

# Machine Learning for the Materials Scientist

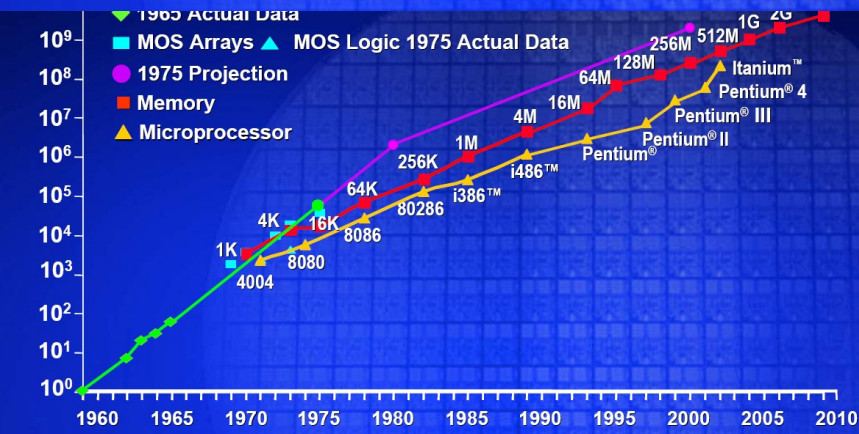
Chris Fischer\*, Kevin Tibbetts, Gerbrand Ceder  
Massachusetts Institute of Technology, Cambridge, MA

Dane Morgan  
University of Wisconsin, Madison, WI

# Motivation: materials design through calculation



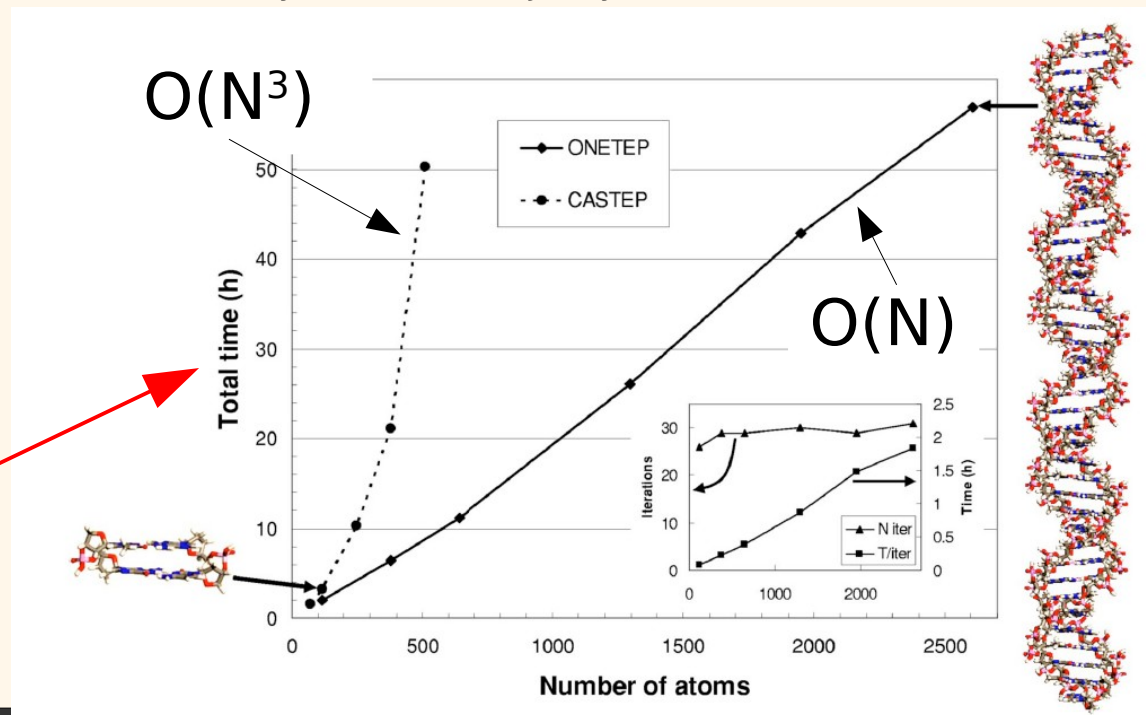
computing power:  
exponential scaling with  
time



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

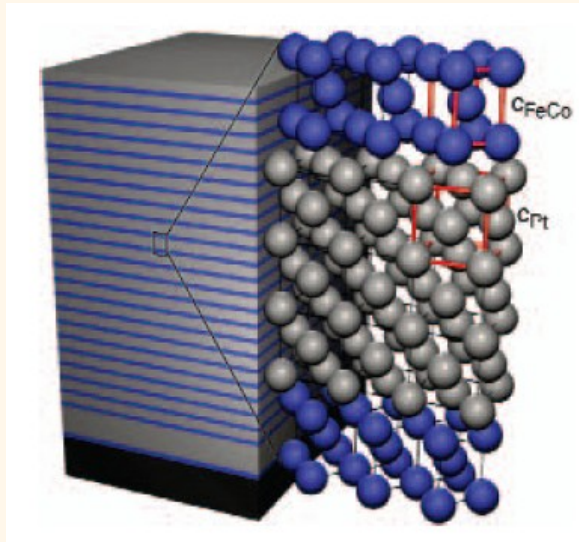
Run-time: polynomial  
scaling with number of  
atoms

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

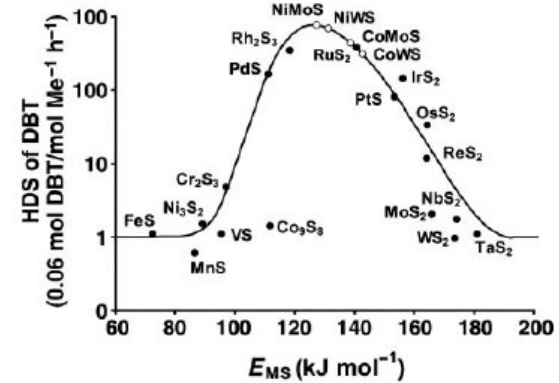
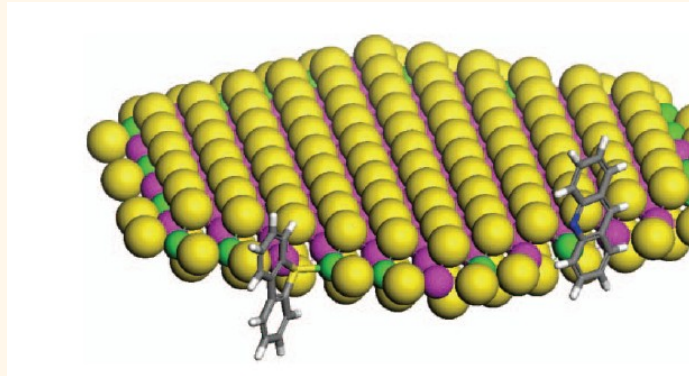


# DFT as a predictive tool

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



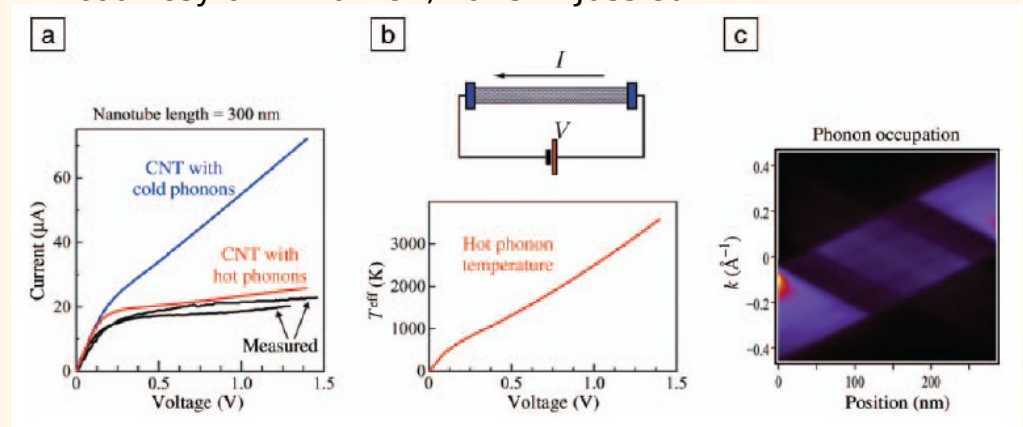
Norskov, J. et. al. MRS Bulletin **31** (2006)



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

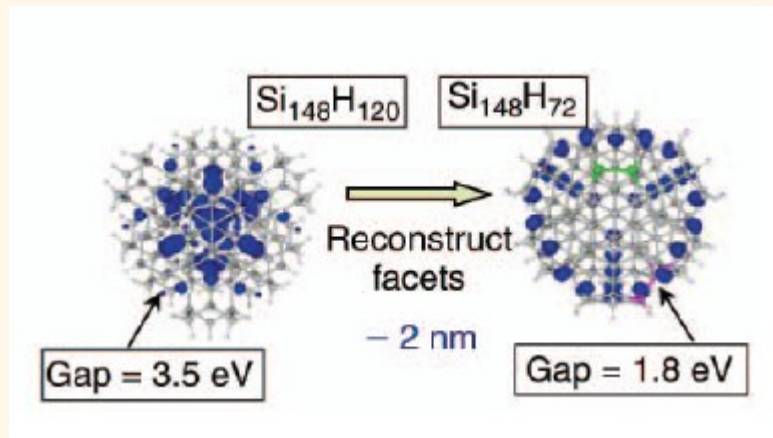


Marzari, N. MRS Bulletin **31** (2006)  
courtesy of M. Lazzeri, Paris VI Jussieu

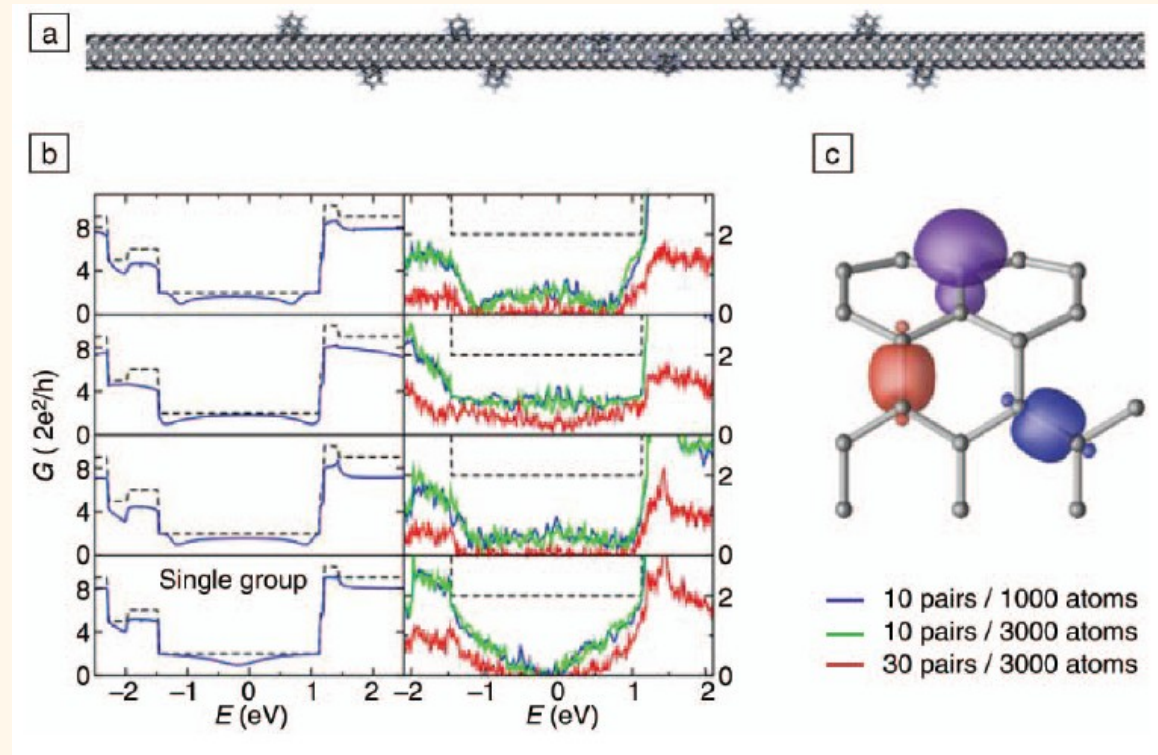


# computational materials design strategies

Calculating properties of realistic nanostructures  
*ab initio*



Galli, G. University of California, Davis



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

# computational materials design strategies

Which combinations yield the optimal material ?

