Materials Data in Action

Predicting the formation of binary compounds using machine learning

CITRINE INFORMATICS

5 September 2019

Chris Borg, Research Scientist

Agenda

0. **Is this a good problem for ML?**

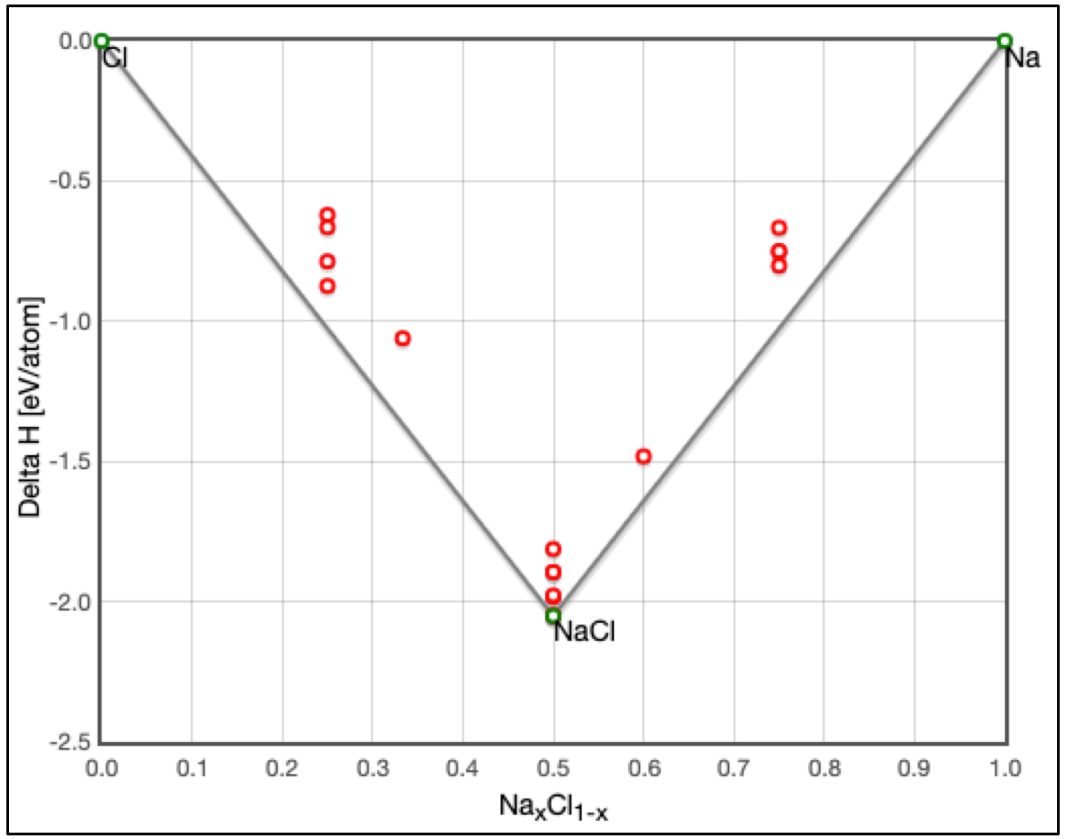
1. **Data preparation**

2. **Modeling**
   a. Classification with random forest
   b. Model quality metrics

3. **Compare to DFT**
   a. Model performance vs. “ground-truth”

4. **Further considerations**

Task: Predict the stability of binary compounds



http://oqmd.org/materials/composition/NaCl

Typically, chemists use thermodynamics to determine the relative stability of a binary (AB) compound.

stabilityVec: [1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1]

1 = stable
0 = unstable

Na, NaCl, Na2Cl3, Cl

Task: build a machine learning model in Python to predict the full stability vector

Is this a reasonable problem for ML to solve?

Machine learning (ML): A set of data-driven algorithms that discern patterns from high-dimensional spaces.

