Big Data Summer
Platja d'Aro, Spain, September 9 - 13, 2019

Recommender system for materials discovery

Isao Tanaka[1,2,3,4]

[1] Department of Materials Science and Engineering, Kyoto University, JAPAN
[2] Elements Strategy Initiative for Structural Materials, Kyoto University, JAPAN
[3] Center for Materials Research by Information Integration, NIMS, JAPAN
[4] Nanostructure Research Laboratory, Japan Fine Ceramics Center, JAPAN

Inorganic Crystal Structure Database (ICSD)

World largest database
for known inorganic crystals.

187,000 crystal structures
→ 82,000 structures
excluding duplicates
incompletes, etc.

Number of chemical elements	Number of chemical combinations (only for simple composition ratio)
1	~100
2	~100,000
3	~10,000,000
4	~1,000,000,000 (1billion)

Many systems are yet-unexplored !

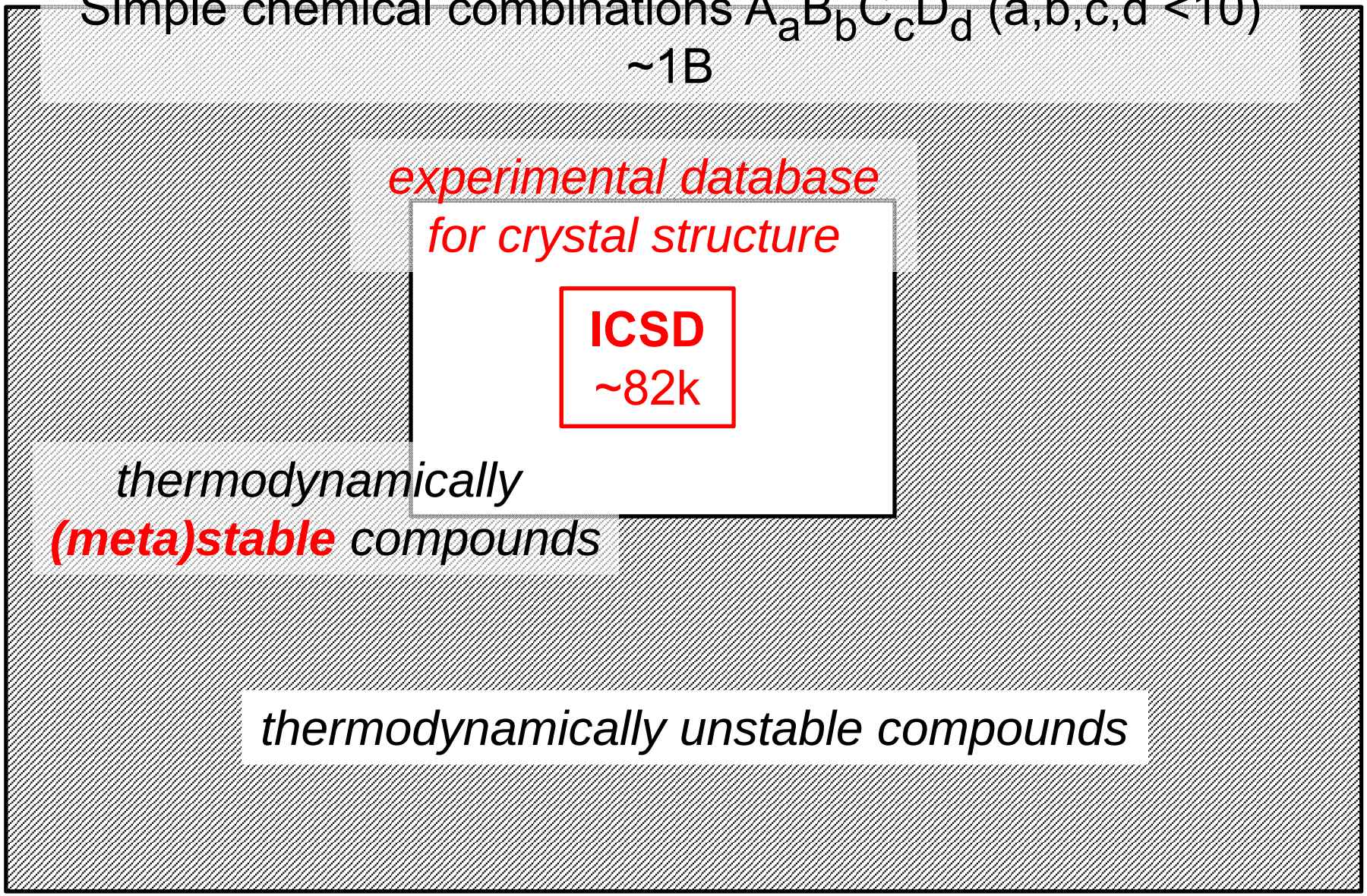
Vast chemistry space to explore

Simple chemical combinations $A_aB_bC_cD_d$ (a,b,c,d <10)
~1B

experimental database
for crystal structure

ICSD
~82k

Vast chemistry space to explore



Discovery of a novel Sn(II)-based oxide for daylight-driven photocatalyst

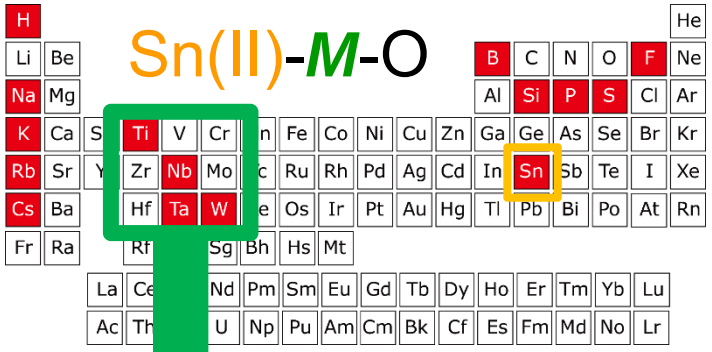
DFT calcs + Experiments

Hiroyuki Hayashi, Shota Katayama, Takahiro Komura, Yoyo Hinuma, Tomoyasu Yokoyama, Kou Mibu, Fumiyasu Oba and IT

Hiroyuki Hayashi

Advanced Science 9, (2016) 1600246

Target compounds of interests; Sn(II)-M-oxides



4A – 6A transition metal oxides

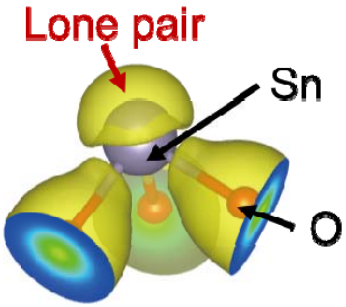
widely used for photocatalysts

ex. TiO_2, WO_3, $NaTaO_3$, TaON, …

Wide band-gaps

Sn(II) oxides

Narrow band-gaps



$SnO\text{-}MO_{q/2}$ pseudobinary

q	*M*	known compounds
4	Ti, Zr, Hf	$SnTiO_3$, Sn_2TiO_4
5	V, Nb, Ta	$SnNb_2O_6$, $Sn_2Nb_2O_7$, $SnTa_2O_6$, $Sn_2Ta_2O_7$, $SnTa_4O_{11}$
6	Cr, Mo, W	$SnWO_4$, Sn_2WO_5, Sn_3WO_6

Only 10 compounds are known

Reported high visible-light photocatalytic activity

Inorganic Crystal Structure Database (ICSD)

World largest database
for known inorganic crystals.

177,000 crystal structures
→ 82,000 structures
excluding duplicates,
incompletes, etc.

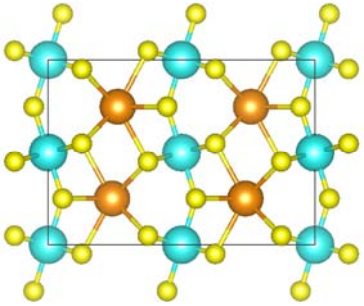
9,100 structure prototypes
(e.g. rock-salt, perovskite, ...)