## PERIODIC TABLE OF THE ELEMENTS

<http://www.kkf-split.hr/periodni/en/>

Metal	Semimetal	Nonmetal
1 Alkali metal	10 Chalcogens element	17 Halogens element
2 Alkaline earth metal	11 Transition metals	18 Noble gas
Lanthanide	STANDARD STATE (100 °C; 101 kPa)	
Actinide	Ne - gas	Fe - solid
	Ga - liquid	T - synthetic

RELATIVE ATOMIC MASS (A<sub>r</sub>)

GROUP IUPAC

GROUP CAS

ATOMIC NUMBER

SYMBOL

ELEMENT NAME

PERIOD	1 IA	2 IIA	3	4	5	6	7	8	9	10	11	12 IIB	13 IIIA	14 IVA	15 VA	16 VIA	17 VIIA	18 VIIIA
1	1.0079 <b>H</b> HYDROGEN																	4.0026 <b>He</b> HELIUM
2	3 6.941 <b>Li</b> LITHIUM	4 9.0122 <b>Be</b> BERYLLIUM											5 10.811 <b>B</b> BORON	6 12.011 <b>C</b> CARBON	7 14.007 <b>N</b> NITROGEN	8 15.999 <b>O</b> OXYGEN	9 18.998 <b>F</b> FLUORINE	10 20.180 <b>Ne</b> NEON
3	11 22.990 <b>Na</b> SODIUM	12 24.305 <b>Mg</b> MAGNESIUM											13 26.982 <b>Al</b> ALUMINIUM	14 28.086 <b>Si</b> SILICON	15 30.974 <b>P</b> PHOSPHORUS	16 32.065 <b>S</b> SULPHUR	17 35.453 <b>Cl</b> CHLORINE	18 39.948 <b>Ar</b> ARGON
4	19 39.098 <b>K</b> POTASSIUM	20 40.078 <b>Ca</b> CALCIUM	21 44.956 <b>Sc</b> SCANDIUM	22 47.867 <b>Ti</b> TITANIUM	23 50.942 <b>V</b> VANADIUM	24 51.996 <b>Cr</b> CHROMIUM	25 54.938 <b>Mn</b> MANGANESE	26 55.845 <b>Fe</b> IRON	27 58.933 <b>Co</b> COBALT	28 58.693 <b>Ni</b> NICKEL	29 63.546 <b>Cu</b> COPPER	30 65.39 <b>Zn</b> ZINC	31 69.723 <b>Ga</b> GALLIUM	32 72.64 <b>Ge</b> GERMANIUM	33 74.922 <b>As</b> ARSENIC	34 78.96 <b>Se</b> SELENIUM	35 79.904 <b>Br</b> BROMINE	36 83.80 <b>Kr</b> KRYPTON
5	37 85.468 <b>Rb</b> RUBIDIUM	38 87.62 <b>Sr</b> STRONTIUM	39 88.906 <b>Y</b> YTRIUM	40 91.224 <b>Zr</b> ZIRCONIUM	41 92.906 <b>Nb</b> NIOBIUM	42 95.94 <b>Mo</b> MOLYBDENUM	43 (98) <b>Tc</b> TECHNETIUM	44 101.07 <b>Ru</b> RUTHENIUM	45 102.91 <b>Rh</b> RHODIUM	46 106.42 <b>Pd</b> PALLADIUM	47 107.87 <b>Ag</b> SILVER	48 112.41 <b>Cd</b> CADMIUM	49 114.82 <b>In</b> INDIUM	50 118.71 <b>Sn</b> TIN	51 121.76 <b>Sb</b> ANTIMONY	52 127.60 <b>Te</b> TELLURIUM	53 126.90 <b>I</b> IODINE	54 131.29 <b>Xe</b> XENON
6	55 132.91 <b>Cs</b> CAESIUM	56 137.33 <b>Ba</b> BARIUM	57-71 <b>La-Lu</b> Lanthanide	72 178.49 <b>Hf</b> HAFNIUM	73 180.95 <b>Ta</b> TANTALUM	74 183.84 <b>W</b> TUNGSTEN	75 186.21 <b>Re</b> RHENIUM	76 190.23 <b>Os</b> OSMIUM	77 192.22 <b>Ir</b> IRIDIUM	78 195.08 <b>Pt</b> PLATINUM	79 196.97 <b>Au</b> GOLD	80 200.59 <b>Hg</b> MERCURY	81 204.38 <b>Tl</b> THALLIUM	82 207.2 <b>Pb</b> LEAD	83 208.98 <b>Bi</b> BISMUTH	84 (209) <b>Po</b> POLONIUM	85 (210) <b>At</b> ASTATINE	86 (222) <b>Rn</b> RADON
7	87 (223) <b>Fr</b> FRANCIUM	88 (226) <b>Ra</b> RADIUM	89-103 <b>Ac-Lr</b> Actinide	104 (261) <b>Rf</b> RUTHERFORDIUM	105 (262) <b>Db</b> DUBNIUM	106 (266) <b>Sg</b> SEABORGIUM	107 (264) <b>Bh</b> BOHRIUM	108 (277) <b>Hs</b> HASSIUM	109 (268) <b>Mt</b> MEITNERIUM	110 (281) <b>Uun</b> UNUNNIUM	111 (272) <b>Uuu</b> UNUNUNIUM	112 (285) <b>Uub</b> UNUNBIUM	114 (289) <b>Uuq</b> UNUNQUADIUM					

**LANTHANIDE**

57 138.91 <b>La</b> LANTHANUM	58 140.12 <b>Ce</b> CERIUM	59 140.91 <b>Pr</b> PRASEODYMIUM	60 144.24 <b>Nd</b> NEODYMIUM	61 (145) <b>Pm</b> PROMETHIUM	62 150.36 <b>Sm</b> SAMARIUM	63 151.96 <b>Eu</b> EUROPIUM	64 157.25 <b>Gd</b> GADOLINIUM	65 158.93 <b>Tb</b> TERBIUM	66 162.50 <b>Dy</b> DYSPROSIUM	67 164.93 <b>Ho</b> HOLMIUM	68 167.26 <b>Er</b> ERBIUM	69 168.93 <b>Tm</b> THULIUM	70 173.04 <b>Yb</b> YTTERIUM	71 174.97 <b>Lu</b> LUTETIUM
-------------------------------------	----------------------------------	--	-------------------------------------	-------------------------------------	------------------------------------	------------------------------------	--------------------------------------	-----------------------------------	--------------------------------------	-----------------------------------	----------------------------------	-----------------------------------	------------------------------------	------------------------------------

**ACTINIDE**

89 (227) <b>Ac</b> ACTINIUM	90 232.04 <b>Th</b> THORIUM	91 231.04 <b>Pa</b> PROTACTINIUM	92 238.03 <b>U</b> URANIUM	93 (237) <b>Np</b> NEPTUNIUM	94 (244) <b>Pu</b> PLUTONIUM	95 (243) <b>Am</b> AMERICIUM	96 (247) <b>Cm</b> CURIUM	97 (247) <b>Bk</b> BERKELIUM	98 (251) <b>Cf</b> CALIFORNIUM	99 (252) <b>Es</b> EINSTEINIUM	100 (257) <b>Fm</b> FERMIUM	101 (258) <b>Md</b> MENDELEVIUM	102 (259) <b>No</b> NOBELIUM	103 (262) <b>Lr</b> LAWRENCIUM
-----------------------------------	-----------------------------------	--	----------------------------------	------------------------------------	------------------------------------	------------------------------------	---------------------------------	------------------------------------	--------------------------------------	--------------------------------------	-----------------------------------	---------------------------------------	------------------------------------	--------------------------------------

(1) Pure Appl. Chem., 73, No. 4, 657-683 (2001)  
Relative atomic mass is shown with five significant figures. For elements with no stable nuclides, the value enclosed in brackets indicates the mass number of the longest-lived isotope of the element.  
However there are such elements (Th, Pa, and U) do have a characteristic terrestrial isotopic composition, and for these an atomic weight is tabulated.

Editor: Aditya Vardhan (adivar@netlinx.com)

Copyright © 1998-2003 EniG. (eni@kf-split.hr)

# 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

# computational materials design strategies

Which combinations yield the optimal material ?

## PERIODIC TABLE OF THE ELEMENTS

<http://www.kkf-split.hr/periodni/en/>

GROUP IUPAC

GROUP CAS

ATOMIC NUMBER

SYMBOL

ELEMENT NAME

RELATIVE ATOMIC MASS (A<sub>r</sub>)

13 10.811

**B**

BORON

Metal										Semimetal		Nonmetal			
Alkali metal		Alkaline earth metal		Transition metals		Lanthanide		Actinide		Chalcogens element		Halogens element		Noble gas	

STANDARD STATE (100 °C; 101 kPa)

Ne - gas      Fe - solid

Ga - liquid      T<sub>e</sub> - synthetic

PERIOD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
GROUP	IA	IIA	IIIB	IVB	VB	VIB	VIB	VIB	VIB	VIB	VIB	VIB	IIIA	IVA	VA	VIA	VIA	VIA	VIIIA
1	1.0079 <b>H</b> HYDROGEN																		4.0026 <b>He</b> HELIUM
2	3 6.941 <b>Li</b> LITHIUM	4 9.0122 <b>Be</b> BERYLLIUM											5 10.811 <b>B</b> BORON	6 12.011 <b>C</b> CARBON	7 14.007 <b>N</b> NITROGEN	8 15.999 <b>O</b> OXYGEN	9 18.998 <b>F</b> FLUORINE	10 20.180 <b>Ne</b> NEON	
3	11 22.990 <b>Na</b> SODIUM	12 24.305 <b>Mg</b> MAGNESIUM											13 26.982 <b>Al</b> ALUMINIUM	14 28.086 <b>Si</b> SILICON	15 30.974 <b>P</b> PHOSPHORUS	16 32.065 <b>S</b> SULPHUR	17 35.453 <b>Cl</b> CHLORINE	18 39.948 <b>Ar</b> ARGON	
4	19 39.098 <b>K</b> POTASSIUM	20 40.078 <b>Ca</b> CALCIUM	21 44.956 <b>Sc</b> SCANDIUM	22 47.867 <b>Ti</b> TITANIUM	23 50.942 <b>V</b> VANADIUM	24 51.996 <b>Cr</b> CHROMIUM	25 54.938 <b>Mn</b> MANGANESE	26 55.845 <b>Fe</b> IRON	27 58.933 <b>Co</b> COBALT	28 58.693 <b>Ni</b> NICKEL	29 63.546 <b>Cu</b> COPPER	30 65.39 <b>Zn</b> ZINC	31 69.723 <b>Ga</b> GALLIUM	32 72.64 <b>Ge</b> GERMANIUM	33 74.922 <b>As</b> ARSENIC	34 78.96 <b>Se</b> SELENIUM	35 79.904 <b>Br</b> BROMINE	36 83.80 <b>Kr</b> KRYPTON	
5	37 85.468 <b>Rb</b> RUBIDIUM	38 87.62 <b>Sr</b> STRONTIUM	39 88.906 <b>Y</b> YTRITIUM	40 91.224 <b>Zr</b> ZIRCONIUM	41 92.906 <b>Nb</b> NIOBIUM	42 95.94 <b>Mo</b> MOLYBDENUM	43 (98) <b>Tc</b> TECHNETIUM	44 101.07 <b>Ru</b> RUTHENIUM	45 102.91 <b>Rh</b> RHODIUM	46 106.42 <b>Pd</b> PALLADIUM	47 107.87 <b>Ag</b> SILVER	48 112.41 <b>Cd</b> CADMIUM	49 114.82 <b>In</b> INDIUM	50 118.71 <b>Sn</b> TIN	51 121.76 <b>Sb</b> ANTIMONY	52 127.60 <b>Te</b> TELLURIUM	53 126.90 <b>I</b> IODINE	54 131.29 <b>Xe</b> XENON	
6	55 132.91 <b>Cs</b> CAESIUM	56 137.33 <b>Ba</b> BARIUM	57-71 <b>La-Lu</b> Lanthanide	72 178.49 <b>Hf</b> HAFNIUM	73 180.95 <b>Ta</b> TANTALUM	74 183.84 <b>W</b> TUNGSTEN	75 186.21 <b>Re</b> RHENIUM	76 190.23 <b>Os</b> OSMIUM	77 192.22 <b>Ir</b> IRIDIUM	78 195.08 <b>Pt</b> PLATINUM	79 196.97 <b>Au</b> GOLD	80 200.59 <b>Hg</b> MERCURY	81 204.38 <b>Tl</b> THALLIUM	82 207.2 <b>Pb</b> LEAD	83 208.98 <b>Bi</b> BISMUTH	84 (209) <b>Po</b> POLONIUM	85 (210) <b>At</b> ASTATINE	86 (222) <b>Rn</b> RADON	
7	87 (223) <b>Fr</b> FRANCIUM	88 (226) <b>Ra</b> RADIUM	89-103 <b>Ac-Lr</b> Actinide	104 (261) <b>Rf</b> RUTHERFORDIUM	105 (262) <b>Db</b> DUBNIUM	106 (266) <b>Sg</b> SEABORGIUM	107 (264) <b>Bh</b> BOHRIUM	108 (277) <b>Hs</b> HASSIUM	109 (268) <b>Mt</b> MEITNERIUM	110 (281) <b>Uun</b> UNUNNIUM	111 (272) <b>Uuu</b> UNUNUNIUM	112 (285) <b>Uub</b> UNUNBIUM		114 (289) <b>Uuq</b> UNUNQUADIUM					

**LANTHANIDE**

57 138.91 <b>La</b> LANTHANUM	58 140.12 <b>Ce</b> CERIUM	59 140.91 <b>Pr</b> PRASEODYMIUM	60 144.24 <b>Nd</b> NEODYMIUM	61 (145) <b>Pm</b> PROMETHIUM	62 150.36 <b>Sm</b> SAMARIUM	63 151.96 <b>Eu</b> EUROPIUM	64 157.25 <b>Gd</b> GADOLINIUM	65 158.93 <b>Tb</b> TERBIUM	66 162.50 <b>Dy</b> DYSPROSIUM	67 164.93 <b>Ho</b> HOLMIUM	68 167.26 <b>Er</b> ERBIUM	69 168.93 <b>Tm</b> THULIUM	70 173.04 <b>Yb</b> YTTERIUM	71 174.97 <b>Lu</b> LUTETIUM
-------------------------------------	----------------------------------	--	-------------------------------------	-------------------------------------	------------------------------------	------------------------------------	--------------------------------------	-----------------------------------	--------------------------------------	-----------------------------------	----------------------------------	-----------------------------------	------------------------------------	------------------------------------

(1) Pure Appl. Chem., 73, No. 4, 657-683 (2001)  
Relative atomic mass is shown with five significant figures. For elements with no stable nuclides, the value enclosed in brackets indicates the mass number of the longest-lived isotope of the element.

