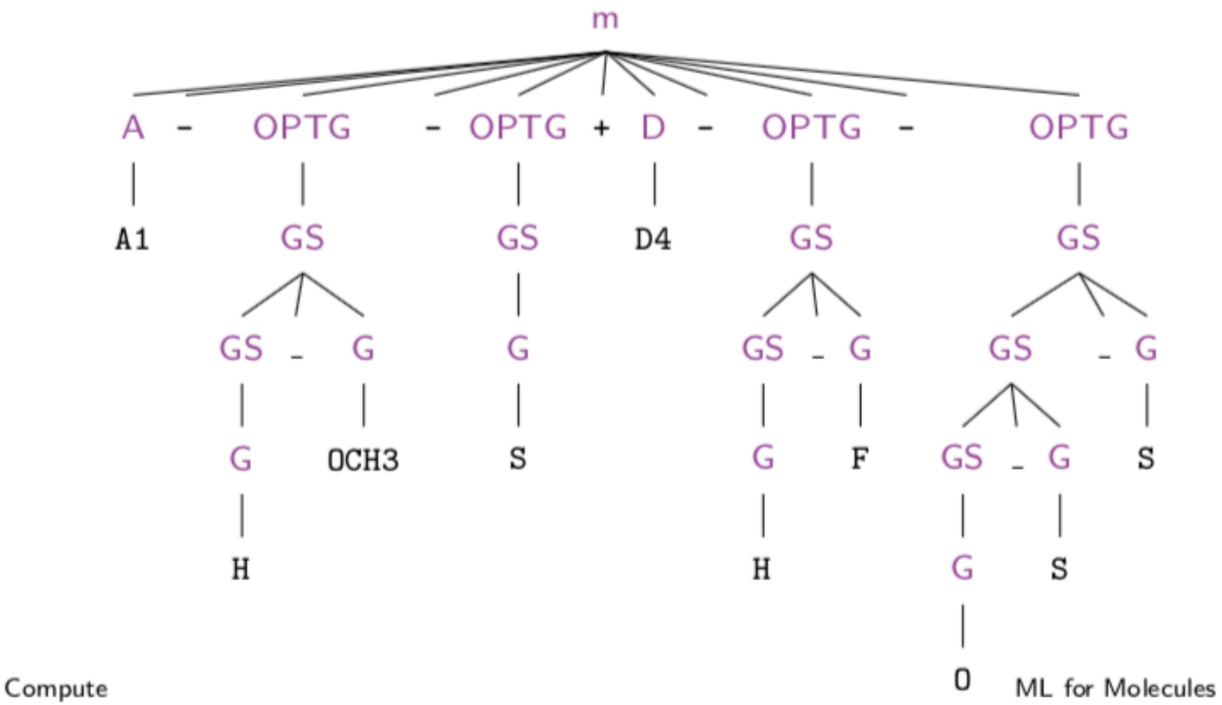
A → A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13