High-dimensional input space: chemical composition (can be mapped to 100s of physically meaningful features)

Binary output: stable/unstable labels for 68k compositions (82 elements, 2572 AB combinations)

Validation / ”ground-truth”: DFT calculated enthalpies of formation

Preparing dataset

Inputs:

formulaA	formulaB
Ac	Ag
Ac	Al
Ac	As
Ac	Ba
Ac	Bi

...

Output:

stabilityVec
[1.0,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0,1.0]
[1.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0]
[1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0]
[1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0]
[1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0]

2572 rows, 99 columns: element A, element B, stability vector, 48 features per element (A, B)

```
df = df.explode('stabilityVec')
```

Input: Output:

FORMULA	stabilityVec
Ac1	1.0
Ac0.9 Ag0.1	0.0
Ac0.8 Ag0.2	0.0
Ac0.7 Ag0.3	1.0
Ac0.6 Ag0.4	0.0
Ac0.5 Ag0.5	1.0
Ac0.4 Ag0.6	0.0
Ac0.3 Ag0.7	0.0
Ac0.2 Ag0.8	0.0
Ac0.1 Ag0.9	0.0
Ag1	1.0

Each row is now a single composition (input) and stability classification (output)

Deriving physically-informed features

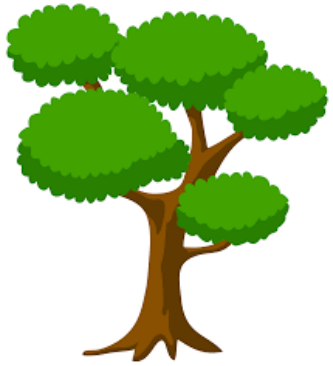
formulaA	formulaB
Ac	Ag
Ac	Al
Ac	As
Ac	Ba
Ac	Bi

FORMULA
Ac1
Ac0.9 Ag0.1
Ac0.8 Ag0.2
Ac0.7 Ag0.3

```
There are 120 possible descriptors:

['MagpieData minimum Number' 'MagpieDa
 'MagpieData range Number' 'MagpieData
 'MagpieData avg_dev Number' 'MagpieDa
 'MagpieData minimum MendeleevNumber'
 'MagpieData range MendeleevNumber' 'M
 'MagpieData avg_dev MendeleevNumber'
```

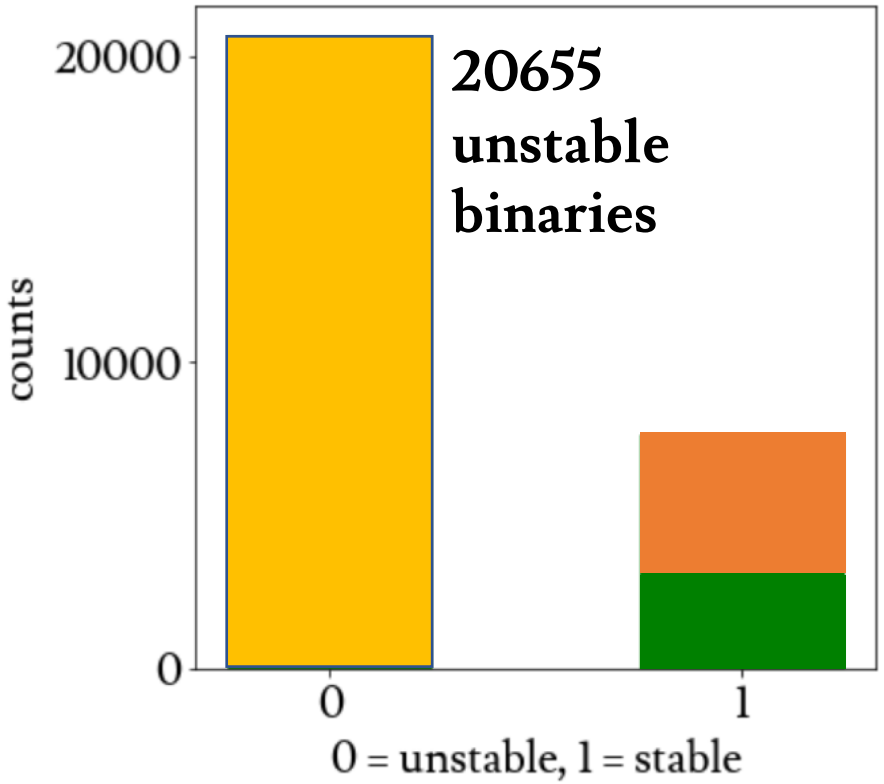
```python
from matminer.featurizers.conversions import StrToComposition
from matminer.featurizers.composition import ElementProperty

df = StrToComposition().featurize_dataframe(df, "FORMULA")
ep_feat = ElementProperty.from_preset(preset_name="magpie")
df = ep_feat.featurize_dataframe(df, col_id="composition")
```