However three such elements (Th, Pa, and U) do have a characteristic terrestrial isotopic composition, and for these an atomic weight is tabulated.

**ACTINIDE**

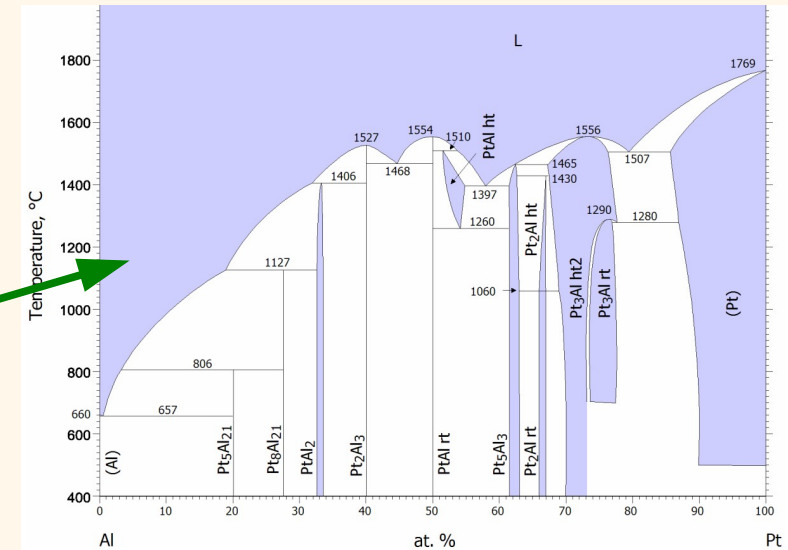
89 (227) <b>Ac</b> ACTINIUM	90 232.04 <b>Th</b> THORIUM	91 231.04 <b>Pa</b> PROTACTINIUM	92 238.03 <b>U</b> URANIUM	93 (237) <b>Np</b> NEPTUNIUM	94 (244) <b>Pu</b> PLUTONIUM	95 (243) <b>Am</b> AMERICIUM	96 (247) <b>Cm</b> CURIUM	97 (247) <b>Bk</b> BERKELIUM	98 (251) <b>Cf</b> CALIFORNIUM	99 (252) <b>Es</b> EINSTEINIUM	100 (257) <b>Fm</b> FERMIUM	101 (258) <b>Md</b> MENDELEVIUM	102 (259) <b>No</b> NOBELIUM	103 (262) <b>Lr</b> LAWRENCIUM
-----------------------------------	-----------------------------------	--	----------------------------------	------------------------------------	------------------------------------	------------------------------------	---------------------------------	------------------------------------	--------------------------------------	--------------------------------------	-----------------------------------	---------------------------------------	------------------------------------	--------------------------------------

Editor: Aditya Vardhan (advan@netlinx.com)

Copyright © 1998-2003 EniG. (eni@kf-split.hr)

# 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) {
```

```
...
```

```
...
```

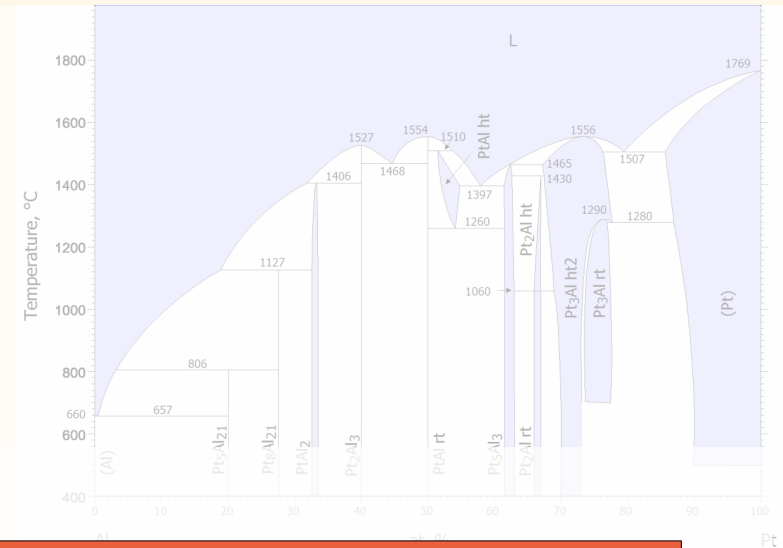
```
  getStablePhases(i);
```

```
...
```

```
...
```

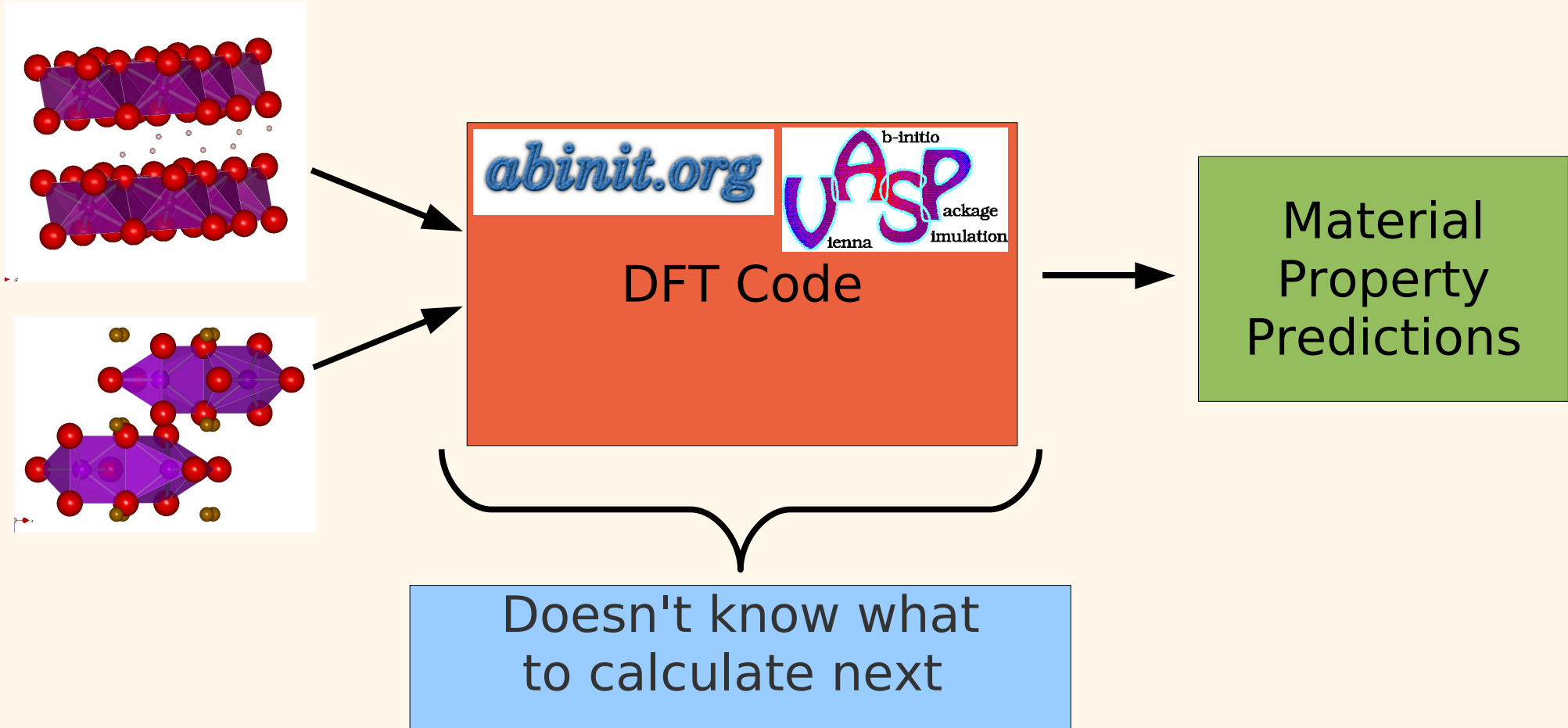
```
  calculateProperty(i);  
  i = nextChemistry();
```