Number of chemical elements	Number of structure prototypes in ICSD
1	120
2	1,700
3	4,700
4	4,300

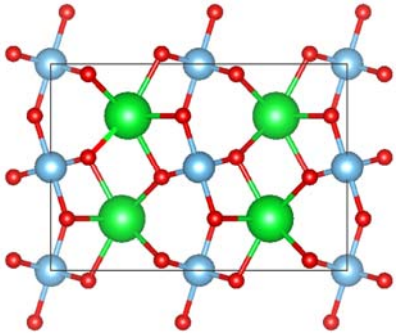
Hypothetical compounds with prototype structures

ICSD prototype

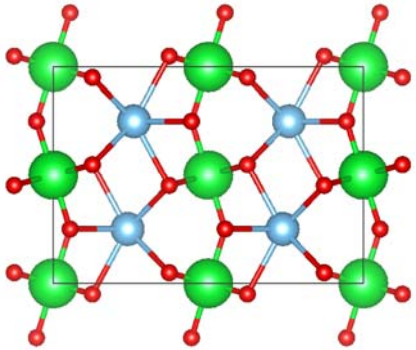
$NdYbS_3$ type



$NdYbS_3$ type $SnTiO_3$



$NdYbS_3$ type $TiSnO_3$



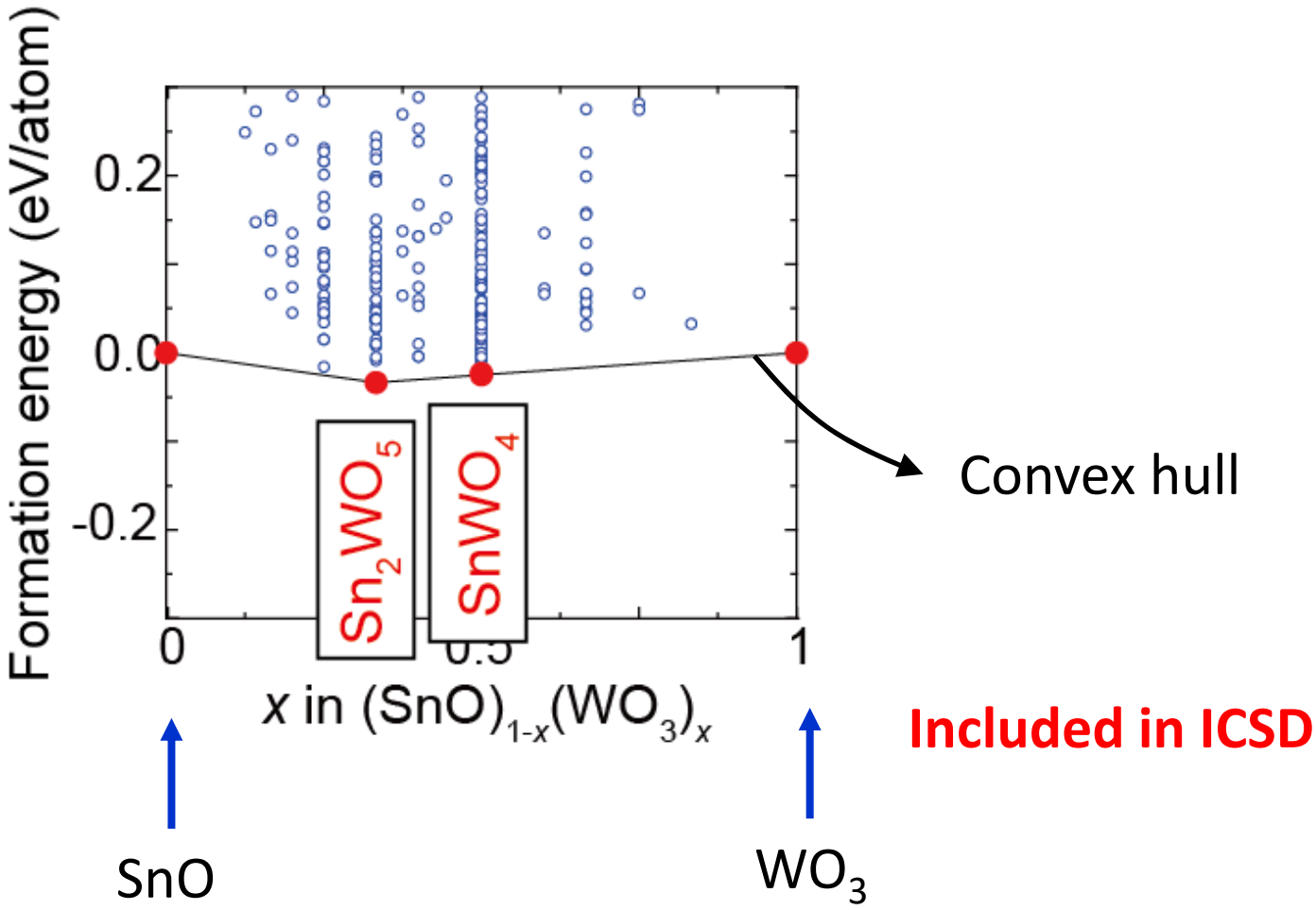
hypothetical compounds

Formal ionic charge

Formal ionic charge	1	2	3	4	5	6
1	154	122	359	209	438	251
2	454	258	663	220	409	
3	500	184	297	109		
4	444	52	149			
5	72	45				
6	78					

