

# Using artificial intelligence to discover new materials

Chris Wolverton

*Dept. of Materials Science and Eng.*

*Northwestern University*



# How to discover new materials?

***DATA CREATION!***

***COLLECTION & CLASSIFICATION!***

***DATA MINING & PREDICTION!***

Calculating  
many known  
materials!

Solving  
unknown  
materials  
structures!

Databases  
of materials  
properties!

Materials  
discovery!

- Open Quantum Materials Database (OQMD)
- Machine Learning of materials datasets to accelerate Materials Discovery

# AI as a tool to accelerate discovery

THE VERGE

TECH ▾

SCIENCE ▾

CULTURE ▾

CARS ▾

REVIEWS ▾

LONGFORM

VIDEO

MORE ▾



SCIENCE \ TECH \ ARTIFICIAL INTELLIGENCE \

## How AI is helping us discover materials faster than ever

*We can predict which compounds can create materials before setting foot in a lab*

By [Angela Chen](#) | [@chengela](#) | Apr 25, 2018, 2:45pm EDT

APR 22, 2018 @ 07:35 AM

2,772 👁

The Little Black Book of Billionaire Secrets

## Scientists Use Artificial Intelligence To Discover New Materials



☰ Forbes



TECHNOLOGY

OTHER VOICES

# 3 Technologies That Could Create Trillion-Dollar Markets Over the Next Decade

By Greg Satell Updated April 21, 2019 9:00 a.m. ET



# Algorithms vs. Learning



# Machine Learning in Real Life: Netflix



Independence Day

17,700 Movies  
in the  
**Netflix Competition**

Todd.Holloway@gmail.com 03/25/2007

Exercise  
Bonus Material  
IMAX  
Documentary  
Slam  
Slam  
Super Si Me  
Con  
When We Were Kings  
A Beautiful Mind  
Brave Heart



**BellKor's Pragmatic Chaos wins the Netflix Prize**

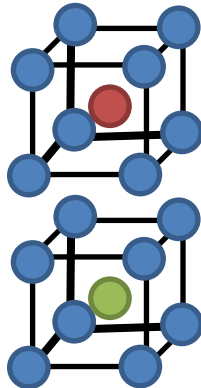
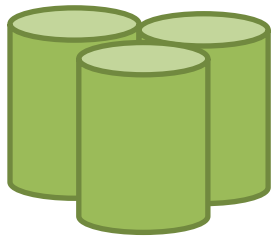
# Materials Informatics Workflow

Collect (Data)

Process  
(Data)

Represent  
(Material)

Learn  
(Property)



$$\Delta H_f = -1.0$$

$$\Delta H_f = -0.5$$

 $\vec{X}$  $\vec{y}$ 

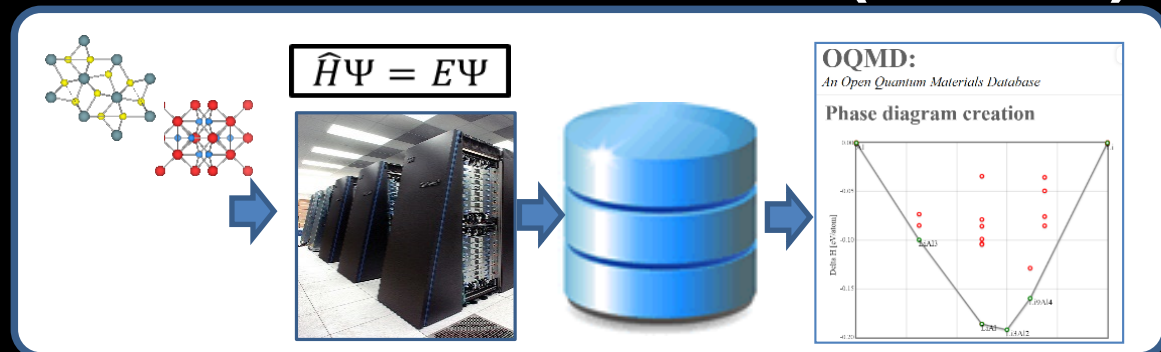
$Z_A$	$Z_B$	$\Delta H_f$
3	4	-1.0
3	5	-0.5

$$\Delta H_f = f(Z_A, Z_B)$$

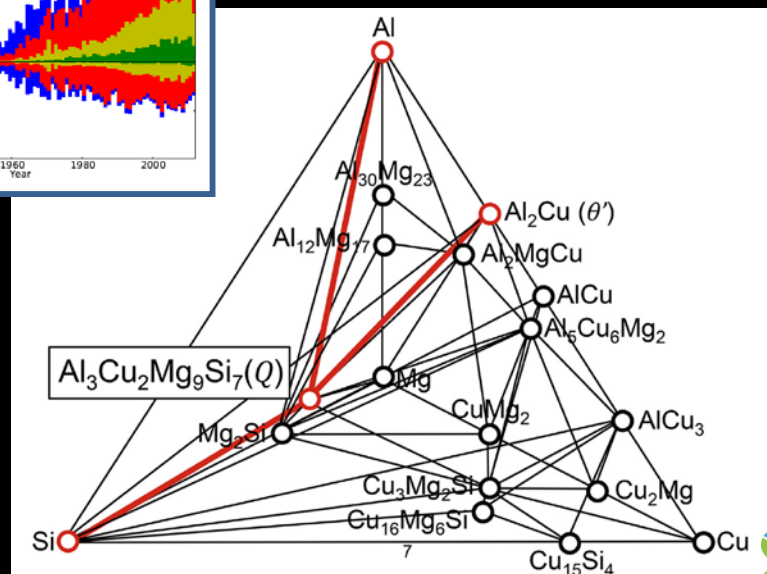
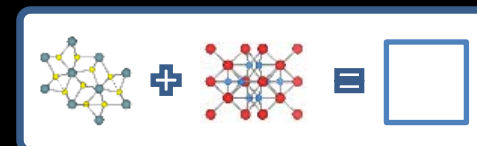
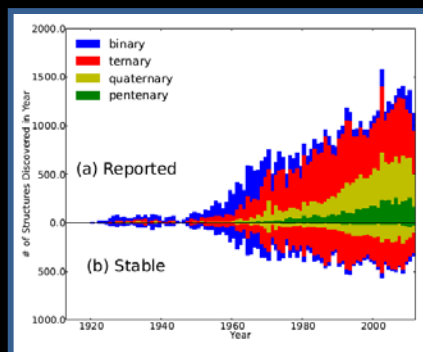
- 1) Data: Training and Test Datasets
- 2) Materials Representation: How do we tell our machine what a material is?
- 3) Machine Learning Algorithm: Many options in available toolkits (Weka, scikit-learn, etc.). For this talk, mostly ensembles of decision trees and (convolution) neural nets.

# High-Throughput Computational Approaches: The Open Quantum Materials Database (OQMD)

- Large-scale DFT database of known (~50K) and hypothetical (~500K) inorganic crystalline compounds
- Open, online, freely available (oqmd.org)
- Automatic computation of phase stability (arbitrary # of components)



Credit: Wikipedia for computer images





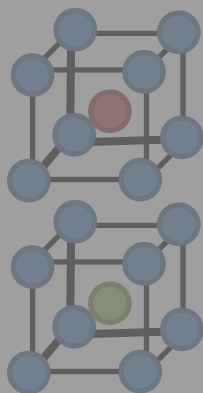
# Materials Informatics Workflow

Collect

Process

Represent

Learn



$$\Delta H_f = -1.0$$

$$\Delta H_f = -0.5$$

 $\vec{X}$  $\vec{y}$ 

$Z_A$	$Z_B$	$\Delta H_f$
3	4	-1.0
3	5	-0.5

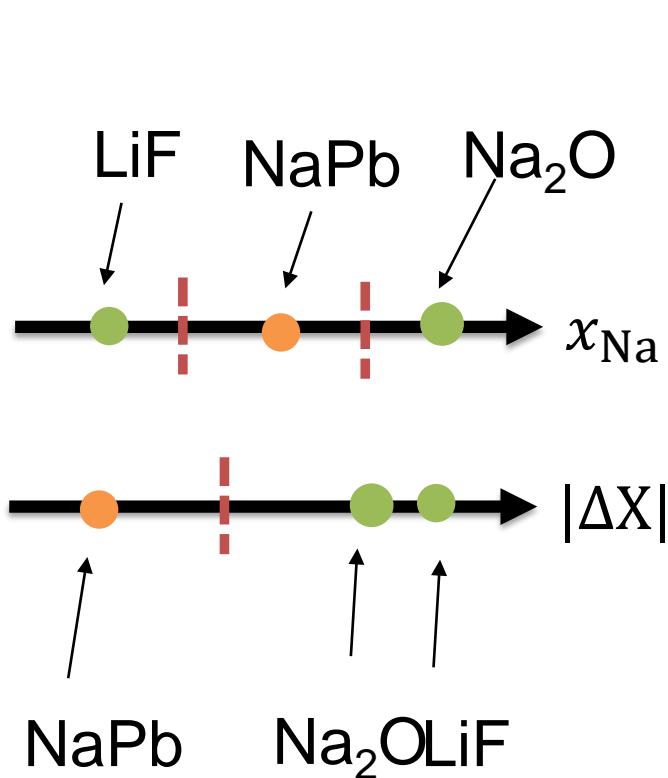
$$\Delta H_f = f(Z_A, Z_B)$$

How can one create problem-independent representations?