D → D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10

OPTG → * | GS

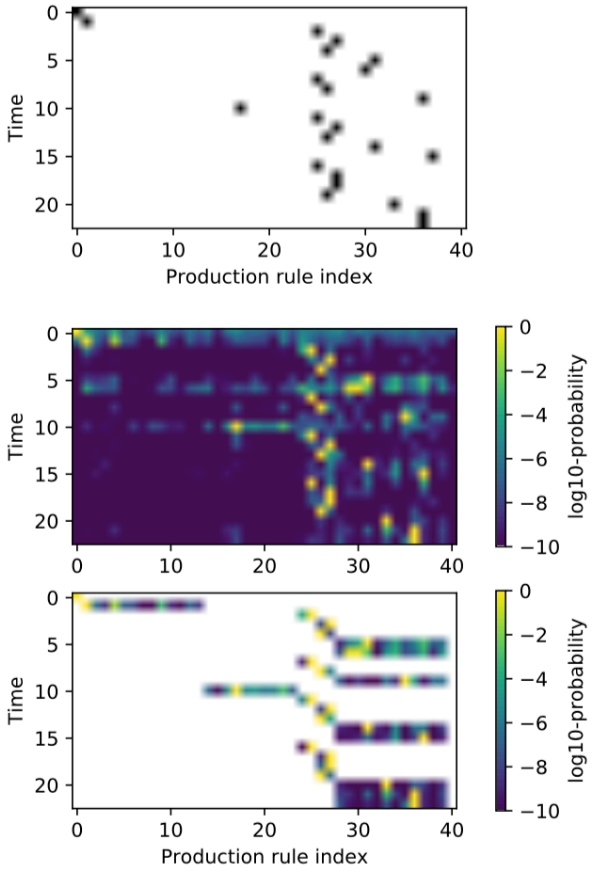
GS → G | GS _ G

G → Ge | CH3 | OCH3 | H | C | O | SCH3 | NCH3 | S | F | Si | Se



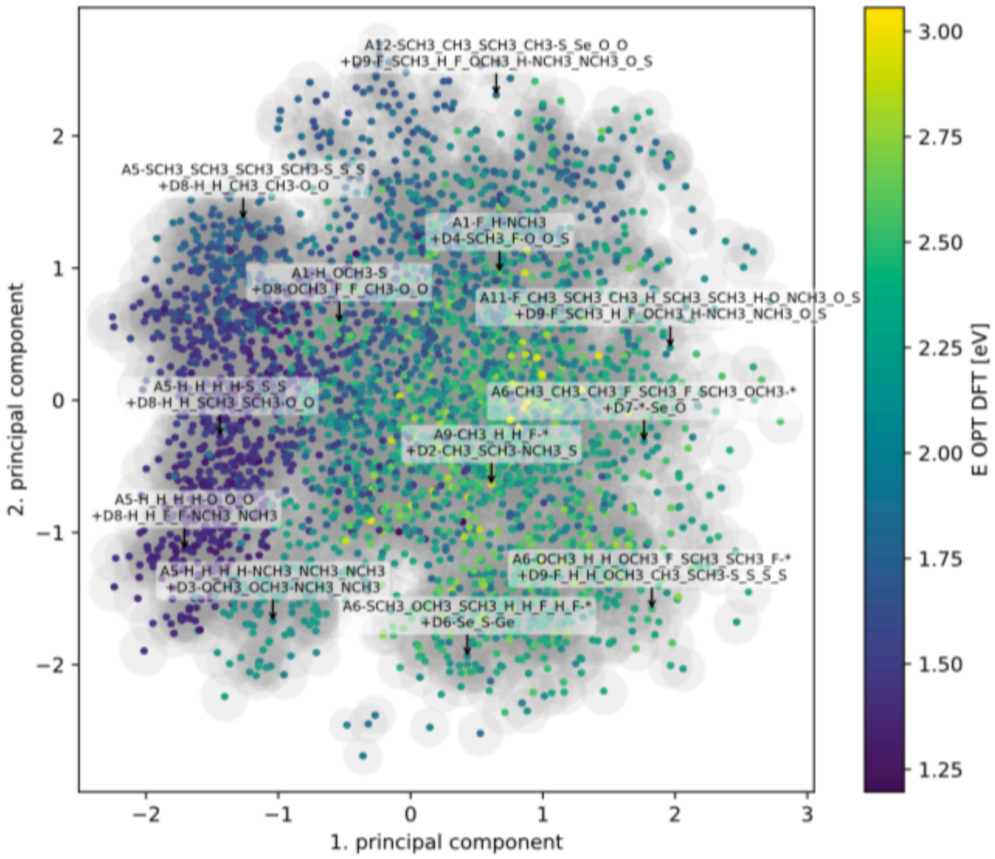
Production rule matrix encoding

m → A - OPTG - OPTG + D - OPTG - OPTG
A → A1
OPTG → GS
GS → GS _ G
GS → G
G → H
G → OCH3
OPTG → GS
GS → G
G → S
D → D4
OPTG → GS
GS → GS _ G
GS → G
G → H
G → F
OPTG → GS
GS → GS _ G
GS → GS _ G
GS → G
G → O
G → S
G → S