Choosing an estimator



Random forest

- Interpretable (ranked feature importance)
- Can capture non-linear correlations



Establishing a baseline:

1. random guess = 0.5
2. ”intuition” = 0.82

5144 stable monatomic

2493 stable binaries

Evaluating predictions

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=1)
rf = RandomForestClassifier(n_estimators=20, random_state=1)
rf.fit(X_train, y_train)
```

	N (predicted)	P (predicted)
N (actual)	TN = 6665	FP = 173
P (actual)	FN = 583	TP = 1916

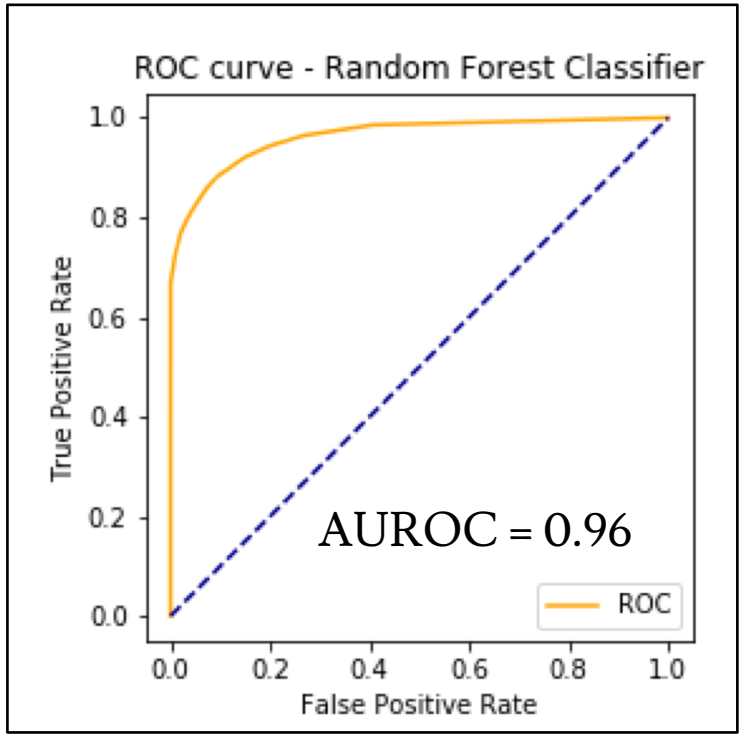
$$\text{Accuracy} = \frac{6665 + 1916}{6665 + 1916 + 583 + 173} = \mathbf{0.92}$$

F1 score = **0.84**

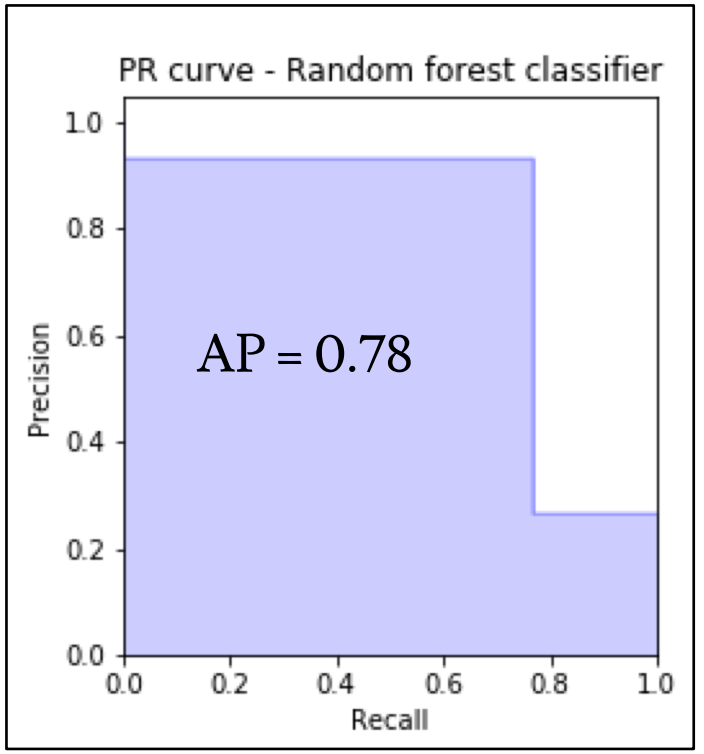
3-fold CV avg. accuracy = 0.963

10-fold CV avg. accuracy = 0.967

Graphical model accuracy metrics



Typically, an ROC curve is a good measure of the ability of a binary classifier.



However this is an overly optimistic view due to the large class imbalance between stable and unstable compounds. PR curve better captures the classification ability of the model.

Comparing against DFT

unit_comp	is_stable
Ac1	True
Ac0.9 Be0.1	False
Ac0.8 Be0.2	False

Added DFT calculated stability to test data.

While accuracy is high, FP is more than 2x TP.

If a researcher were to attempt synthesis of predicted compounds, they may likely succeed 50% of the time.

	N (predicted)	P (predicted)
N (actual)	TN = 6230	FP = 218
P (actual)	FN = 196	TP = 97

Accuracy = 0.94

F1 score = 0.32

Important features

	importance
MagpieData range MendeleevNumber	0.071601
MagpieData avg_dev AtomicWeight	0.068320
MagpieData range GSvolume_pa	0.068017
MagpieData avg_dev MendeleevNumber	0.068004
MagpieData avg_dev GSvolume_pa	0.057959

ELSEVIER

Journal of Alloys and Compounds 367 (2004) 167–175

Journal of ALLOYS AND COMPOUNDS

www.elsevier.com/locate/jallcom

Data-driven atomic environment prediction for binaries using the Mendeleev number Part 1. Composition AB

P. Villars[a,*], K. Cenzual[b], J. Daams[a], Y. Chen[c], S. Iwata[c]

[a] Materials Phases Data System, Schwanden 400, Vitznau CH-6354, Switzerland
[b] Department of Inorganic, Analytical and Applied Chemistry, Geneva University, Geneva, Switzerland
[c] QUEST, The University of Tokyo, Tokyo, Japan

Abstract