# What is a representation?

*Set of quantitative attributes that describe a material*

$$\text{Property} = f(\text{Attributes})$$



Representation of material

$$\text{Ex: Attributes} = \mathbf{g}(x_H, x_{He}, \dots)$$

**What does a representation need?**

*Completeness:* Differentiate materials

*Efficiency:* Quick to compute

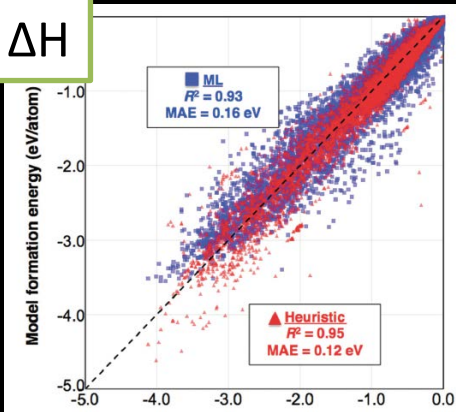
*Accuracy:* Capture important effects

*Diversity:* Many possible properties

How do we create “general-purpose” representations?

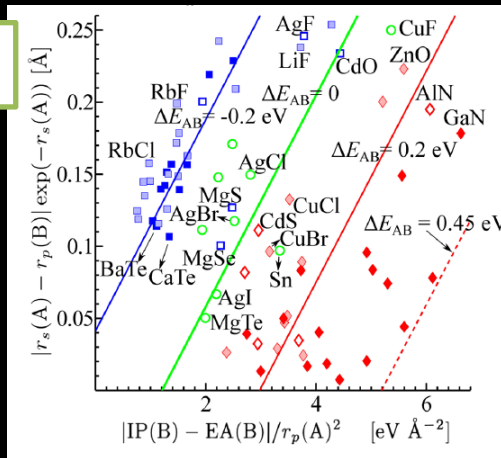
# ML + Materials = “Materials Informatics”

$\Delta H$



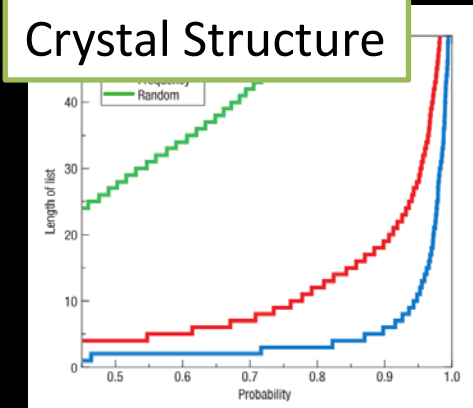
Meredig *et al.* PRB (2014), 094104

$\Delta H$



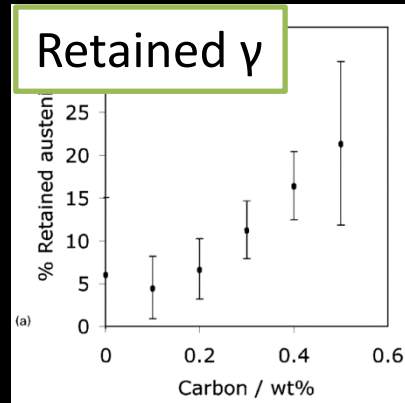
Ghiringhelli *et al.* PRL (2015), 105503

Crystal Structure



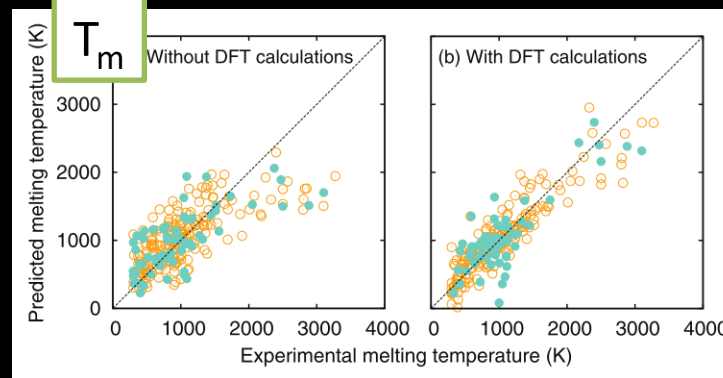
Fischer *et al.* Nat. Mat. (2006), 641

Retained  $\gamma$



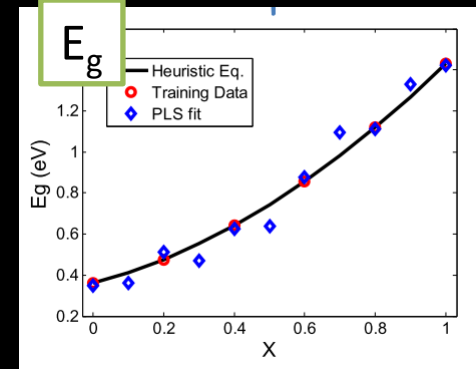
Chatterjee *et al.* MS&T (2007), 819

$T_m$



Seko *et al.* PRB (2014), 054303

$E_g$



Srinivasan, Rajan. Materials (2013), 279

# Focus #1: Representations based on composition alone

Property	Attributes	Reference
Crystal Structure	$VE, \Delta X, n_{av}, \Delta n_{ws}^{1/3}$	Kong et al., 2012
Band Gap	$\Delta X, Z, T_m, R, n_{av}$	Srinivasan & Rajan, 2013
Formation Energy	$\Delta X, Z, n_{s p d f}, \text{row, col}$	Meredig <i>et al.</i> , 2014
Melting Point	$Z, m, n, r^{cov}, l, X, \dots$	Seko <i>et al.</i> , 2014
$\Delta H_f$ : Rocksalt – Wurtzsite	IP, EA, $r_s, r_p, \dots$	Ghiringhelli <i>et al.</i> , 2015

## Observations:

- Different properties, different attributes
- All based on elemental property statistics

**Our Strategy:** Create set that includes all of these and more

# Machine Learning Strategy

- Recall basic calculation recipe:
  - Composition
  - Structure
- People focus on predicting/solving structure, but what if we could predict properties without it?
- Application: **Discovery of new ternary compounds  $A_xB_yC_z$**

# Structure-Independent Model

Instead of mapping an atomic configuration to properties, i.e.,

$$C(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_n) \rightarrow P$$

we instead train a formation energy model on composition only:

$$M(x_H, x_{He}, x_{Li} \dots x_{Pu}) \rightarrow \Delta E_f$$

# General-Use Attributes

**Elemental Property Stats.:** *Mean  $T_m$ , Range  $Z$ , ...*

*6 Statistics:* Mean, variance, max, min, range, mode

*22 Elemental Properties:*  $Z$ , EN, Row, Column, Radius, ...

**Stoichiometric:** *# Components,  $\|x_Z\|_p$*

**Electronic Structure Based:** *Fraction  $p$  Electrons, ...*

**Ionicity:** *Can form Ionic, % Ionic Character, ...*

<https://bitbucket.org/wolverton/magpie>



# Predictions for Discovery: 4500 new stable compounds

Machine learning model can predict the thermodynamic stability of arbitrary compositions without any other input (i.e., without the structure).

Six orders of magnitude less computer time than DFT.

We scan ~1.6 million candidate compositions for novel ternary compounds ( $A_xB_yC_z$ ),

Predict 4500 new stable materials (would represent a ~10% increase in the total number of known ternary compounds).