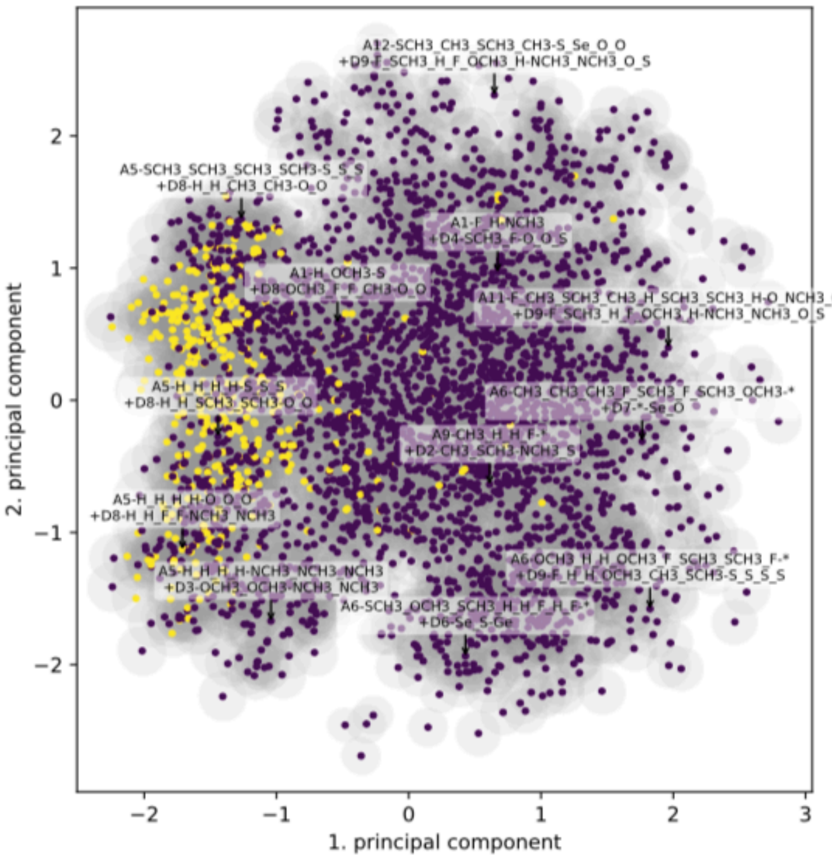


Latent space

First 2 principal components of 32-dimensional space



Colored according to optical gap



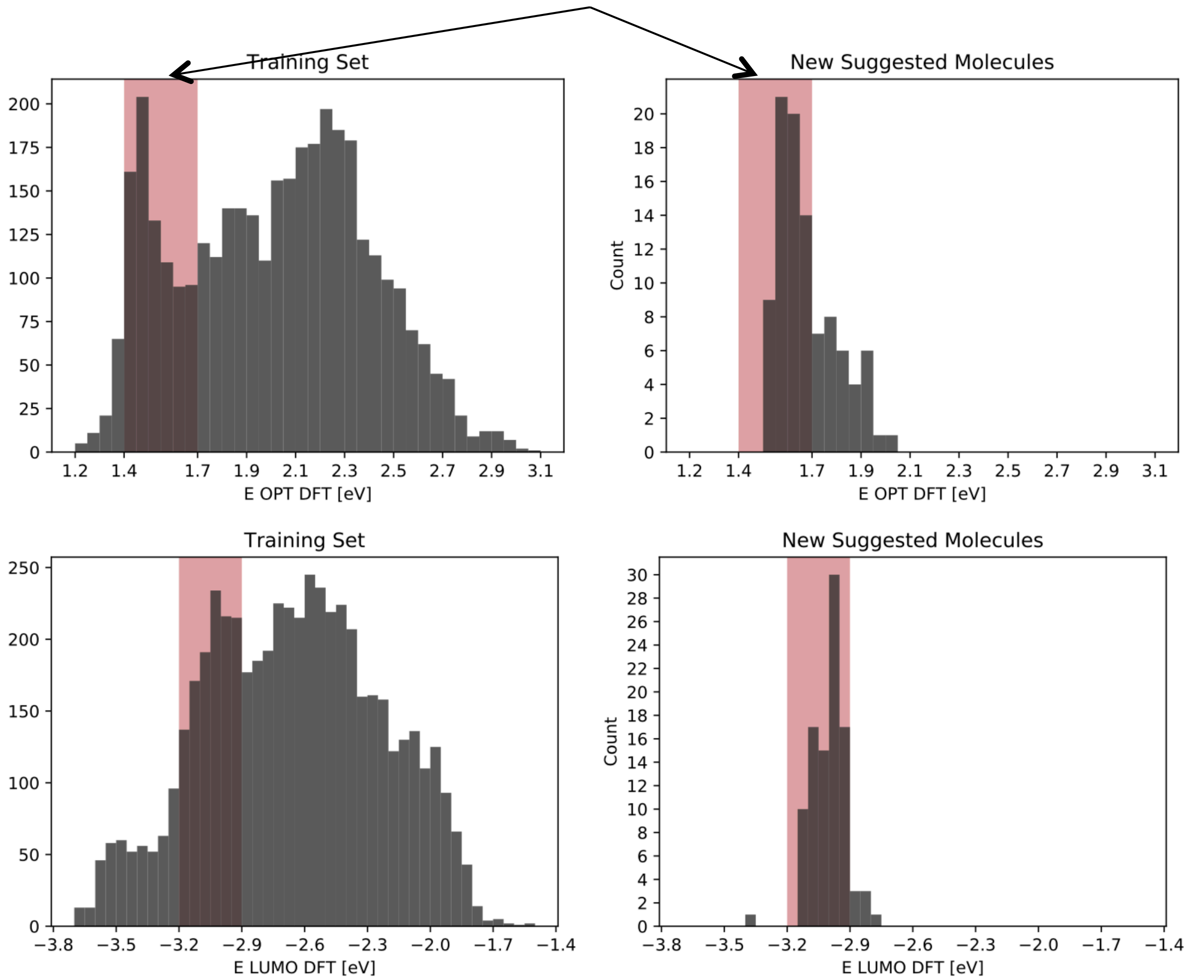
Bright points are within target range

Prediction of new molecules

Target region

Optical band gap

LUMO energy



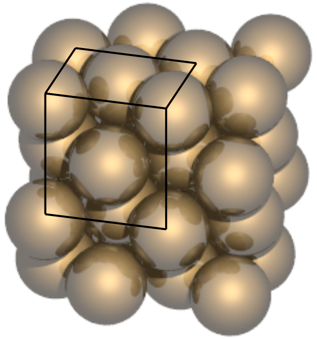
100 new molecules predicted

How do we classify materials without using atomic positions?

Composition, symmetry and prototypes:

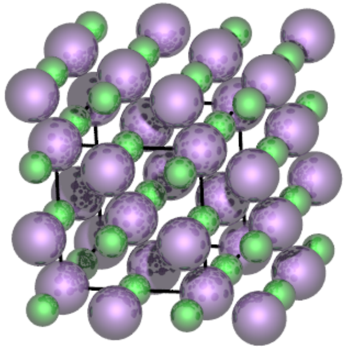
FCC Cu

Space group 225
Only variable is
lattice parameter



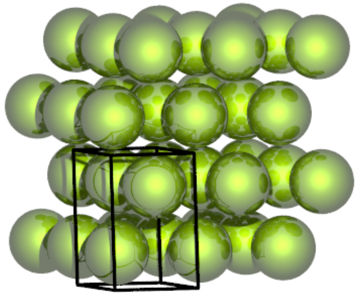
Rocksalt NaCl

Space group 225
Only variable is
lattice parameter



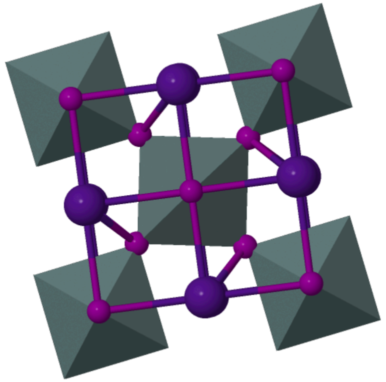
HCP Mg

Space group 194
Two lattice
parameters
c = 1.624*a



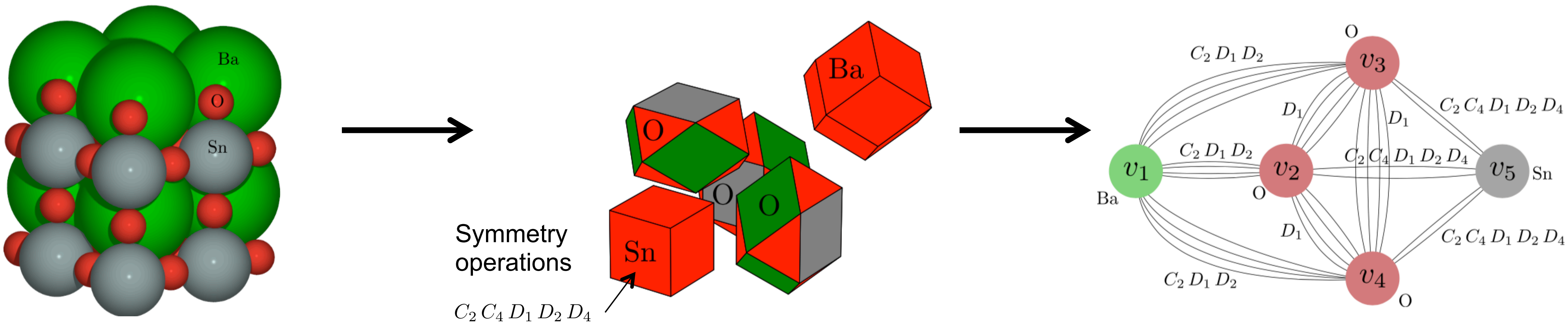
$CsSnI_3$

Space group 127
Two lattice
parameters
Rotation angle



Voronoi cells and graphs

$BaSnO_3$ cubic perovskite Voronoi (Wigner-Seitz) cells Symmetry-labeled graph



Symmetry operations

$C_2\,C_4\,D_1\,D_2\,D_4$

Peter B. Jørgensen, Estefanía Garijo del Río, Mikkel N. Schmidt, Karsten W. Jacobsen, arXiv.org (2018)