The atomic environment types (AETs) (coordination polyhedra) realized by each chemical element in binary compounds at the equi-atomic composition were analyzed based on a comprehensive set of literature data. The Mendeleev number (*MN*) (ordering number listing the chemical elements column by column through the periodic system) was successfully used to classify the chemical systems. An atomic environment type map, using as coordinates the maximum Mendeleev number versus the ratio between the minimum and the maximum Mendeleev number, sub-divided the chemical systems where different atomic environment types occur in distinct stability domains. The same maps also showed a clear separation between chemical systems where intermediate compounds form and those where no compounds form. These maps make it possible to predict the existence of compound that have not yet been investigated with a particular atomic environment.

Keywords: Data mining; Atomic environment; Mendeleev number

1. Introduction

The large amount of experimental data collected over the years contains information and correlations that can be used for a rational semi-empirical approach to materials design. The ideal situation is to be able to relate any kind of "compound property" to one or several parameters characteristic of the constituent elements. In a previous study [1], concerning the formation of binary compounds, a very clear separation between chemical systems where compounds form and those where no compounds form was achieved using maps based on combinations of the Mendeleev numbers (*MN*) of the constituent elements. Very encouraging was the fact that the results obtained on binaries could be extended to ternaries and quaternaries. This is an extremely important point because materials design is nowadays more and more focused on multinaries, but the experimental knowledge available is only substantial for binaries (approximately 70% of all binary systems have been studied, but less than 5% of the ternary and less than 0.5% of the quaternary systems).