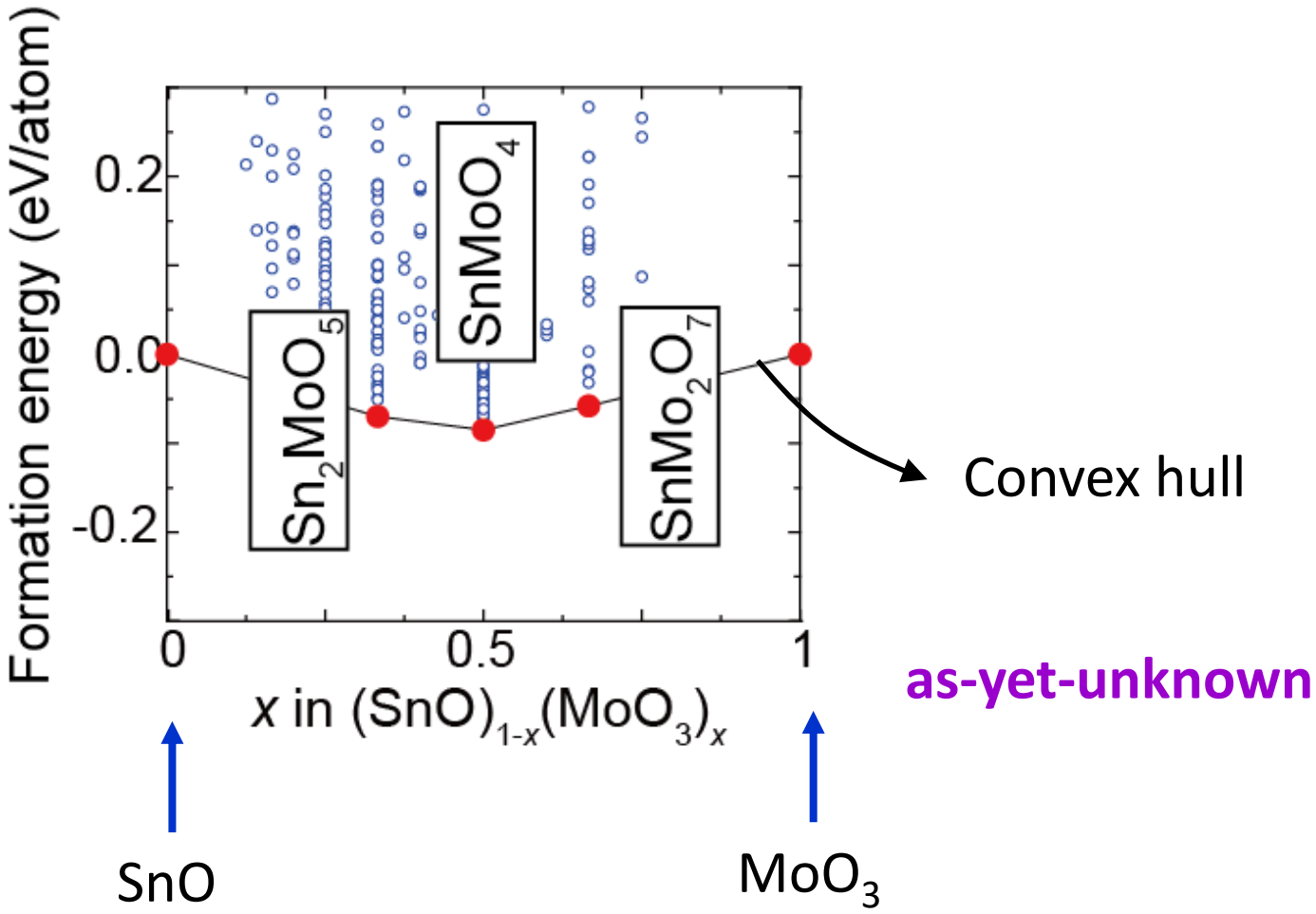
Formation energy by DFT calcs

SnO-WO_3 pseudo binary system



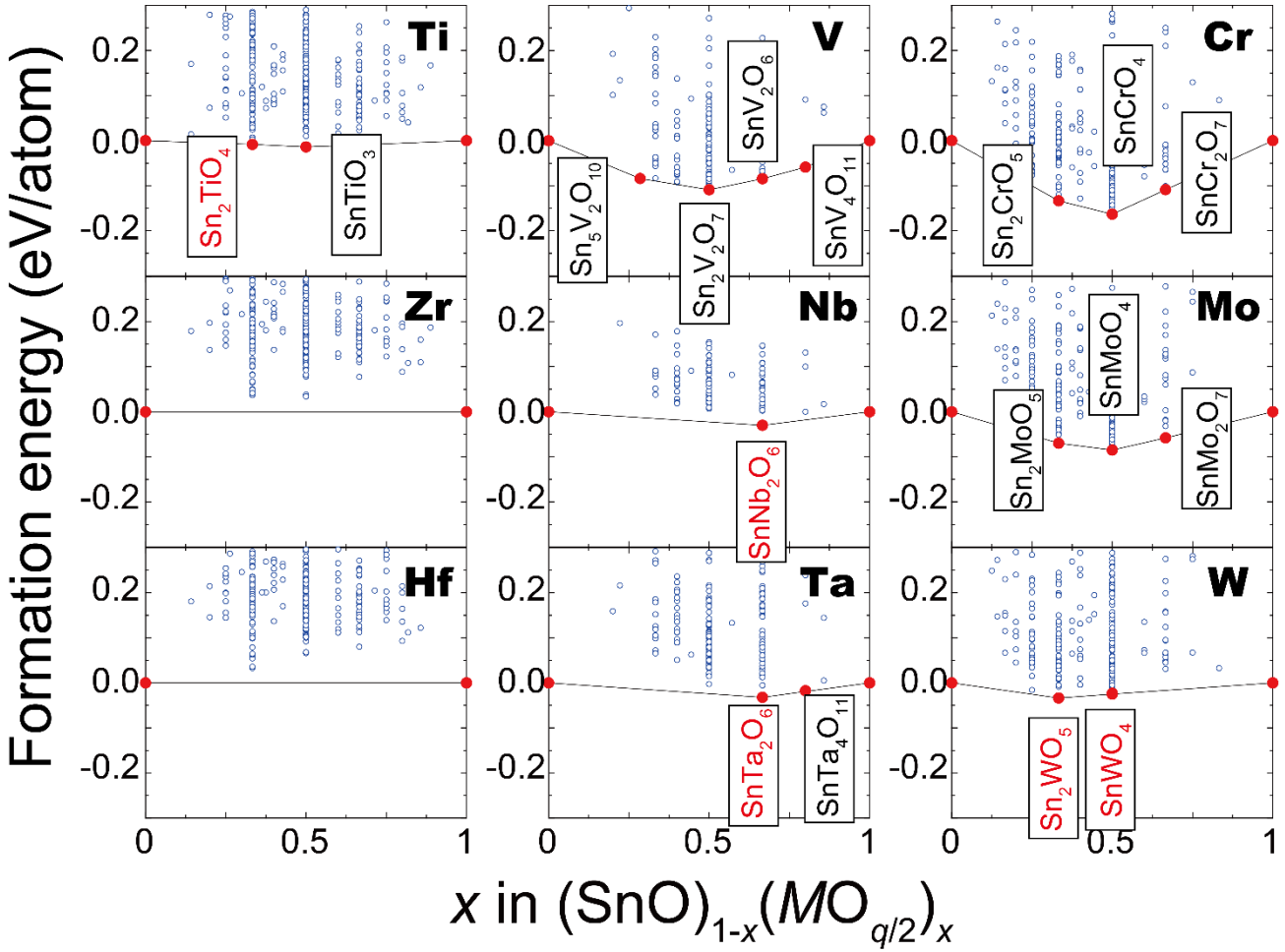
Formation energy by DFT calcs

SnO-MoO_3 pseudo binary system



Formation energy by DFT calcs

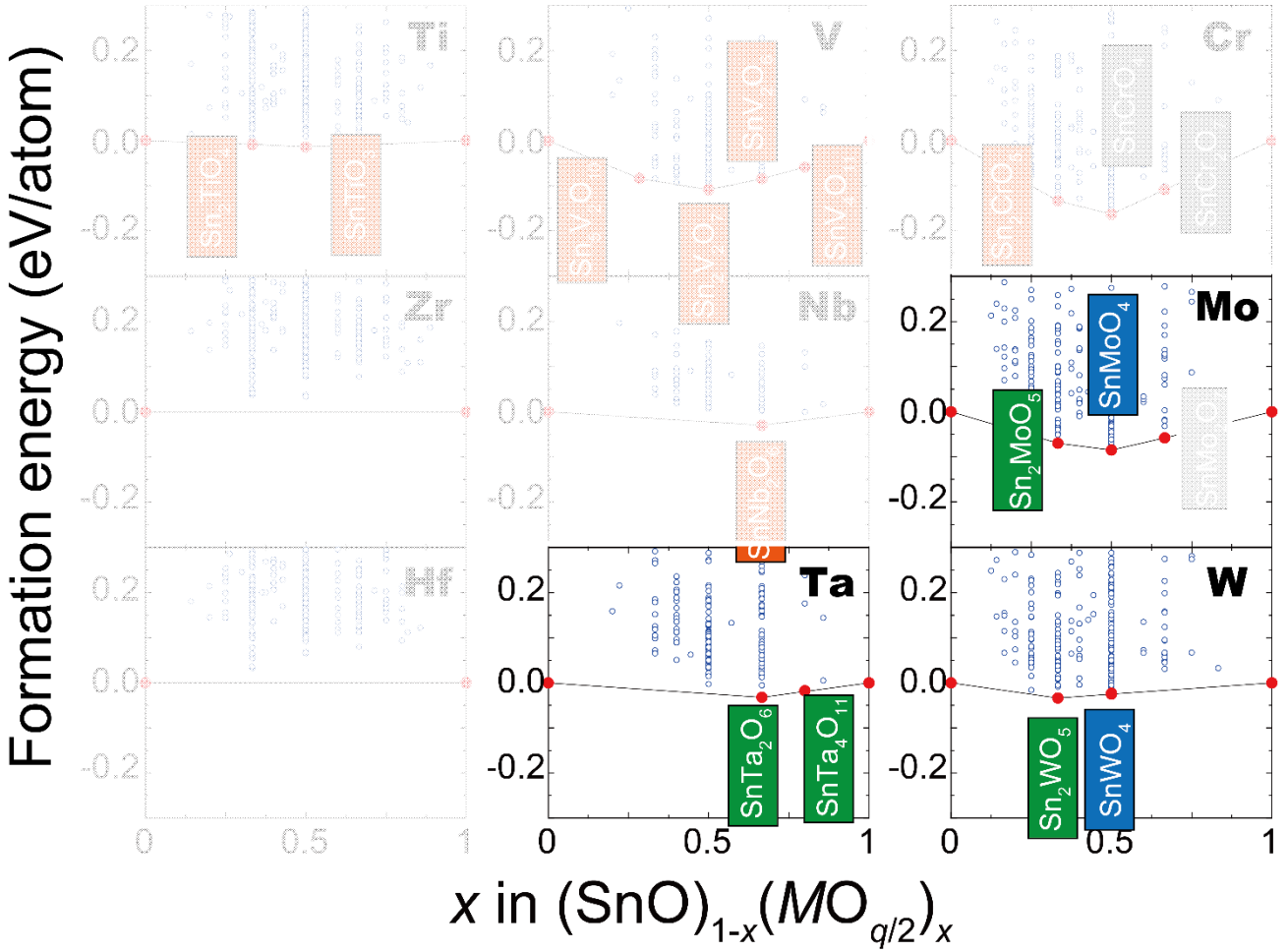
Convex hull of SnO-$MO_{q/2}$ pseudo binary systems