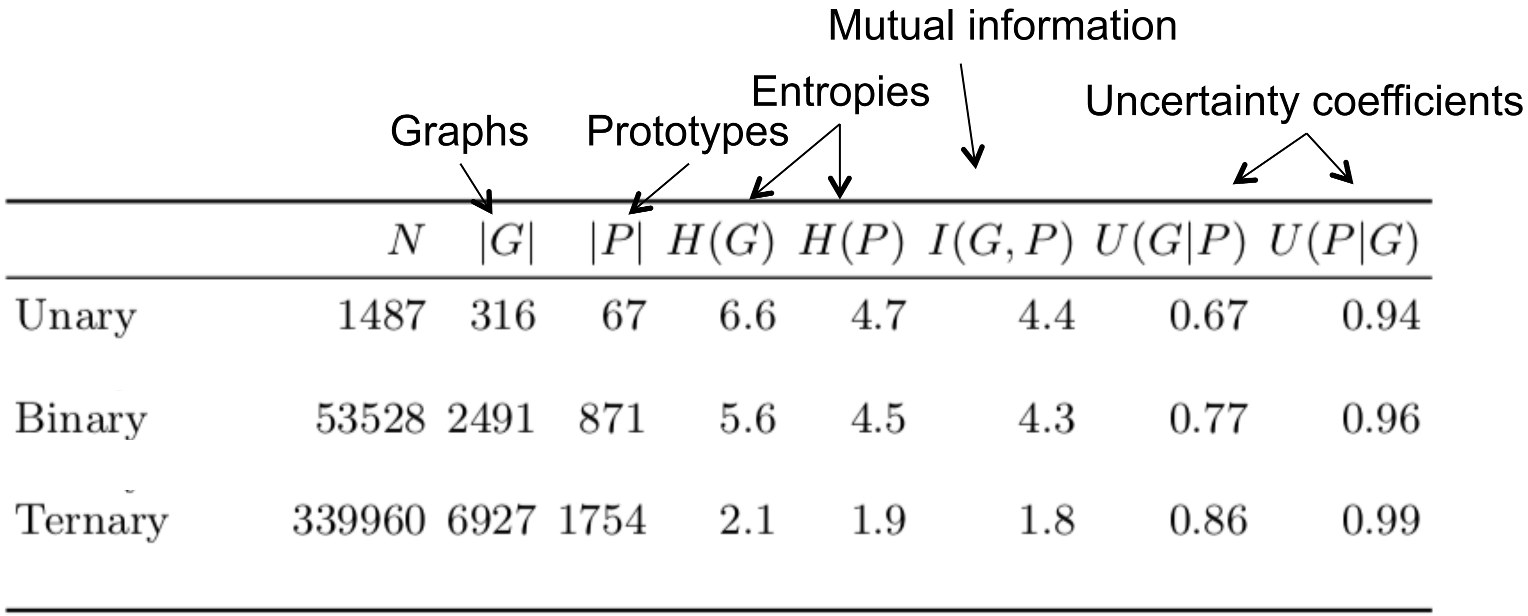
Quotient graph

Graph vs. prototypes

Do we know the prototype if we know the graph (and vice versa)?

Uncertainty coefficient: $$U(P|G) = 1 - \frac{H(P|G)}{H(P)}$$ Entropies

Symmetry labeled graphs provide a more detailed description than prototypes



Mutual information

Entropies

Uncertainty coefficients

Graphs

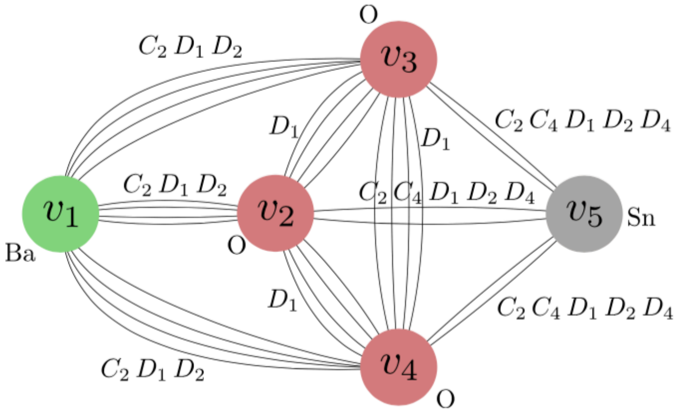
Prototypes

	N	$\|G\|$	$\|P\|$	$H(G)$	$H(P)$	$I(G,P)$	$U(G\|P)$	$U(P\|G)$
Unary	1487	316	67	6.6	4.7	4.4	0.67	0.94
Binary	53528	2491	871	5.6	4.5	4.3	0.77	0.96
Ternary	339960	6927	1754	2.1	1.9	1.8	0.86	0.99

Materials from OQMD database

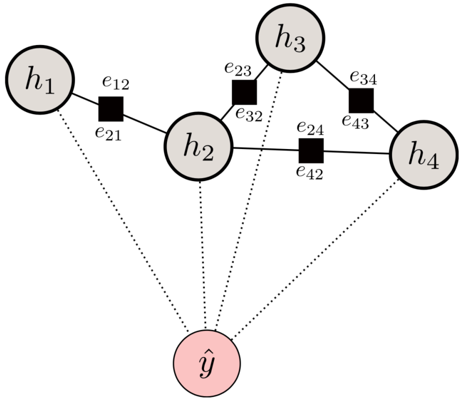
Message passing neural network

Only input:
Atomic numbers Z
Symmetry-labeled quotient graph



atom -> node
bond -> edge

Message passing neural network



$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}^t)$$

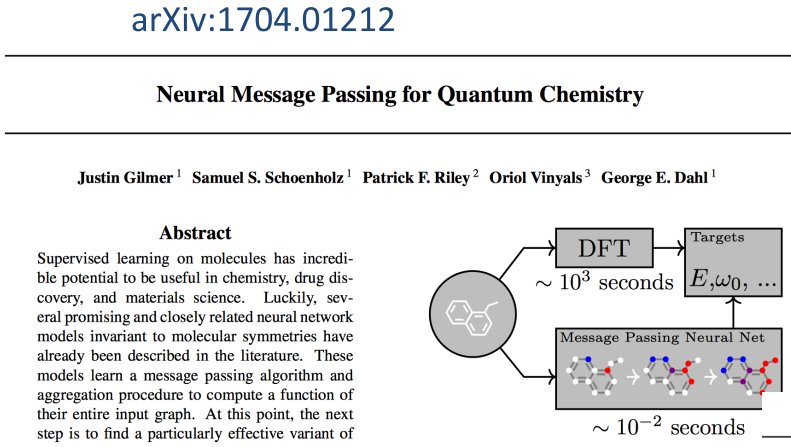
$$h_v^{t+1} = S_t\left(h_v^t, m_v^{t+1}\right)$$

$$e_{vw}^{t+1} = E_t\left(h_v^{t+1}, h_w^{t+1}, e_{vw}^t\right)$$

$$\hat{y} = R\left(\{h_v^T \in G\}\right)$$

Peter B. Jørgensen, Estefanía Garijo del Río, Mikkel N. Schmidt, Karsten W. Jacobsen, arXiv.org (2018)

Neural networks for materials



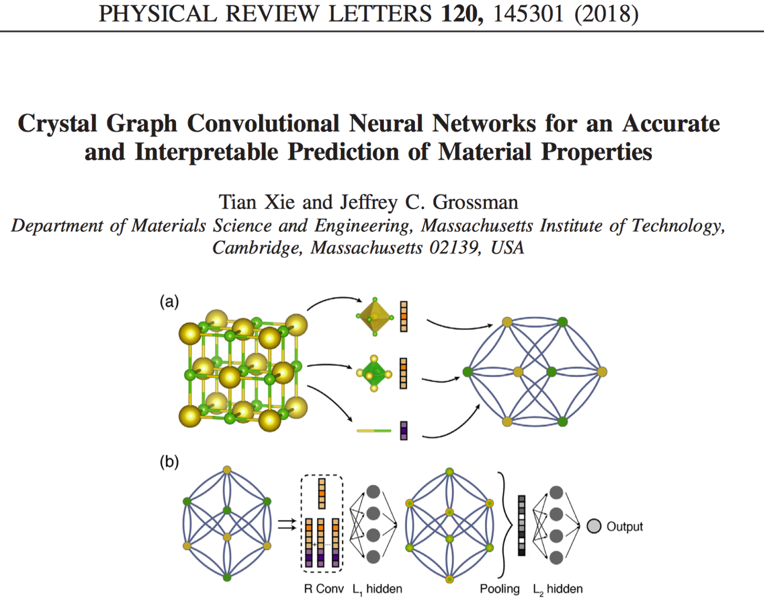
arXiv:1704.01212

Neural Message Passing for Quantum Chemistry

Justin Gilmer [1] Samuel S. Schoenholz [1] Patrick F. Riley [2] Oriol Vinyals [3] George E. Dahl [1]

