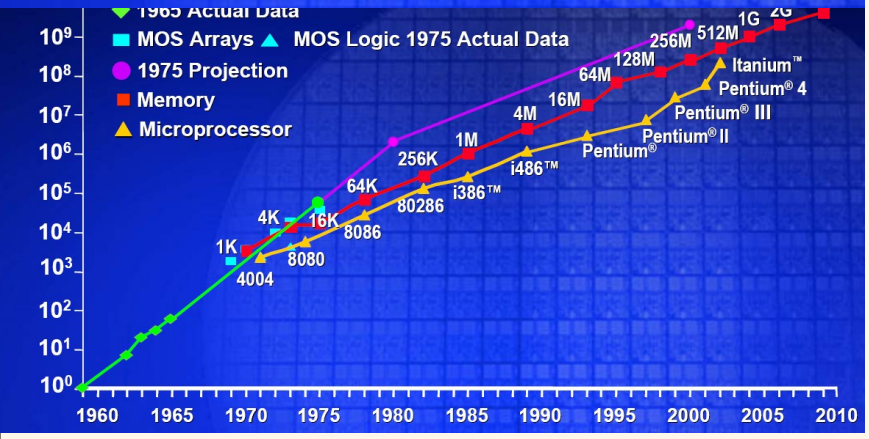
Machine Learning for the Materials Scientist

Chris Fischer*, Kevin Tibbetts, Gerbrand Ceder
Massachusetts Institute of Technology, Cambridge, MA

Dane Morgan
University of Wisconsin, Madison, WI

Motivation: materials design through calculation

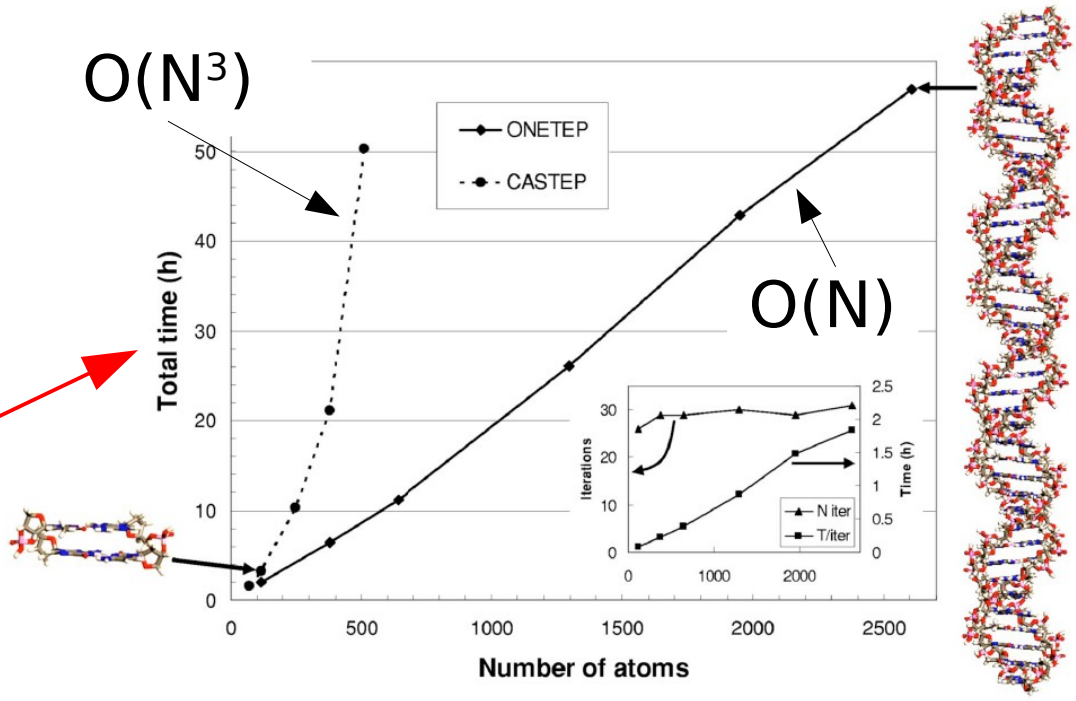


Moore, G. ISSCC 2003 slides (http://www.intel.com)

computing power: *exponential* scaling with time

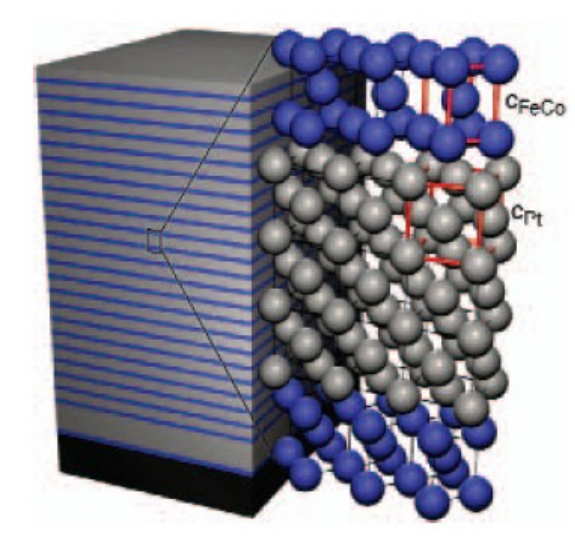
Run-time: *polynomial* scaling with number of atoms

Skylaris, C. et. al. J. Phys. Chem. **122**, 084119 (2005)

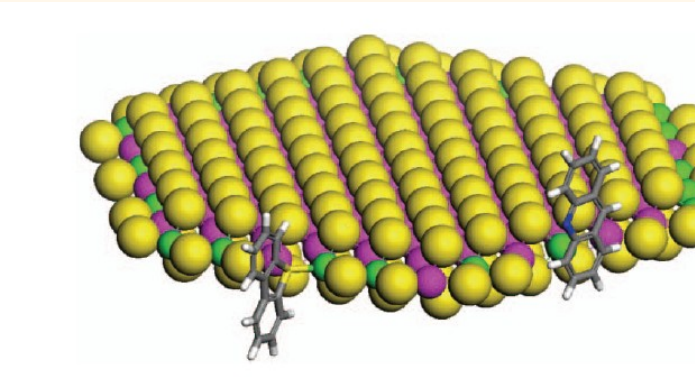


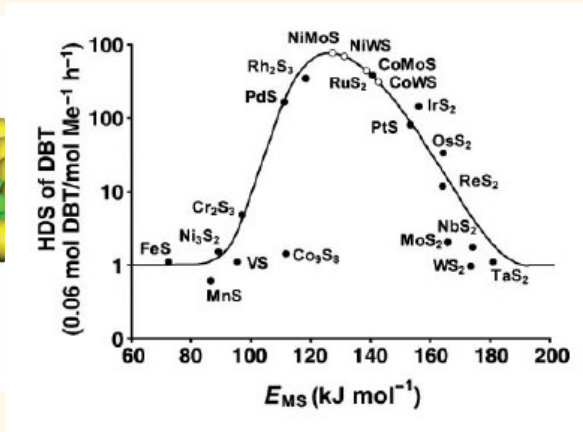
DFT as a predictive tool

Burkett, T. et. al. Phys. Rev. Lett. **93** (2004)

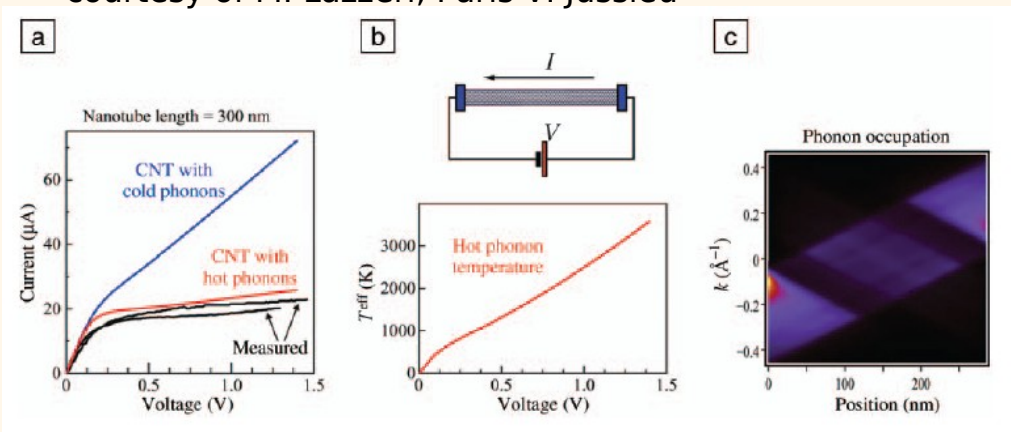


Norskov, J. et. al. MRS Bulletin **31** (2006)

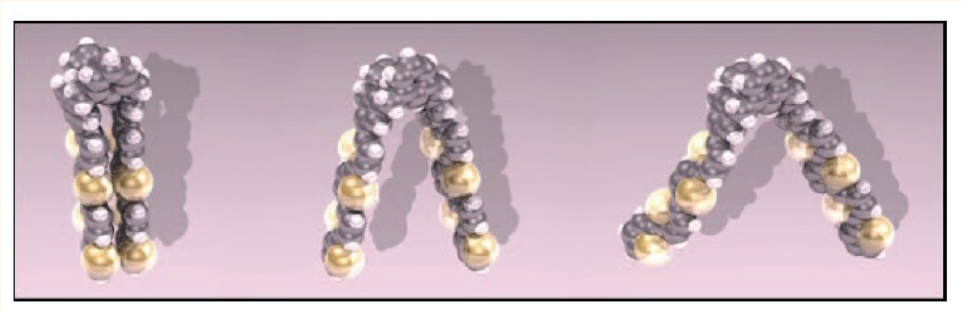




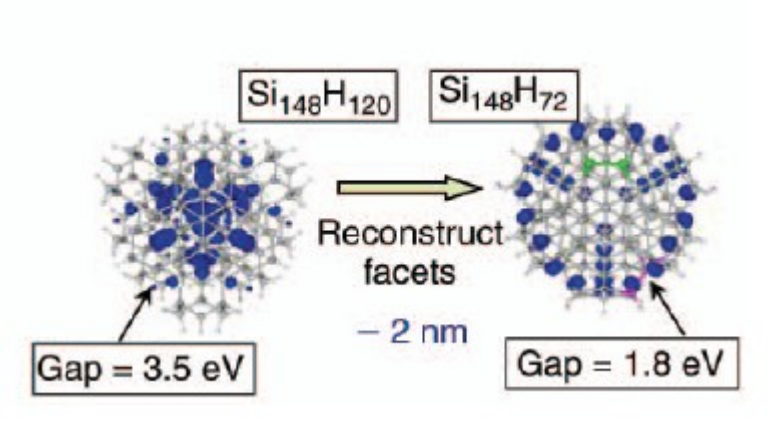
Marzari, N. MRS Bulletin **31** (2006)
courtesy of M. Lazzeri, Paris VI Jussieu