```
}
```

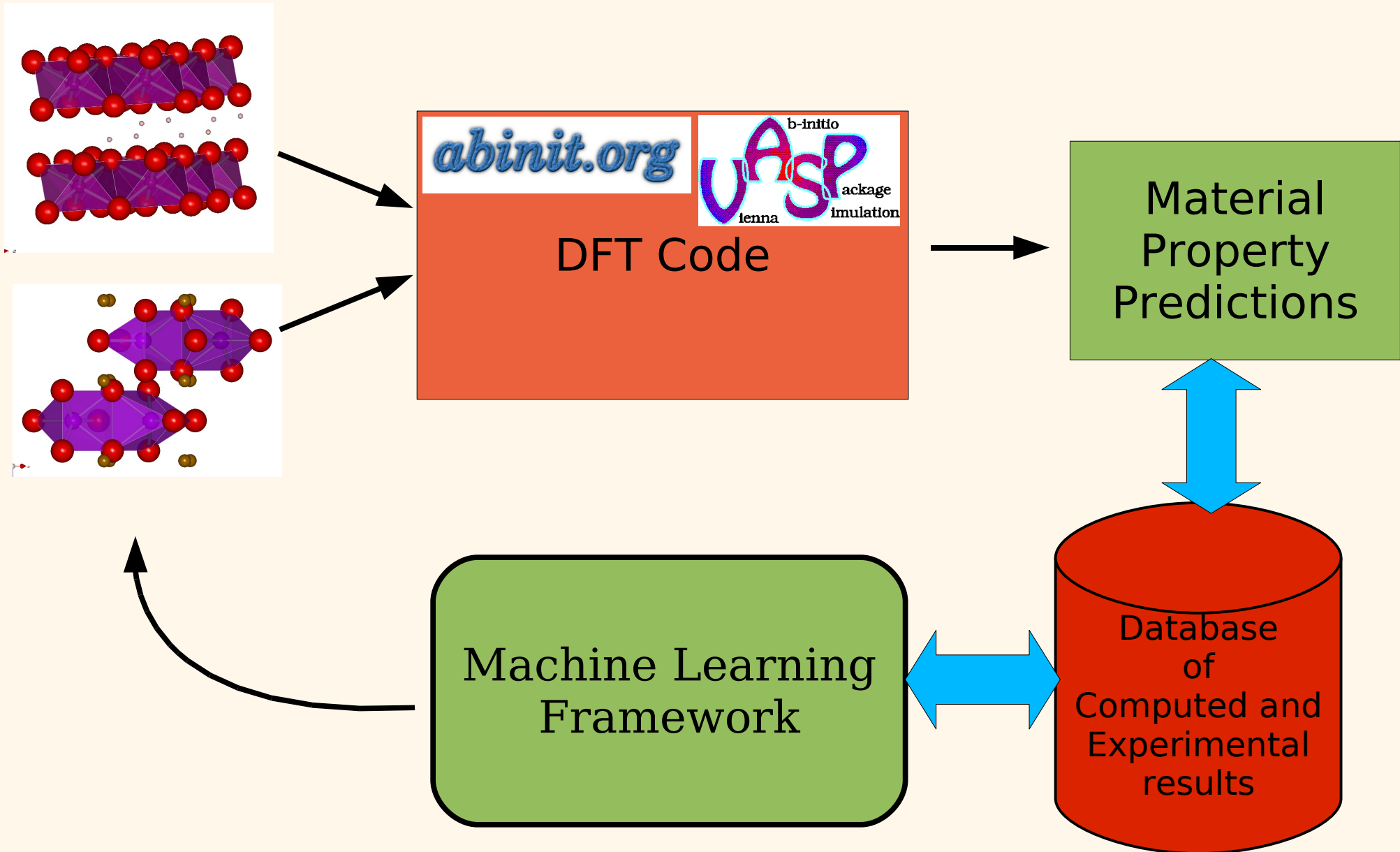


Machine Learning  
needed here !!

# The need for machine learning



# The need for machine learning



# 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**

The screenshot shows the ICSD website search interface. The browser window title is "ICSD for WWW - Mozilla Firefox". The address bar shows "http://icsdweb.fiz-karlsruhe.de/index.php". The search form includes fields for Authors/Code, Years, Journal, Title/Comment, Elements, Element Count, Chem/Mineral Name, ANX/Pearson/S.Type, System, Laue Class, Centering, Space Group, Wyckoff Sequence, Remarks, Min. Distance, Distance Select, Distance Range, and Co-ordin. There are also "Search" and "Reset" buttons. Below the form, it says "Welcome to the Inorganic Crystal Structure Database. Click the blue heading links for help and examples." The large "ICSD" logo is visible in the background. At the bottom, it says "Demo database (The Full database will be used if available after the first query is entered) Copyright: 2003-2007 Fachinformationszentrum (FIZ) Karlsruhe PHP/MySQL Interface V07-09-16 copyright: 2003-2007 by Peter Hewat email: hewat@ill.fr".

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**

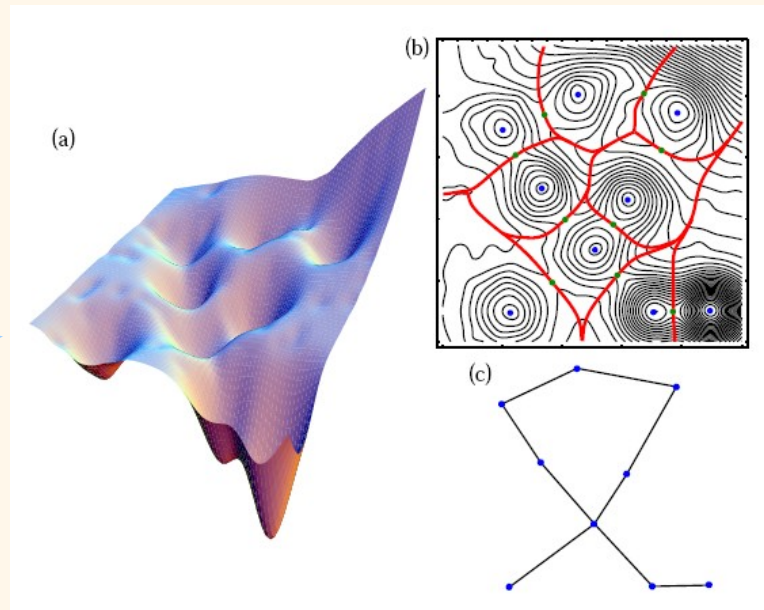
# Methods to search for structure

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

$$\text{GroundState} \equiv \arg \min_{\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N} E(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$$

$$\# \text{ of dimensions} = 3N - 3 + \dim(a, b, c, \alpha, \beta, \gamma)$$

complex energy landscape



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



# Methods to search for structure

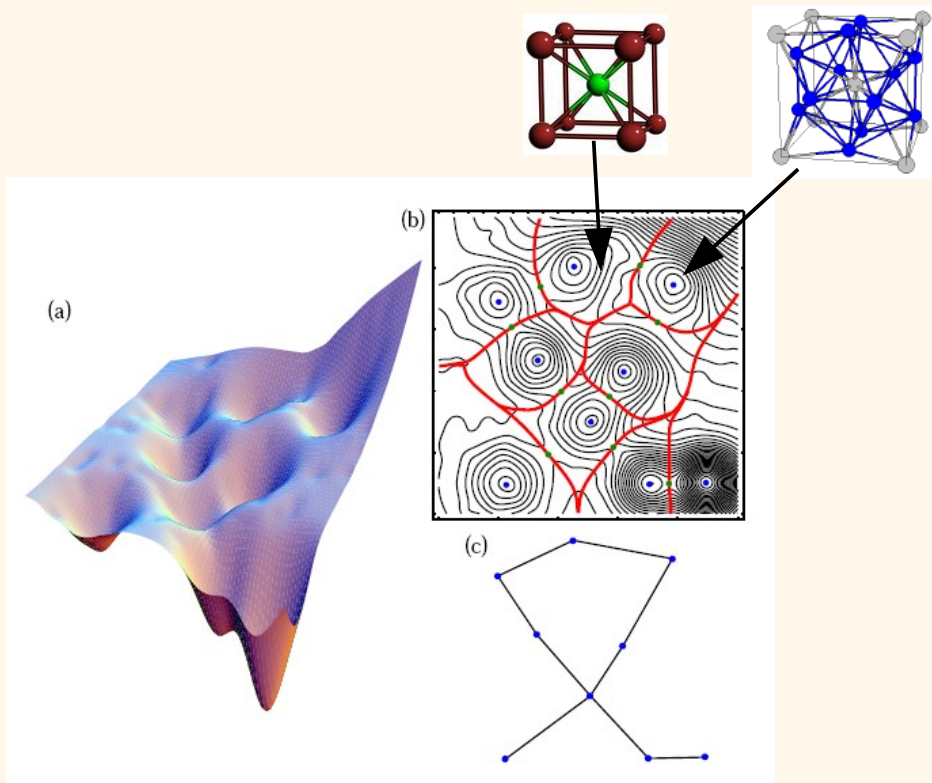
**Coordinate Search:**  
Optimize energy (or free energy) directly in the space of atomic coordinates