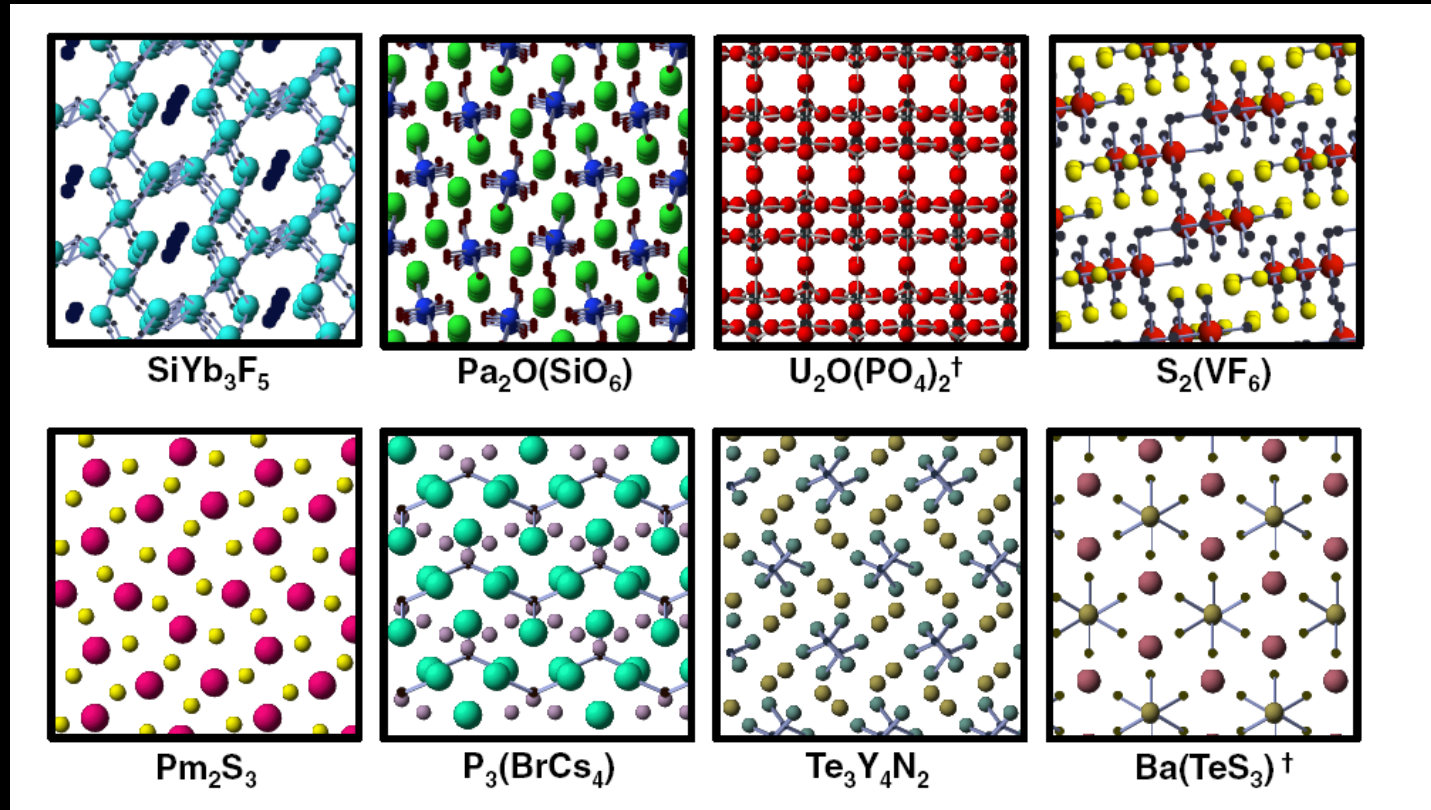
Complete list of predicted compounds:

[http://journals.aps.org/prb/supplemental/10.1103/PhysRevB.89.094104/predictions\\_dat.pdf](http://journals.aps.org/prb/supplemental/10.1103/PhysRevB.89.094104/predictions_dat.pdf)

Meredig et al., Phys. Rev. B **89**, 094104 (2014).



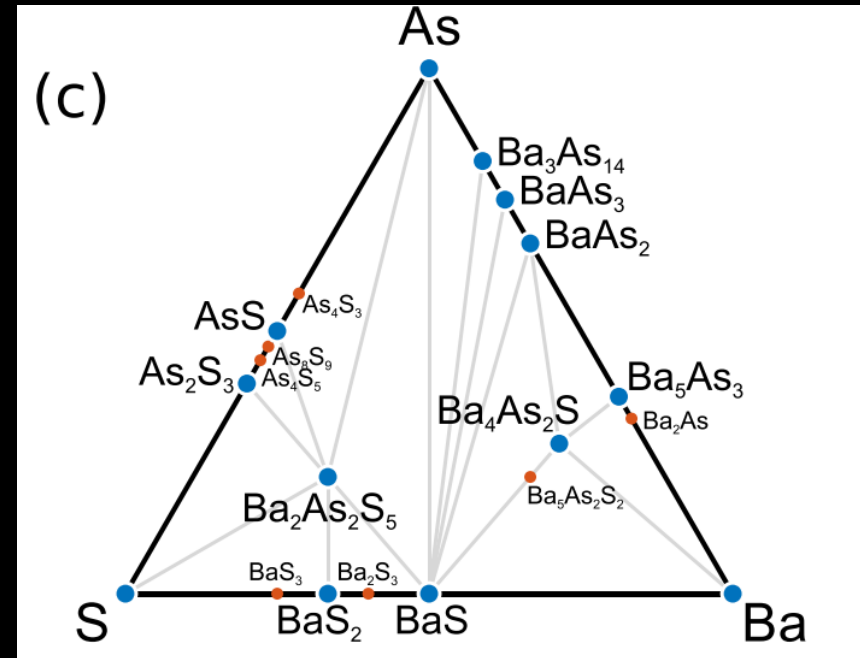
# Validating high-ranking compositions with crystal structure prediction



Tested 9 predicted stoichiometries. In 8 cases, crystal structure prediction methods found a structure with DFT energy lower than all combinations of existing known phases.

# Using this ML model to find new energy materials

- ML pointed to novel compounds in Ba-As-S system
- Minima Hopping Method (structure prediction method), to find structures:  $\text{Ba}_4\text{As}_2\text{S}$  and  $\text{Ba}_2\text{As}_2\text{S}_5$
- Discovered entire *families* of these  $X_4Y_2Z$  and  $X_2Y_2Z_5$  compounds
- Promising solar cell (band gap and absorption) and thermoelectrics (power factor and thermal conductivity)



SHALL WE PLAY A GAME? ■

TIC-TAC-TOE

BLACK JACK

GIN RUMMY

HEARTS

BRIDGE

CHECKERS

CHESS

POKER

FIGHTER COMBAT

GUERRILLA ENGAGEMENT

DESERT WARFARE

AIR-TO-GROUND ACTIONS

THEATERWIDE TACTICAL WARFARE

THEATERWIDE BIOTOXIC AND CHEMICAL WARFARE

GLOBAL THERMONUCLEAR WAR

## SHALL WE PLAY A GAME?

### How good is your chemical intuition?

Welcome to the Wolverton group's metal detection challenge. The point of this game is to test whether the intuition of scientists is better than a model produced using machine learning.

#### Rules of the Game

---

1. Decide whether the given compound is a metal or non metal
2. All compounds are from the ICSD
3. Only the lowest-energy compound at a given composition is used
4. The electronic structure of each compound was determined using Density Functional Theory (DFT)
5. Metals are defined as compounds with a DFT band gap energy of 0
6. Band gaps energies were determined from the total electronic DOS
7. Calculations were performed with tetrahedral integration

#### Your Opponent

---

You will be competing against a model trained against 3000 randomly-selected compounds from the ICSD. See above for the rules governing which compounds were used. To create this model, 81 attributes were calculated for each compound and used to create decision rules with the rotation forest algorithm. These attributes include things like the average electronegativity of each element and the maximum difference between their melting temperatures.

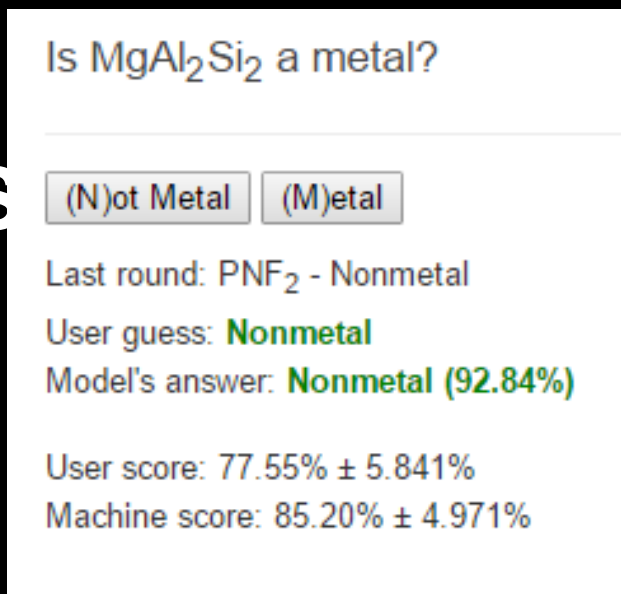
<http://palestrina.northwestern.edu/metal-detection/>

# Simple Example: Is it a Metal?

**Task:** Given composition,  $E_g > 0$ ?

**Training Set Dataset:** 3000 entries from the OQMD

**Score:** Accuracy ~90%



Is  $\text{MgAl}_2\text{Si}_2$  a metal?