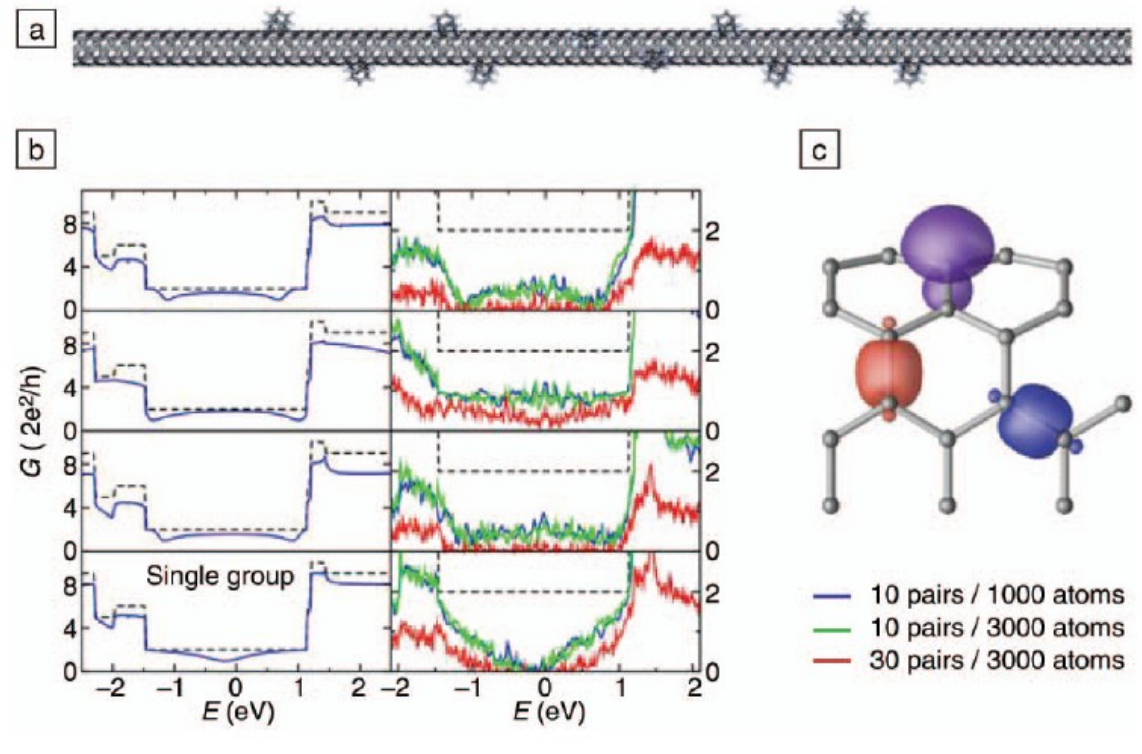
Marzari, N. MRS Bulletin **31** (2006)
courtesy of D. Scherlis, MIT



Calculating properties of realistic nanostructures *ab initio*

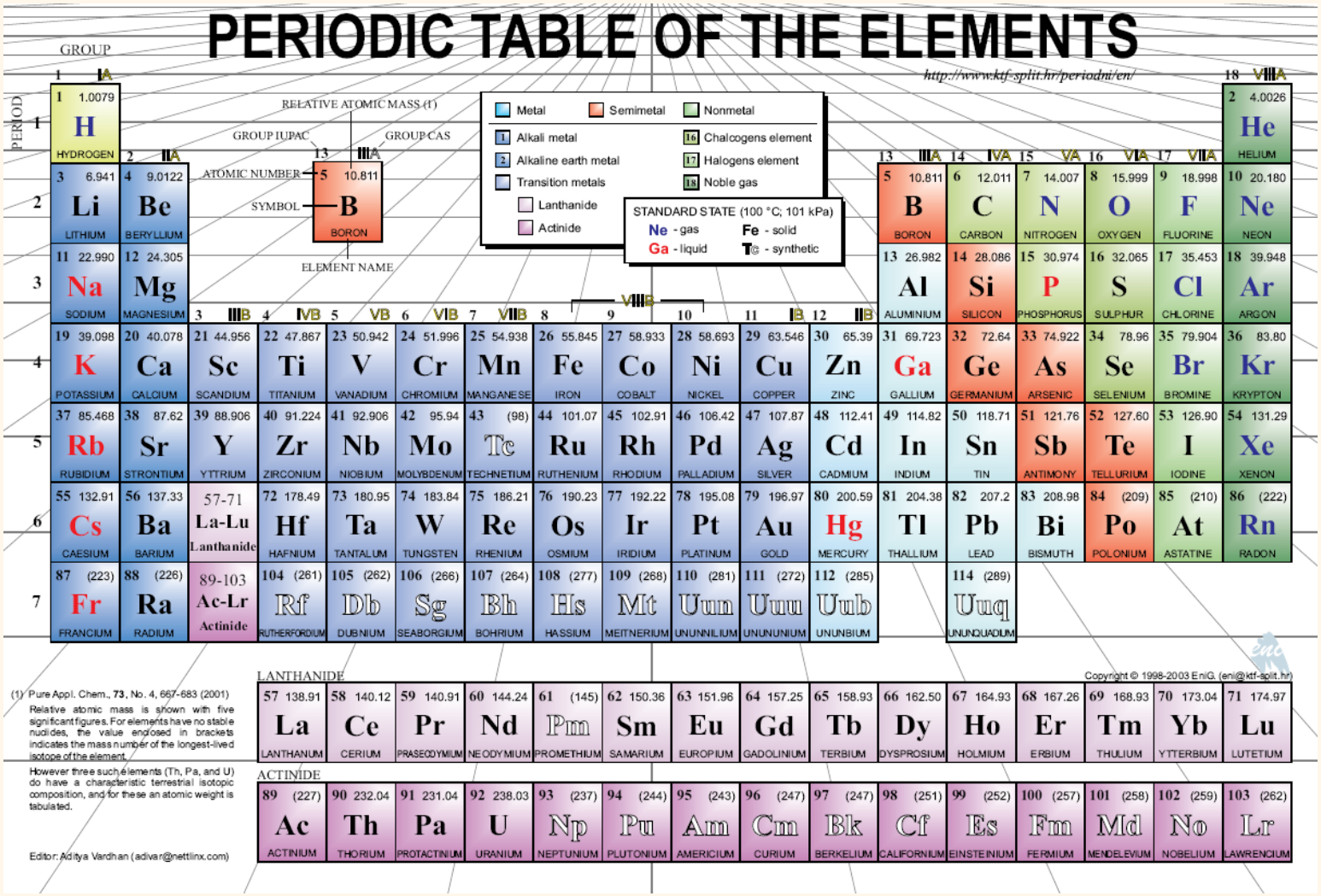


Galli, G. University of California, Davis



Lee, Y. S. et al. PRL **95** 076804 (2005)

Which combinations yield the optimal material ?



Outline

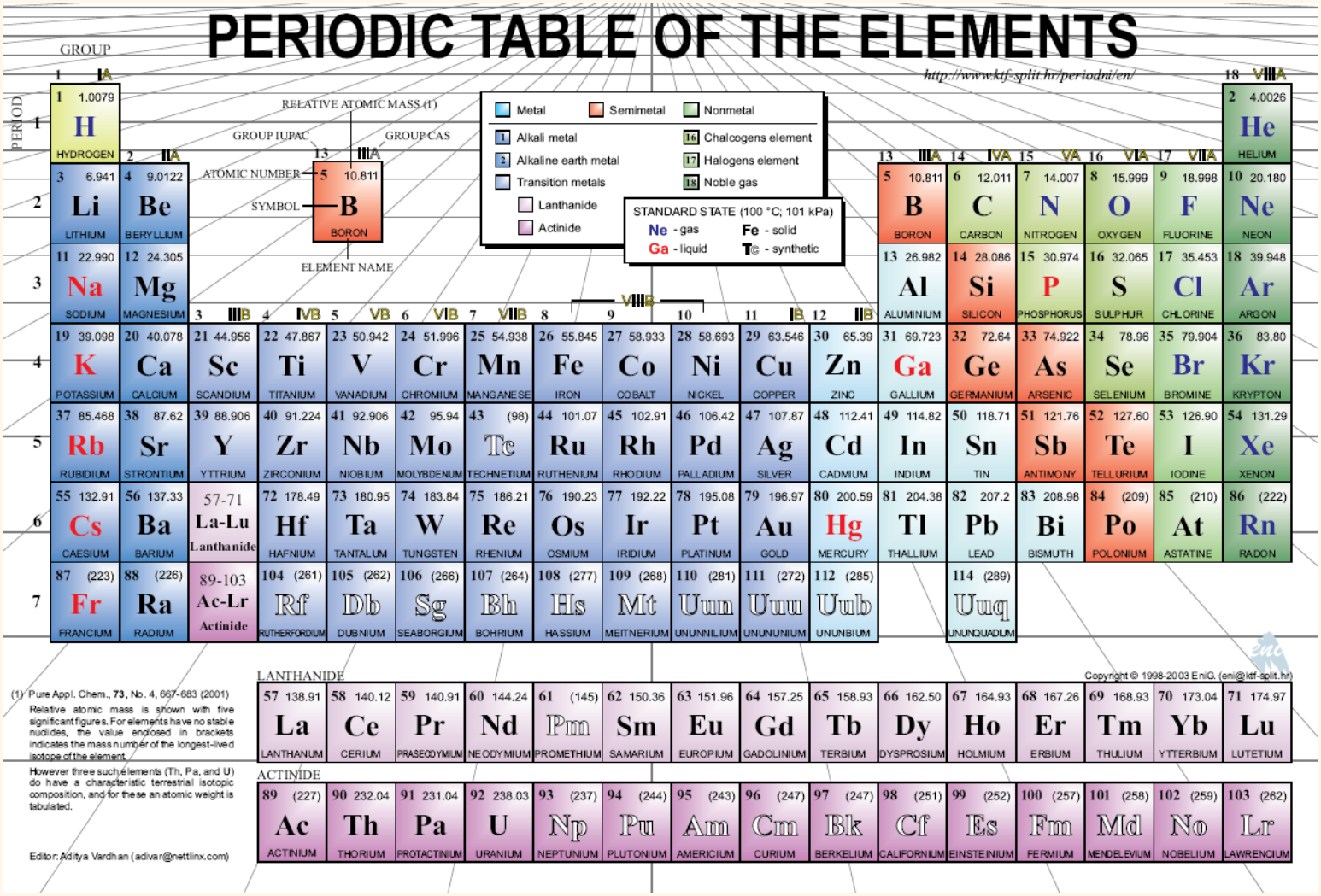
Machine learning in Computational Materials Design

Searching for Structure: combining historical information with Density Functional Theory

Data Mining the High-Throughput engine

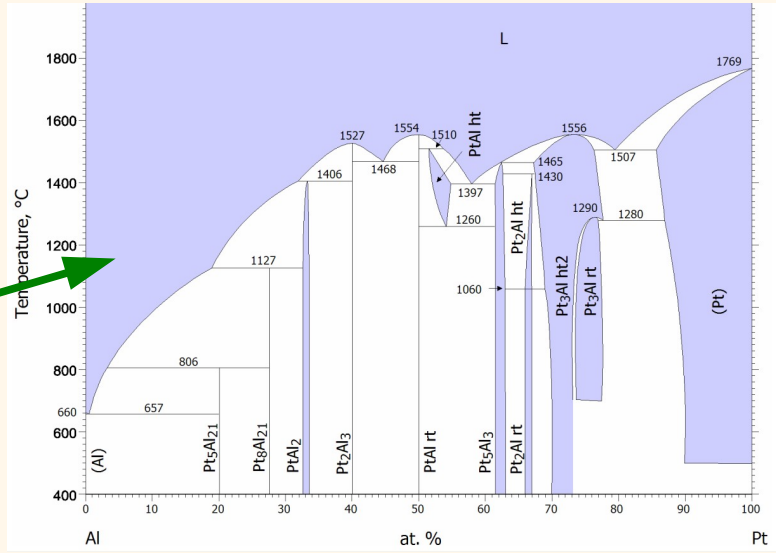
wrap-up

Which combinations yield the optimal material ?



