This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2022-8352C

MLDL

Machine Learning and Deep Learning Conference 2022

Graph Neural Network Modeling of Vacancy Formation for Materials Discovery in Solar Thermochemical Water Splitting

Matthew Witman / Sandia National Laboratories
Anuj Goyal / National Renewable Energy Laboratory
Tadashi Ogitsu / Lawrence Livermore National Laboratory
Anthony McDaniel / Sandia National Laboratories
Stephan Lany / National Renewable Energy Laboratory

Funded by HydroGEN program (DOE-EERE-HFTO)

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

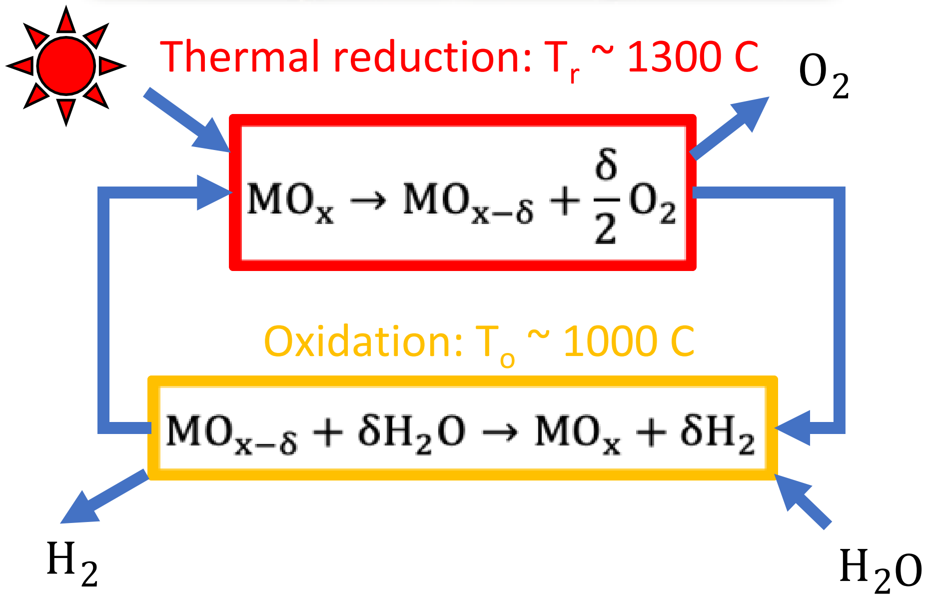
Abstract

We present a graph neural network modeling approach that fully automates the prediction of the DFT-relaxed vacancy formation enthalpy of any crystallographic site from its DFT-relaxed host structure. Applicable to arbitrary structures with an accuracy limited principally by the amount/diversity of the data on which it is trained, this model accelerates the screening of vacancy defects by many orders of magnitude by replacing the (up to 100s of) DFT supercell relaxations required for each symmetrically unique crystal site. It can thus be used off-the-shelf to rapidly screen 10,000s of crystal structures (which can contain millions of unique defects) from existing databases of DFT-relaxed crystal structures. We demonstrate the model's practical utility by high-throughput screening metal oxides from the Materials Project to identify high potential candidates for solar thermochemical water splitting. Ultimately, this modeling approach provides a significant screening and discovery capability for any application in which vacancy defects are the primary driver of a material's utility.

- Develop a novel, generalizable graph neural network approach for predicting vacancy defect properties
- Achieve many orders of magnitude faster prediction than first-principles calculations, e.g. DFT
- Rapidly screen for new, optimal materials for clean energy applications, e.g. H_2 generation
- Future work: significant potential for further development and application to other materials science domains

Solar thermochemical water splitting (STCH) generates *green* (CO_2-free) H_2

Direct 2 step redox cycle (*nb*. >300 proposed cycles…) [1]

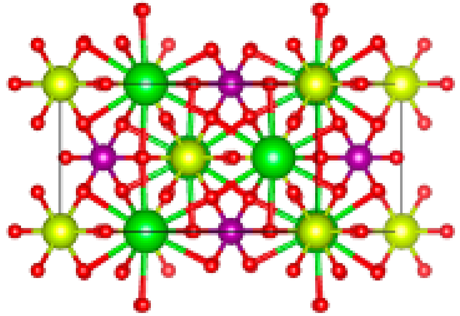


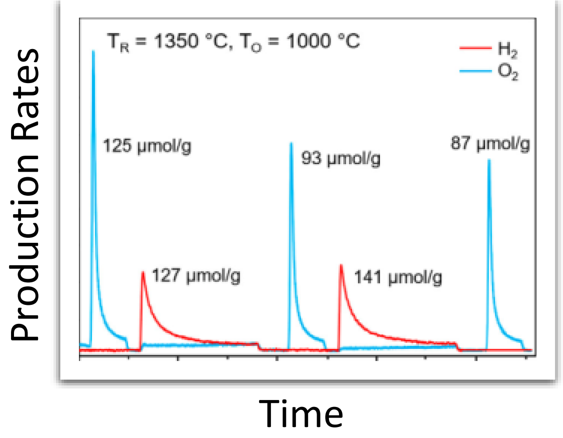
No MO_x solution yet that meets all requirements for T_r, stability, kinetics, etc.