Last round:  $\text{PNF}_2$  - Nonmetal

User guess: **Nonmetal**

Model's answer: **Nonmetal (92.84%)**

User score: 77.55%  $\pm$  5.841%

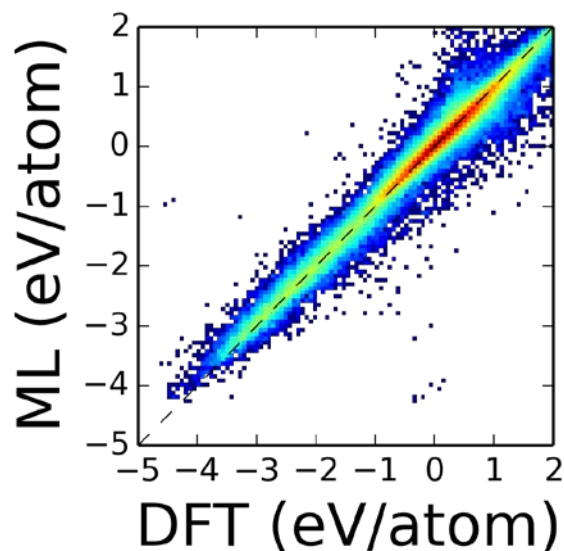
Machine score: 85.20%  $\pm$  4.971%

Game: [palestrina.northwestern.edu/metal-detection/](http://palestrina.northwestern.edu/metal-detection/)

# Application to the OQMD

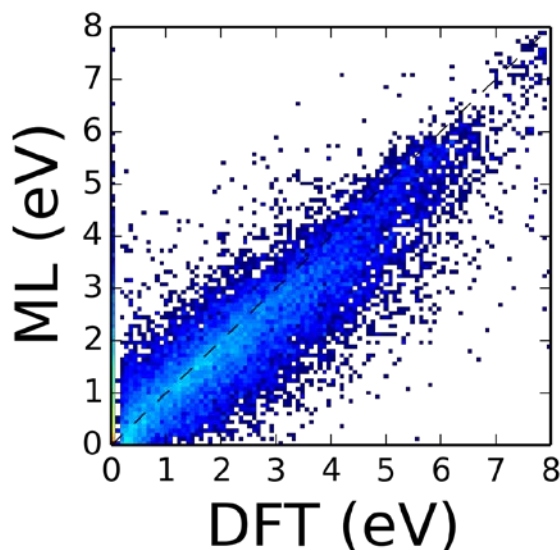
**Dataset:** 240000 DFT Calculations (OQMD.org)

$\Delta H_f$



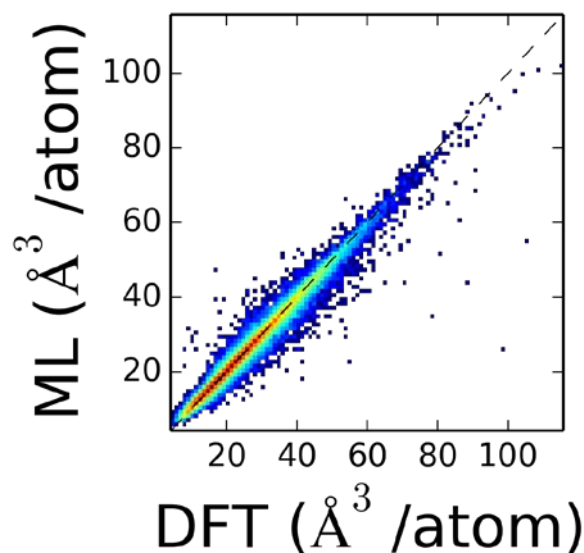
**R:** 0.944  
**MAE:** 80.5 meV/atom

$E_g$



**R:** 0.924  
**MAE:** 0.21 eV

$V$

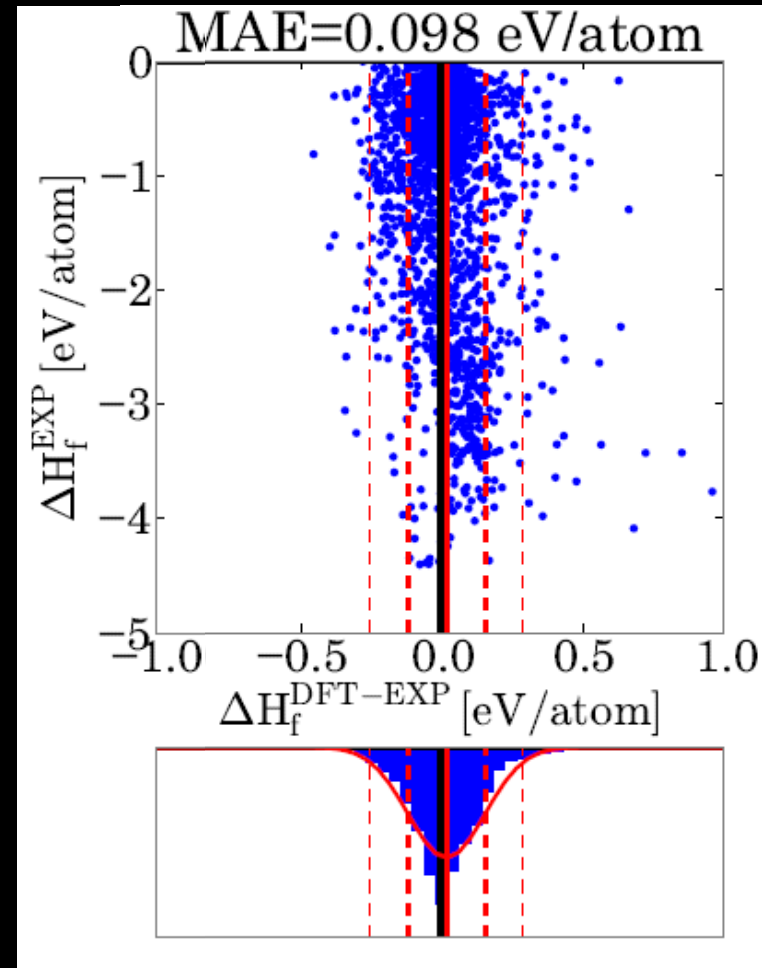


**R:** 0.993  
**MAE:** 0.452 Å³/atom

# Accuracy of DFT Formation Energies

(comparison with a large number of ~1670 experimentally measured points)

$$\Delta H_f(\sigma) = E(\sigma) - \sum x_i E_i$$



J. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM 65, 1501 (2013).

FERE: V. Stevanovic, S. Lany, X. Zhang, and A. Zunger, Phys. Rev. B 85, 115104 (2012).

Mixing GGA/GGA+U: A. Jain et al., Comput. Mater. Sci. 50, 2295 (2011).

# Predicting Glass Forming Ability

**Application:** Metallic Glasses

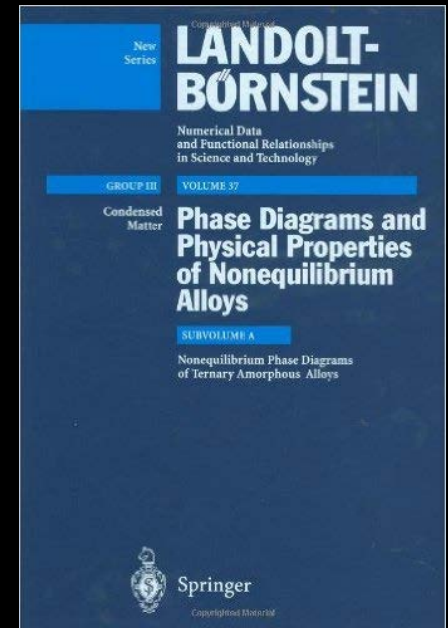
**Goal:** Predict glass-forming ability

**Dataset:** Landolt-Börnstein

- 6836 experimental measurements
- 295 ternary systems
- Binary property: **[Can Form Glass]** | **[Cannot Form]**