Motivation: searching for new materials

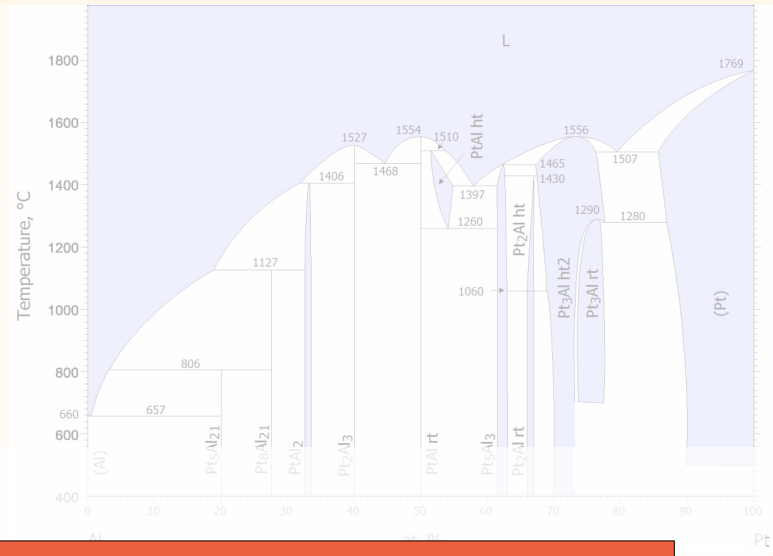
```
for i in (relevant chemistries) {
    ...
    ...

    getStablePhases(i);

    ...
    ...

    calculateProperty(i);
    i = nextChemistry();
}
```



Depends on which phases are stable **and** their *structure*

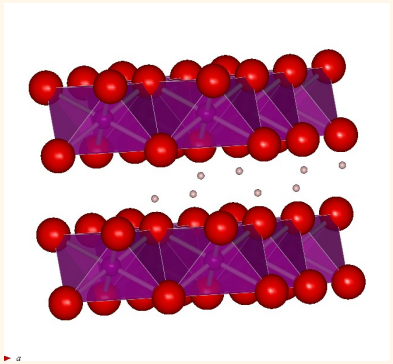
Motivation: materials by design

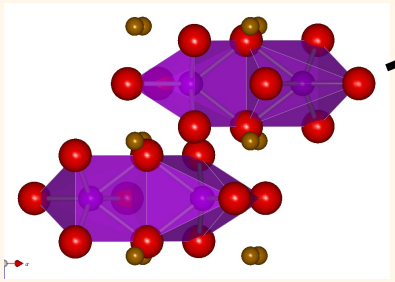
```
for i in (relevant chemistries) {
    ...
    ...
    getStablePhases(i);
    ...
    ...
    calculateProperty(i);
    i = nextChemistry();
}
```



Machine Learning needed here !!

The need for machine learning





abinit.org

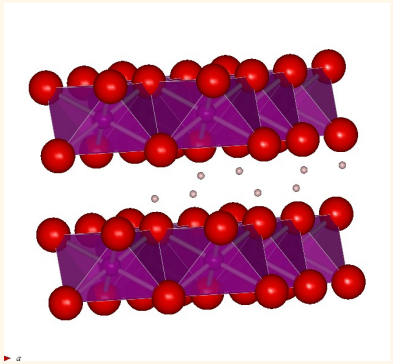
b-initio VASP ackage imulation tenna

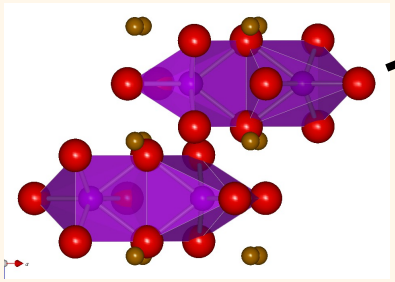
DFT Code

Material Property Predictions

Doesn't know what to calculate next

The need for machine learning





abinit.org

b-initio VASP ackage imulation tenna

DFT Code

Material Property Predictions

Machine Learning Framework

Database of Computed and Experimental results

Computational Materials Design poised for impact

'Commodity' computational resources

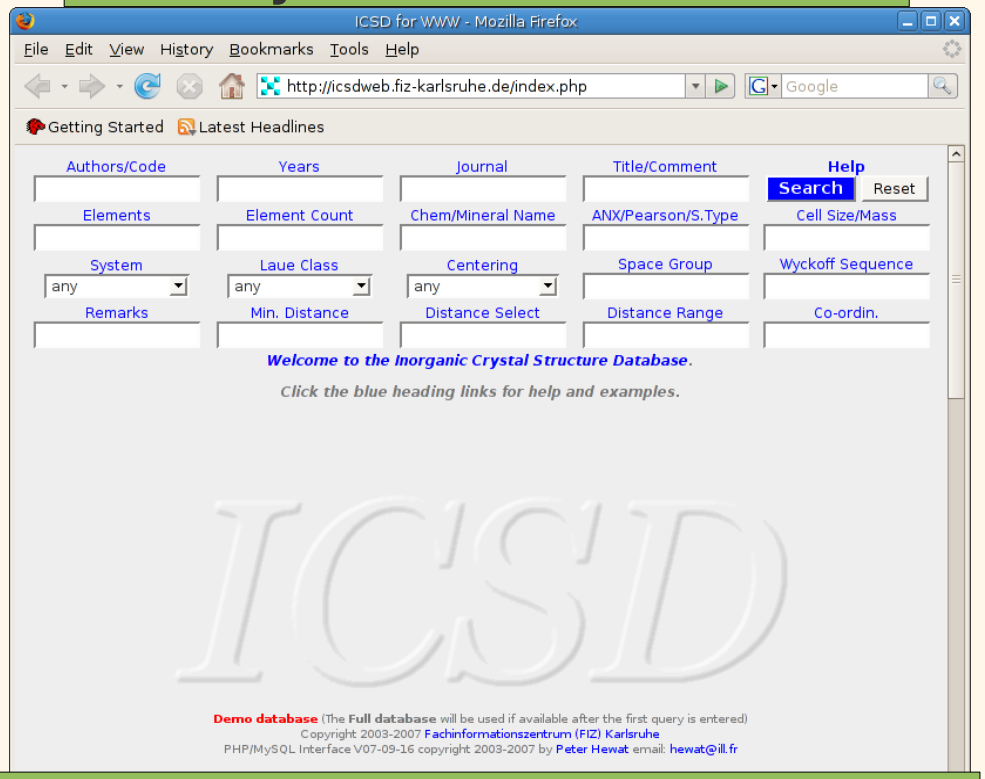
+

Open source electronic structure software

~$200-250k capital investment → Computing budget ~50k compounds/year

Computational Materials Design poised for impact

ICSD: World's Largest database of inorganic crystal structures



First Entry: 1913
# of entries: 100,243
usable compounds: 29,962

Computing budget
~50k compounds/year

The structure search problem

```
for i in (relevant chemistries) {
    ...
    ...
    getStablePhases(i);
    ...
    ...
    calculateProperty(i);
    i = nextChemistry();
}
```

Where do we put the atoms if no experimental structure is known ??

Depends on which phases are stable **and** their *structure*

Strategies to search for structure

Coordinate Search:
Optimize energy (or free energy) directly in the space of atomic coordinates

Heuristic Rules or Chemical Intuition

Methods to search for structure

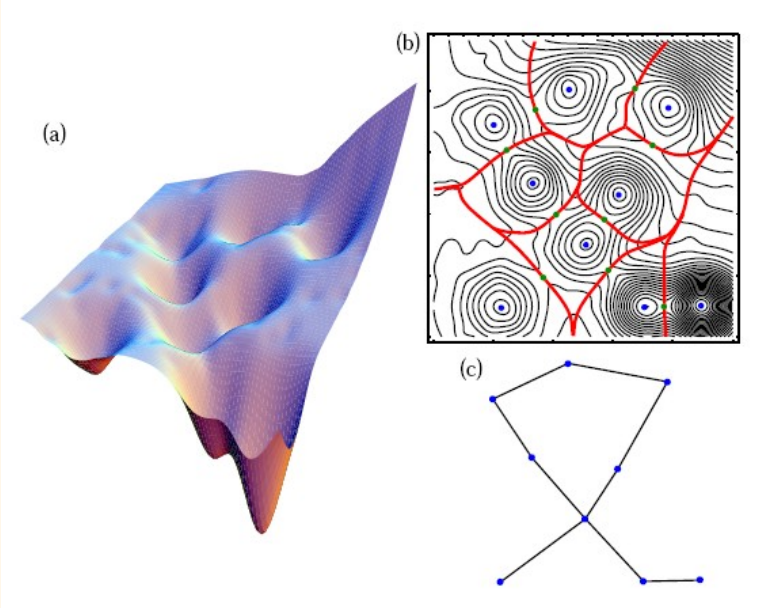
Coordinate Search:

Optimize energy (or free energy) directly in the space of atomic coordinates

$$GroundState \equiv arg\ min_{\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N} E\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right)$$

# of dimensions = 3N – 3 + dim(a,b,c,α,β,γ)

complex energy landscape



Doye, J. PRL, **88**, 238701, (2002)

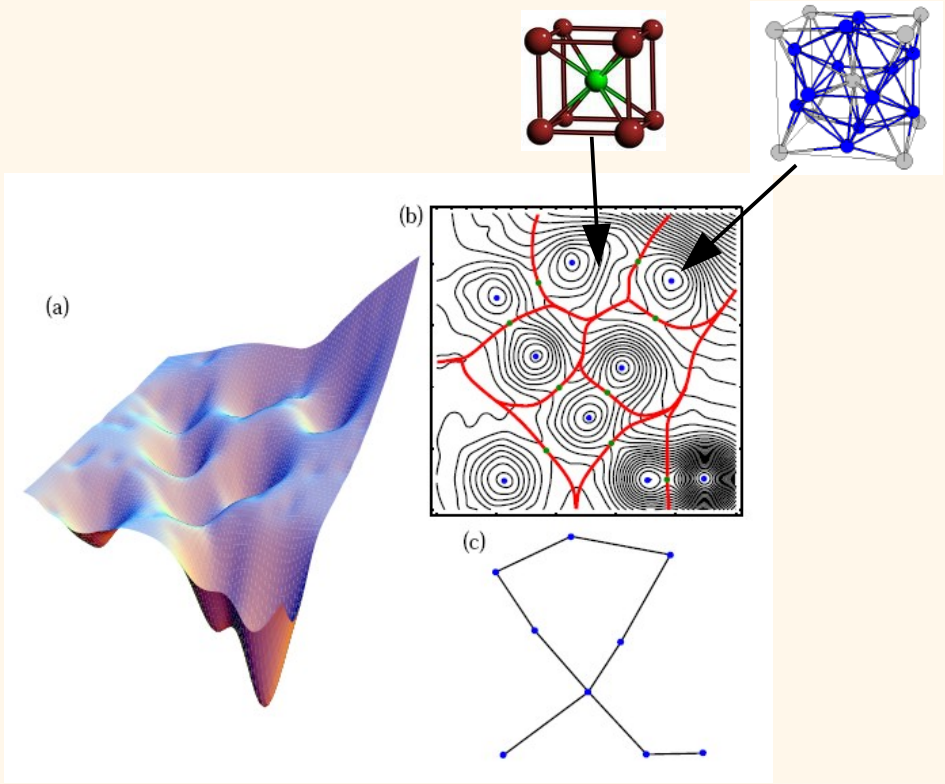
Methods to search for structure

Coordinate Search:

Optimize energy (or free energy) directly in the space of atomic coordinates

$$GroundState \equiv arg\ min_{\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N} E\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right)$$