[1] www.energy.gov/eere/fuelcells/hydrogen-production-thermochemical-water-splitting

Top candidates (BCM-12R) are well-studied & characterized

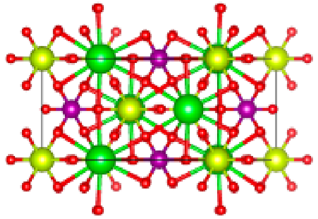
Experiments: Directly measure and evaluate H_2 and O_2 production rates





~Month to synthesize, characterize, test 1 material

1st principles (DFT): Compute oxygen vacancy formation enthalpy (e.g., ΔH_d^O) of all sites



Thermodynamic "sweet-spot":

At least one $\Delta H_d^O \in [2.3, 4.0]$ eV

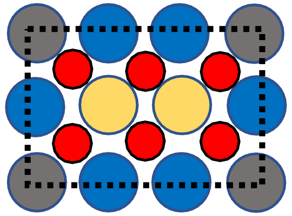
All $\Delta H_d^O > 2.3$ eV

~Month to compute this proxy for handful of materials

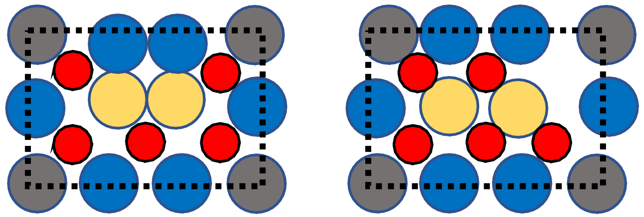
Computational search for $\Delta H_d^o \in [2.3, 4.0]$ eV rapidly encounters scaling issues

Need the vacancy formation enthalpy, ΔH_d, *of all N symmetry sites*:

Relaxed host energy, E_h



Relaxed defect energy, E_d

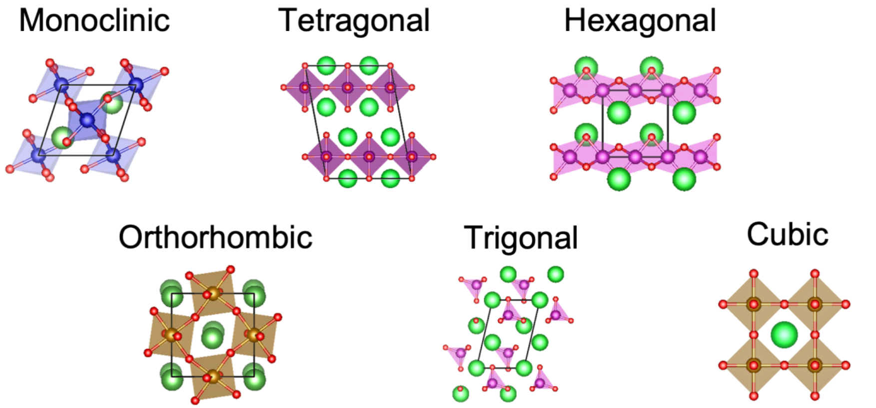


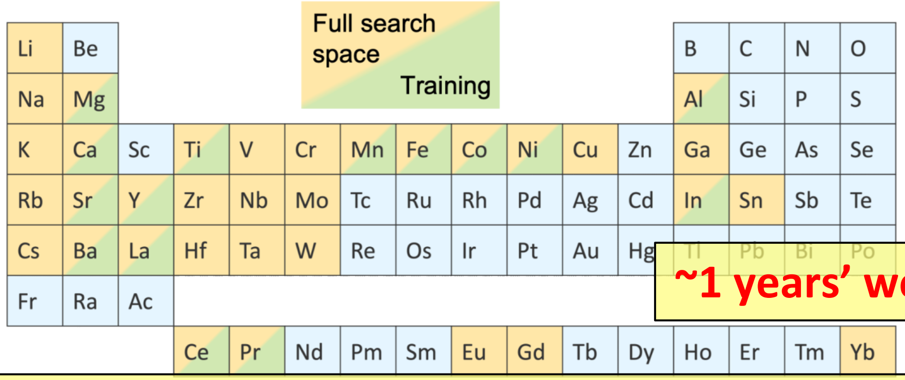
$$\Delta H_d = E_d - E_H + \sum_i n_i \mu_i^{ref}$$

- ΔH_d: Defect Formation Energy
- E_d, E_H: Relaxed Defect and Host Supercell Energy
- $\sum_i n_i \mu_i^{ref}$: Atomic Reference Energy

Requires N+1 DFT relaxations

First-principles DFT workflow is robust but costly (using NRELMatDb hosts)