**Model:** Random Forest

- 90% accurate in 10-fold cross-validation



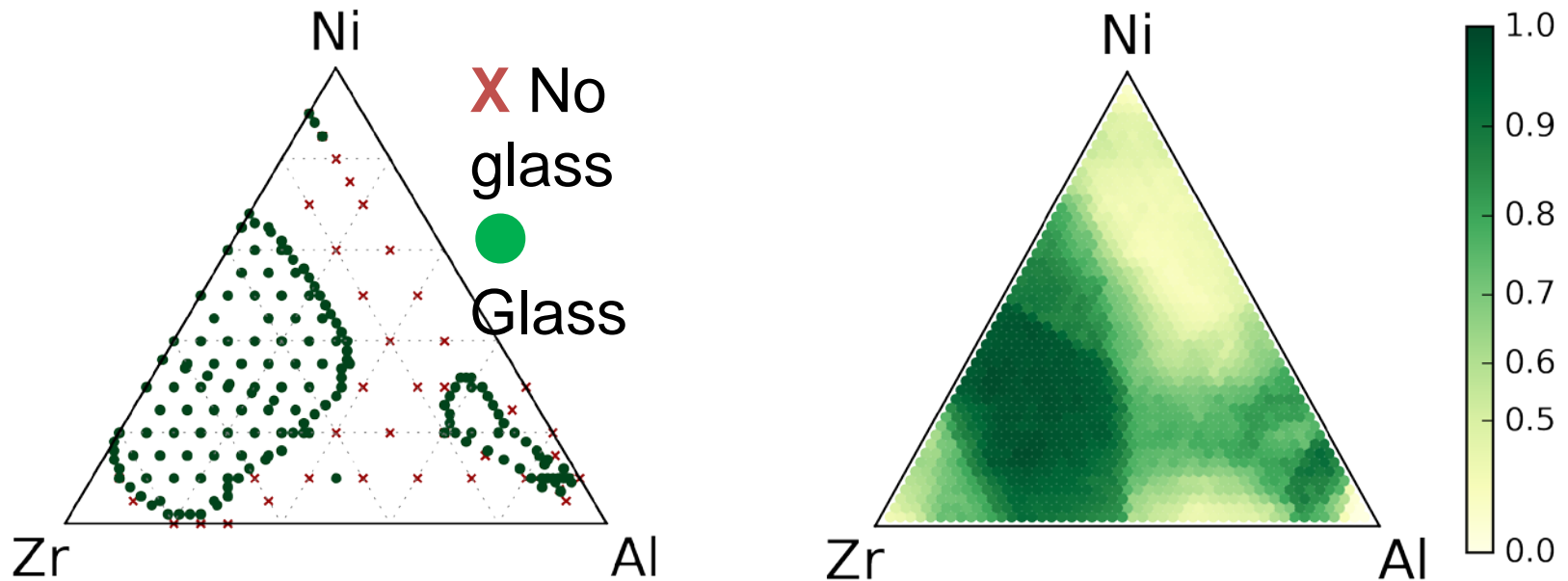


# Predicting Glass-Forming Ability

Test: Remove Al-Ni-Zr data from training data, try to predict

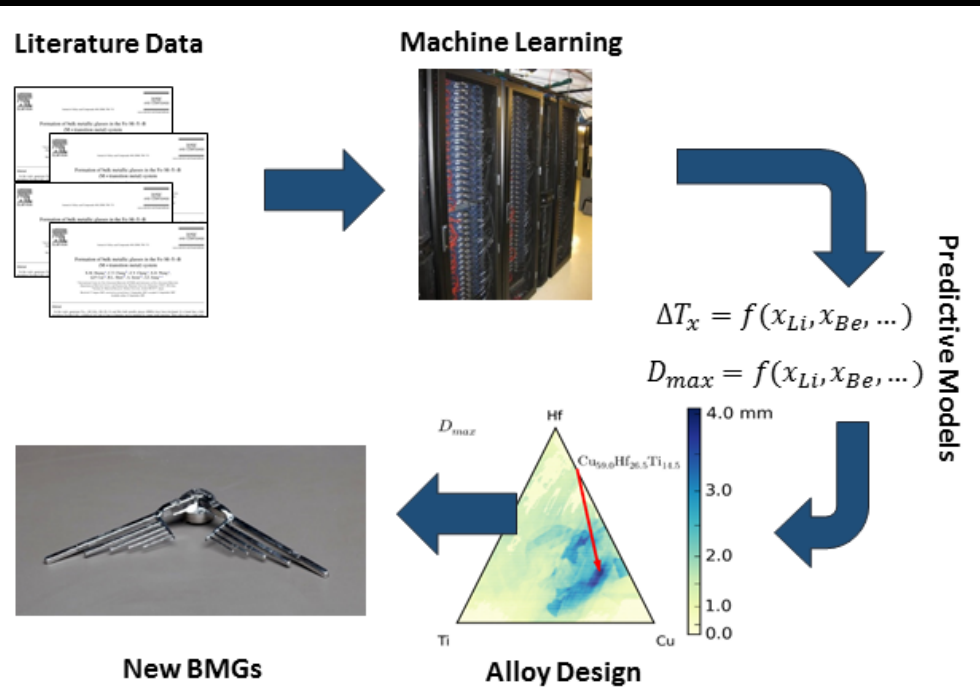
Measured

Predicted



Same representation, very different material property

# ML Prediction of New BMG Compositions



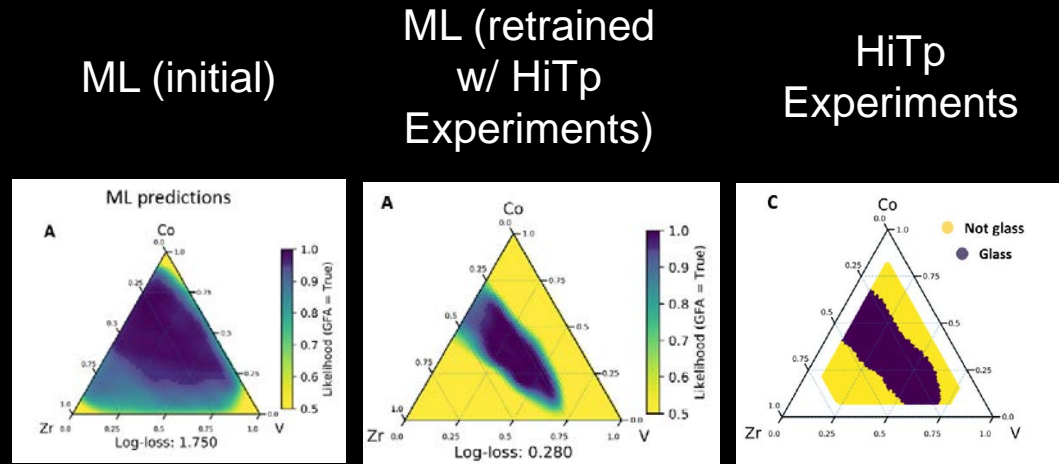
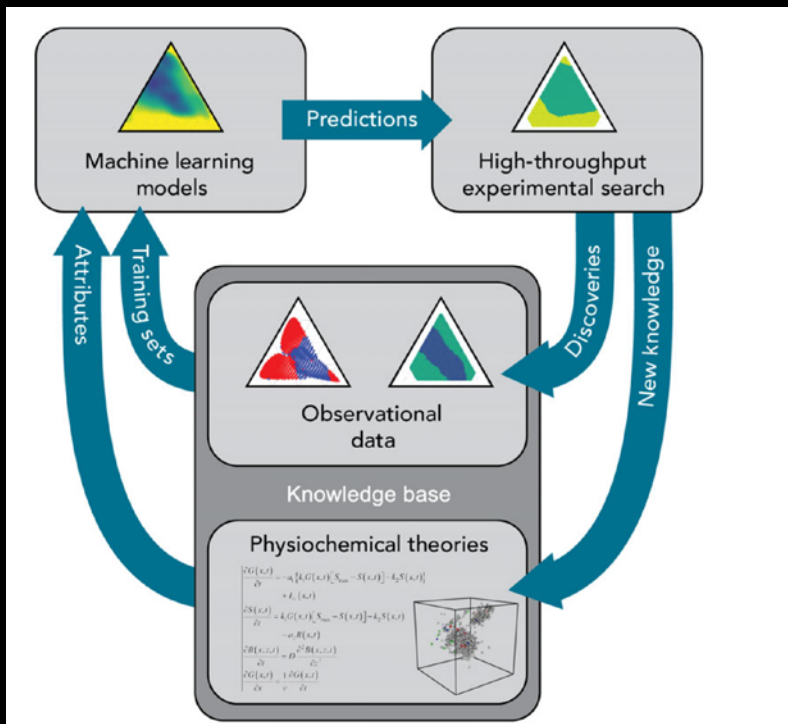
## Search Space:

- 53 Elements
- 27 Million Compositions
- Not “near” any known existing BMG

Top Alloys	
Zr <sub>0.4</sub> Cr <sub>0.16</sub> Cu <sub>0.44</sub>	Hf <sub>0.66</sub> Fe <sub>0.32</sub> Co <sub>0.02</sub>
Zr <sub>0.5</sub> Cr <sub>0.04</sub> Fe <sub>0.46</sub>	Hf <sub>0.58</sub> Fe <sub>0.42</sub>
Hf <sub>0.52</sub> Fe <sub>0.28</sub> Re <sub>0.2</sub>	Zr <sub>0.32</sub> Cr <sub>0.3</sub> Ni <sub>0.38</sub>
Hf <sub>0.42</sub> Ni <sub>0.42</sub> Ag <sub>0.16</sub>	Hf <sub>0.52</sub> Fe <sub>0.28</sub> Os <sub>0.2</sub>
Zr <sub>0.58</sub> Ni <sub>0.26</sub> Ir <sub>0.16</sub>	V <sub>0.18</sub> Ni <sub>0.62</sub> B <sub>0.2</sub>