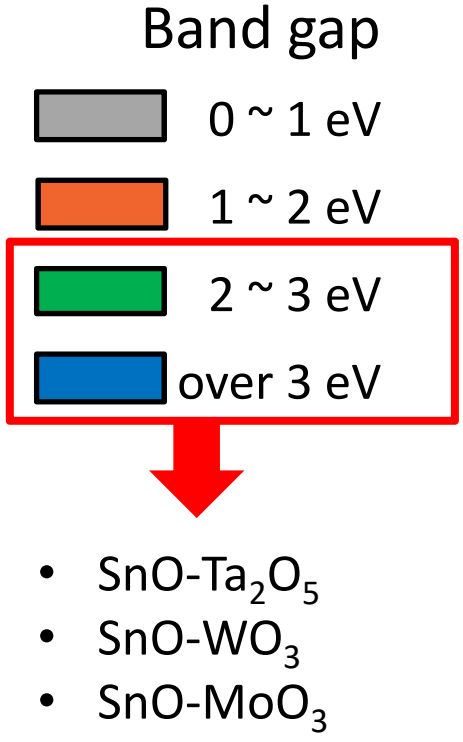


Reported oxides in ICSD (Red characters) are located on the convex hull.

Band Gap

Band gap of actual photocatalysts ≥ 2 eV (GGA)