DFT workflow

A. Material Space

Diversity in Training:
40 unique crystal structures and
38 unique compositions

B. Host Calculations

Relaxed structure,
Atomic spin and oxidation state,
Enthalpy of formation,
Bandgap, Electron effective mass

C. Defect Calculations

Calculated oxide space:
~200 host structures
~1500 defect relaxations

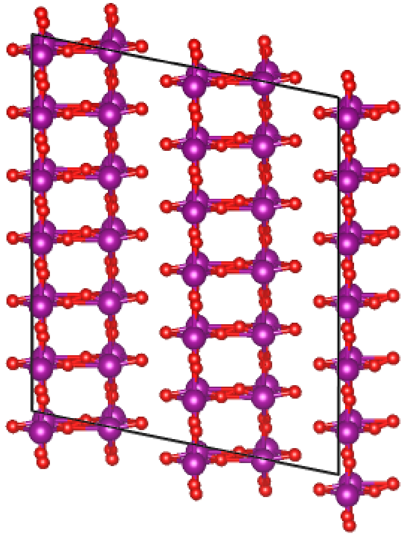
~1 years' work

Existing oxide space:
~10,000s host structures
~1Ms defect relaxations

100+ years' work… so more efficient model needed

A more generalizable approach is needed to model vacancy defects

(1) Compositional features can't differentiate symmetry sites



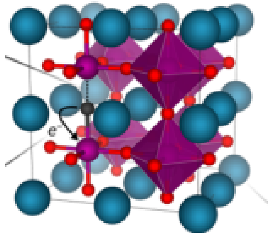
Wyckoff site	ΔH_d [eV]
Mn1	12.2
Mn2	12.1
O1	2.1
O2	2.2
O3	2.6
O4	2.7

$$MnO_{1.5} \mapsto \boldsymbol{x}_{O1} = \boldsymbol{x}_{O4} = \{\bar{v}_{pa}, \bar{r}_{cov}, \bar{\chi}, \ldots\}$$

Problem:
- $\boldsymbol{x}_{O1}$ and $\boldsymbol{x}_{O4}$ are identical but…
- Local structures quite different
- Target value differs by 0.6 eV

(2) Hand-engineered $\boldsymbol{x}$ used for specific material classes

Wexler et al. *J. Am. Chem. Soc.* **2021** ΔH_d MAE = 0.45 eV



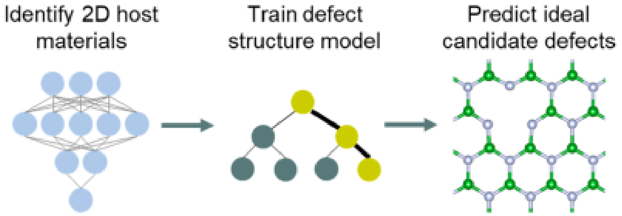
Linear Model works well for ABO3 perovskites:

$$0.1\Sigma E_b - 1.5V_r + 0.4E_g - 55.8E_{hull} + 0.4 \text{ (eV)}$$

Crystal bond dissociation; Crystal reduction; Band gap; Stability ← $\boldsymbol{x}$

Limitations:
- Model only validated for O sites in ABO_3 perovskites
- Needs re-derivation for other structure classes

Frey et al. *ACS Nano*. **2020** ΔH_d MAE = 0.67 eV



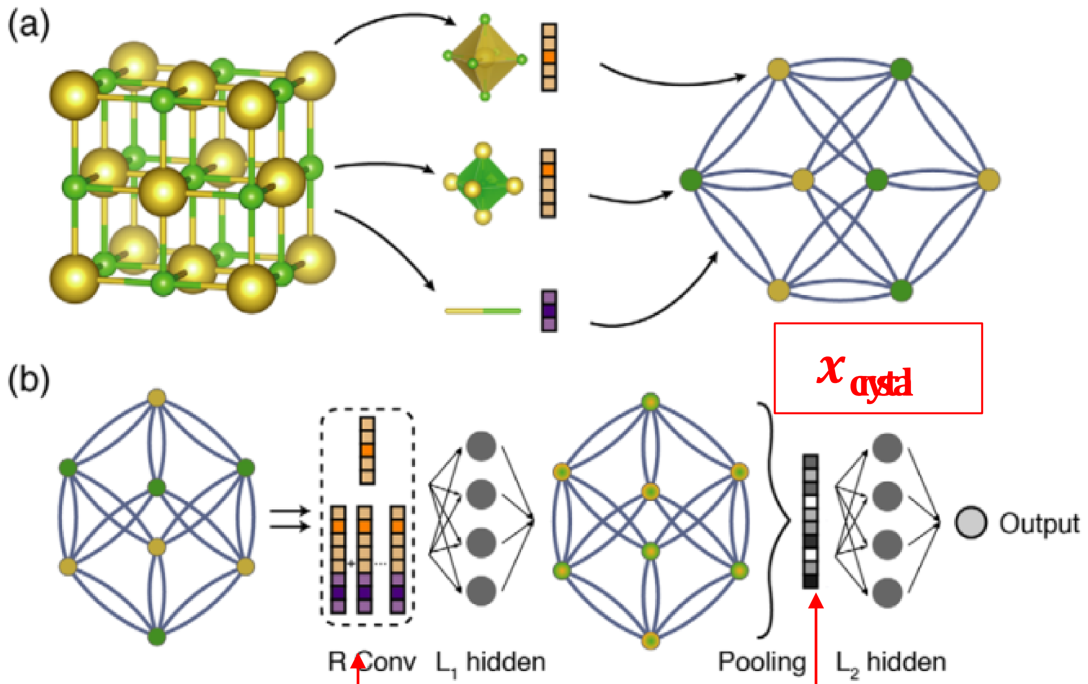
Random forest model works well for certain vacancies in 2D materials