# of dimensions = 3N – 3 + dim(a,b,c,α,β,γ)



Doye, J. PRL, **88**, 238701, (2002)

Proposed Solutions

Calculate energy of a finite set of structure prototypes

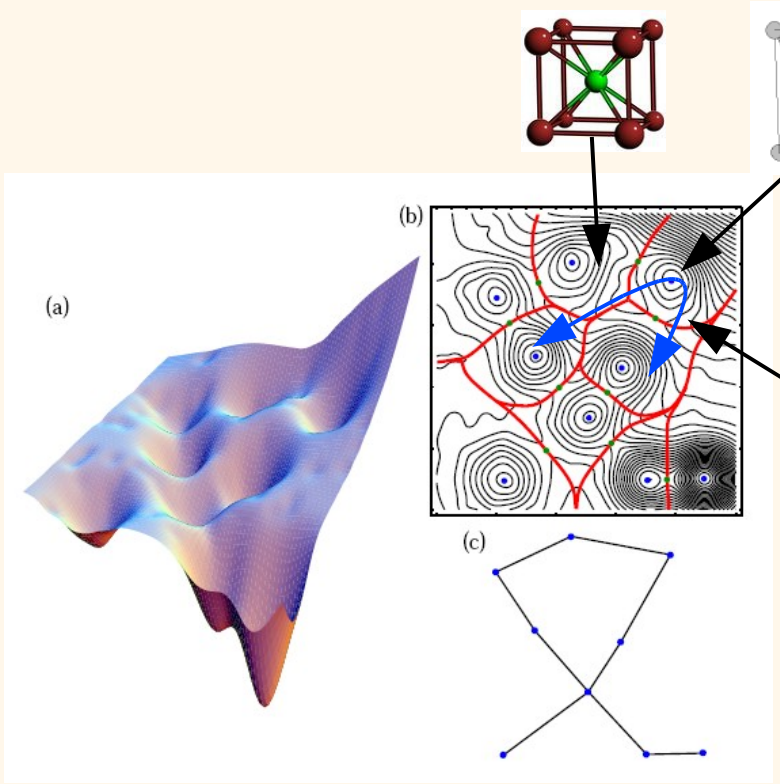
Methods to search for structure

Coordinate Search:

Optimize energy (or free energy) directly in the space of atomic coordinates

$$GroundState \equiv arg\ min_{\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N} E\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right)$$

# of dimensions = 3N – 3 + dim(a,b,c,α,β,γ)



Doye, J. PRL, **88**, 238701, (2002)

Proposed Solutions

Calculate energy of a finite set of structure prototypes

Use a stochastic optimization procedure (hop from basin to basin)

e.g., Simulated Annealing
Genetic Algorithms

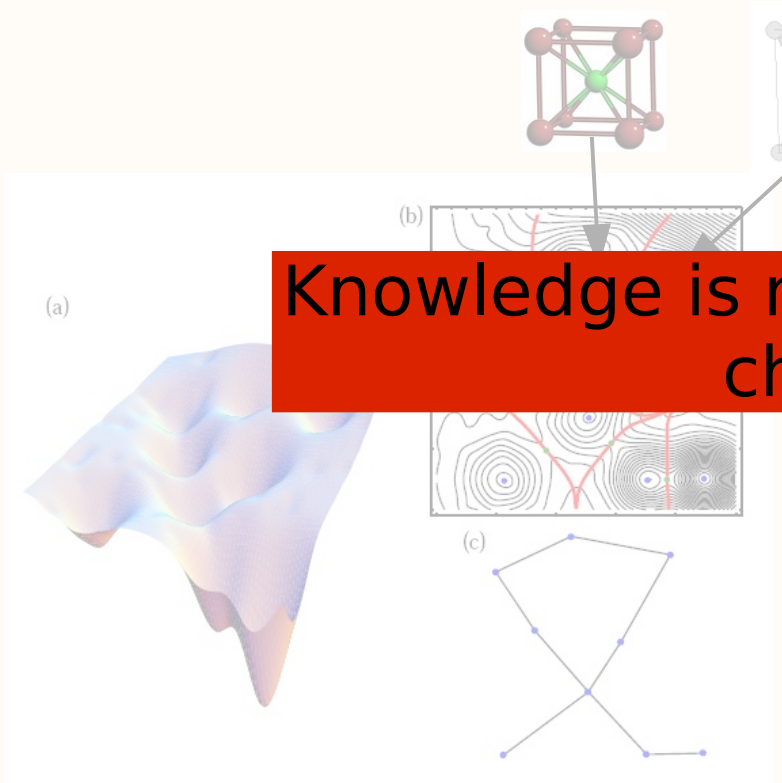
Methods to search for structure

Coordinate Search:

Optimize energy (or free energy) directly in the space of atomic coordinates

$$GroundState \equiv arg\ min_{\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N} E\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right)$$

# of dimensions = 3N – 3 + dim(a,b,c,α,β,γ)



Doye, J. PRL, **88**, 238701, (2002)

Knowledge is not *transferred* across chemistries

Proposed Solutions

Calculate energy of a finite set of structure prototypes

Use a stochastic optimization procedure (hop from basin to basin)

e.g., Simulated Annealing
Genetic Algorithms

Methods to search for structure

Heuristic Rules

Use previous experiments to *suggest* what to calculate

How ?

Identify a set of simple parameters based on alloy constituents

1932: Pauling electronegativity	$\Delta\chi$
1935: Laves & Witte	$\Delta r_{A,B}$
1926,1936-7: Hume-Rothery, Mott & Jones	$n_{at}^{(e)}$
1976: Miedema	$\Delta n_{ws}^{(e)}$

Methods to search for structure

Heuristic Rules

Plot stable structures in space of parameters

1986: Pettifor

1983: Villars

$\Delta\chi$ $\Delta r_{A,B}$ $n_{at}^{(e)}$

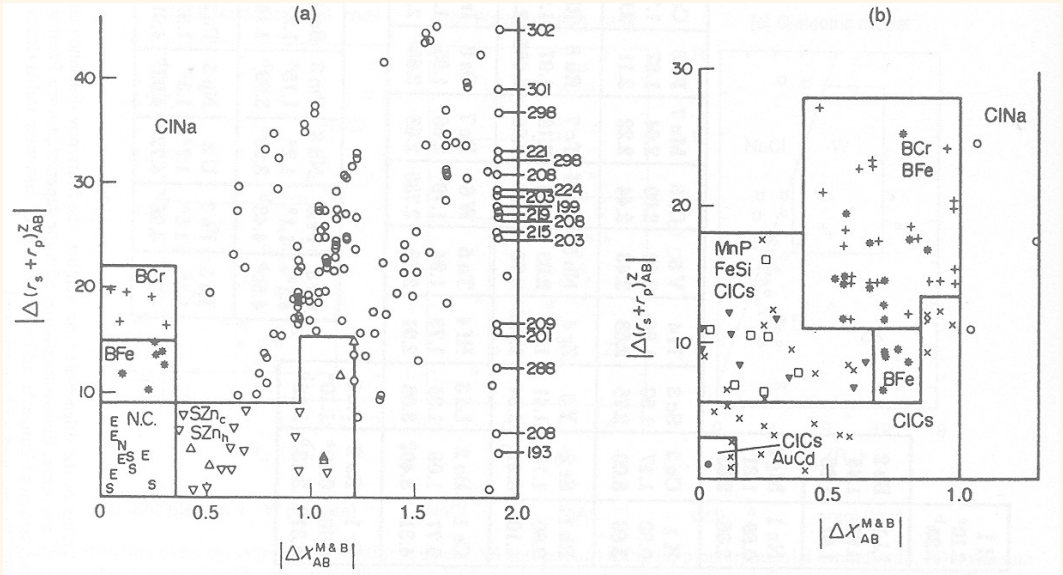
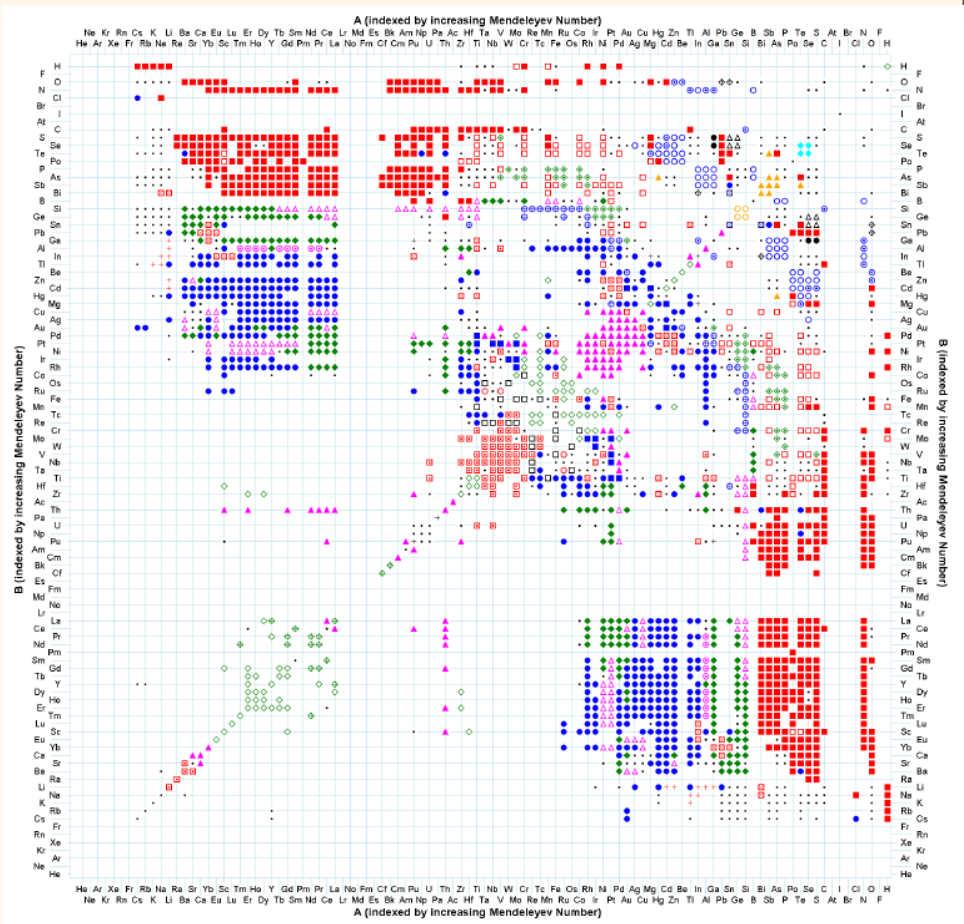


Figure 5. The Villars maps for AB compounds corresponding to the average electron-per-atom ratio of $\bar{N}=4$ and 6.5 respectively (From Villars, 1983. Reproduced with permission). (a) $\bar{N}=4$; (b) $\bar{N}=6.5$