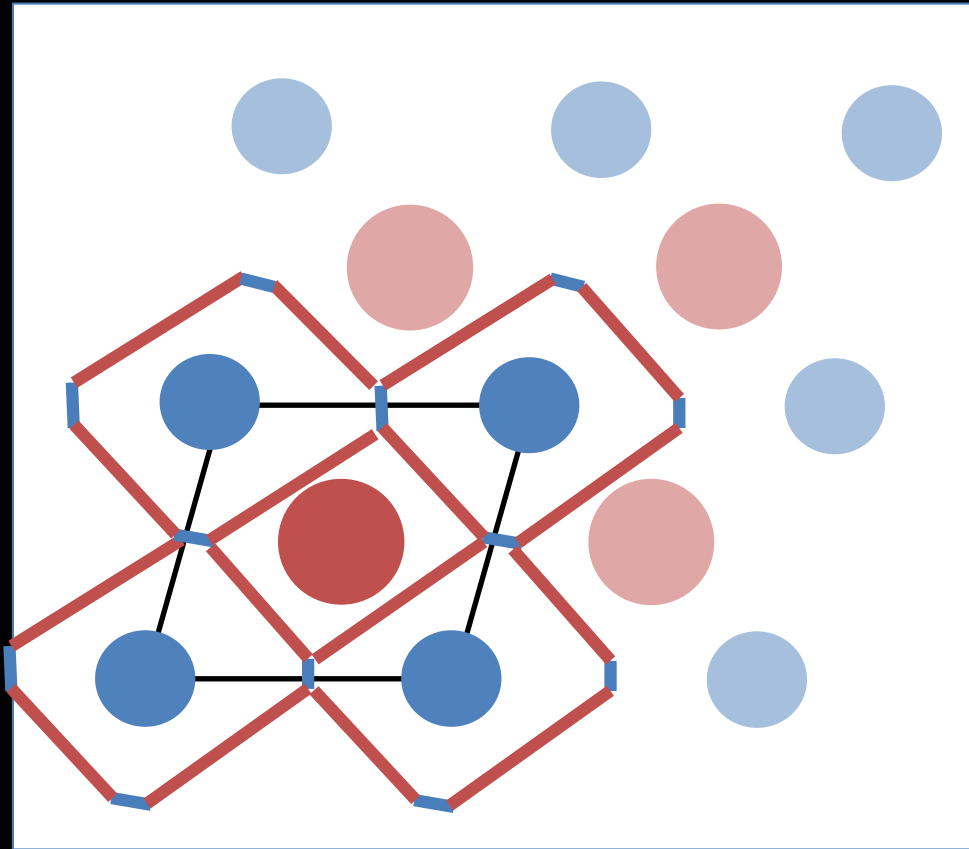
[http://oqmd.org/static/analytics/glass\\_search.html](http://oqmd.org/static/analytics/glass_search.html)

# Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments



- *Discovered a new glass-forming ternary system (Co-V-Zr)*
- *Include processing-dependent conditions in ML model*

# Focus #2: Adding Crystal Structure Information to Representation



## Our Approach:

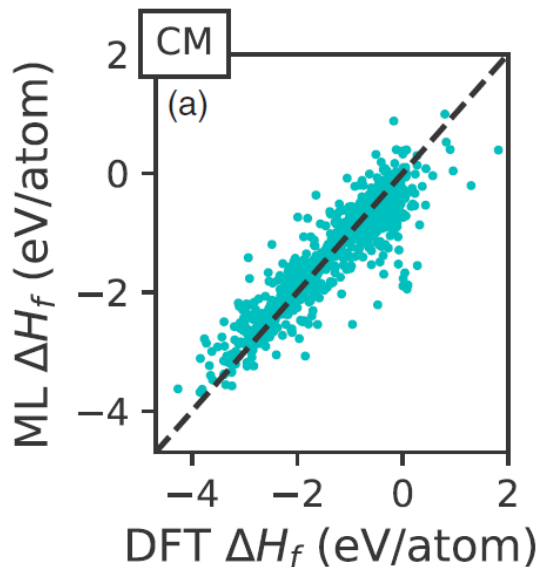
Voronoi-tessellation-based attributes

## Atomic Characteristics:

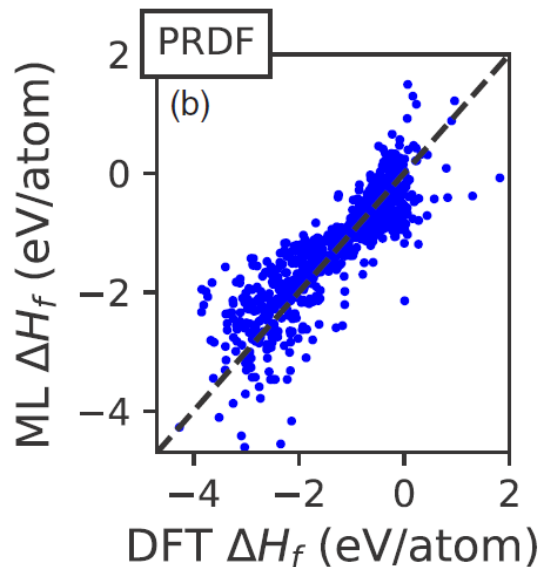
1. Element identity
2. Coordination number
3. Bond length
4. Cell size
- ...

Atomic Characteristics + Descriptive Statistics = 275 Attributes

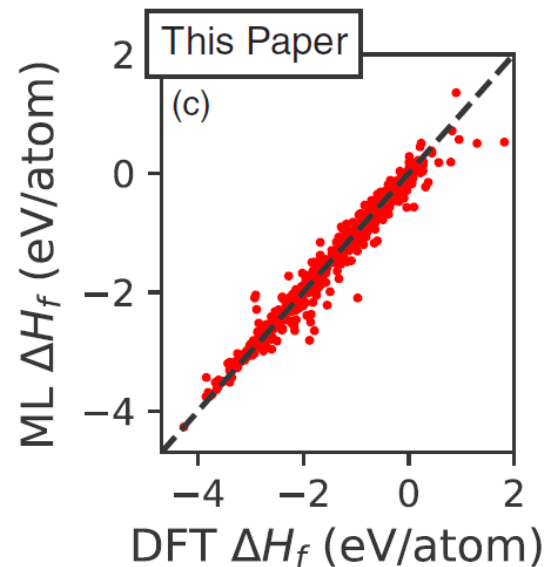
# Formation Energy ML Models: Comparison of Representations



R: 0.923  
MAE: 0.25 eV/atom



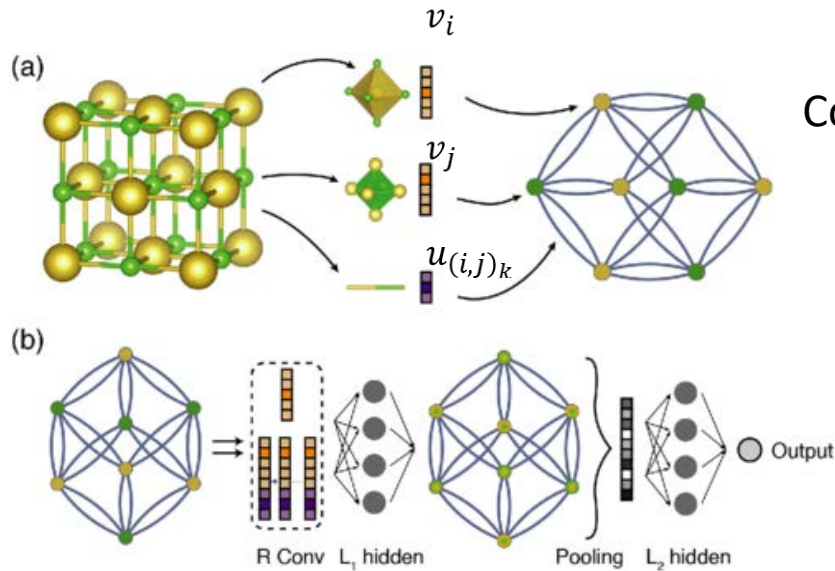
R: 0.858  
MAE: 0.37 eV/atom



R: 0.988  
MAE: 0.09 eV/atom

Composition  
Attributes +  
Voronoi

# Crystal Graph Convolutional Networks (CGCNNs)



Convolution Function:

$$v_i^{(t+1)} = v_i^{(t)} + \sum_{j,k} \sigma \left( z_{(i,j)_k}^{(t)} \mathbf{W}_f^{(t)} + \mathbf{b}_f^{(t)} \right) \odot g \left( z_{(i,j)_k}^{(t)} \mathbf{W}_s^{(t)} + \mathbf{b}_s^{(t)} \right)$$

$$z_{(i,j)_k}^{(t)} = v_i^{(t)} \oplus v_j^{(t)} \oplus u_{(i,j)_k}^{(t)}$$

Pooling Function:

$$v_c = \text{Mean}(v_0^{(T)}, v_1^{(T)}, \dots, v_N^{(T)})$$

$$v_i^{(t+1)} = \text{Conv}(v_i^{(t)}, v_j^{(t)}, u_{(i,j)_k}) \quad v_c = \text{Pool}(v_0^{(T)}, v_1^{(T)}, \dots, v_N^{(T)})$$