Limitations:
- Model and features specific to 2D material classes
- Needs re-derivation for other structure classes

Automated feature extraction of graph neural networks (GNNs) enables efficient and generic modeling of vacancy formation enthalpy

Automated feature extraction with GNNs [1]

Interpret crystal as a graph (nodes = atoms, "bonds" = edges)



Repeated passing of info. between neighbors

Pool atom features to create crystal feature vector

[1] Xie, et al. *P.R.L.* 120 (14), **2018**
[2] Witman, et al. *Submitted*

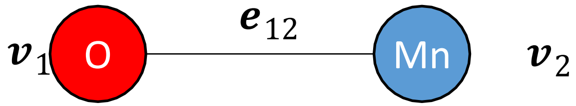
Deriving a "defect GNN" approach[2]

Predict using only host structure, $\boldsymbol{X}_h$, and defect atom index, i'

DFT: $\Delta H_d = E_{\mathrm{DFT}}(\boldsymbol{X}_d) - E_{\mathrm{DFT}}(\boldsymbol{X}_h) + \mathrm{ref}$

ML: $\Delta \widehat{H}_d = f_{\mathrm{GNN}}(\boldsymbol{X}_h, i'; \theta)$

➢ Example graph:



➢ Encode the graph (step $t = 0$):

$$\boldsymbol{v}_1^{t=0} = \{r_O, \chi_O, \ldots, s_1\}$$

Accuracy boosting, site-specific inputs (i.e. oxidation state)

➢ Convolutions ($t = 1 \ldots T$)

$$\boldsymbol{v}_i^{(t+1)} = g\left(\boldsymbol{v}_i^{(t)} + \sum_j \sigma\left(\boldsymbol{z}_{ij}^{(t)} \boldsymbol{W}_1^{(t)} + \boldsymbol{b}_1^{(t)}\right) \odot g\left(\boldsymbol{z}_{ij}^{(t)} \boldsymbol{W}_2^{(t)} + \boldsymbol{b}_2^{(t)}\right)\right)$$

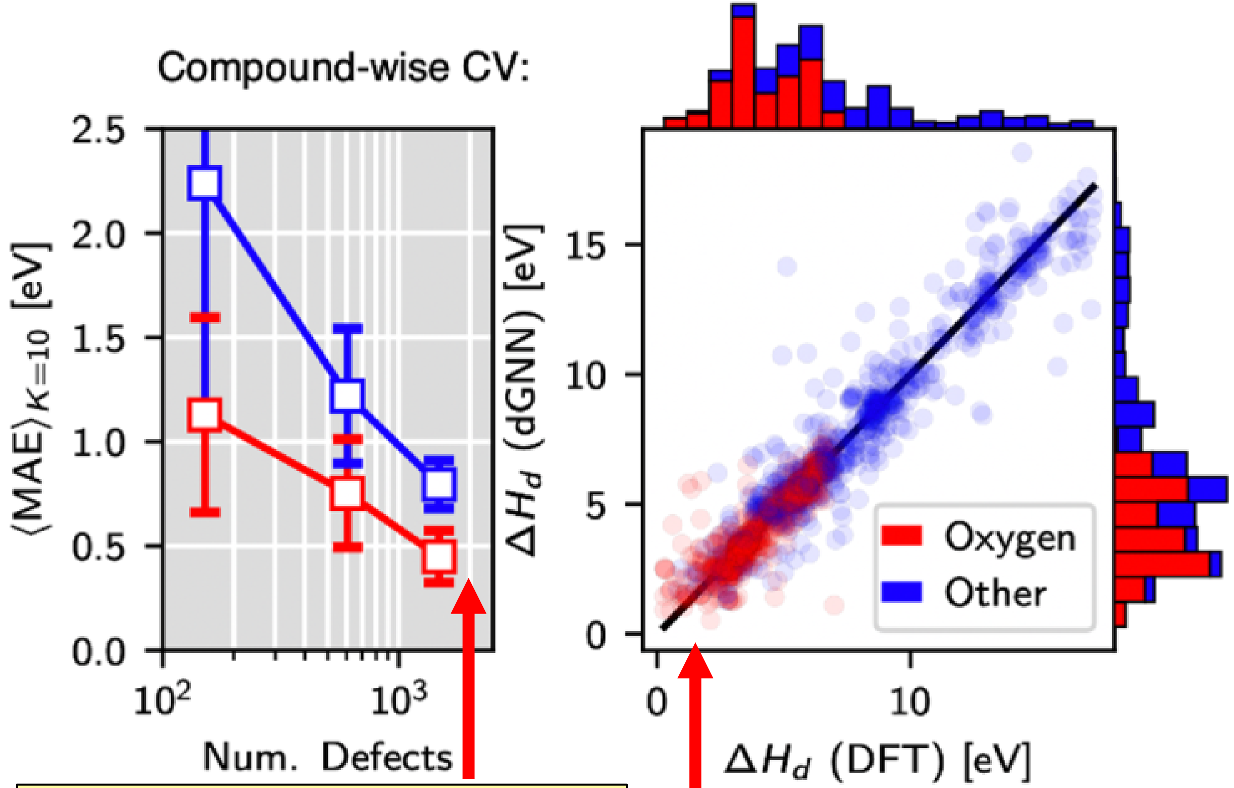
➢ Property prediction (no pooling):