Methods to search for structure

Heuristic Rules

Plot stable structures in space of parameters

1986: Pettifor

1983: Villars

$\Delta\chi$ $\Delta r_{A,B}$ $n_{at}^{(e)}$

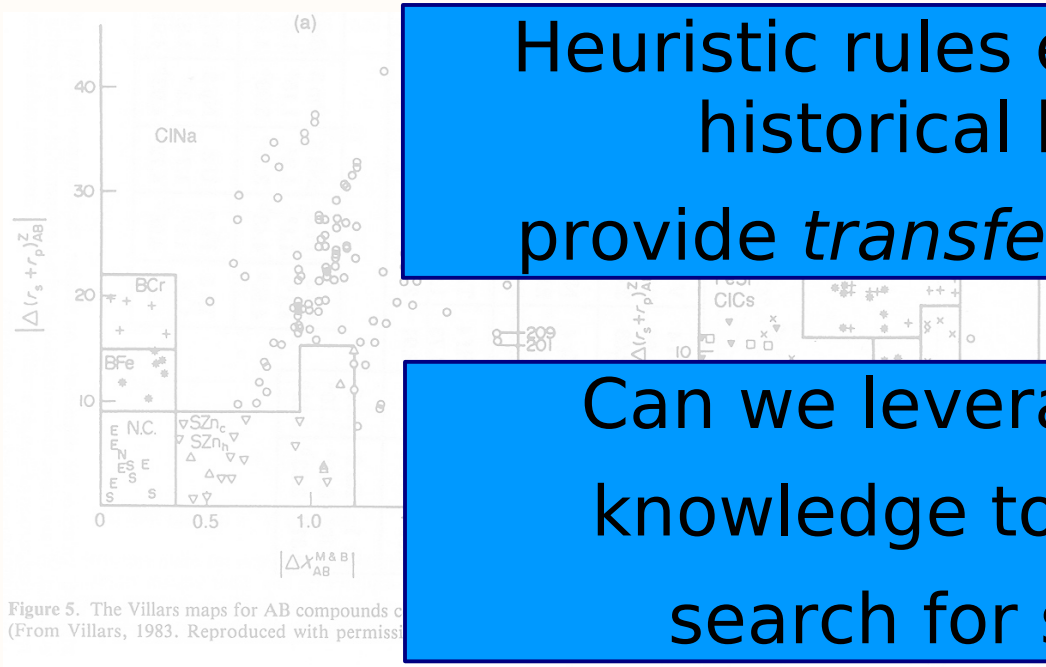
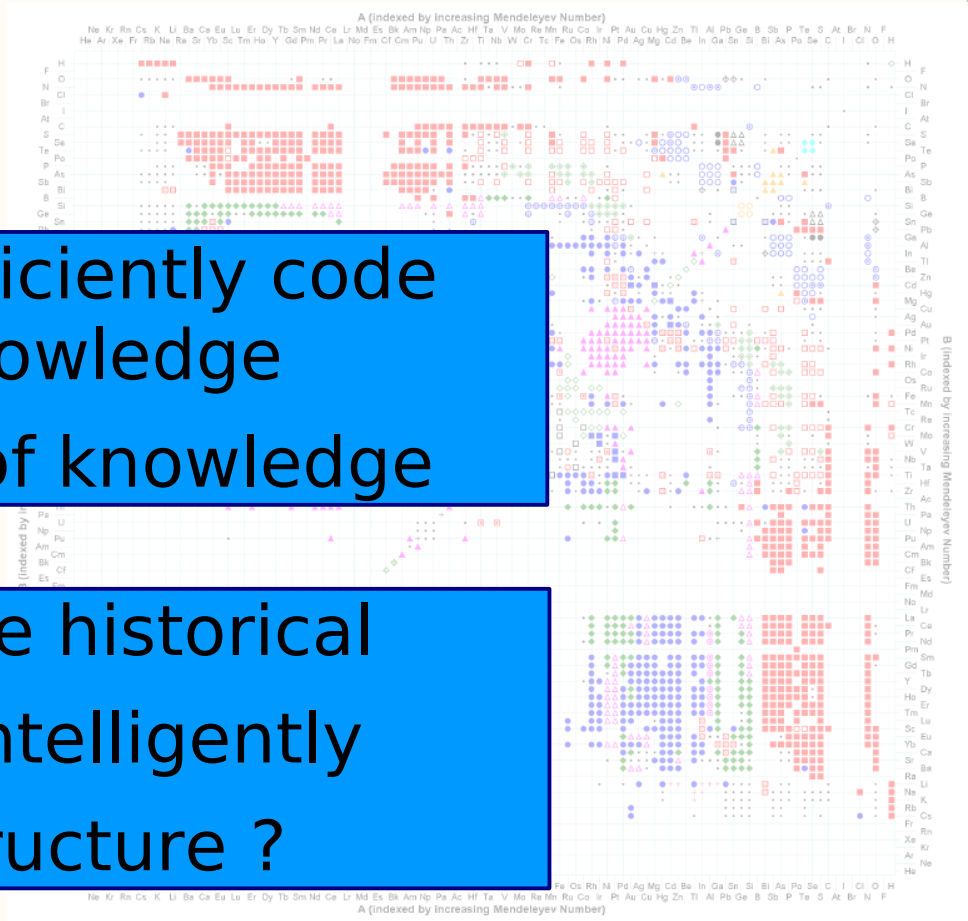


Figure 5. The Villars maps for AB compounds c
(From Villars, 1983. Reproduced with permissi



Heuristic rules efficiently code historical knowledge

provide *transfer* of knowledge

Can we leverage historical knowledge to intelligently search for structure ?

description of knowledge base

Knowledge Base

Experimental Data

Pauling File binaries edition (Villars, P. et. al. J. of Alloys and Compounds, (2004))

1335 binary alloys

3975 non-unique compounds

4263 compounds total

alloys not containing elements:
He, B, C, N, O, F, Ne, Si, P, S, Cl, Ar, As, Se, Br, Kr, Te, I, Xe, At, Rn

Machine learning framework: concepts

$$\boldsymbol{x} = \left(x_A, x_0, \ldots, x_{\frac{1}{2}}, \ldots, x_B \right)$$

Low temperature state of alloy

$$Data \equiv \left\{ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \right\}$$

database of N binary alloys

Machine learning framework: concepts

$$\boldsymbol{x} = \left(x_A, x_0, \ldots, x_{\frac{1}{2}}, \ldots, x_B \right)$$

Low temperature state of alloy

$$Data \equiv \left\{ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \right\}$$

database of N binary alloys

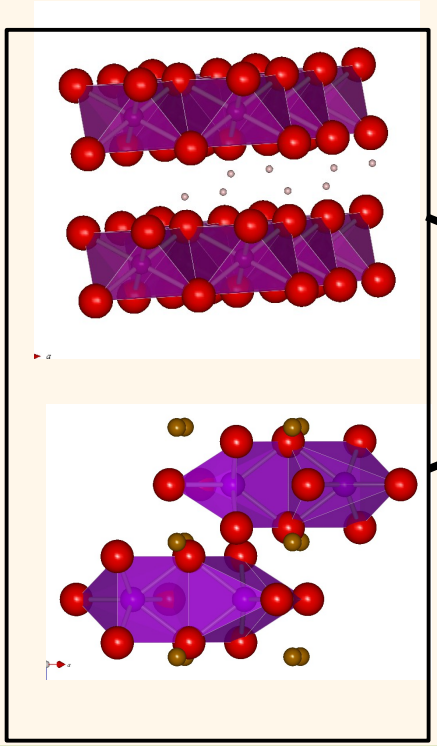
$$p(\boldsymbol{x})$$

Probability of low temperature state (fitted to data)

$$p(\tilde{x} | \boldsymbol{e})$$

Probability of low temperature state conditioned on evidence 'e'

how to use the machine learning framework



Set of likely structure candidates

abinit.org

b-initio Vienna Package for Simulation

DFT Code

Material Property Predictions

Machine Learning Framework

$p(\tilde{x}|\boldsymbol{e})$

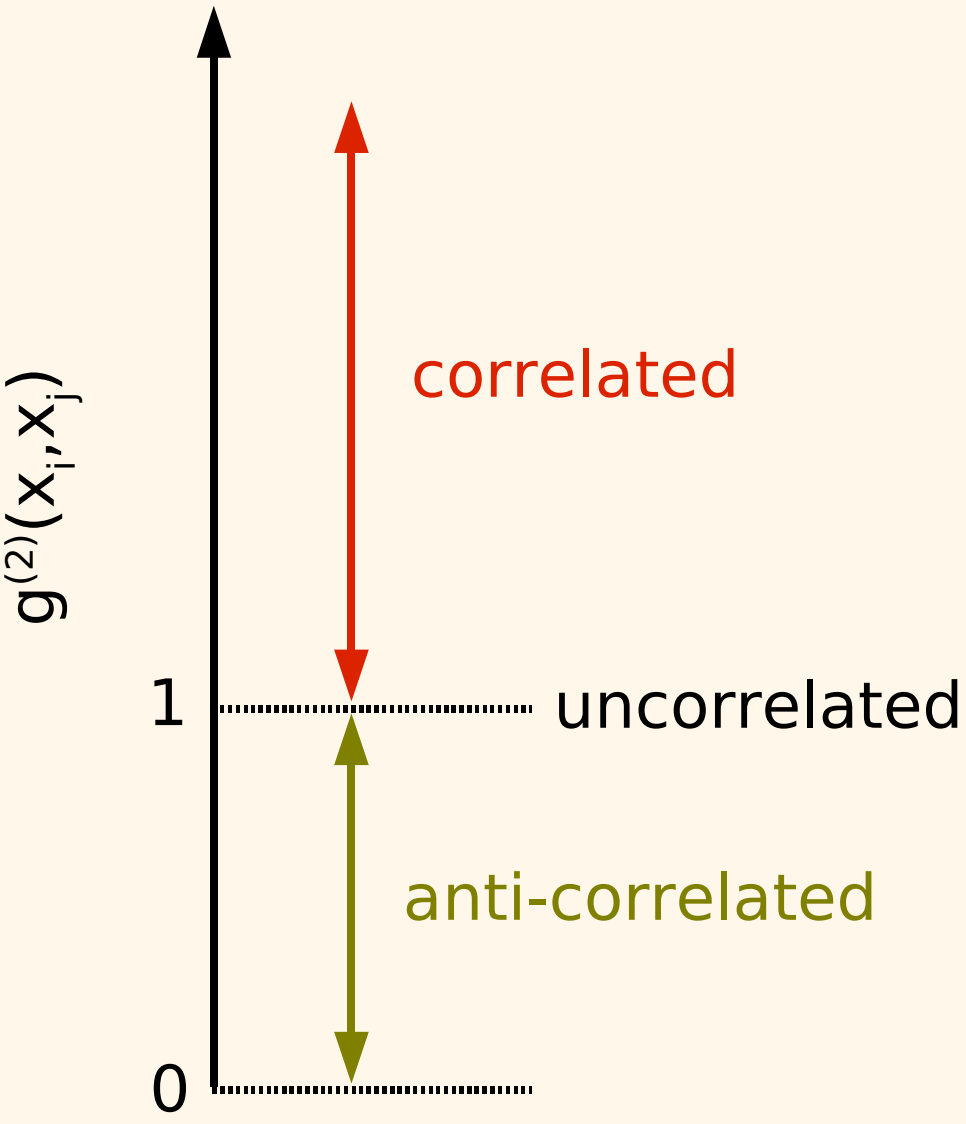
Database of Computed and Experimental results

Preliminaries and open questions

Are probabilities consistent with physical intuition ?

Do probabilities encode the physics of structure stability ?

quantifying correlation in probabilistic framework



probability that *both* structures occur in *same* system estimated from database

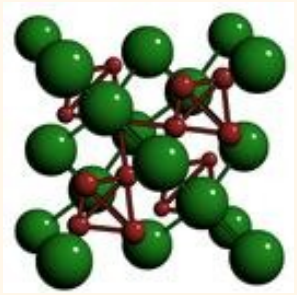
Pair Cumulant

$$g_{ij}(x_i, x_j) = \frac{p(x_i, x_j)}{p(x_i)\,p(x_j)}$$

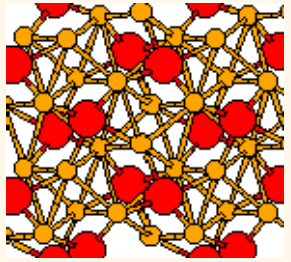
probability that only x_i occurs

Do probabilities embody real physical effects ?