$$\text{GroundState} \equiv \arg \min_{\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N} E(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$$

$$\# \text{ of dimensions} = 3N - 3 + \dim(a, b, c, \alpha, \beta, \gamma)$$

## Proposed Solutions

Calculate energy of a finite set of structure prototypes



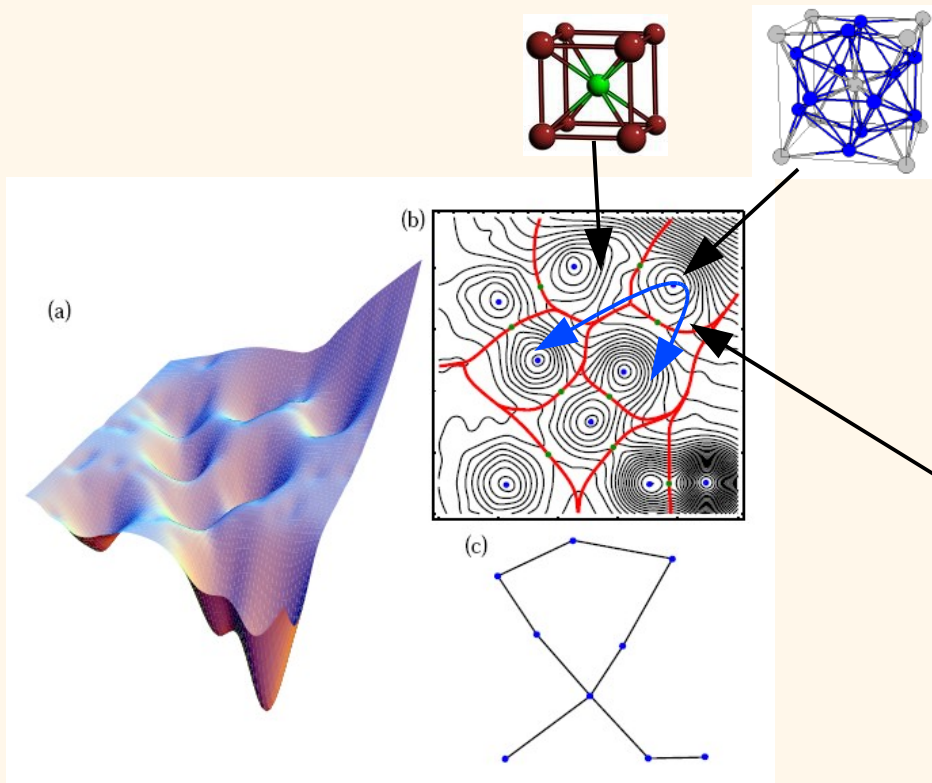
Doye, J. PRL, **88**, 238701, (2002)

# Methods to search for structure

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

$$\text{GroundState} \equiv \arg \min_{\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N} E(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$$

$$\# \text{ of dimensions} = 3N - 3 + \dim(a, b, c, \alpha, \beta, \gamma)$$



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

$$\text{GroundState} \equiv \arg \min_{\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N} E(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$$

# of dimensions =  $3N - 3 + \dim(a, b, c, \alpha, \beta, \gamma)$

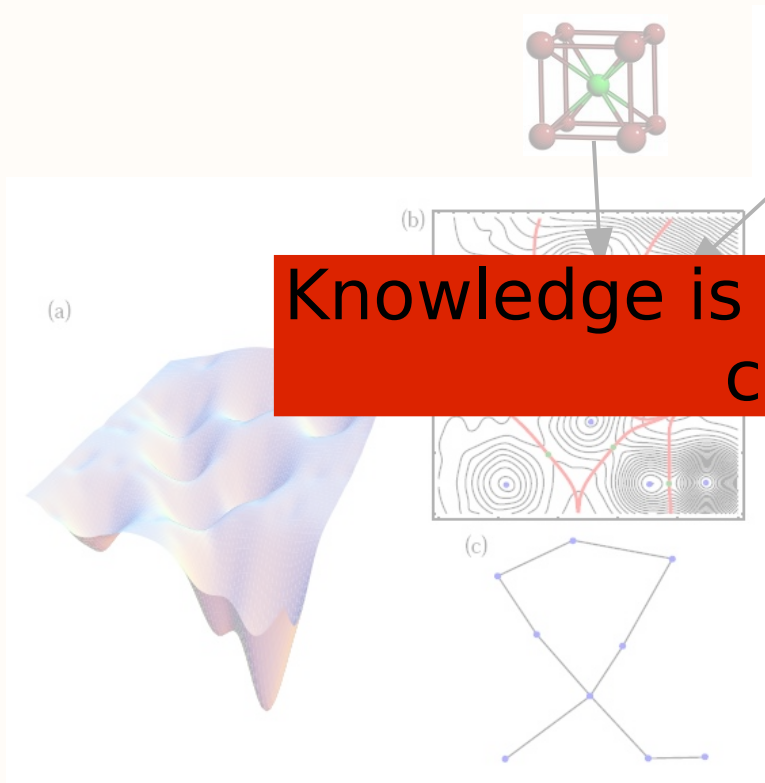
## Proposed Solutions

Calculate energy of a finite set of structure prototypes

**Knowledge is not *transferred* across chemistries**

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

e.g., Simulated Annealing  
Genetic Algorithms



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

# 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

$$\Delta \chi$$

$$\Delta r_{A,B}$$

$$n_{at}^{(e)}$$

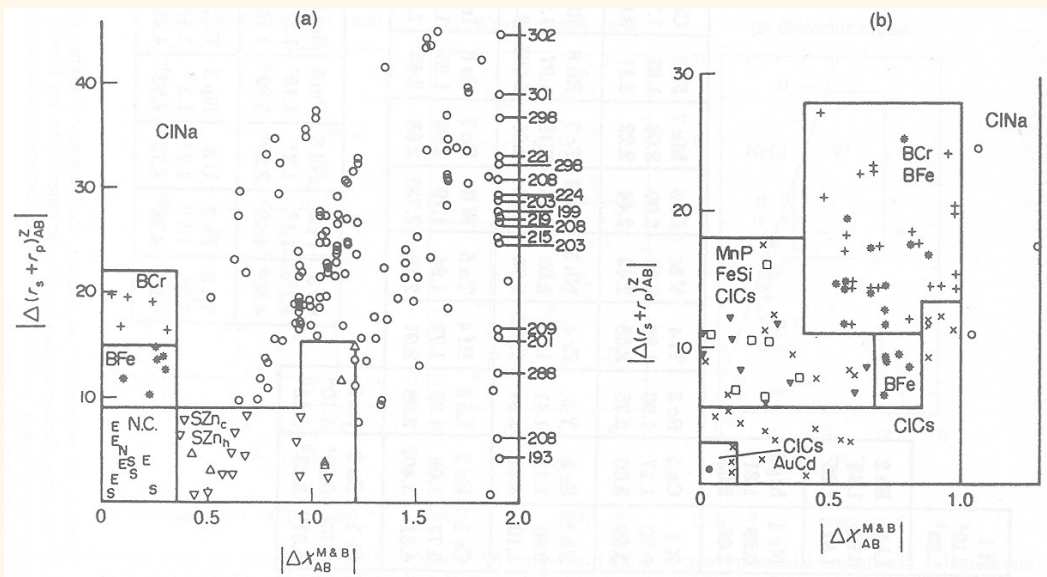
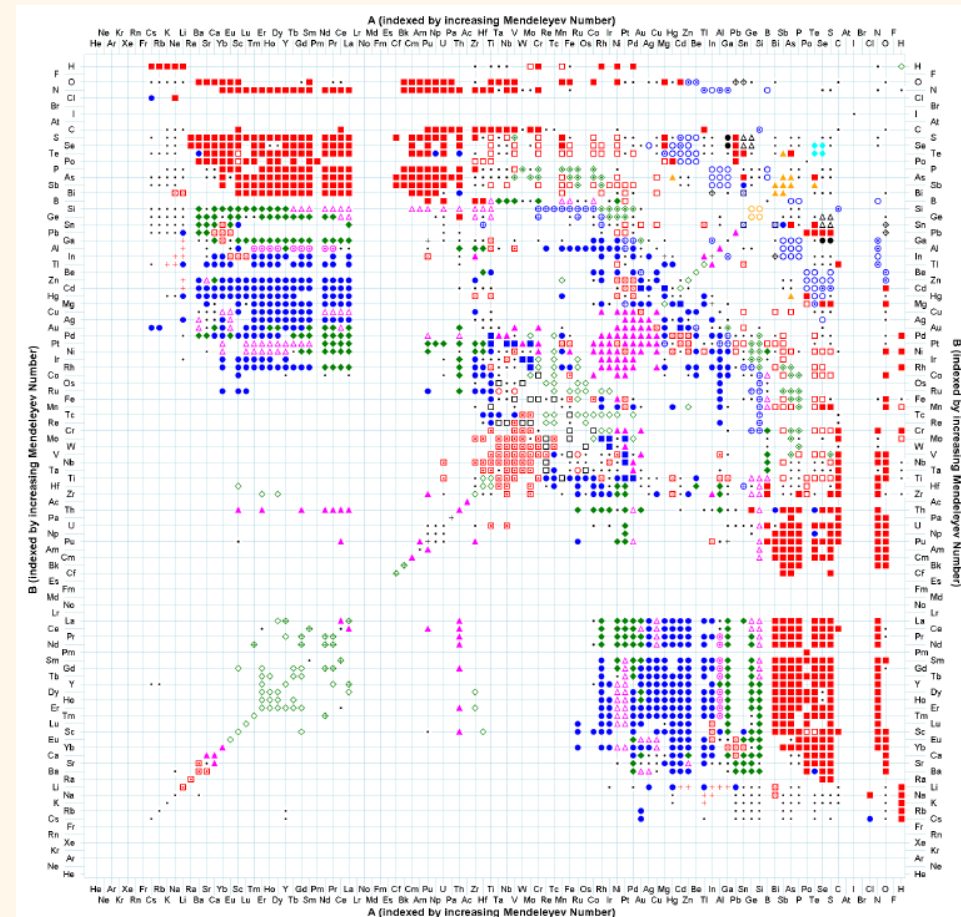


Figure 5. The Villars maps for AB compounds corresponding to the average electron-per-atom ratio of  $\bar{N}=4$  and  $6.5$  respectively (From Villars, 1983. Reproduced with permission). (a)  $\bar{N}=4$ ; (b)  $\bar{N}=6.5$



# Methods to search for structure

## Heuristic Rules

Plot stable structures in space of parameters

1986: Pettifor

1983: Villars

$\Delta\chi$

$\Delta r_{A,B}$

$n_{at}^{(e)}$