Xie and Grossman, Phys. Rev. Lett., 2018)

# Measuring the performance of the model

- Predict DFT formation energy,
- Open Quantum Materials Database<sup>1</sup>
  - Training data set: 200,000 entries
  - Testing data set: 20,000 entries
- Benchmark using the 3D CNN & Voronoi tessellation models<sup>2</sup>
- Single validation test is done

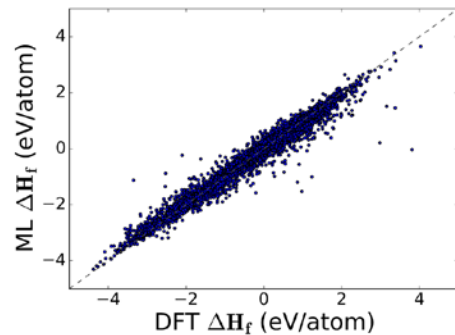


***Performance is measured by model prediction accuracy of the testing data set***

1. J. E. Saal *et al.* JOM **65**, 1501 (2013).  
2. L. Ward *et al.* Phys. Rev. B **96**, 024104 (2017)

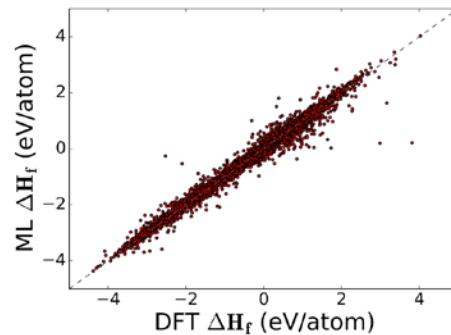
# Results: Formation energy predictions

(a) Voronoi + RF



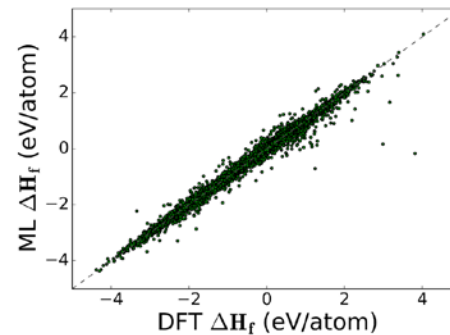
MAE: 86.2 meV/atom  
RMSE: 144.5 meV/atom

(b) Voronoi + NN



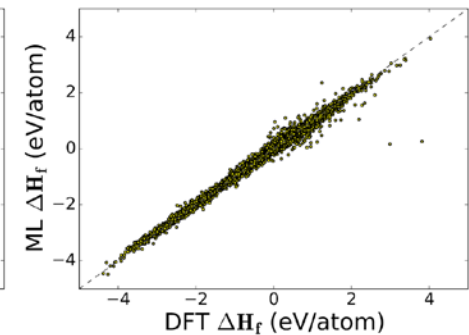
MAE: 61.0 meV/atom  
RMSE: 107.0 meV/atom

(c) CNN



MAE: 48.7 meV/atom  
RMSE: 91.2 meV/atom

(d) CGCNN



MAE: 41.1 meV/atom  
RMSE: 74.5 meV/atom

- Error of CGCNN (41 meV/atom) is much less than difference between DFT and experimental formation energies ( $\sim 100$  meV/atom)<sup>1</sup>
- CGCNN model outperforms all other models
- CGCNN has less outliers

1. J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM 65, 1501 (2013).



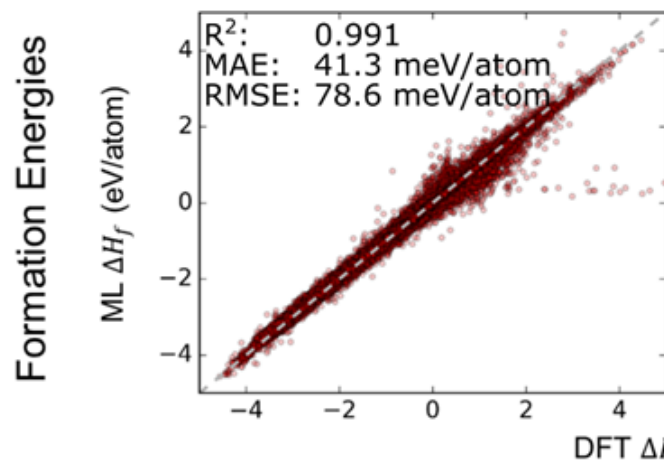
# Improving the CGCNN method

## Training on OQMD

Training Set: 200,000  
Compounds  
Test Set: 20,000  
Compounds

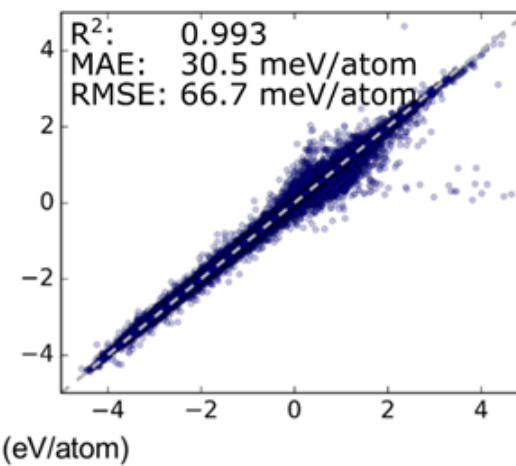
“Original”  
CGCNN

(a)



“Improved”  
iCGCNN

(b)