$$\boldsymbol{x}_{\mathrm{defect}} = \sigma(\boldsymbol{v}_{i'}^T \oplus \boldsymbol{v}_g \cdot \boldsymbol{W} + \boldsymbol{b})$$
$$\Delta \widehat{H}_d = \boldsymbol{x}_{\mathrm{defect}} \cdot \boldsymbol{W} + b$$

- ➢ Extract defect feature vector
- ➢ Use host's global properties, $\boldsymbol{v}_g$ = {band gap, …}

Defect GNN approach validated for use in high-throughput screening exercise

Benchmark accuracy has been met for HT screening



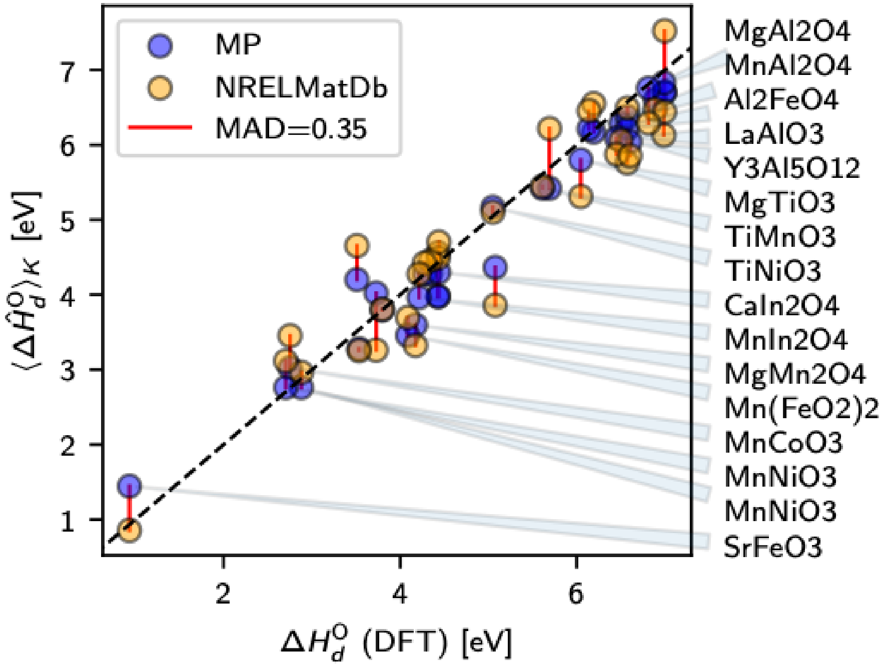
- ➢ MAE < 450 meV
- ➢ Still have log-linear decrease w/more data
- ➢ Minutes to train on CPUs

- ➢ Can predict O *and* non-O vacancies

Effects of encoding strategies, predicting compounds containing cations *not* in training, etc.