Abstract

Supervised learning on molecules has incredible potential to be useful in chemistry, drug discovery, and materials science. Luckily, several promising and closely related neural network models invariant to molecular symmetries have already been described in the literature. These models learn a message passing algorithm and aggregation procedure to compute a function of their entire input graph. At this point, the next step is to find a particularly effective variant of

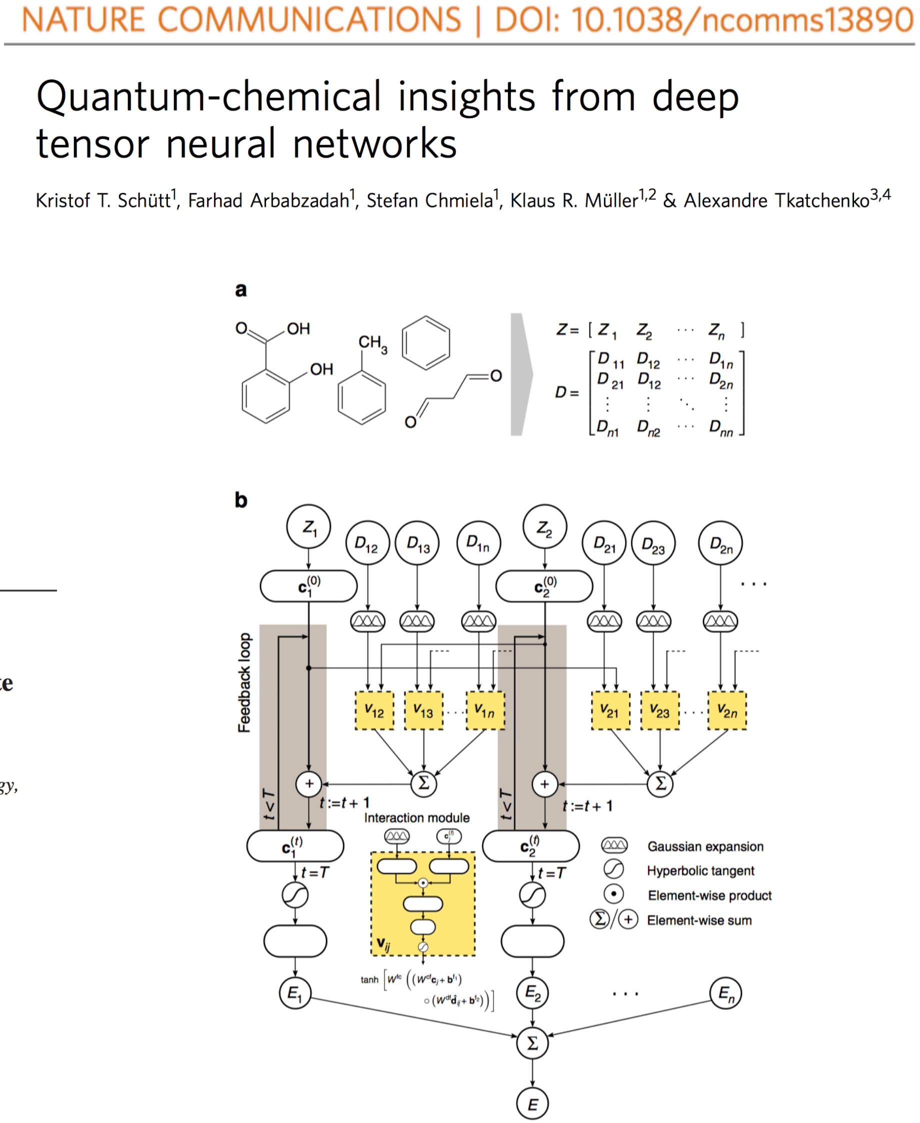


PHYSICAL REVIEW LETTERS **120,** 145301 (2018)

Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties

Tian Xie and Jeffrey C. Grossman

Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA



NATURE COMMUNICATIONS | DOI: 10.1038/ncomms13890

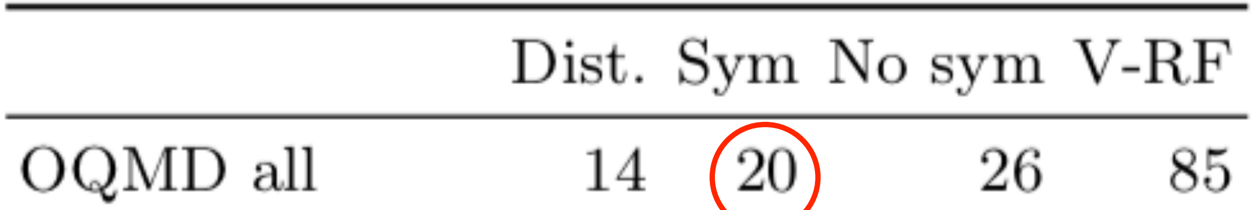
Quantum-chemical insights from deep tensor neural networks

Kristof T. Schütt[1], Farhad Arbabzadeh[1], Stefan Chmiela[1], Klaus R. Müller[1,2] & Alexandre Tkatchenko[3,4]

Predictions on OQMD (5-fold cross validation)