Compounds stabilized by “size” effect:



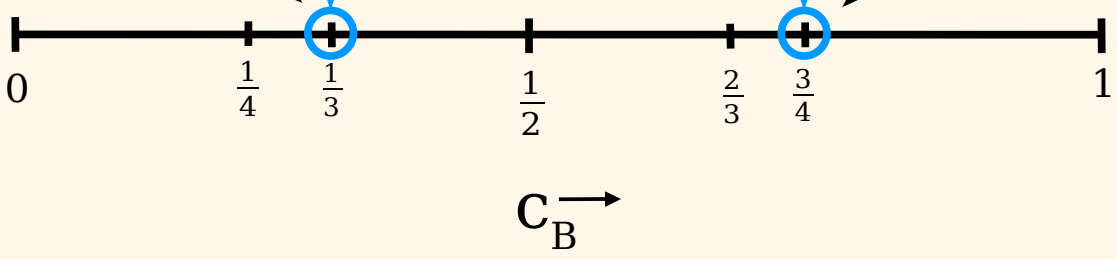
$$g_{ij}(x_i, x_j) = \frac{p(x_i, x_j)}{p(x_i)\,p(x_j)}$$



$MgCu_2$

Fe_3C

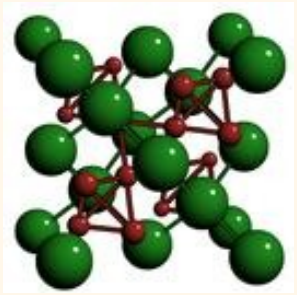
8.48



Data from Pauling File, Binaries Edition

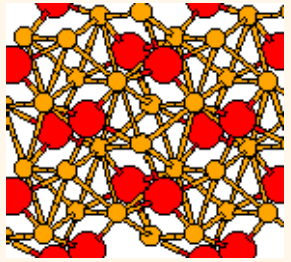
Do probabilities embody real physical effects ?

Compounds stabilized by “size” effect:

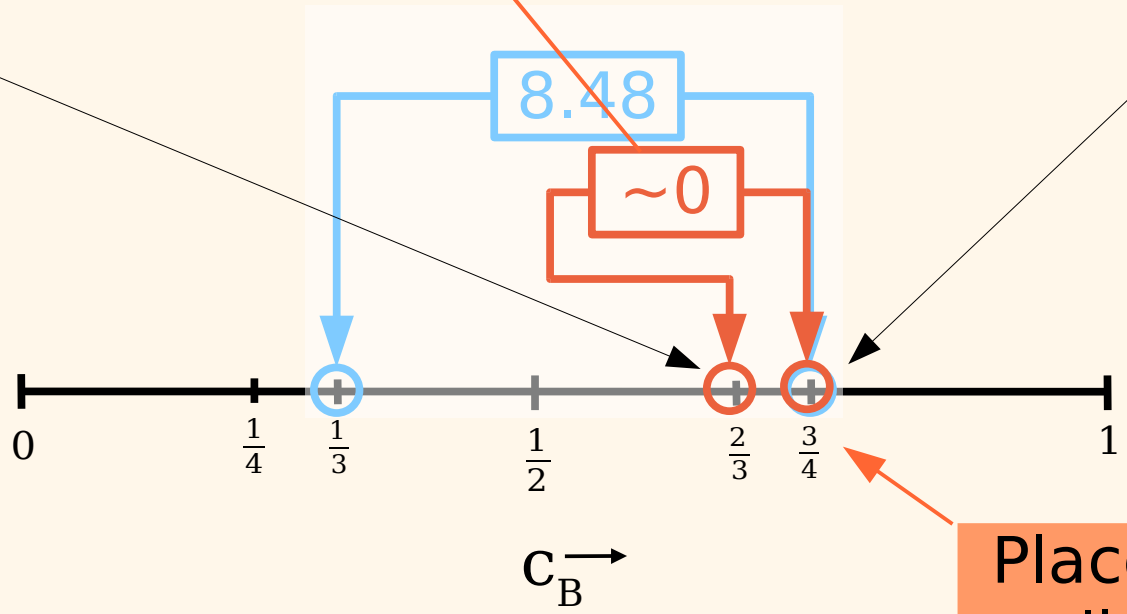


$MgCu_2$

$$g_{ij}(x_i, x_j) = \frac{p(x_i, x_j)}{p(x_i)\,p(x_j)}$$



Fe_3C



Places 'small' atoms on 'large' atom sites

Data from Pauling File, Binaries Edition

G. Ceder

how probabilities represent physics of mixing: more interesting correlations

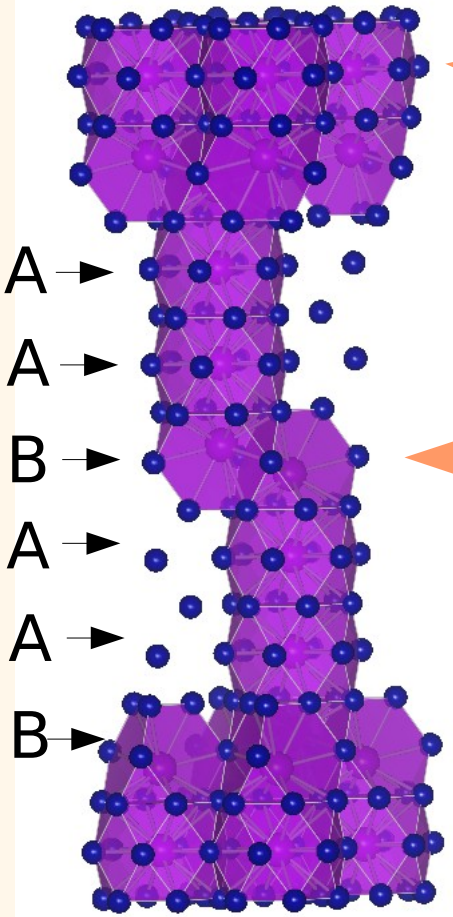
Gd_2Co_7

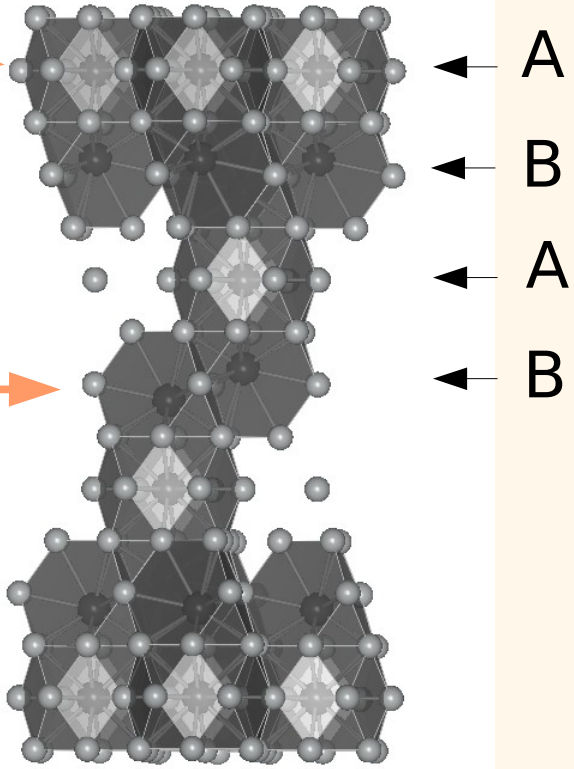
AABAAB... stacking

$g_{ij}(x_i, x_j)=54$

$PuNi_3$

ABAB... stacking





Both structures share the same local environments

Structure correlation observations

Correlation factors are probabilistic analogue of heuristic rules

No *explicit* reference to physics. Physics is *embedded* in experimental data

Information theory for structure stability

Suppose I know Fe_3C forms @ c = ¾, how does this change prediction @ c = ½ ?

How much information is carried by knowledge of structure ?

Mutual Information

$$I_{i,j} = \sum_{x_i, x_j} p(x_i, x_j) \log \left(\frac{p(x_i, x_j)}{p(x_i) p(x_j)} \right)$$

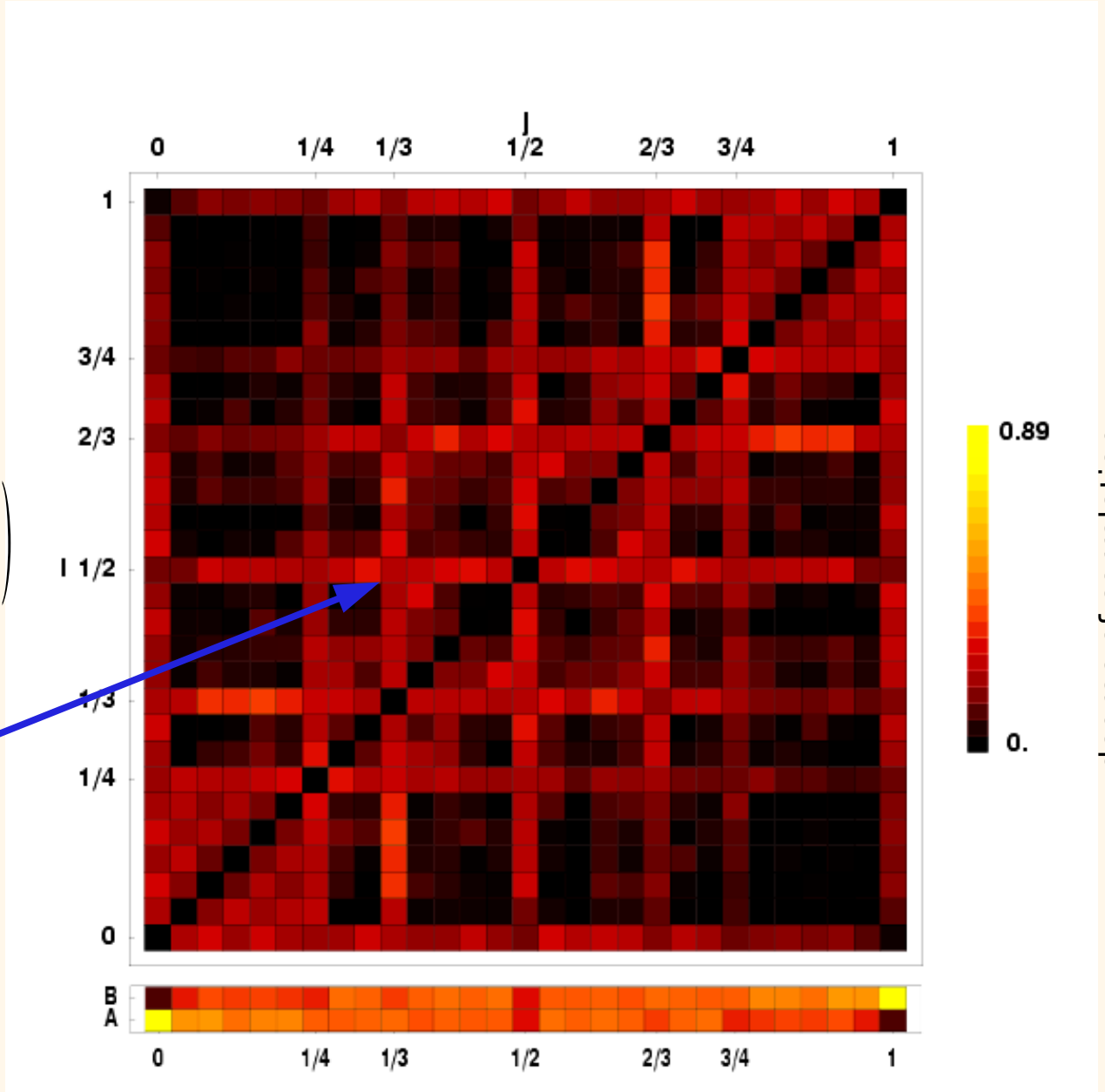
$$I_{i,j} = \left\langle \log \left[g_{ij}(x_i, x_j) \right] \right\rangle$$