➢ See preprint

NRELMatDb vs. Materials Project (MP) structure inputs



- ➢ Robust to small variations in structure
- ➢ Can screen using a different database (MP)

Hight-throughput screening 2,000 oxides (50,000 unique defects) rediscovers known water-splitting oxides and identifies new ones (~10 top candidates)

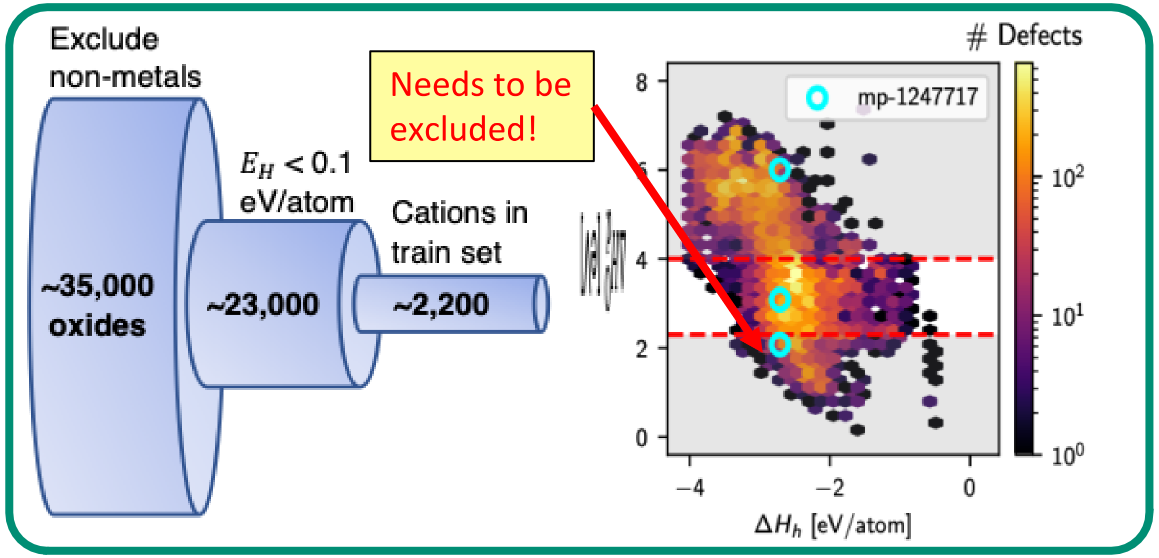
(1) Co-design of host defects and stability for water-splitting

Metric	Requirement
Frac. of defects w/$\Delta H_d^O > 2.3$ eV	$x_{min} = 1$
Frac. of defects w/$\Delta H_d^O \in [2.3, 4.0]$ eV	$x_{rng} > 0$
Host stability criteria (ranges intersect)	$\Delta\mu'_{O_2} \cap \Delta\mu_{O_2}^{\phi_H<X} \neq \emptyset$

Operating range for STCH

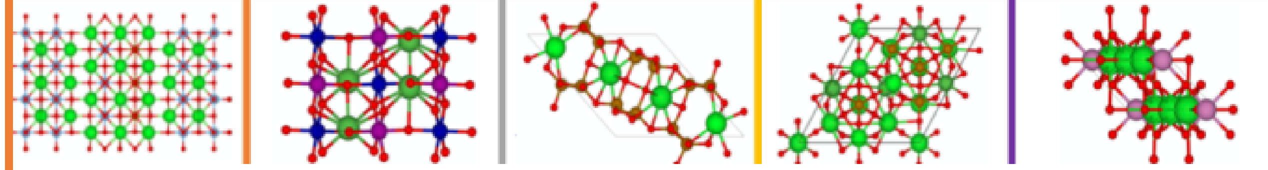
Range where host's grand energy above hull (ϕ_H) is < X

(2) Screen the Materials Project for all defects



(3) Identify and filter increasingly promising targets

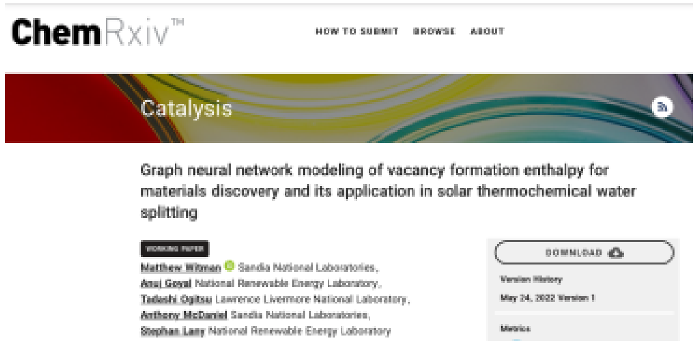
197 formulas (48 training)	114 formulas (33 training)	34 formulas (17 training)	16 formulas (11 training)	9 formulas (9 training)
➢ $x_{min,1} = 1$ ➢ $x_{rng,1} > 0$ ➢ $\Delta\mu_{O_2}^{\phi_H<0.1}$	➢ $x_{min,2} = 1$ ➢ $x_{rng,2} > 0$ ➢ $\Delta\mu_{O_2}^{\phi_H<0.1}$	➢ $x_{min,3} = 1$ ➢ $x_{rng,3} > 0$ ➢ $\Delta\mu_{O_2}^{\phi_H<0.05}$	➢ $x_{min,3} = 1$ ➢ $x_{rng,3} > 0$ ➢ $\Delta\mu_{O_2}^{\phi_H=0}$	➢ $x_{min,3} = 1$ ➢ $x_{rng,3} = 1$ ➢ $\Delta\mu_{O_2}^{\phi_H=0}$
$Sr_6Ti_3FeO_{14}$ (mp-1645141)	La_2MnCoO_6 (mp-19208)	$BaSr(FeO_2)_4$ (mp-1228024)	$Ba_5SrLa_2Fe_4O_{15}$ (mp-698793)	$Ba_3In_2O_6$ (mp-20352)