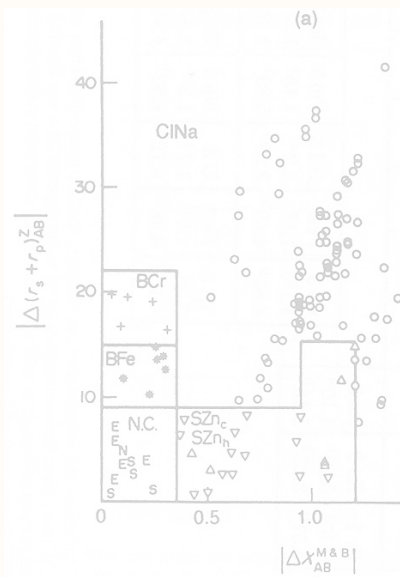
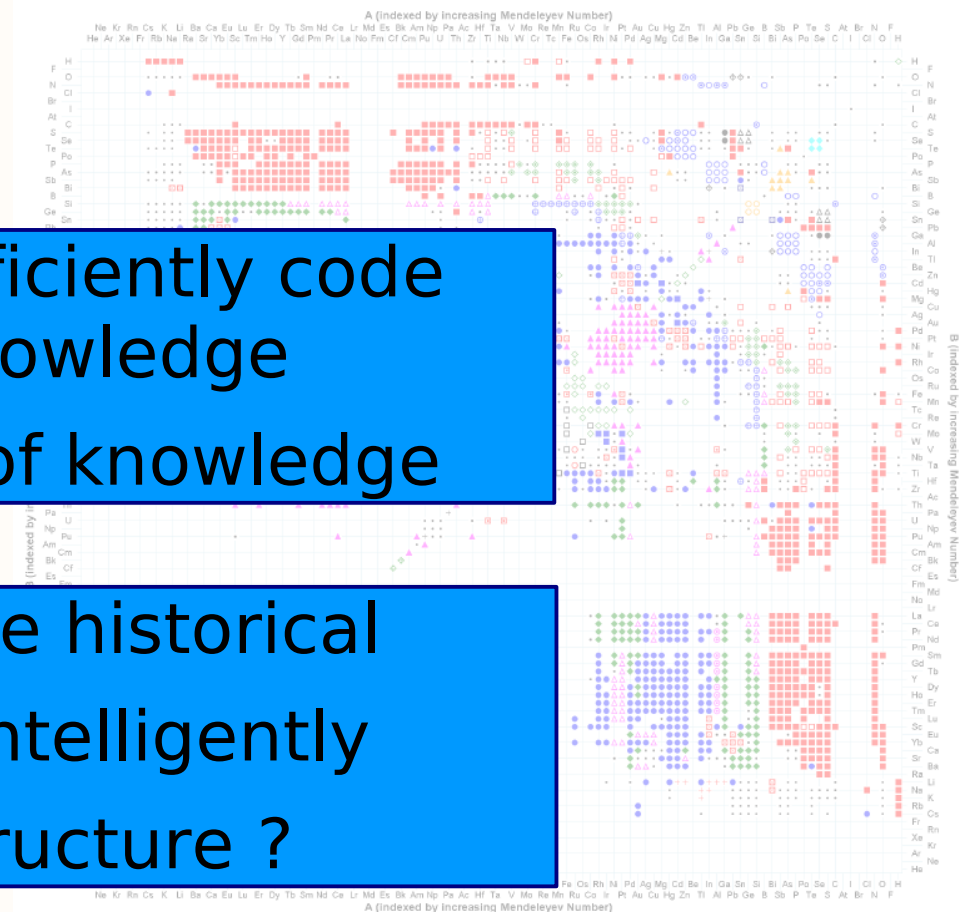


Figure 5. The Villars maps for AB compounds (From Villars, 1983. Reproduced with permission)

Heuristic rules efficiently code historical knowledge  
provide *transfer* of knowledge

Can we leverage historical knowledge to intelligently search for structure ?



## 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

$$\mathbf{x} = \left( X_A, X_0, \dots, X_{\frac{1}{2}}, \dots, X_B \right)$$

Low temperature state of alloy

$$Data \equiv \{ \mathbf{x}_1, \dots, \mathbf{x}_N \}$$

database of N binary alloys



# Machine learning framework: concepts

$$\mathbf{x} = \left( \mathbf{x}_A, \mathbf{x}_0, \dots, \mathbf{x}_{\frac{1}{2}}, \dots, \mathbf{x}_B \right)$$

Low temperature state of alloy

$$Data \equiv \{ \mathbf{x}_1, \dots, \mathbf{x}_N \}$$

database of N binary alloys

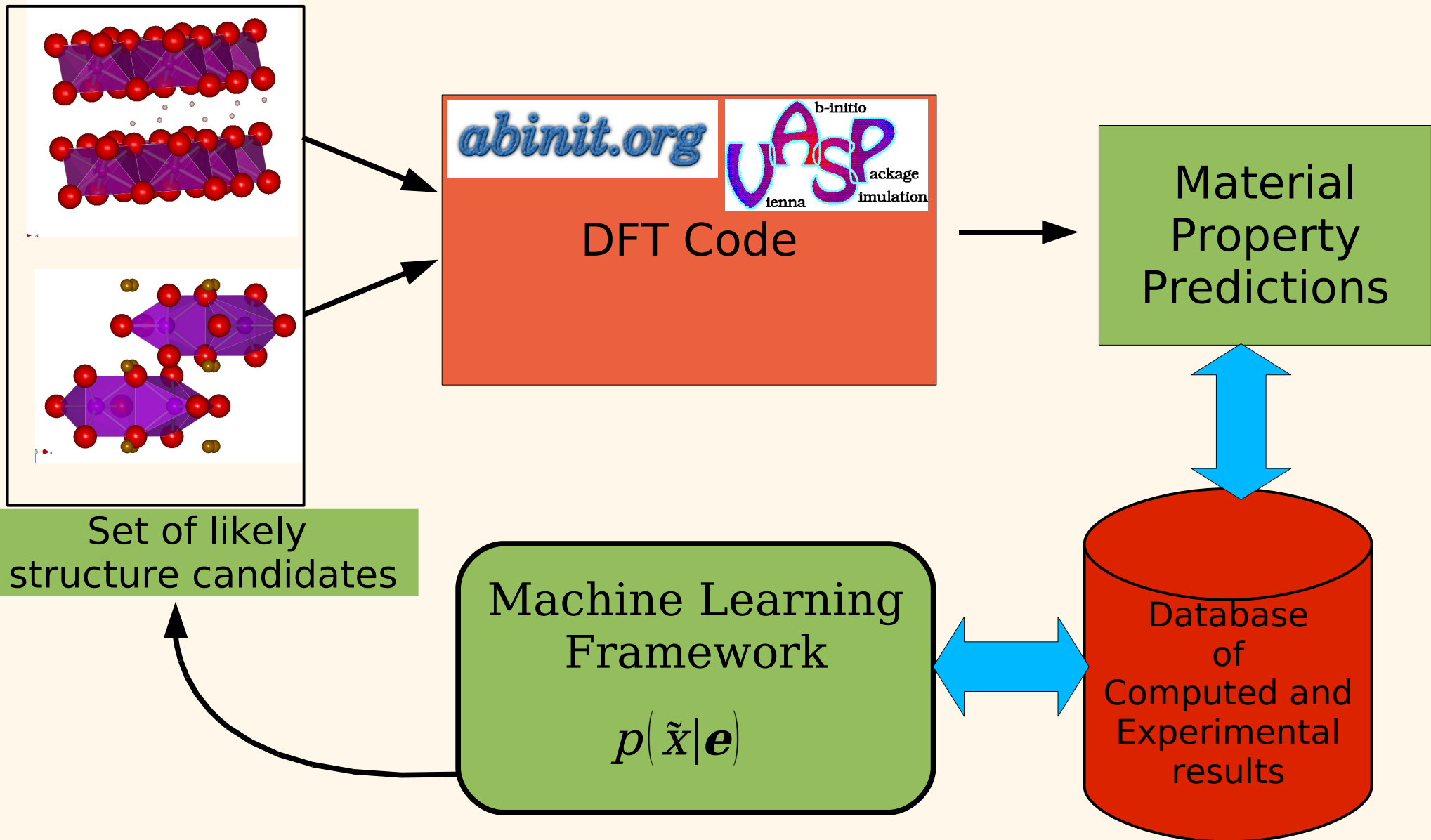
$$p(\mathbf{x})$$

Probability of low temperature state (fitted to data)

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

Probability of low temperature state conditioned on evidence 'e'

# how to use the machine learning framework

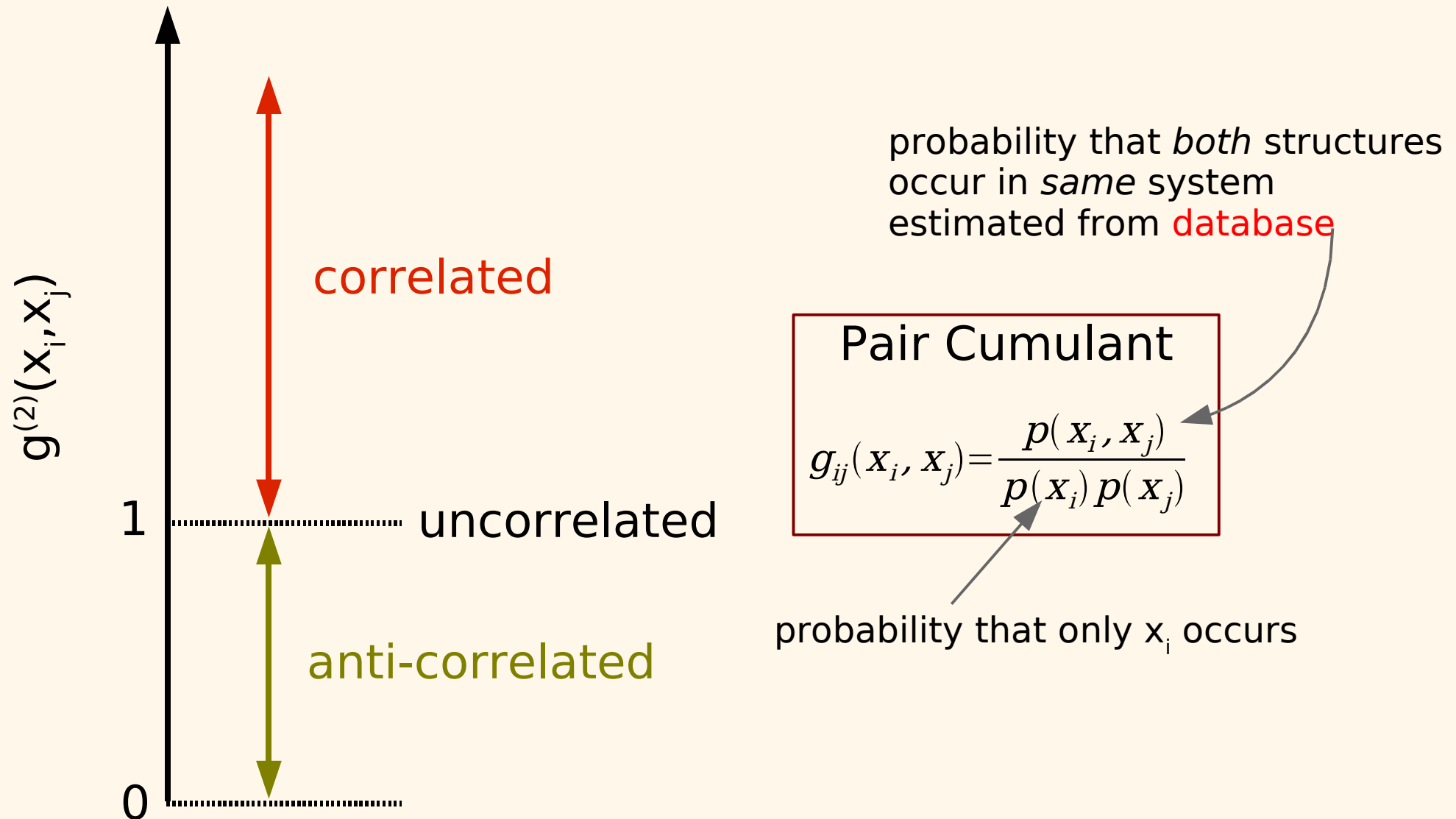


# Preliminaries and open questions

Are probabilities consistent with physical intuition ?

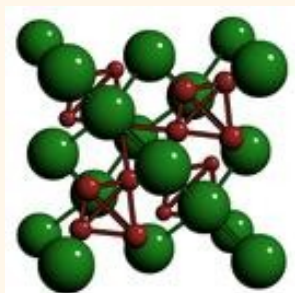
Do probabilities encode the physics of structure stability ?

# quantifying correlation in probabilistic framework

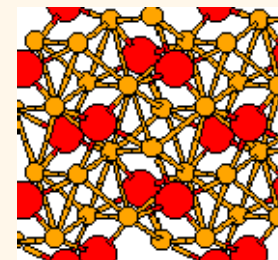


# how probabilities represent physics of mixing

Do probabilities embody real physical effects ?  
Compounds stabilized by “size” effect:



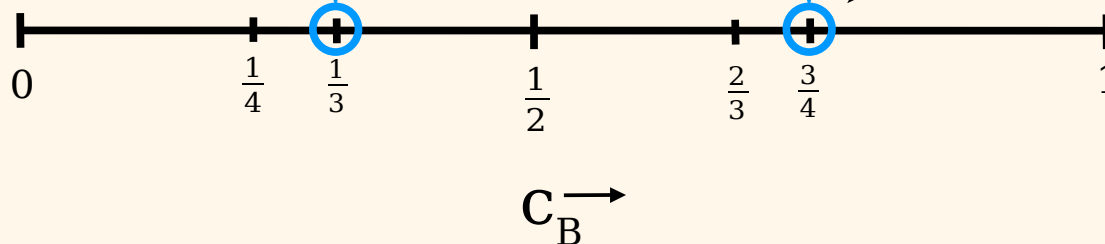
MgCu<sub>2</sub>



Fe<sub>3</sub>C

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

8.48



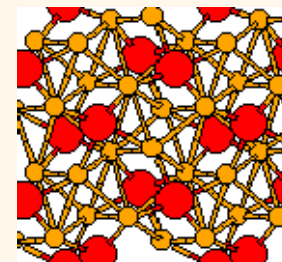
Data from Pauling File, Binaries Edition

# how probabilities represent physics of mixing

Do probabilities embody real physical effects ?  