Information theory for structure stability

Each element of matrix is correlation between X_i and X_j

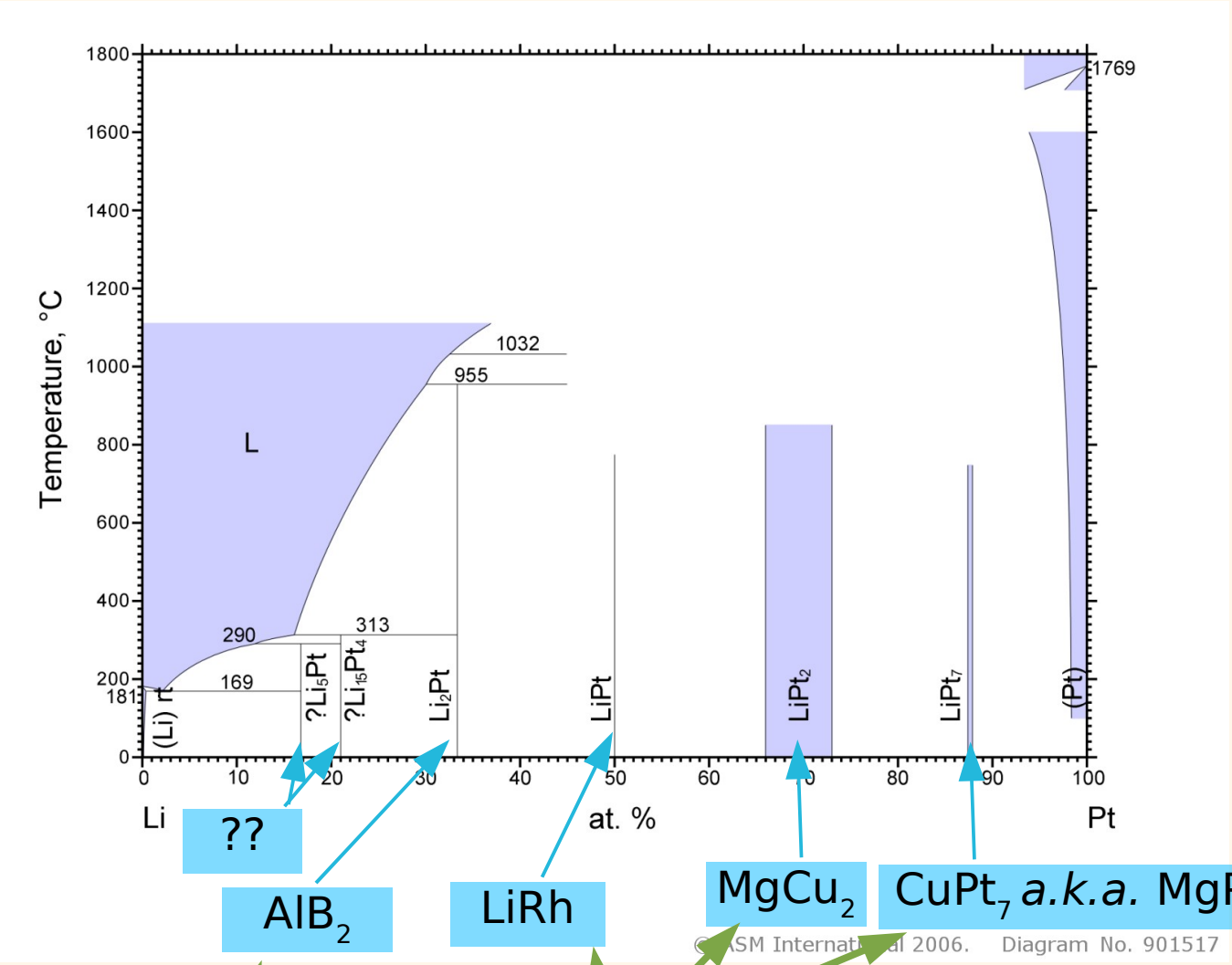
$$I_{i,j} = \sum_{x_i, x_j} p(x_i, x_j) \log \left(\frac{p(x_i, x_j)}{p(x_i) p(x_j)} \right)$$

e.g.,
X_i="AB prototype"
and
X_j="A_2B prototype"