After having defined the conditions for the chemical systems where binary compounds form, it seemed of interest to further study the structural features of these compounds. We have focused on the most frequently occurring stoichiometry for binary compounds, AB, and for each compound we have considered the atomic environment type (AET), or coordinations, of the atoms. The major motivation of the study was to find answers to the following questions: Do particular chemical elements strongly prefer certain AETs? How complex is the distribution of A and B atoms in mixed AETs? Do there exist 'stability criteria' for the different AETs?

1.1. Elemental parameters

The elemental parameters proposed so far for the construction of different compound property maps can be subdivided into different groups: atomic number factor, group number factor, quantum number factor, Mendeleev number factor, cohesion-energy factor (enthalpy of formation, etc.), electrochemical factor (electronegativity, etc.), size factor (Zunger pseudo-potential, covalent radius, atomic volume, etc.), atomic environment (AE) factor. The first four groups are completely independent of the nature of other atoms.

* Corresponding author.
E-mail address: villars@mpds.ch (P. Villars).

0925-8388/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.jallcom.2003.08.060

"The Mendeleev number (MN) (ordering number listing the chemical elements column by column through the periodic system) was successfully used to classify the chemical systems"

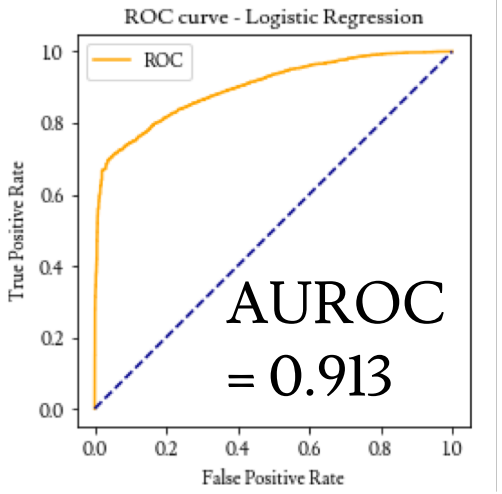
Conclusions

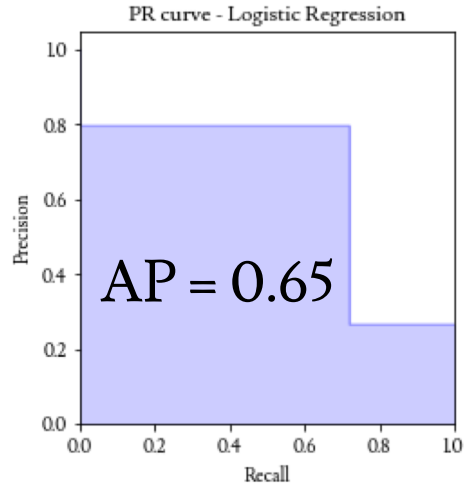
The random forest model trained on the given data is able to discern, above a baseline, between stable and unstable compounds.

However, the class imbalance (low number of stable compounds) skews our perspective on the model's classification ability (as shown in the comparison to DFT)

This imbalance causes an equivalent number of false positives and false negatives and could be address through re-samples (super- or sub- sampling)

Logistic regression results





Held out test set

	N (predicted)	P (predicted)
N (actual)	TN = 6395	FP = 443
P (actual)	FN = 681	TP = 1818

Accuracy = 0.88
F1_score = 0.76

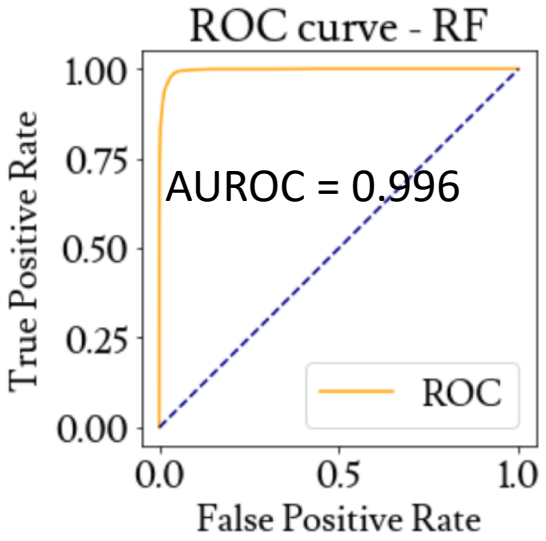
DFT-labeled test_data.csv

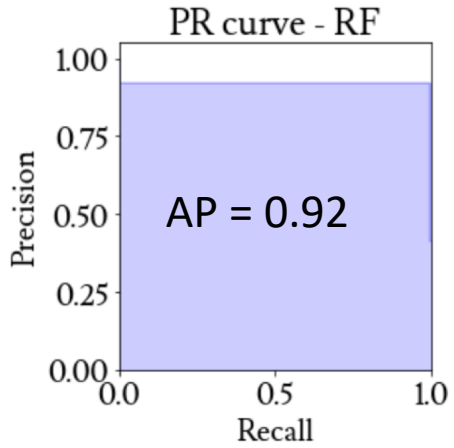
	N (predicted)	P (predicted)
N (actual)	TN = 6046	FP = 402
P (actual)	FN = 199	TP = 94

Accuracy = 0.911
F1_score = 0.24

rf supersampled