- ➢ Filter candidates with increasingly certain performance
- ➢ Mainly identifies known, synthesizable compounds
- ➢ ~100 are not AXO_3, $A_{n+1}X_nO_{3n+1}$, $Fe_{3-n}M_nO_4$, CeO_2, etc.
- ➢ Rediscovers complex, known water-splitting materials (not in training data) like $Ba_4CeMn_3O_{12}$

All training data, code, screening scripts, and finalized predictions are provided open-source for community use and customized filtering before attempting experiments

Open access preprint and summary of screening results are provided in user friendly, customizable csv:

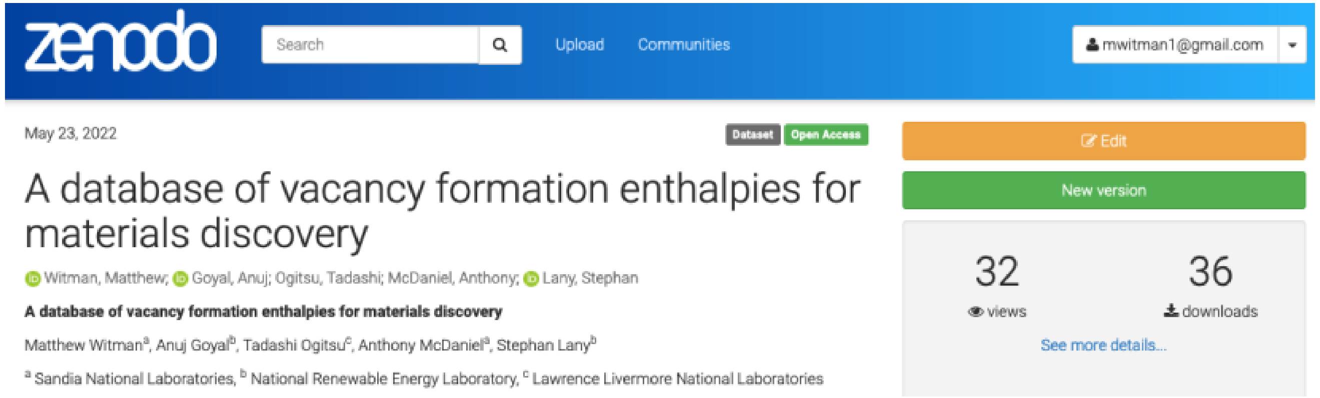


...

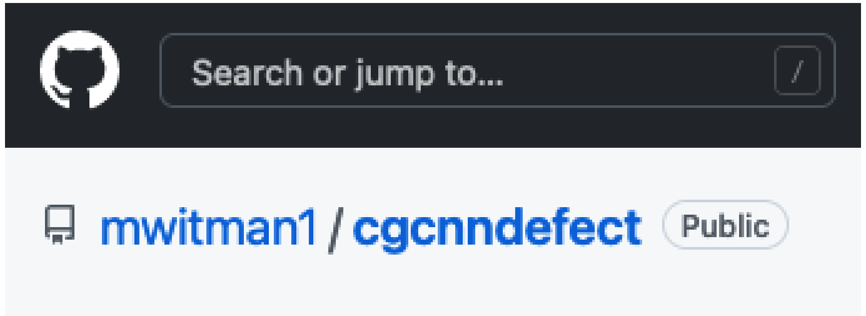
Formula	F	min(H)	max(H)	$\bar{U}$	$x_{\min,1}$	$x_{\min,2}$	$x_{\min,3}$	$x_{\mathrm{rng},1}$	$x_{\mathrm{rng},2}$	$x_{\mathrm{rng},3}$	ΔH_c	$\Delta\mu_O^{\phi_H=0}$	$\Delta\mu_O^{\phi_H<0.05}$	$\Delta\mu_O^{\phi_H<0.1}$
$Ba_2Fe_6O_{11}$ (mp-652683)	4	3.07	3.73	0.21	1.0	1.0	1.0	1.0	1.0	1.0	-2.14	[nan, nan]	[-2.73, -0.50]	[-2.93, -0.11]
Mn_3O_4 (mp-18759)	4	3.33	3.34	0.24	1.0	1.0	1.0	1.0	1.0	1.0	-2.06	[-2.48, -1.50]	[-2.63, -1.20]	[-2.78, -0.90]
Fe_2CoO_4 (mp-753222)	4	3.52	3.52	0.26	1.0	1.0	1.0	1.0	1.0	1.0	-1.76	[nan, nan]	[-2.63, -0.95]	[-2.70, -0.50]
$BaSr(FeO_2)_4$ (mp-1228024)	4	3.32	3.54	0.26	1.0	1.0	1.0	1.0	1.0	1.0	-2.23	[nan, nan]	[-2.68, -0.68]	[-2.98, -0.38]
$BaMnO_3$ (mp-[illegible])	4	2.96	2.98	0.31	1.0	1.0	1.0	1.0	1.0	1.0	-2.47	[nan, nan]	[-2.60, 0.00]	[-2.86, 0.00]

...