- $SnO\text{-}Ta_2O_5$
- $SnO\text{-}WO_3$
- $SnO\text{-}MoO_3$

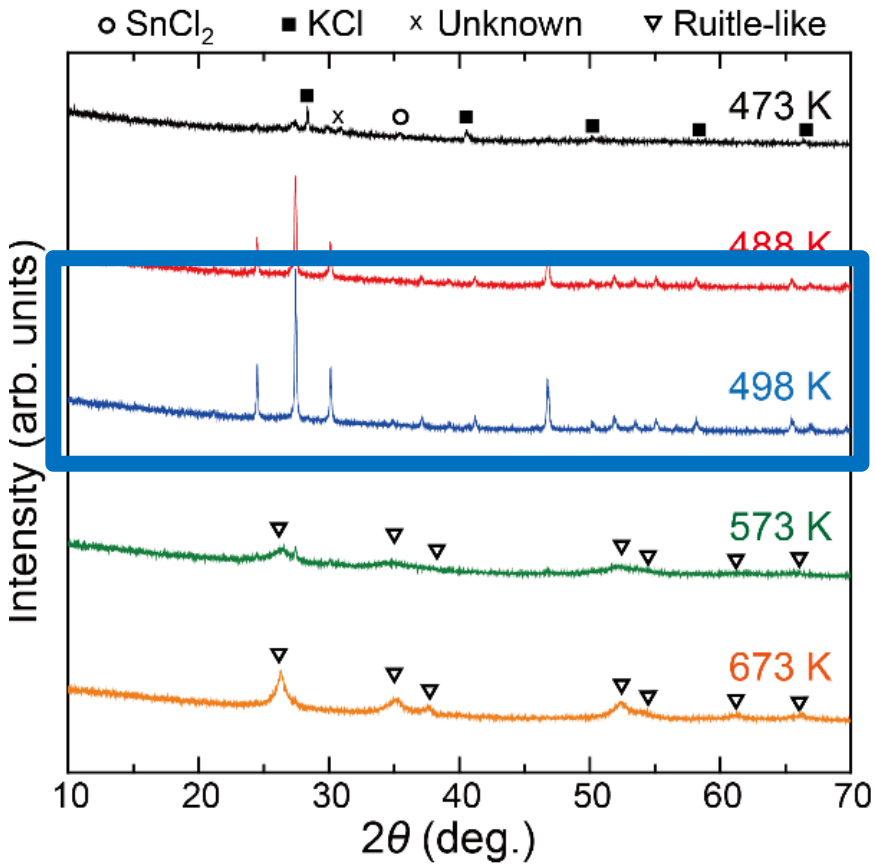
Experimental results

Synthesis of $SnMoO_4$

Mixture of $SnCl_2$ and K_2MoO_4 powders

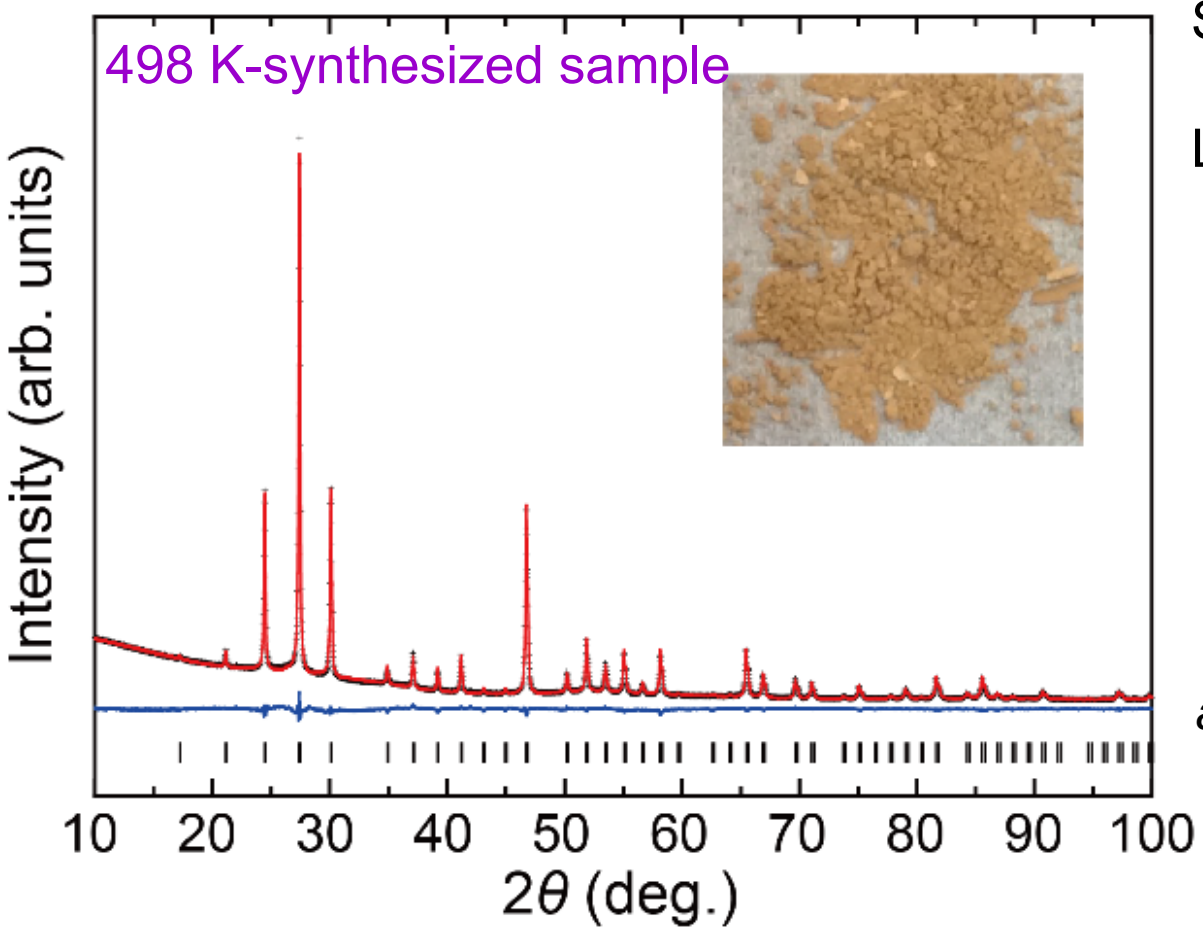
1 hour annealing in Ar gas

Washed and dried



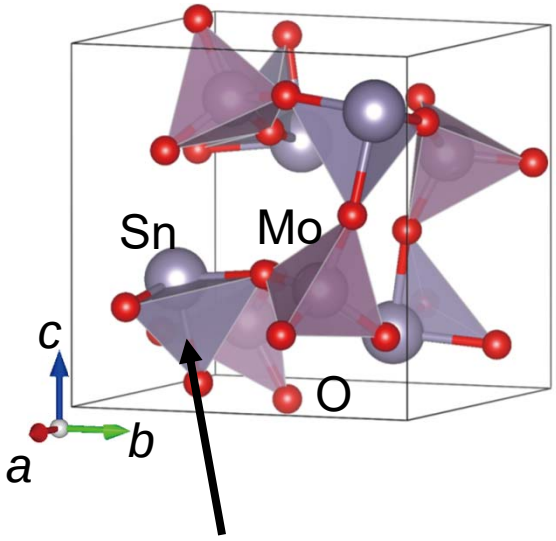
Crystal structure of $SnMoO_4$

Newly discovered compound



Space group type: $P2_13$ (Cubic)

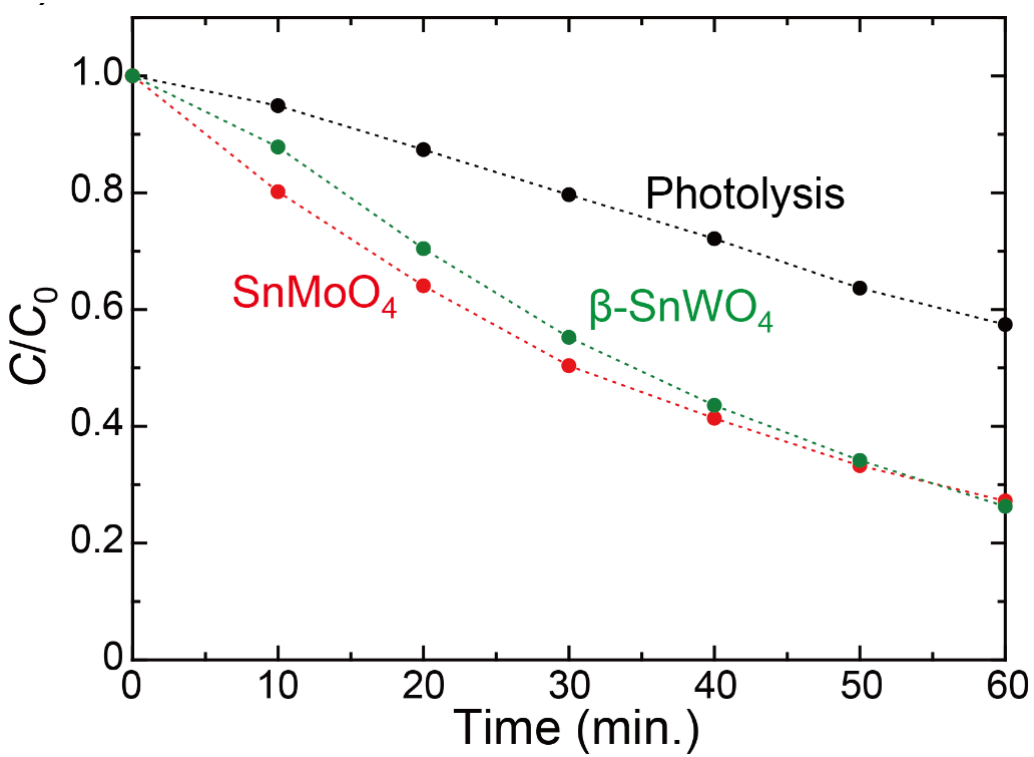
Lattice constant: a = 7.26 Å



Trigonal prism which is characteristic of Sn(II)

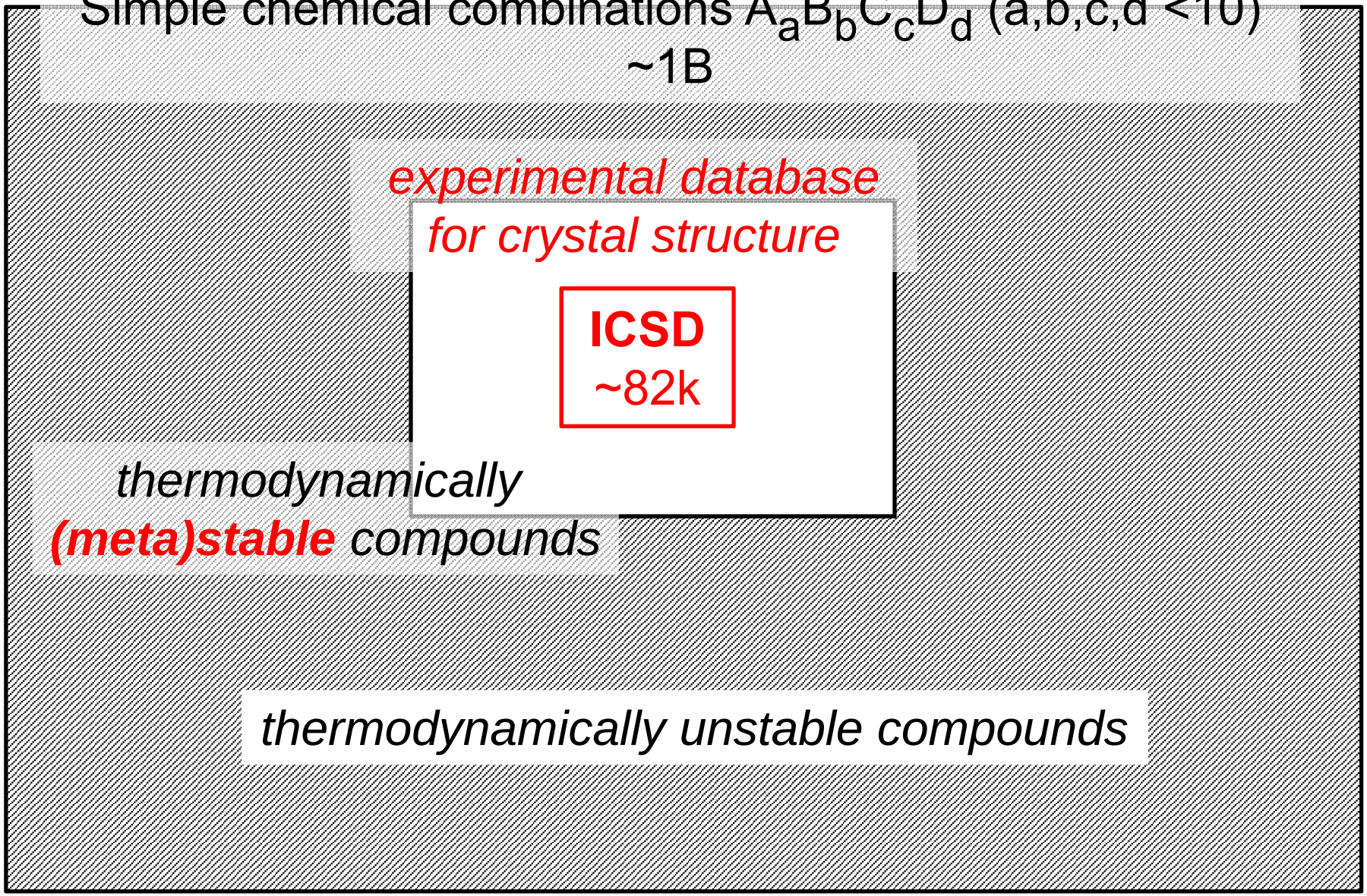
Photocatalytic activity of $SnMoO_4$

Degradation of methylene blue under simulated day-light



Newly-discovered $SnMoO_4$ powder exhibits clear photocatalytic activity.