Compounds stabilized by "size" effect:

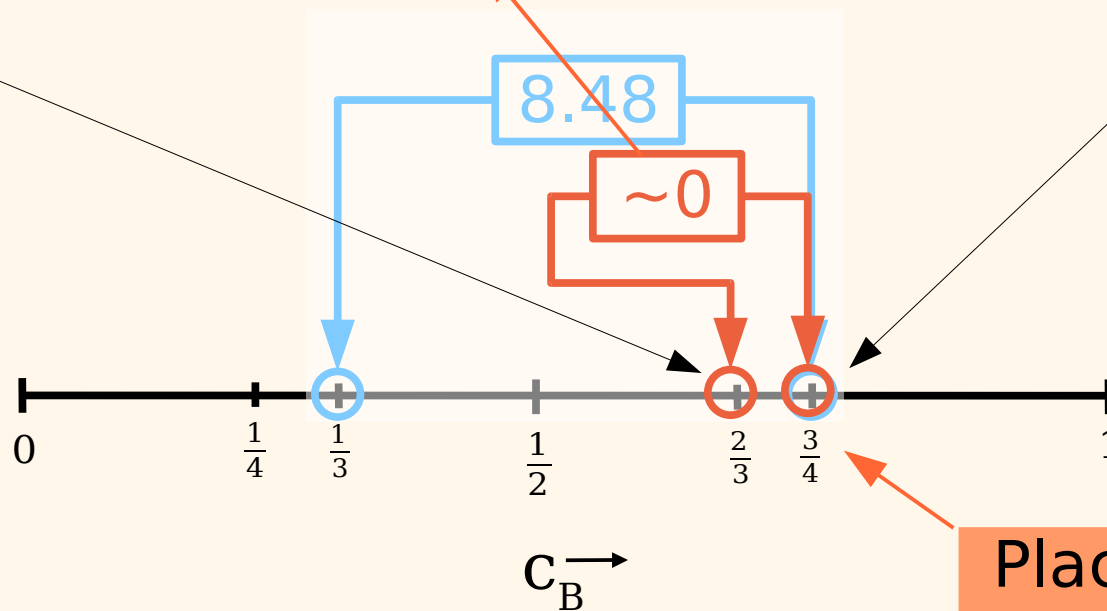


MgCu<sub>2</sub>



Fe<sub>3</sub>C

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



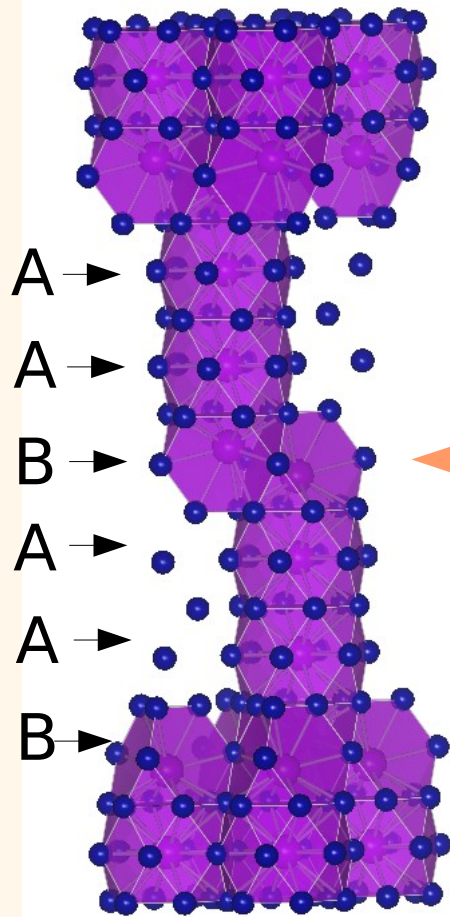
Places 'small' atoms on 'large' atom sites

Data from Pauling File, Binaries Edition

G. Ceder

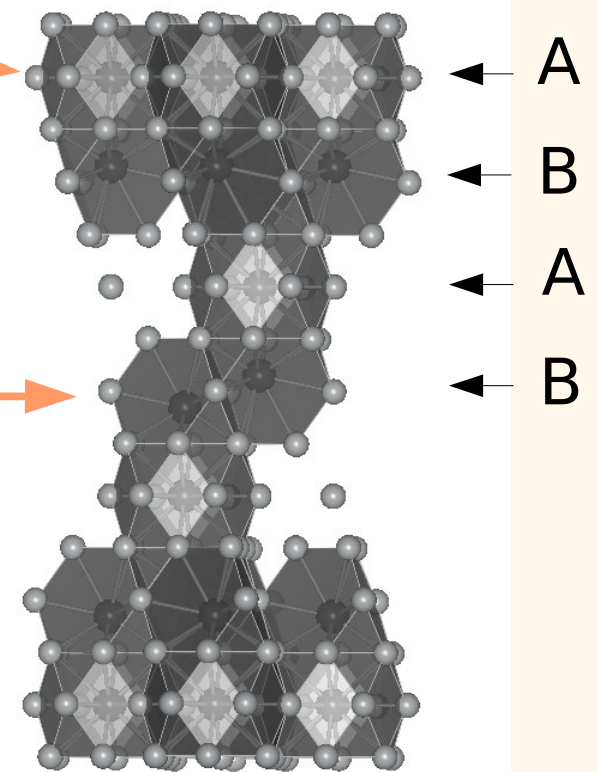
# how probabilities represent physics of mixing: more interesting correlations

$\text{Gd}_2\text{Co}_7$   
AABAAB... stacking



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

$\text{PuNi}_3$   
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  $\text{Fe}_3\text{C}$  forms @  $c = 3/4$ , how does this change prediction @  $c = 1/2$  ?

How much information is carried by knowledge of structure ?

## Mutual Information

$$I_{i,j} = \sum_{\mathbf{x}_i, \mathbf{x}_j} p(\mathbf{x}_i, \mathbf{x}_j) \log \left( \frac{p(\mathbf{x}_i, \mathbf{x}_j)}{p(\mathbf{x}_i)p(\mathbf{x}_j)} \right)$$

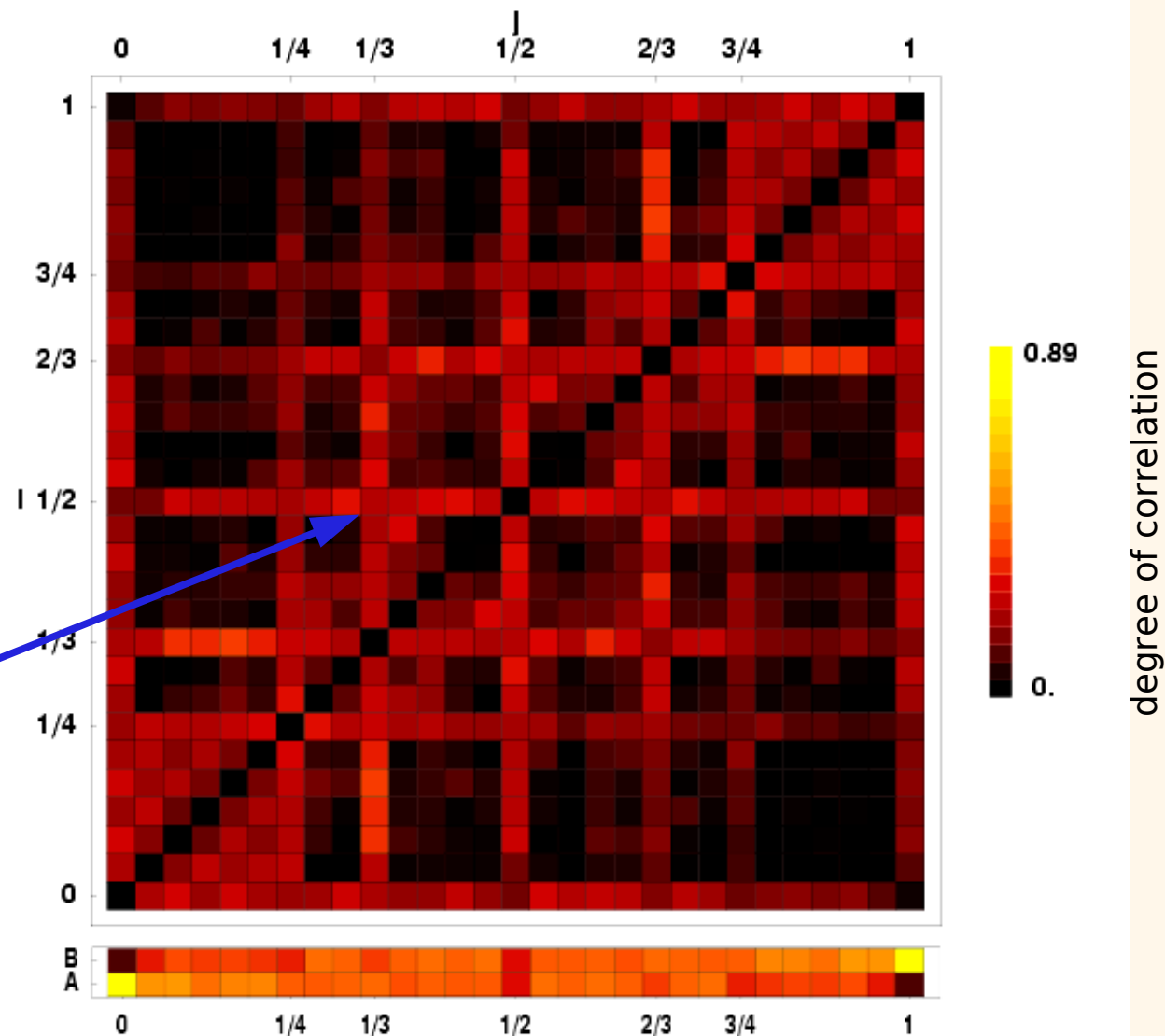
$$I_{i,j} = \left\langle \log [g_{ij}(\mathbf{x}_i, \mathbf{x}_j)] \right\rangle$$

# Information theory for structure stability

Each element of matrix is correlation between  $X_i$  and  $X_j$

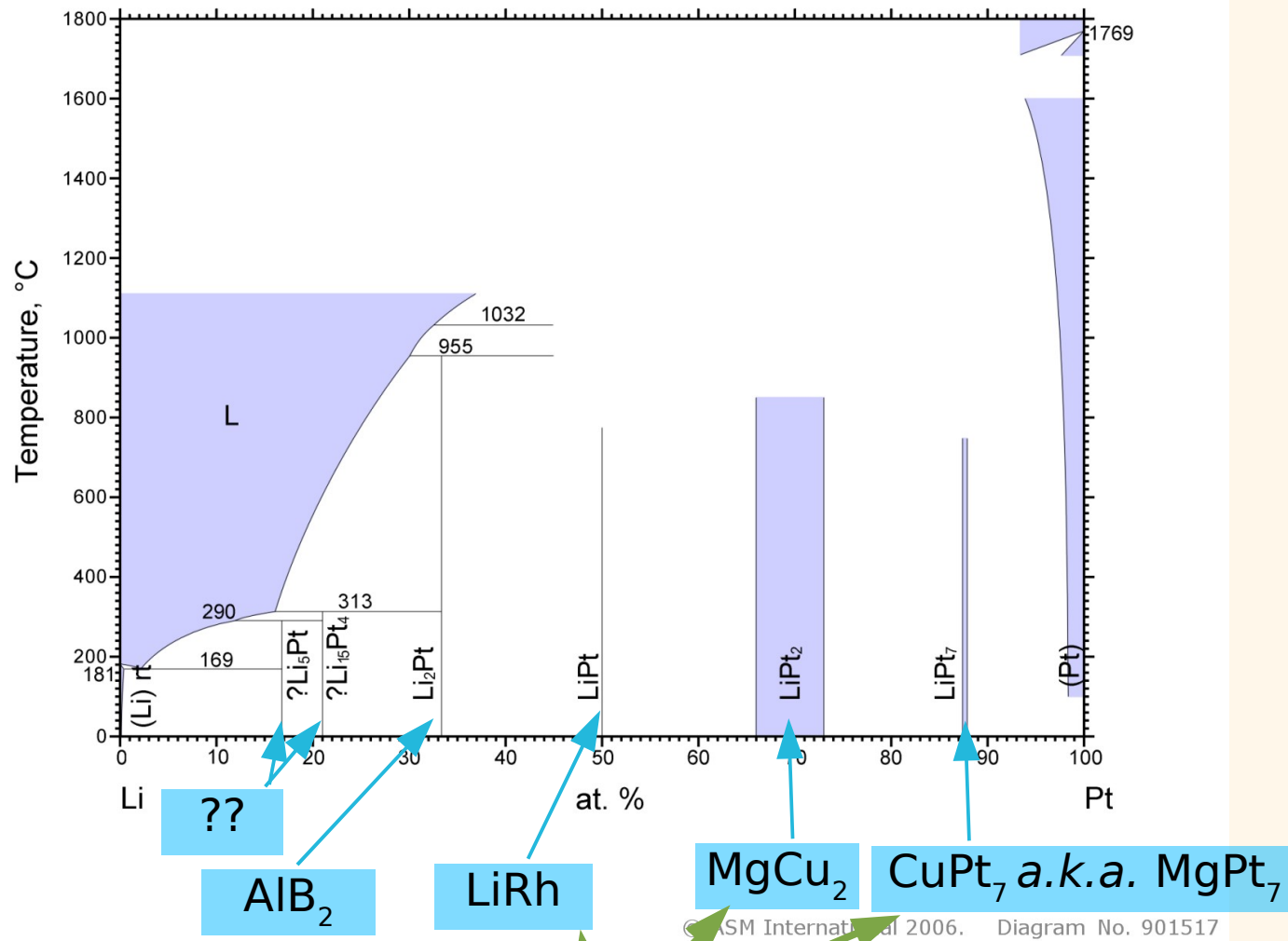
$$I_{i,j} = \sum_{x_i, x_j} p(x_i, x_j) \log \left( \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right)$$

e.g.,  
 $X_i$  = "AB prototype"  
and  
 $X_j$  = "A<sub>2</sub>B prototype"



## Prediction and validation in Li-Pt

# Predicting structures in Li-Pt



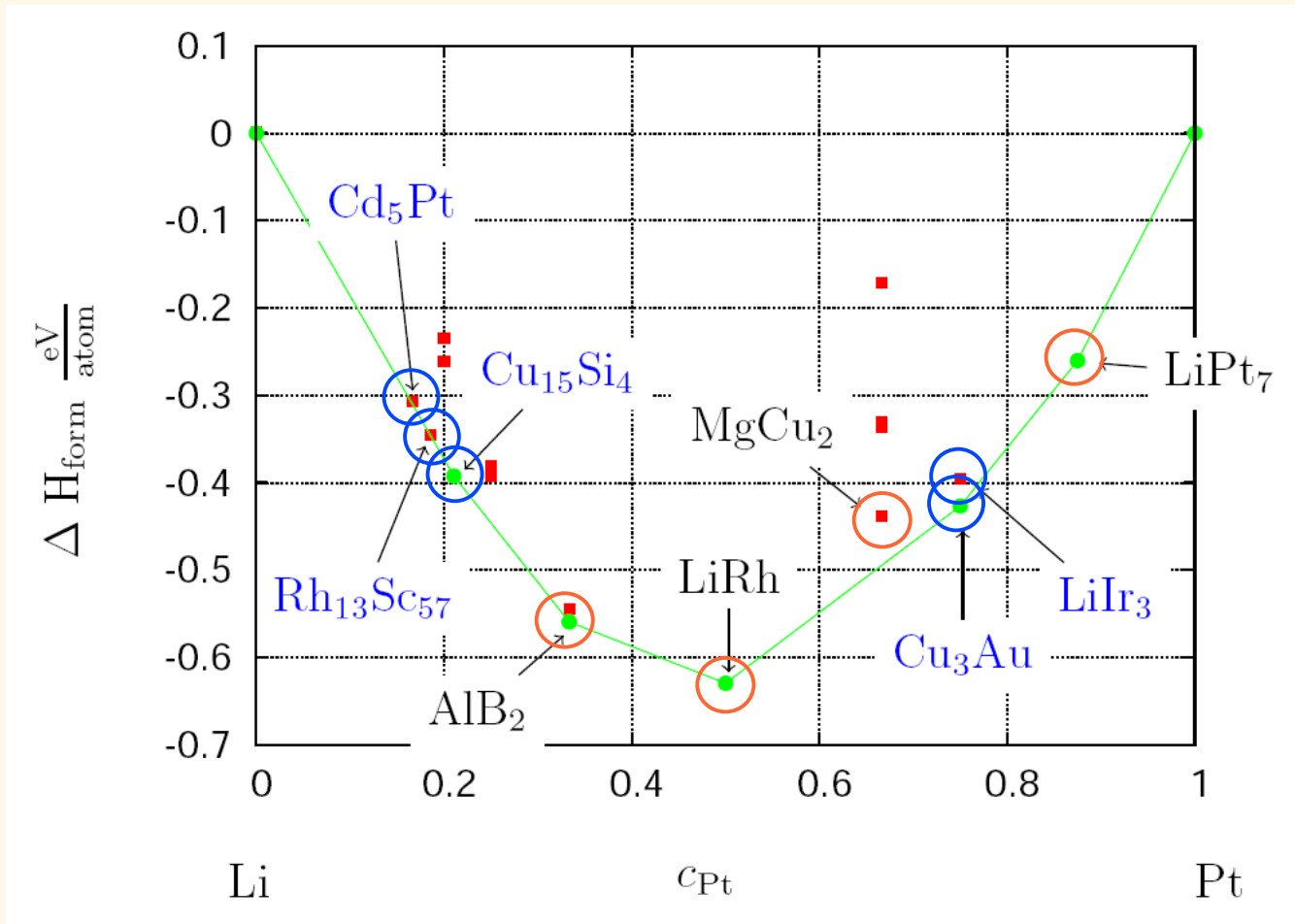
Use these as conditioning evidence for:

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

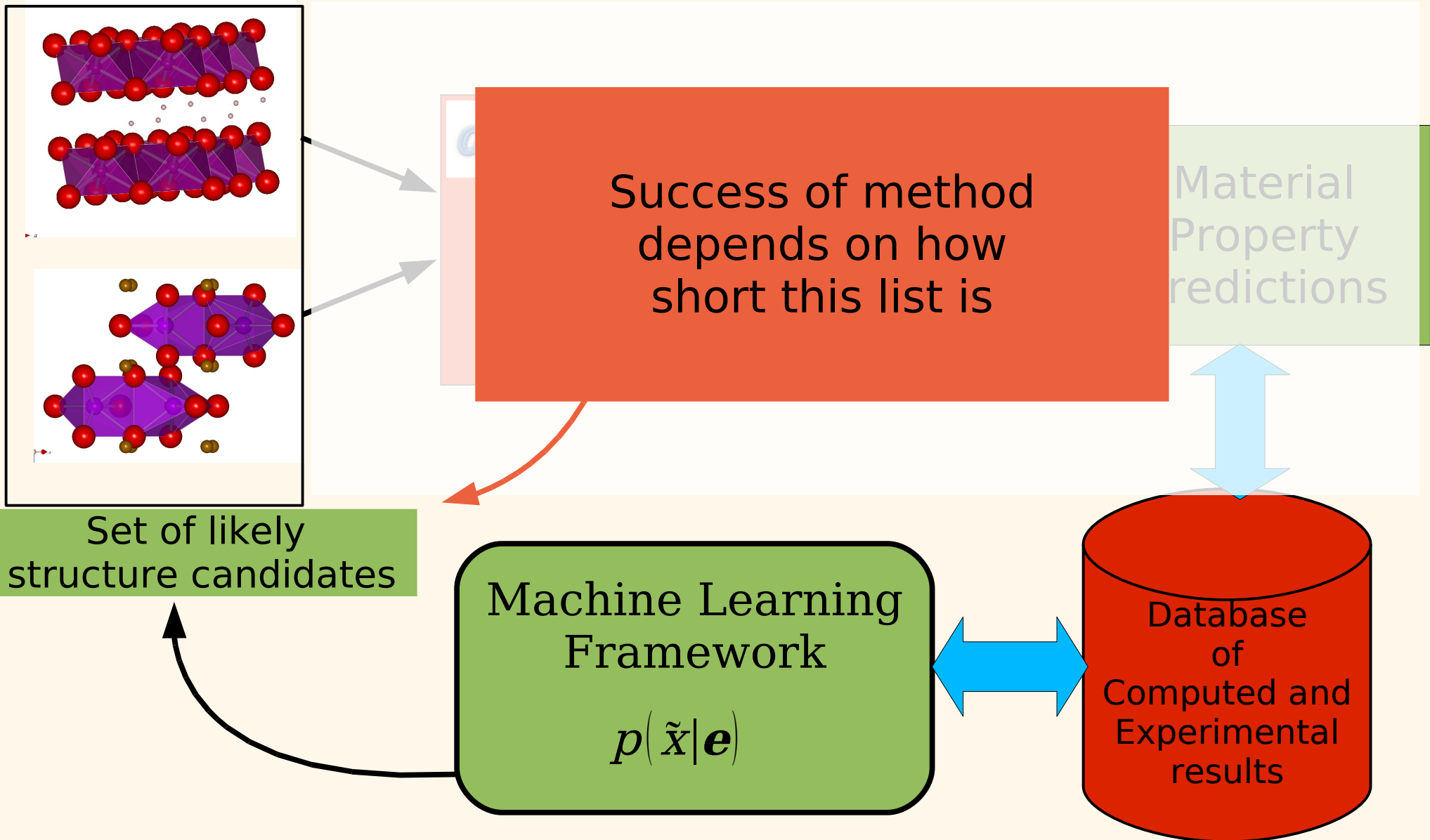
# Predicting structures in Li-Pt

Known phases

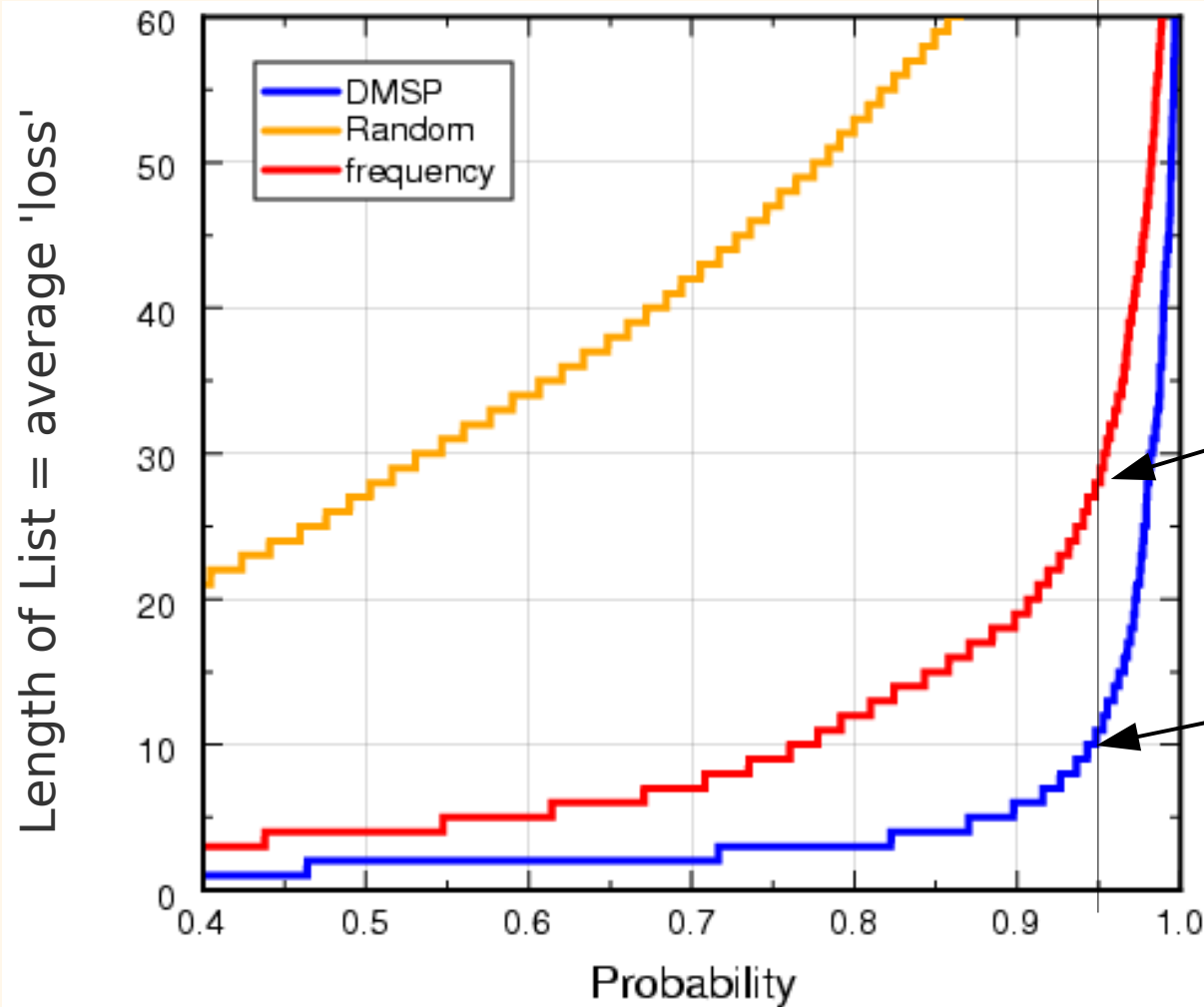
Suggested phases