```
0.0    20655
1.0    15116
Name: stabilityVec, dtype: int64
stable binaries:  9972
unstable binaries:  20655
stable monatomic:  5144
unstable monatomic:  0
```





Held out test set

	N (predicted)	P (predicted)
N (actual)	TN = 6483	FP = 412
P (actual)	FN = 29	TP = 4881

Accuracy = 0.96
F1_score = 0.96

DFT-labeled test_data.csv

	N (predicted)	P (predicted)
N (actual)	TN = 5897	FP = 551
P (actual)	FN = 126	TP = 167

Accuracy = 0.9
F1_score = 0.33

Further considerations

- Investigate design space, low F1 score could be due to inability to extrapolate
- Feature engineering
- More data

THANK YOU!

CITRINE INFORMATICS

Tuning rf hyperparameters

n_estimators = 3	N (predicted)	P (predicted)
N (actual)	TN = 6500	FP = 338
P (actual)	FN = 539	TP = 1960

n_estimators = 10	N (predicted)	P (predicted)
N (actual)	TN = 6665	FP = 173
P (actual)	FN = 583	TP = 1916

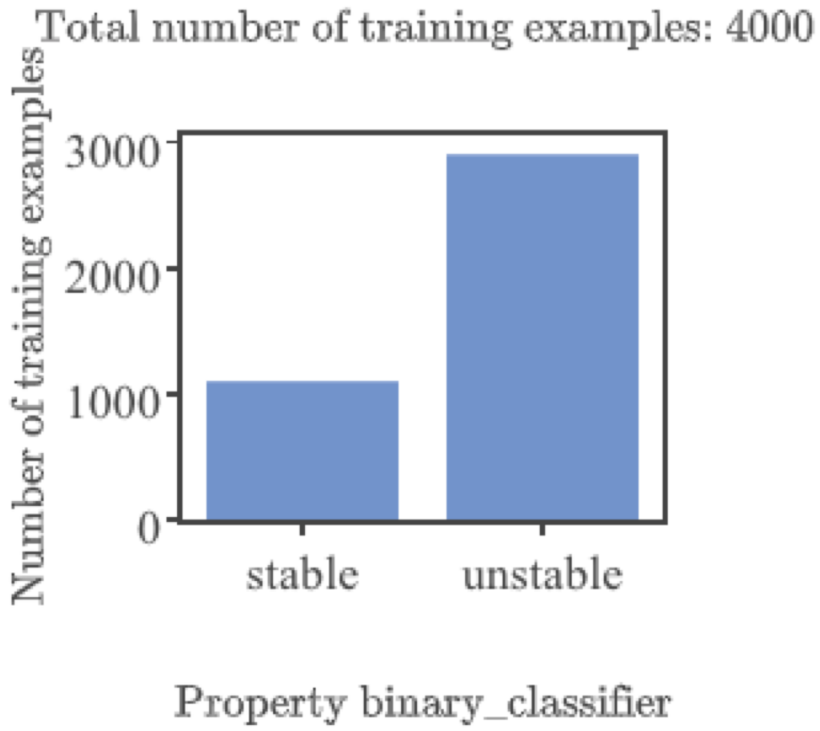
n_estimators = 20	N (predicted)	P (predicted)
N (actual)	TN = 6697	FP = 141
P (actual)	FN = 578	TP = 1921

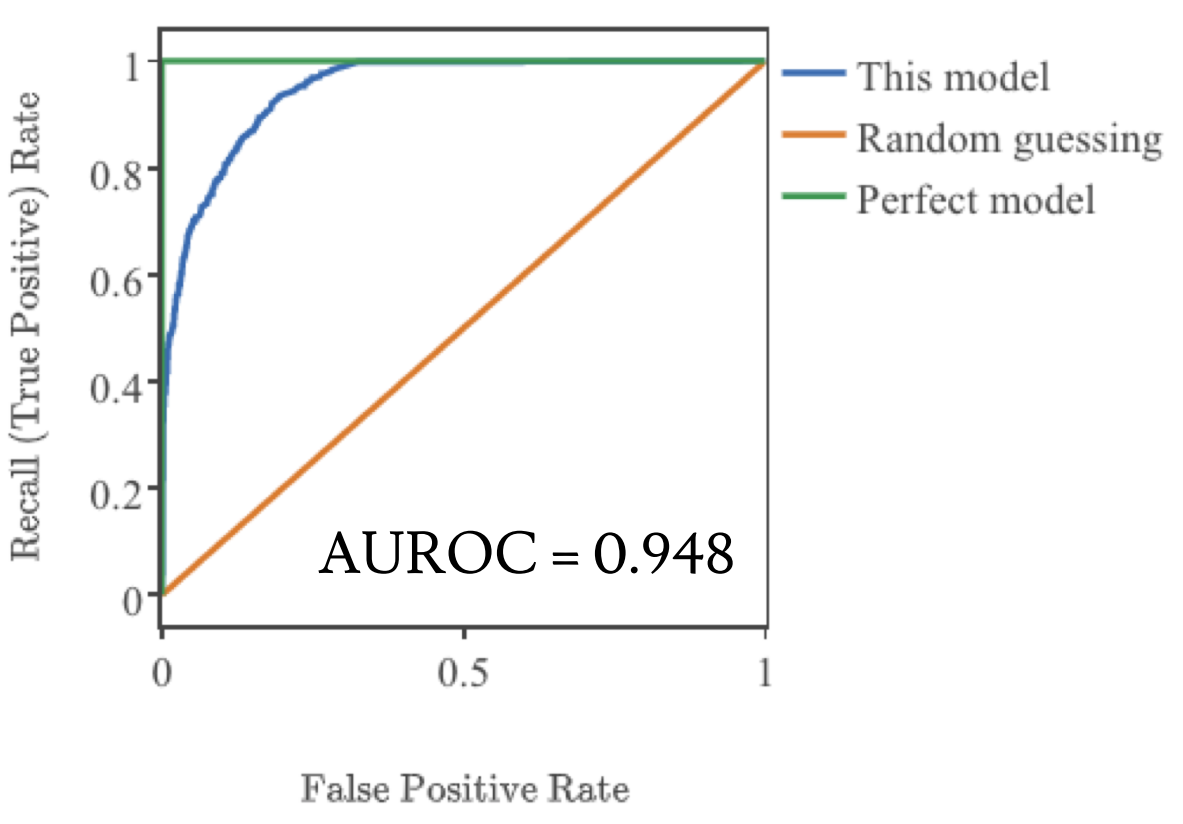
test_size = 0.1	N (predicted)	P (predicted)
N (actual)	TN = 2002	FP = 50
P (actual)	FN = 160	TP = 618

test_size = 0.25	N (predicted)	P (predicted)
N (actual)	TN = 5126	FP = 102
P (actual)	FN = 405	TP = 1440

test_size = 0.33	N (predicted)	P (predicted)