Vast chemistry space to explore



Recommender system for discovery of CRC (Chemically Relevant Composition) using ICSD database



A. Seko, H. Hayashi, H. Kashima and IT

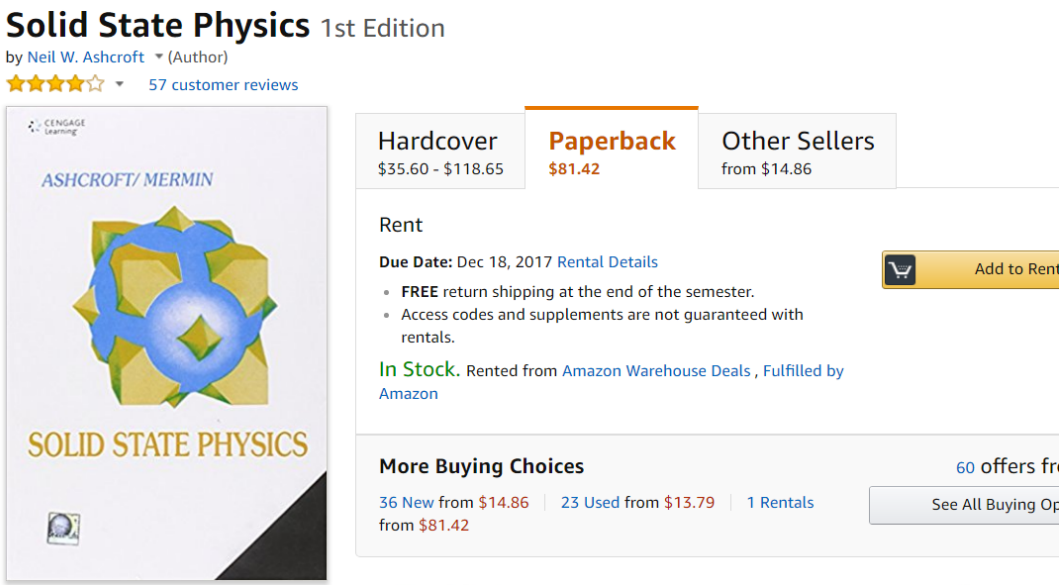
A. Seko, H. Hayashi, H. Kashima, I. Tanaka, *Phys. Rev. Mater*. 2, 013805 (2018)
A. Seko, H. Hayashi, and I. Tanaka, *J. Chem. Phys*. 148, 241719 (2018).

"Recommender system" in E-commerce

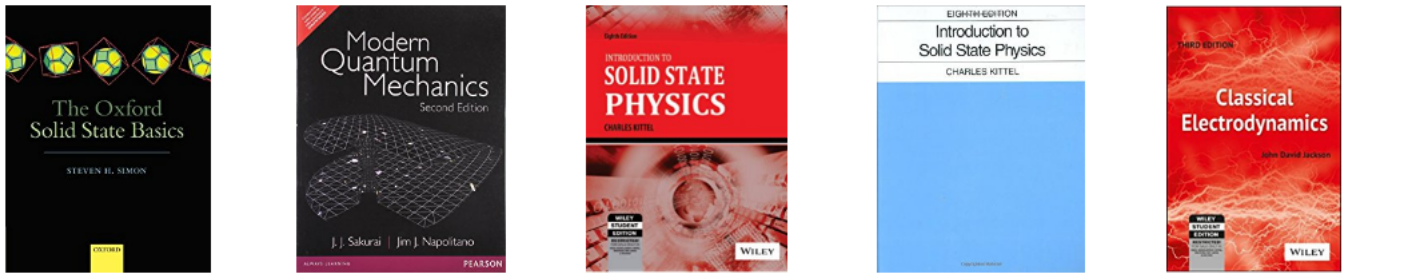
A system that can suggest items to customers, which is sometimes useful.

Amazon.com

Netflix.com



Customers who bought this item also bought = **Recommendation**



The Oxford Solid State Basics
› Steven H. Simon
20
Paperback
$31.99

Modern Quantum Mechanics: PNIE
› J. J. Sakurai
34
Paperback
75 offers from $17.04

Introduction to Solid State Physics
› Charles Kittel
92
Paperback
54 offers from $14.12

Introduction to Solid State Physics
› Charles Kittel
92
Hardcover
$164.93 prime

Classical Electrodynamics
› John David Jackson
171
Paperback
54 offers from $17.29

CRC (Chemically Relevant Composition)

A_2X-BX pseudo-binary (A^{1+}, B^{2+}, X^{2-})

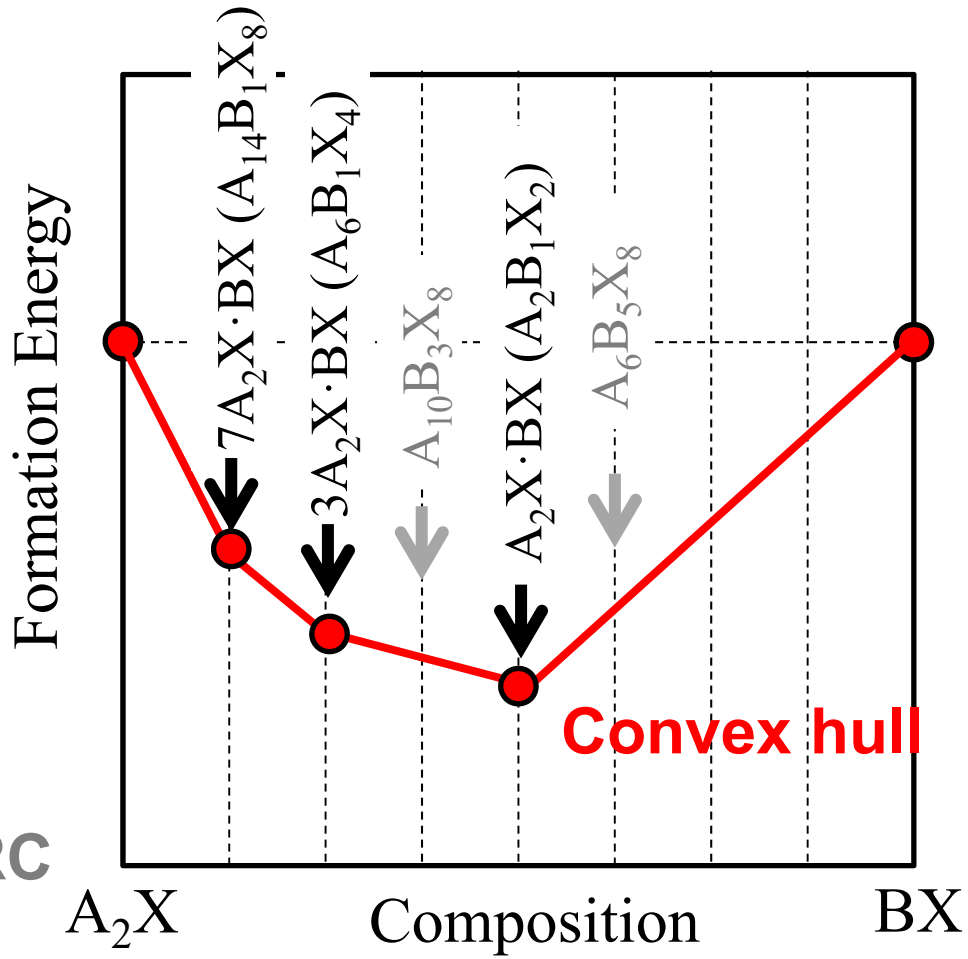
FIZ Karlsruhe ICSD

- $7A_2X{\cdot}1BX$ ($A_{14}B_1X_8$)
- $3A_2X{\cdot}1BX$ ($A_6B_1X_4$)
- $1A_2X{\cdot}1BX$ ($A_2B_1X_2$)