~500000 DFT calculations for inorganic materials

Heat of formation:
Mean absolute error



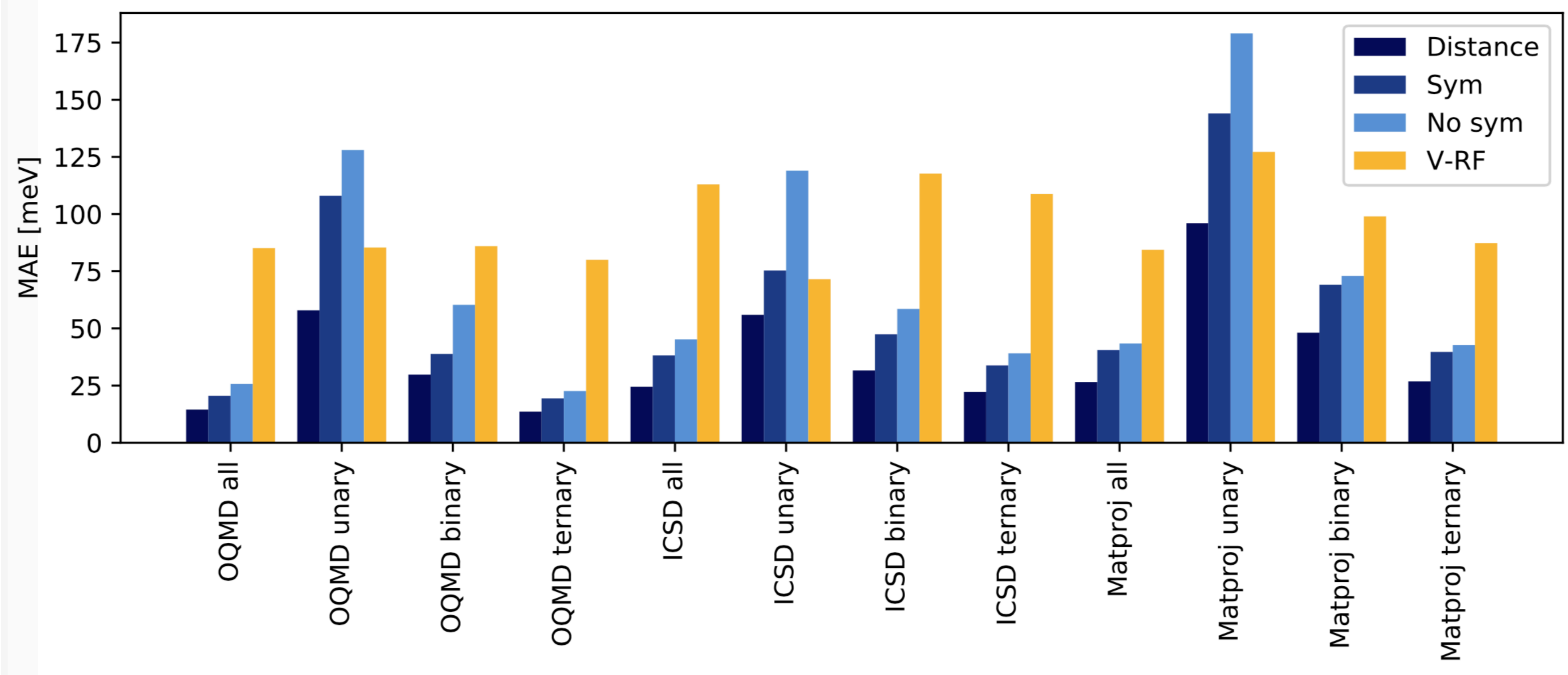
	Dist.	Sym	No sym	V-RF
OQMD all	14	20	26	85

meV!

Only Voronoi graph (+symmetry)

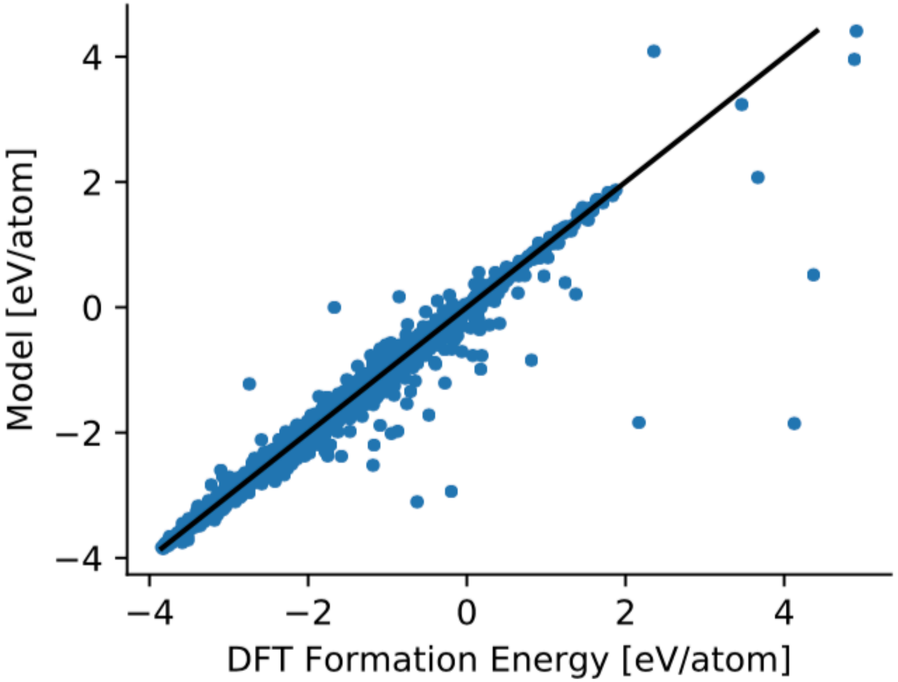
Mean absolute error on formation energies 20 meV!
Accuracy of DFT ~100-200 meV.

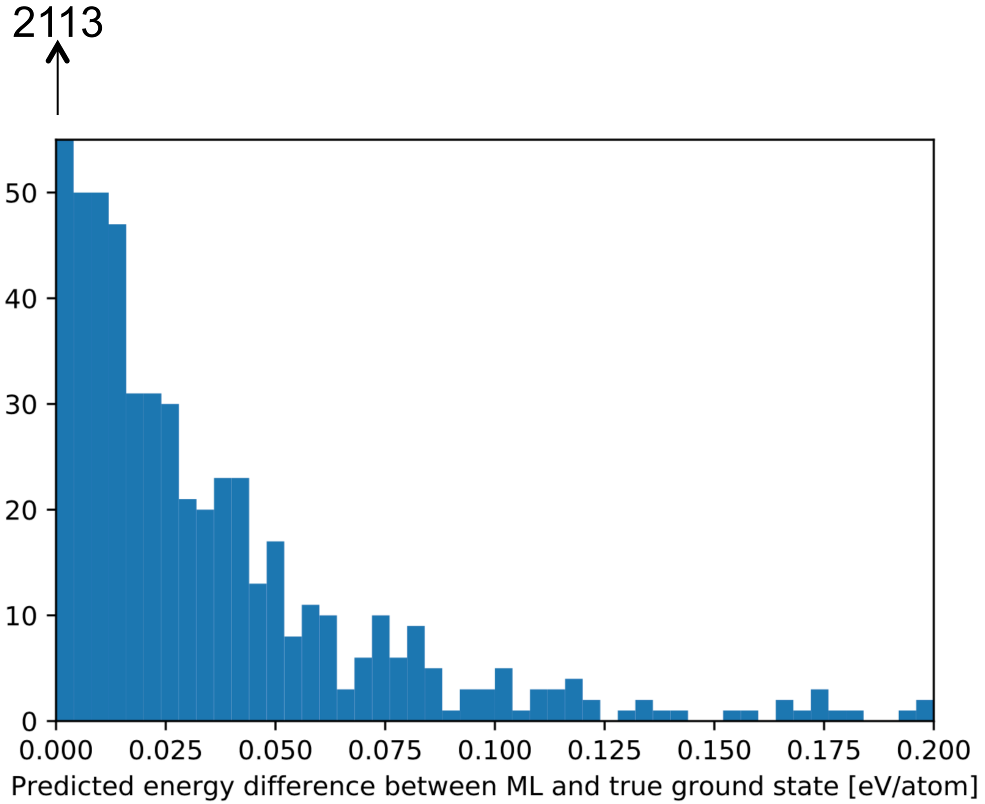
Performance on subsets



The ABO_3 subset of materials

Predictions on ABO_3 oxides
MAE = 35 meV.