# cross validation to evaluate performance



# Cross validation results



**Independent Variables**

**Including structure correlation**

Nature Materials, 6, 641-646, 2006

# Some open questions

**ICSD:** World's Largest database of inorganic crystal structures

The screenshot shows the ICSD website search interface. The browser window title is "ICSD for WWW - Mozilla Firefox". The address bar shows "http://icsdweb.fiz-karlsruhe.de/index.php". The search form includes the following fields and options:

Authors/Code	Years	Journal	Title/Comment	Help
Elements	Element Count	Chem/Mineral Name	ANX/Pearson/S.Type	Search Reset
System	Laue Class	Centering	Space Group	Wyckoff Sequence
Remarks	Min. Distance	Distance Select	Distance Range	Co-ordin.

Below the search form, there is a welcome message: "Welcome to the Inorganic Crystal Structure Database. Click the blue heading links for help and examples." The large "ICSD" logo is visible at the bottom of the page.

What is the information content in a chemical database?

How many 'independent' crystal structures exist in nature ?

First Entry: 1913

# of entries: 100,243

# usable compounds: 29,962

# structure prototypes: 2,485



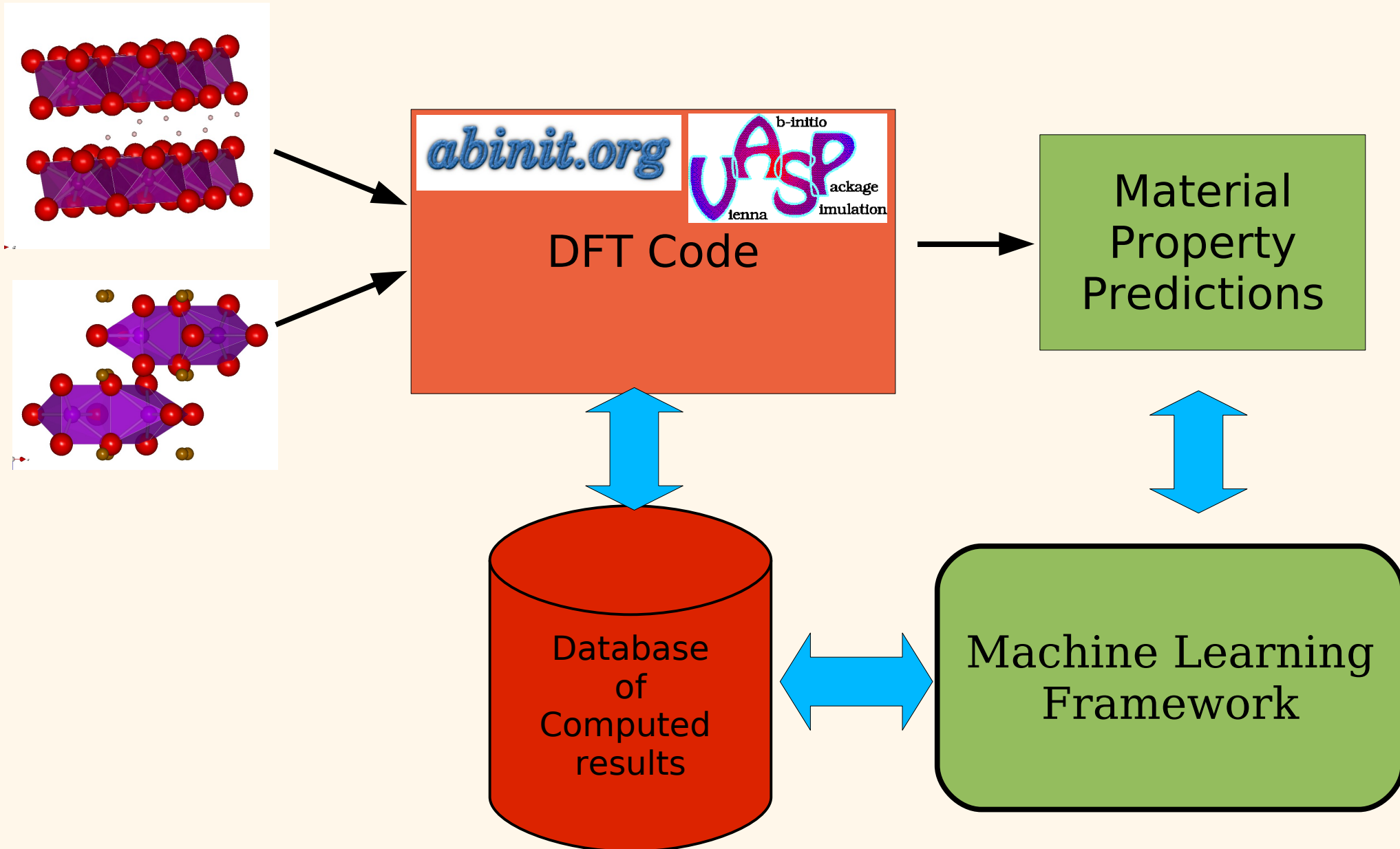
# Structure prediction: wrap-up

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

Much more needed  
here

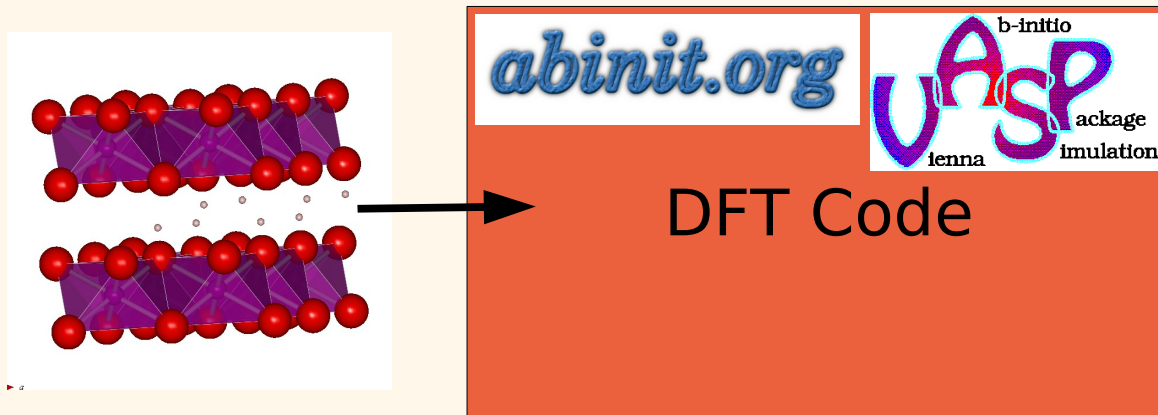
Now have efficient  
tool for this

# Directions for future work/collaboration



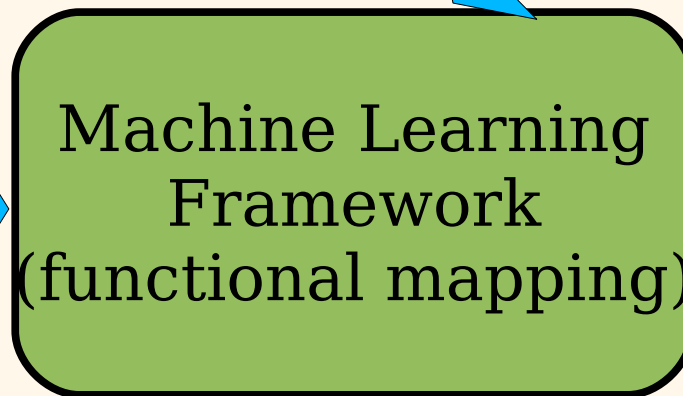
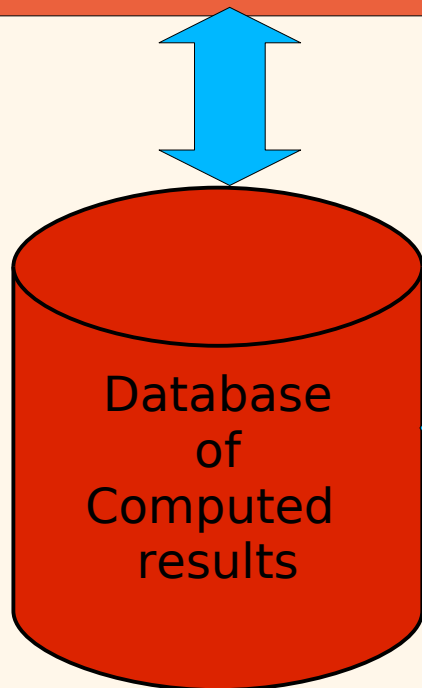
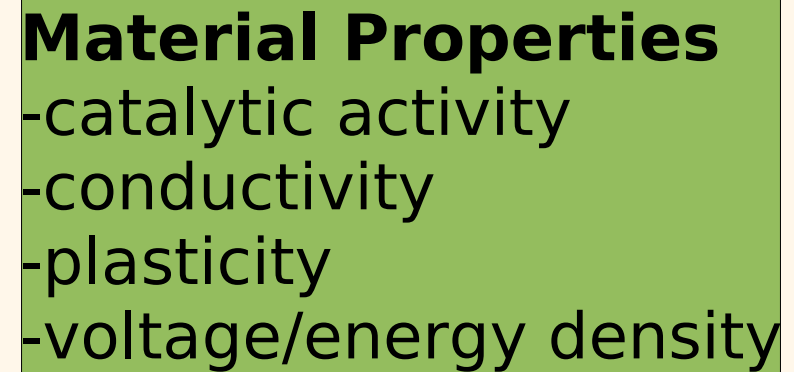
# Directions for future work/collaboration

## Set of features



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

# Directions for future collaboration



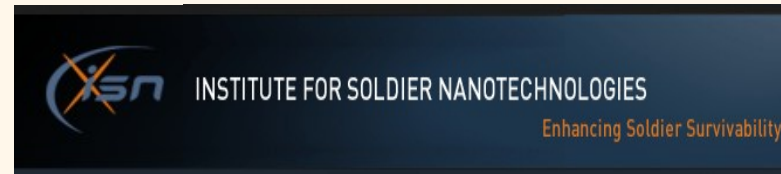
# The End

Data from High Throughput alloy study  
Online structure predictor

<http://datamine.mit.edu>



ITR grant (**DMR-031253**)



- 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