CRC

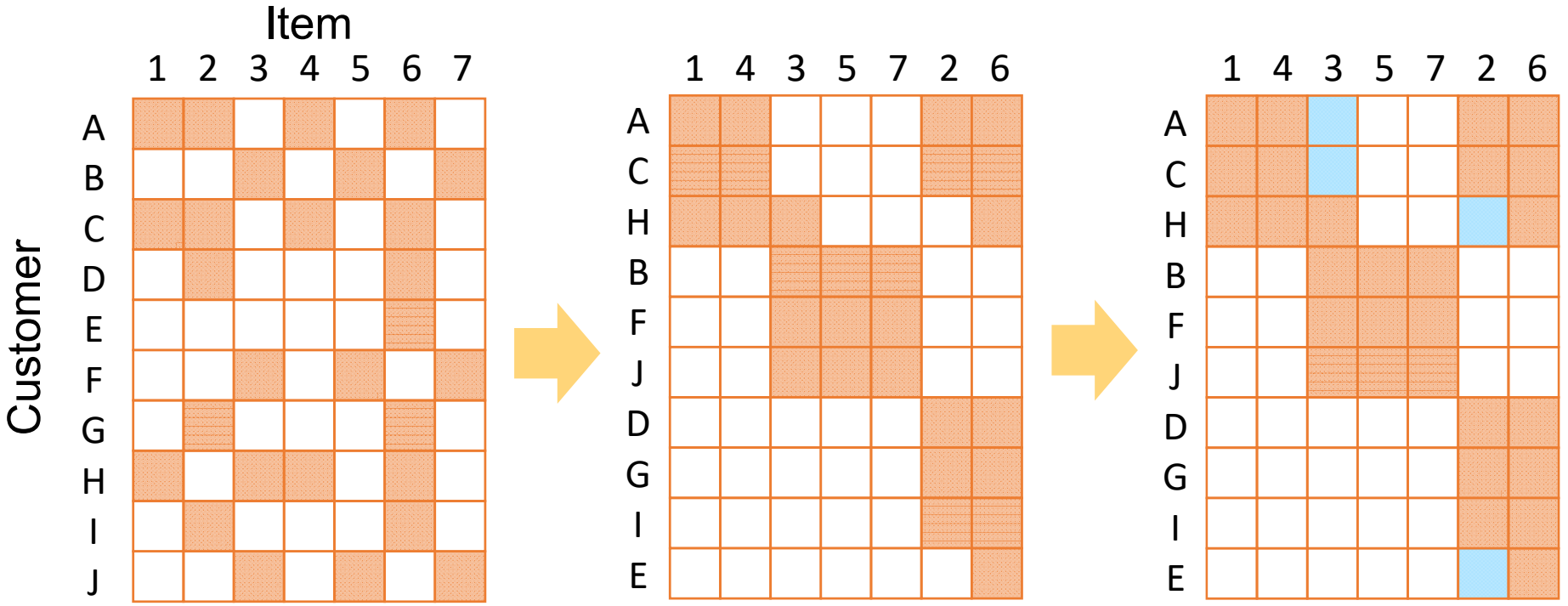
- $5A_2X{\cdot}3BX$ ($A_{10}B_3X_8$)
- $3A_2X{\cdot}5BX$ ($A_6B_5X_8$)

non-CRC



Rating matrix used for recommender system

Rating matrix



Underlying assumption: a low-rank structure of rating matrix.

⇒ **Application to discover new Chemically Relevant Composition (CRC)**

Candidate chemical compositions

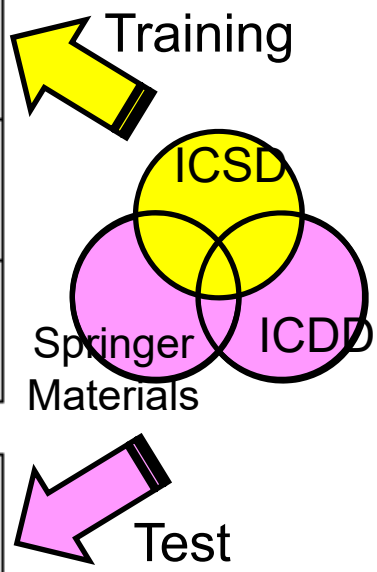
Ternary : $A_aB_bX_x$	max(a, b, x) = 8,	$N = 7.4 \times 10^6$
Quaternary: $A_aB_bC_cX_x$	max(a, b, c, x) = 20,	$N = 1.2 \times 10^9$
Quinary: $A_aB_bC_cD_dX_x$	max(a, b, c, d, x) = 20,	$N = 2.3 \times 10^{10}$

Cations	Li,Na,K,Rb,Cs,Be,Mg,Ca,Sr,Ba,Zn,Cd,Hg,
	B,Al,Sc,Y,La,Ga,In,Tl,Si,Ge,Sn,Pb,P,As,Sb,Bi
	Ti,Zr,Hf,V,Nb,Ta,Cr,Mo,W,Mn,Tc,Re,
	Fe,Ru,Os,Co,Rh,Ir,Ni,Pd,Pt,Cu,Ag,Au,
	Ce,Pr,Nd,Pm,Sm,Eu,Gd,Tb,Dy,Ho,Er,Tm,Yb,Lu
Anions	C,N,O,S,Se,Te,F,Cl,Br,I

Number of entry compounds in three databases

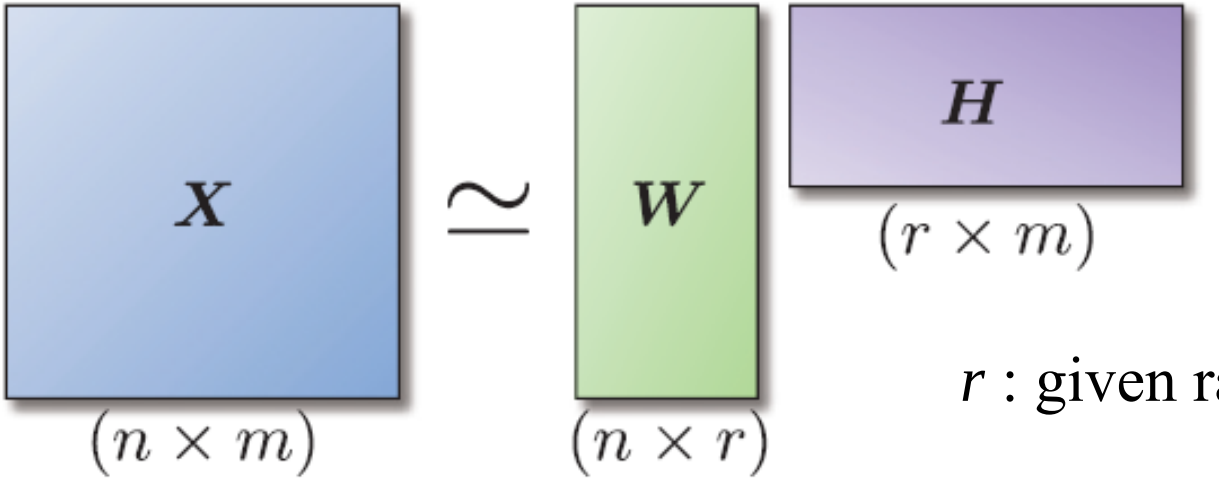
Number of entry compounds

	Ternary	Quaternary	Quinary
FIZ Karlsruhe ICSD	9,313	7,742	1,321
ICDD	9,278	7,864	1,326
Springer Materials Databases	10,461	8,141	1,893
Test data	4,134	4,961	1,616
Candidates	7,405,200	1,188,038,460	23,104,706,560