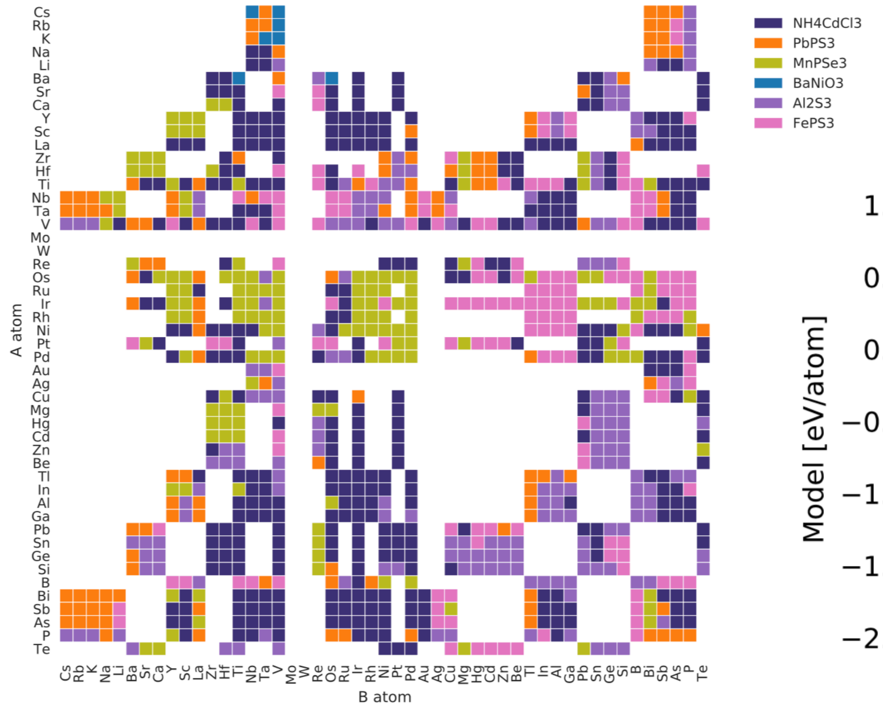




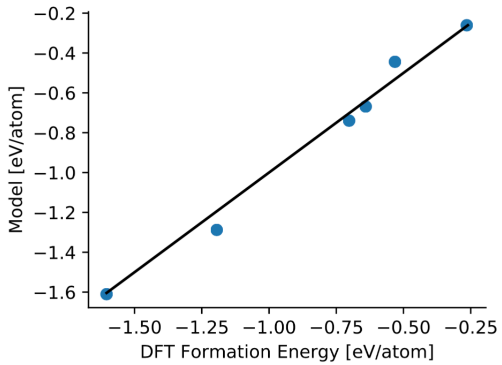
$$\Delta E = E^{ML}(G_{DFT}) - E^{ML}(G_{ML})$$