N (actual)	TN = 6665	FP = 173
P (actual)	FN = 583	TP = 1916

Binary classification on Citrination

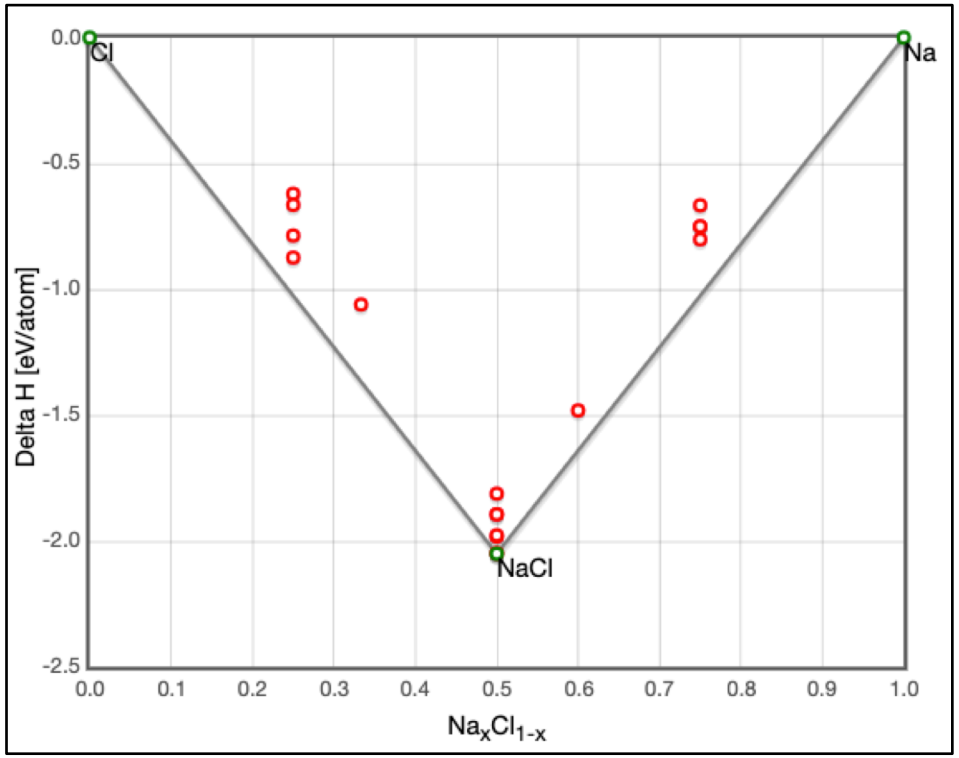




Random subset of data (Citrination limit = 4k rows)

Test case: Na, Cl

Actual: [1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1]



http://oqmd.org/materials/composition/NaCl

Predicted: [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]

compound	stable	unstable
Na	1	0
Na0.9	0.14	0.86
Na0.8	0.22	0.78
Na0.7	0.25	0.75
Na0.6	0.31	0.69
Na0.5	0.35	0.65
Na0.4	0.21	0.79
Na0.3	0.13	0.87
Na0.2	0.07	0.93
Na0.1	0.05	0.95
Cl	1	0

Conclusion: No better than random, worse than intuition