Prediction and validation in Li-Pt

Predicting structures in Li-Pt

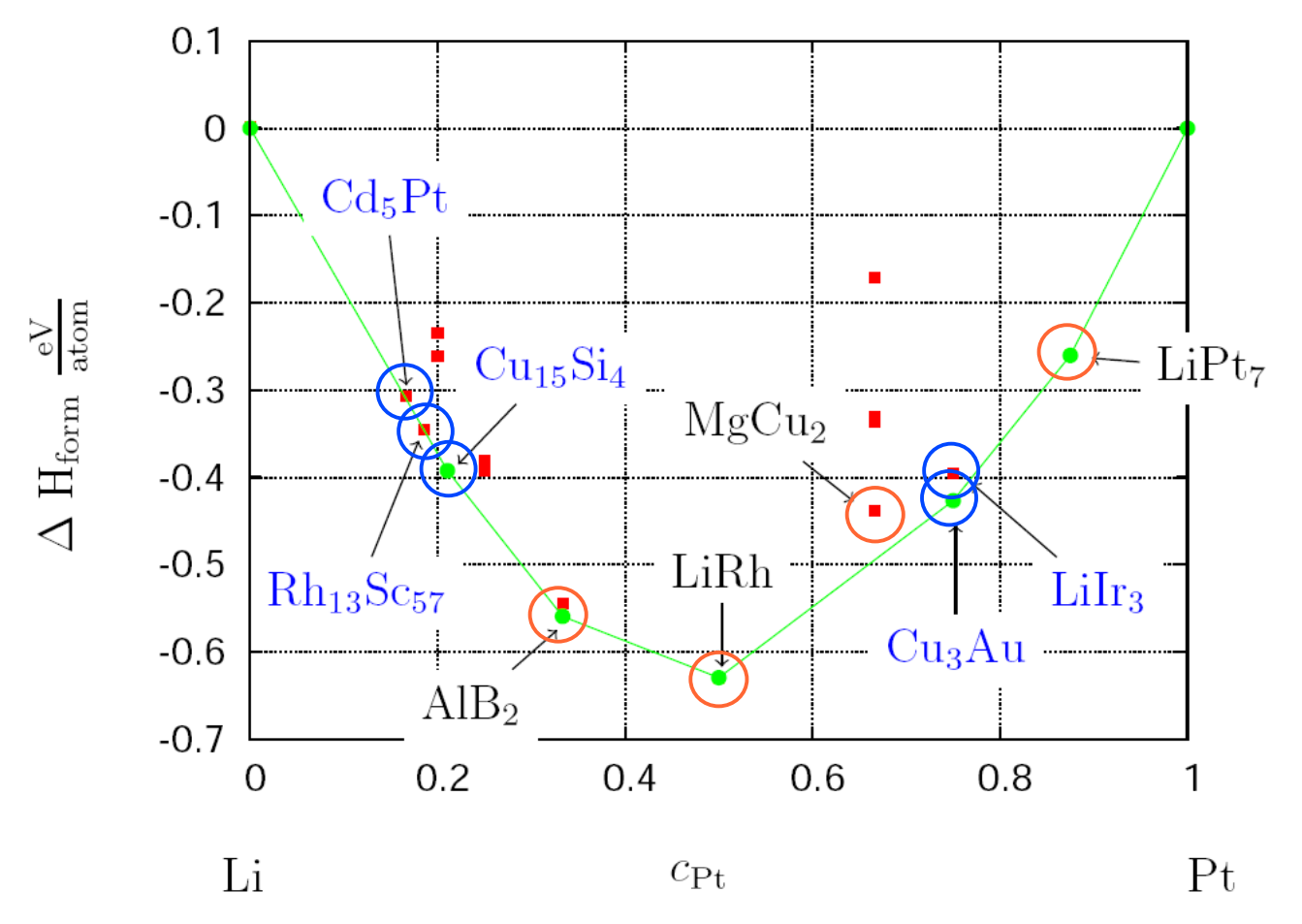


Use these as conditioning evidence for: $p(\tilde{x}|\boldsymbol{e})$

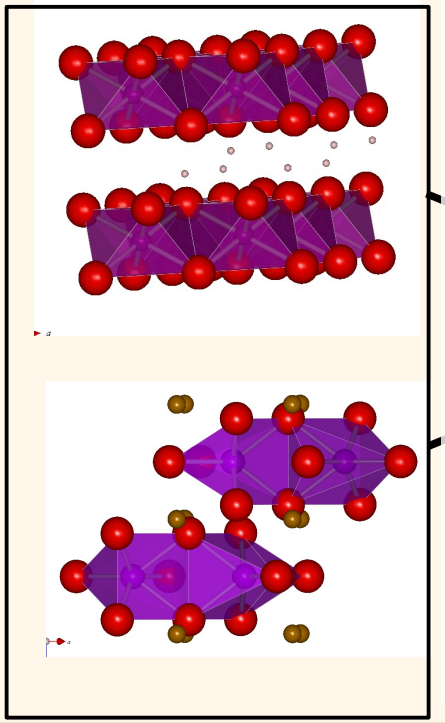
Predicting structures in Li-Pt

Known phases

Suggested phases



cross validation to evaluate performance



Success of method depends on how short this list is

Material Property redictions

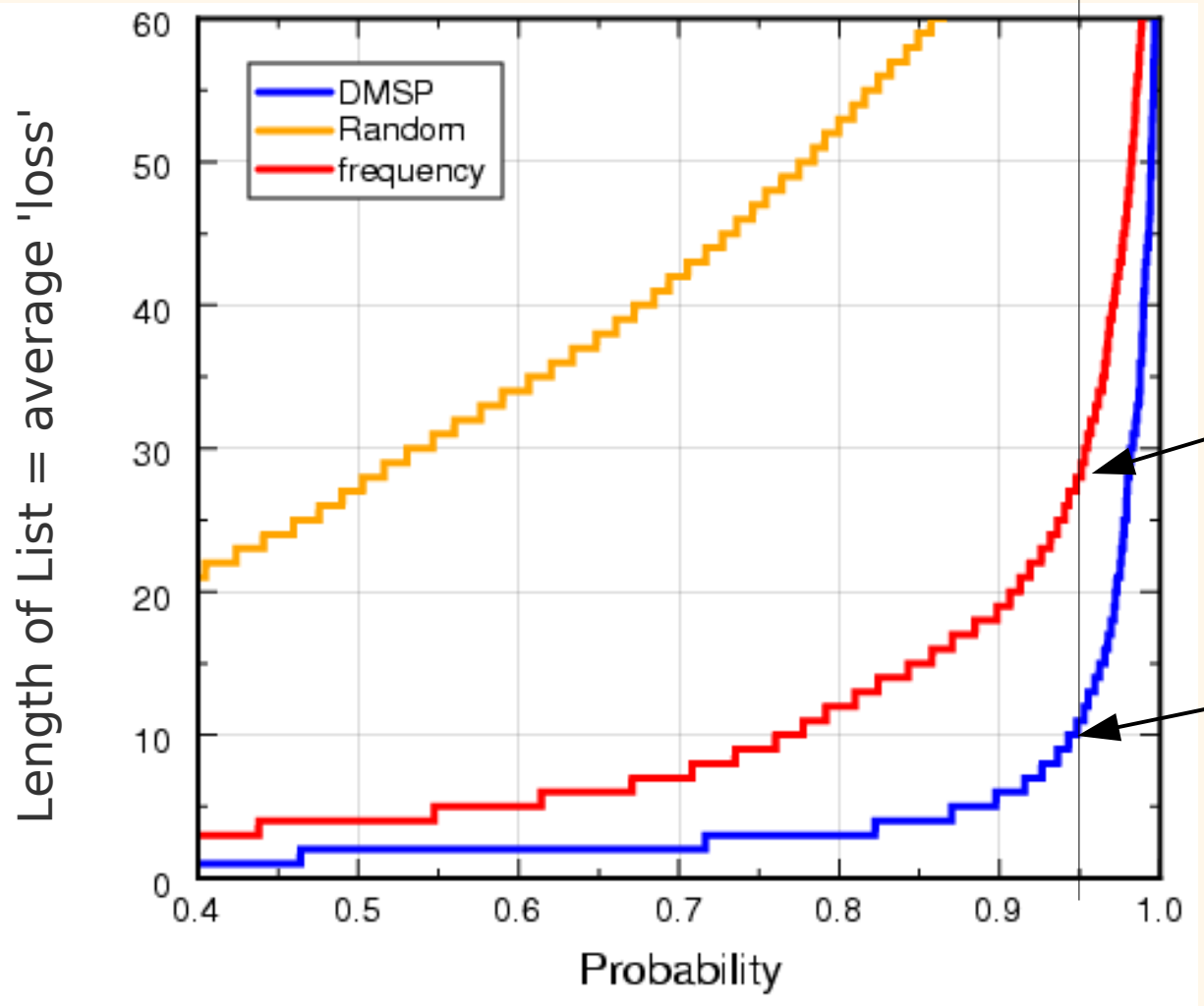
Set of likely structure candidates

Machine Learning Framework

$p(\tilde{x}|\boldsymbol{e})$

Database of Computed and Experimental results

Cross validation results



~28 candidates req'd for freq.

10 candidates --> 95% chance of seeing GS !!

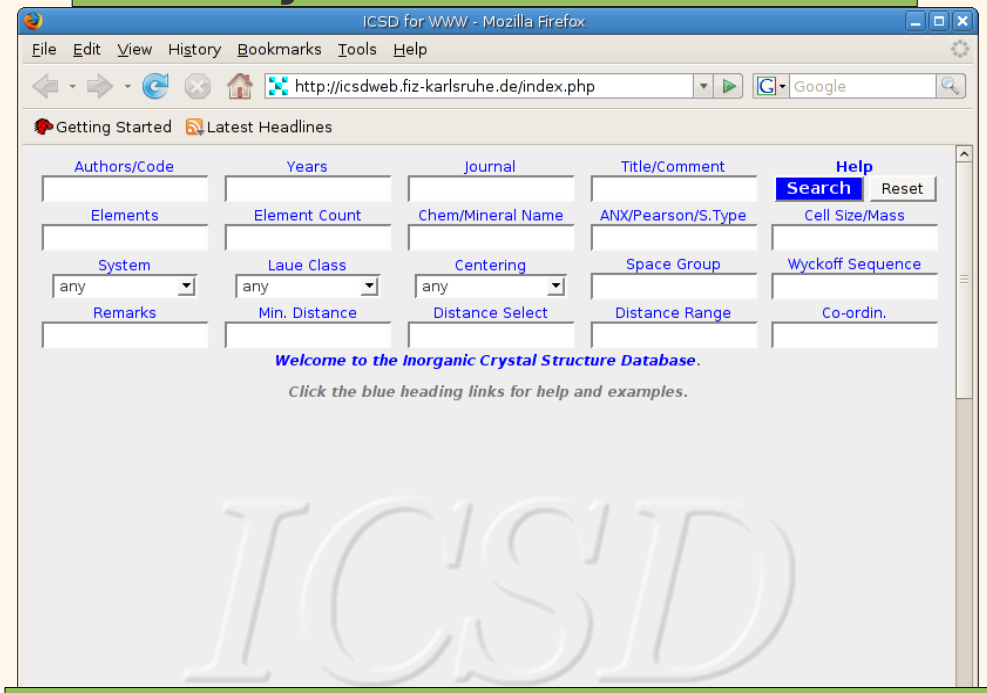
Independent Variables

Including structure correlation

Nature Materials, 6, 641-646, **2006**

Some open questions

ICSD: World's Largest database of inorganic crystal structures



First Entry: 1913
# of entries: 100,243
usable compounds: 29,962
structure prototypes: 2,485

What is the information content in a chemical database?

How many 'independent' crystal structures exist in nature ?

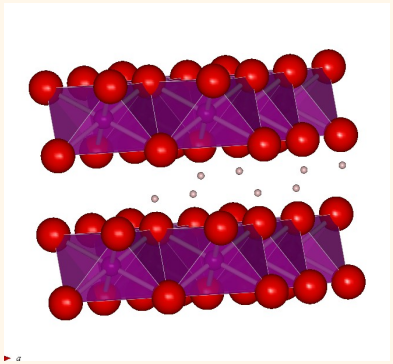
Structure prediction: wrap-up

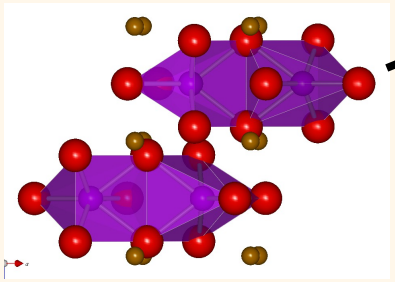
```
for i in (relevant chemistries) {
    ...
    ...
    getStablePhases(i);
    ...
    ...
    calculateProperty(i);
    i = nextChemistry();
}
```

Much more needed here

Now have efficient tool for this

Directions for future work/collaboration





abinit.org

b-initio VASP ackage imulation Vienna

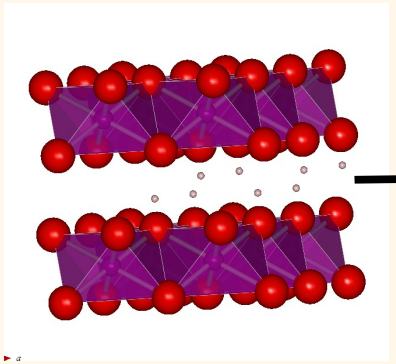
DFT Code

Material Property Predictions

Database of Computed results

Machine Learning Framework

Directions for future work/collaboration



abinit.org

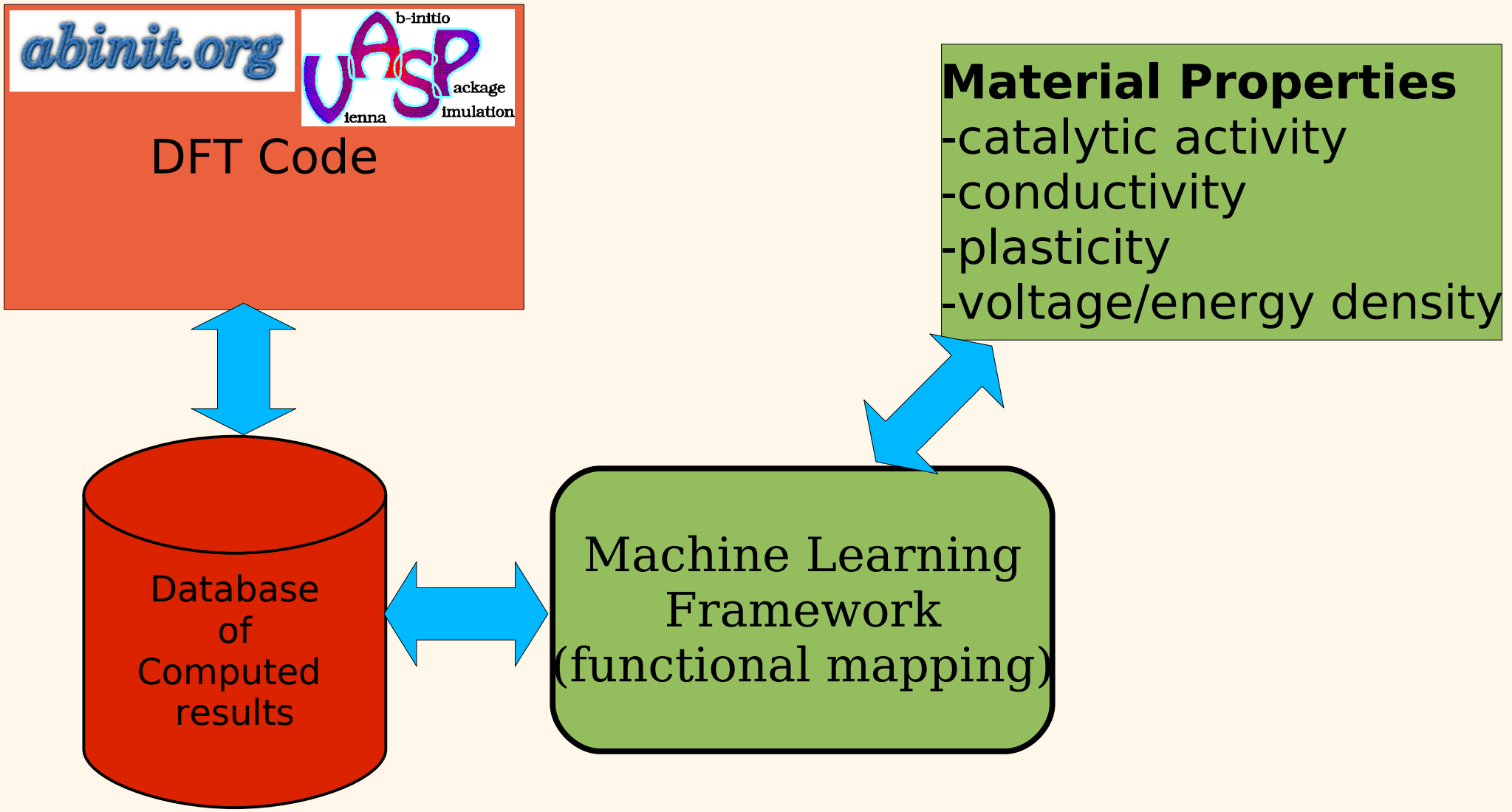
b-initio
VASP
ienna
ackage
imulation

DFT Code

Set of features

- Charge Density
- Total energy
- Bulk moduli
- Coordination
- Bond strength
- Bond character
- Magnetic moments
- Polarization
- ...

Directions for future collaboration



The End

Data from High Throughput alloy study

Online structure predictor

http://datamine.mit.edu

ITR grant (**DMR-031253**)

DELETE ME !!!

- introduce CMS, what is it being applied to ?
- Data mining and materials design – make some outline slide ?
- introduce structure prediction problem, present our solution
- discuss higher order property prediction. data management, dissemination

DATASET NOTES

1335 alloys

3975 non-unique compounds

4263 compounds total

alloys not containing elements:
He, B, C, N, O, F, Ne, Si, P, S, Cl, Ar, As, Se, Br, Kr, Te, I, Xe, At, Rn