Heat of formation for $ABSe_3$ dataset

1000 compositions x 6 structures

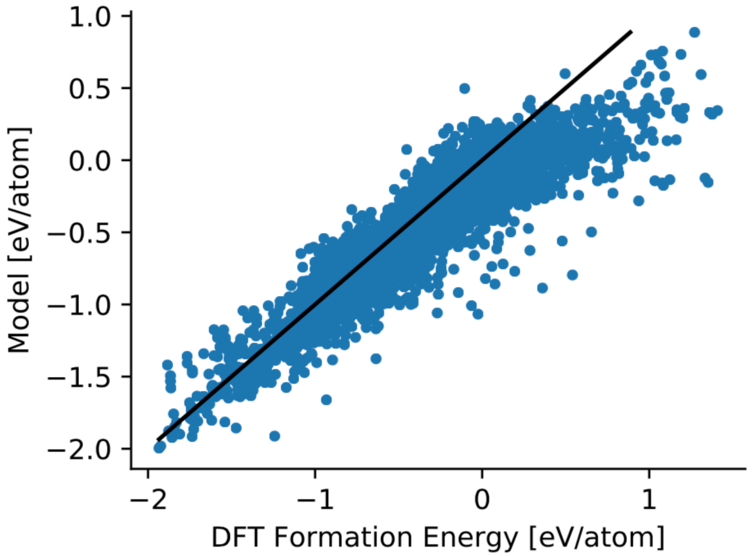


Only 6 $ABSe_3$ selenides in OQMD



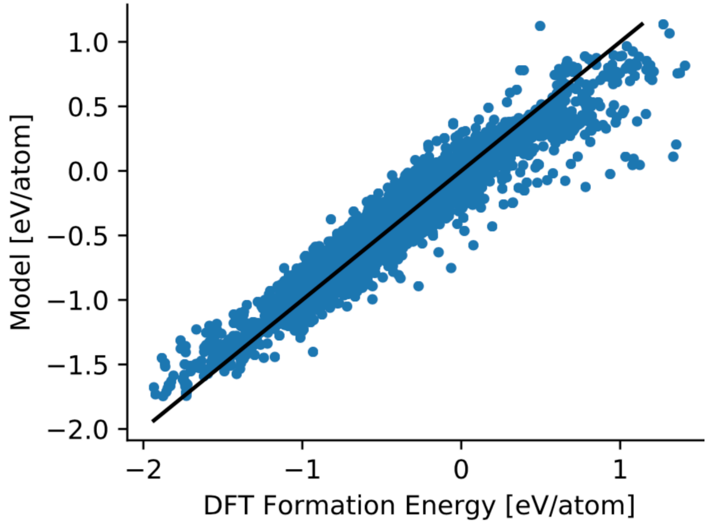
Only training on OQMD

MAE = 0.17 eV



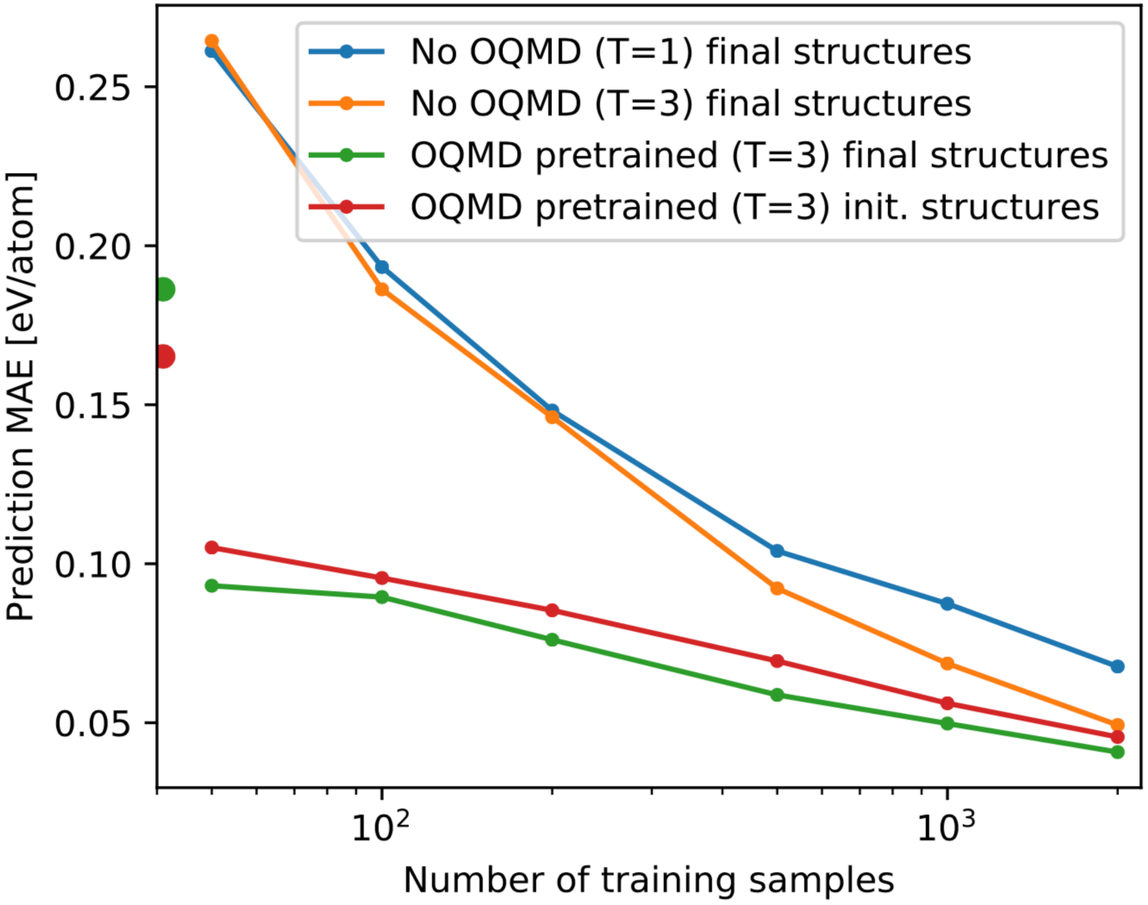
Additional training on 100 $ABSe_3$

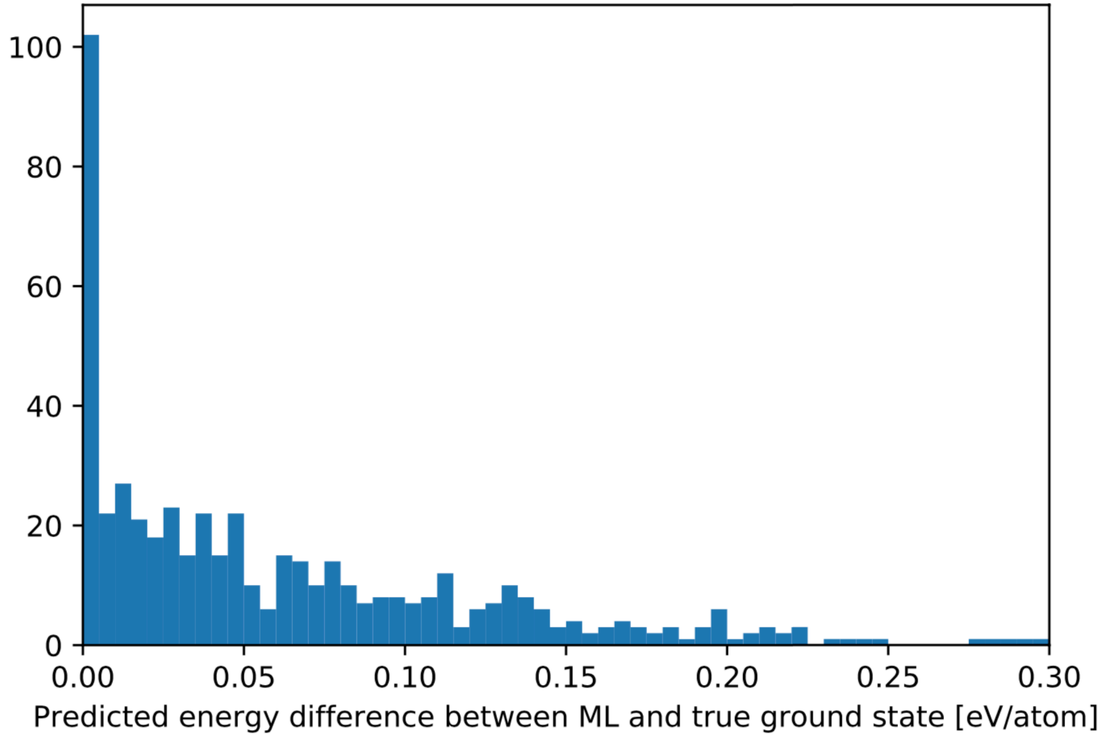
MAE = 0.10 eV



$ABSe_3$ dataset

Additional training on $ABSe_3$ dataset





$$\Delta E = E^{ML}(G_{DFT}) - E^{ML}(G_{ML})$$

ML for computational screening

- Significant speed-up of materials screening with machine learning
- However severe limitations:
  - Better descriptors/understanding of solar cells/water splitting devices
    - Absorption
    - Defects
    - What limits the open-circuit voltage V_{oc}?
  - More accurate calculations beyond DFT (GW/RPA/BSE)

Acknowledgments

- **CAMD/DTU:**
  Mohnish Pandey
  Korina Kuhar
  Estefanía Garijo del Río
  Ivano E. Castelli
  Thomas Olsen
  Kristian S. Thygesen
- **DTU COMPUTE:**
  Peter Bjørn Jørgensen
  Mikkel N. Schmidt
- **SURFCAT/DTU:**
  Andrea Crovetto
  Brian Seger
  Søren Dahl
  Peter Vesborg
  Ole Hansen
  Ib Chorkendorff

CASE
Catalysis for Sustainable Energy

cneec
Center on Nanostructuring for Efficient Energy Conversion

The VILLUM Center for the Science of Sustainable Fuels and Chemicals