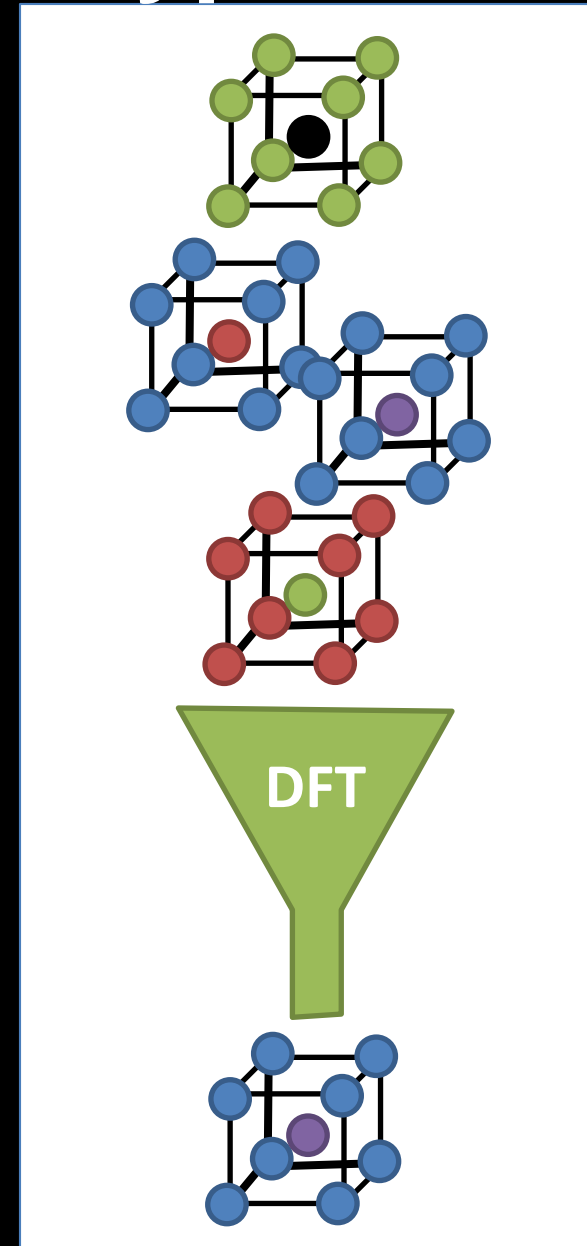
# Application: The Prototype Search

## Common Method: Prototype Search

1. Select a crystal structure
2. Evaluate **all** possibilities with DFT
3. Select only stable ones

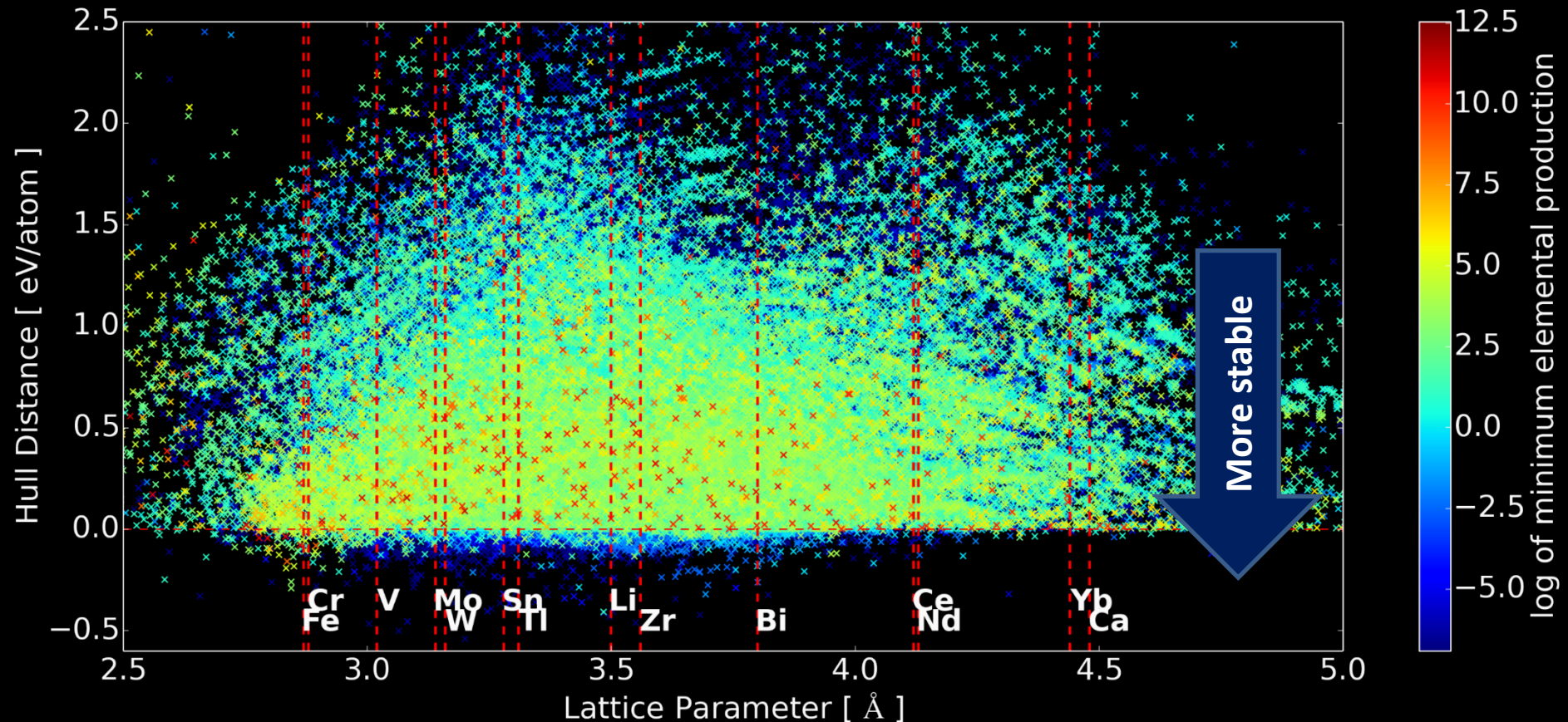
**Challenge:** Computational cost (success rate in finding stable compounds can be very low)

**Possible Solution:** Guide with ML



# High-throughput search for Heusler $X_2YZ$ precipitate strengtheners in BCC metals

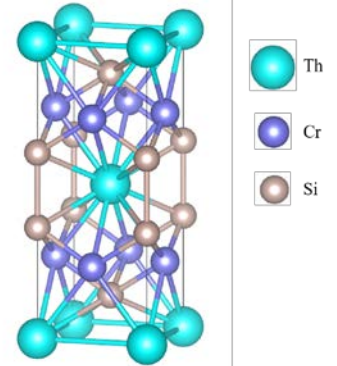
> 180,000 DFT calculations of  $X_2YZ$  Heuslers  
(essentially for all possible X, Y, Z)



S. Kirklin, J. E. Saal, V. Hegde and C. Wolverton, "High-Throughput Combinatorial Screening of Intermetallic Compounds as Strengthening Precipitates" *Acta Mater.* (2016).

# Application: Using iGCNN Deep Learning to accelerate discovery of new stable materials

- ThCr<sub>2</sub>Si<sub>2</sub>-type materials
  - One of the most common prototype structures
  - ~1000 examples of stable compounds in OQMD with this structure type!
- Using combinatorial search method to discover new materials



## Generating new compounds

~120,000 new compounds generated by substituting elements

## Predicting stability

CNN predicted formation energies are used to calculate hull distance

## Validation using DFT

100 compounds with highest predicted stability are cross-checked using DFT

*iGCNN model is **200x** more likely to discover a stable compound than random search (and ~2x more likely than using CGCNN)*

# Summary

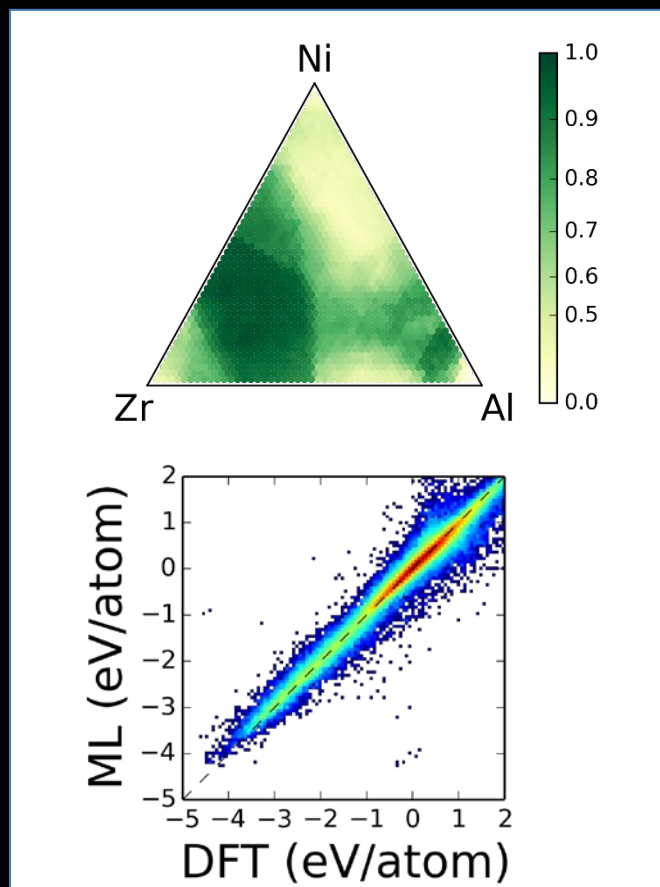
Collect

Process

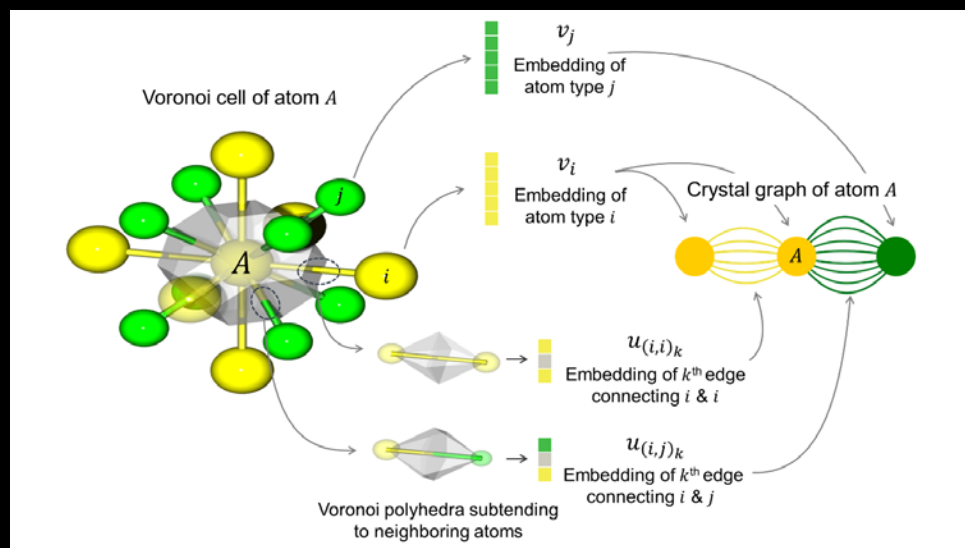
Represent

Learn

Composition-Based Attributes



Crystal Structure Attributes



# More information/resources...

- Machine Learning models



- MAGPIE <https://bitbucket.org/wolverton/magpie>
- B. Meredig et al., "Combinatorial screening for new materials in unconstrained composition space with machine learning", Phys. Rev. B **89**, 094104 (2014).
- L. Ward et al., "A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials" npj Computational Materials 2, 16028 (2016).
- L. Ward, C. Wolverton, "Atomistic calculations and materials informatics: A review", Curr. Opin. Solid State Mater. Sci. 21, 167 (2017).
- L. Ward et al., "Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations", Phys. Rev. B 96, 024104 (2017).
- F. Ren, L. Ward, et al., "Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments", Science Adv. 4, eaaq1566 (2018).