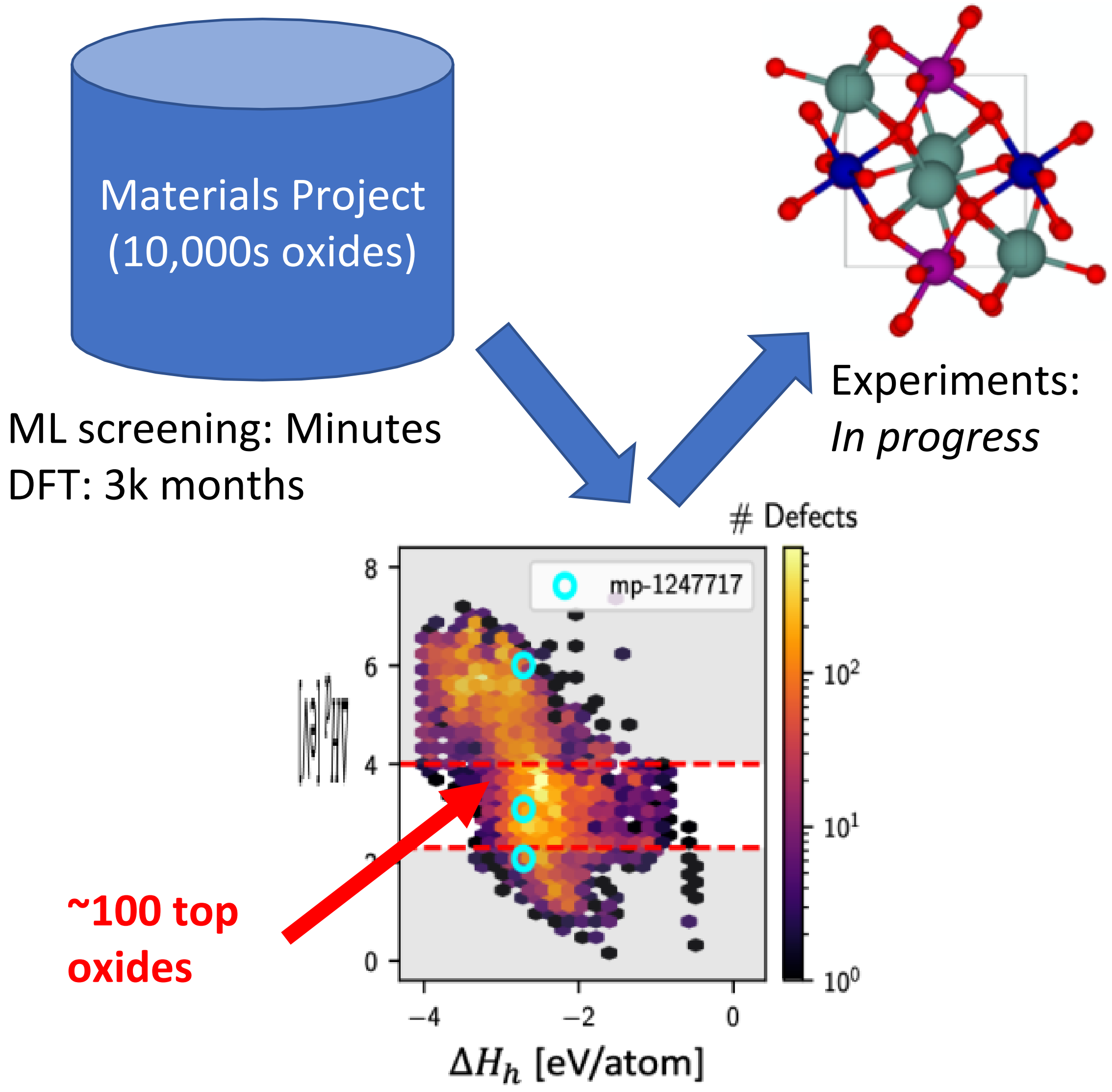
Zenodo repository for training data, analysis, & paper reproducibility:



Github for defect GNN code:



In conclusion, STCH oxides have been screened ~10^6 times faster than brute-force DFT search to identify promising water-splitting oxides



Significant room for future development:

- Expanded cation space
- Expanded anion space (beyond O)
- More structural diversity (2D materials)
- Probe ***other applications*** where vacancies are the primary driver of material utility (or failure)

Thank you for your attention. Questions?

mwitman@sandia.gov