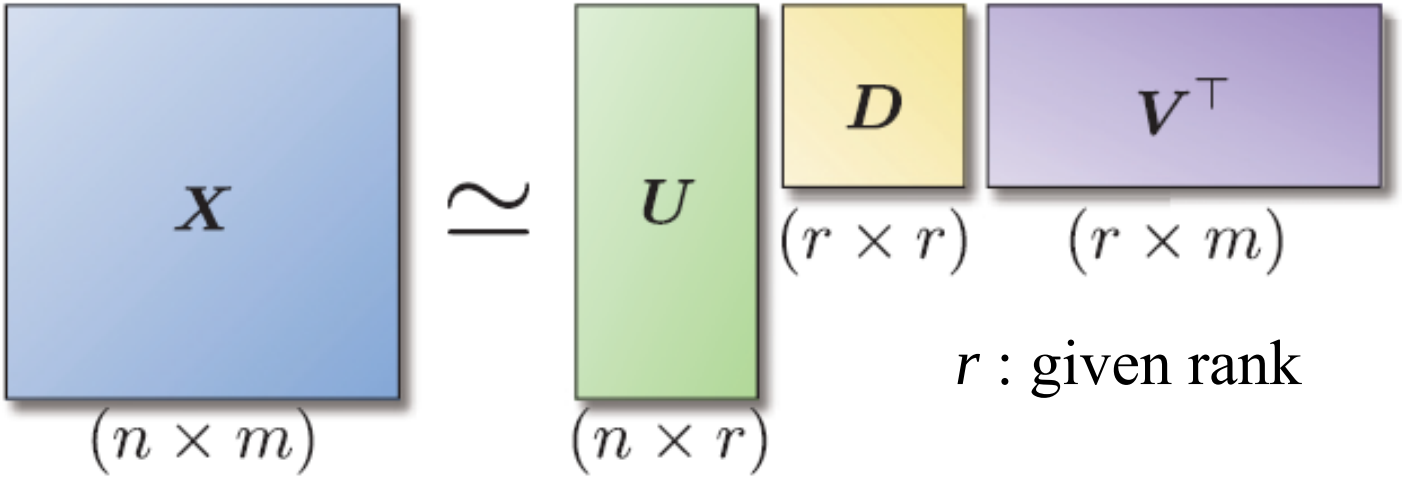
Matrix factorization

(a) NMF Non-negative Matrix Factorization



r : given rank

(b) SVD Singular Value Decomposition



r : given rank

Matrix representation of ternary composition

$A_aB_bX_x$

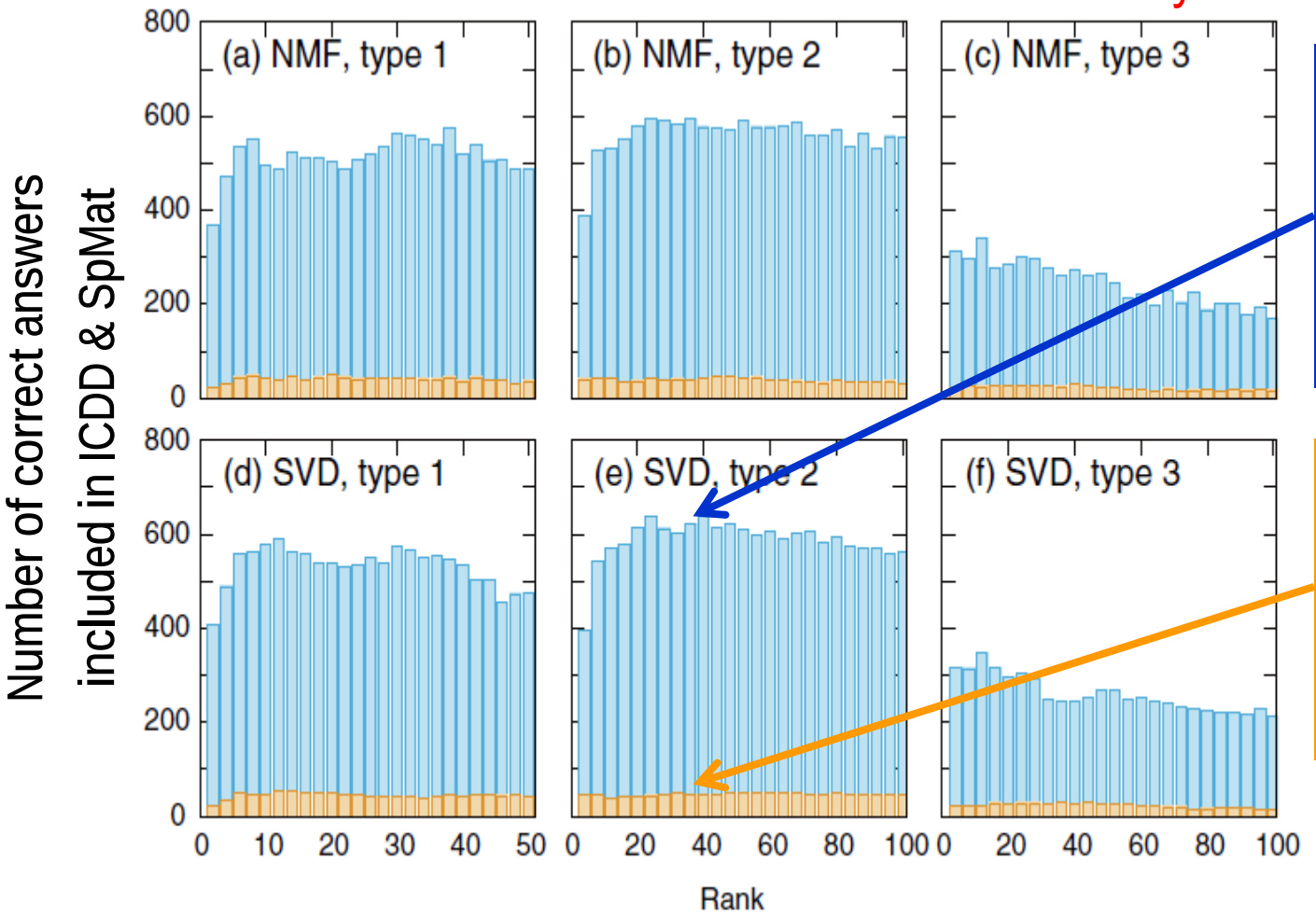
- Type 1 $\{A\}$ and $\{B, X, (a,b,x)\}$
- Type 2 $\{A, X\}$ and $\{B, (a,b,x)\}$
- Type 3 $\{A, B\}$ and $\{X, (a,b,x)\}$

Example of Rating Matrix (Type 1)

$$\boldsymbol{X} = \begin{array}{c} \begin{matrix} \text{Li,O} & \text{Li,O} & \text{Li,O} & \text{Li,O} & \text{Li,O} & & \text{In,O} & \text{In,O} & \text{In,O} & \text{In,O} & \text{In,O} \\ (1,1,1) & (1,1,2) & (1,2,2) & (1,3,3) & (1,4,3) & & (1,1,2) & (1,1,3) & (1,2,4) & (2,2,5) & (3,3,8) \end{matrix} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 & 0 & \cdots & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} \text{Na} \\ \text{Mg} \\ \text{Al} \\ \\ \text{Cu} \\ \text{Zn} \\ \text{Ga} \end{matrix}$$

Validation of CRC prediction by a recommender system for ternary compounds using Tucker decomposition

Ternary # Elements： 7,405,200

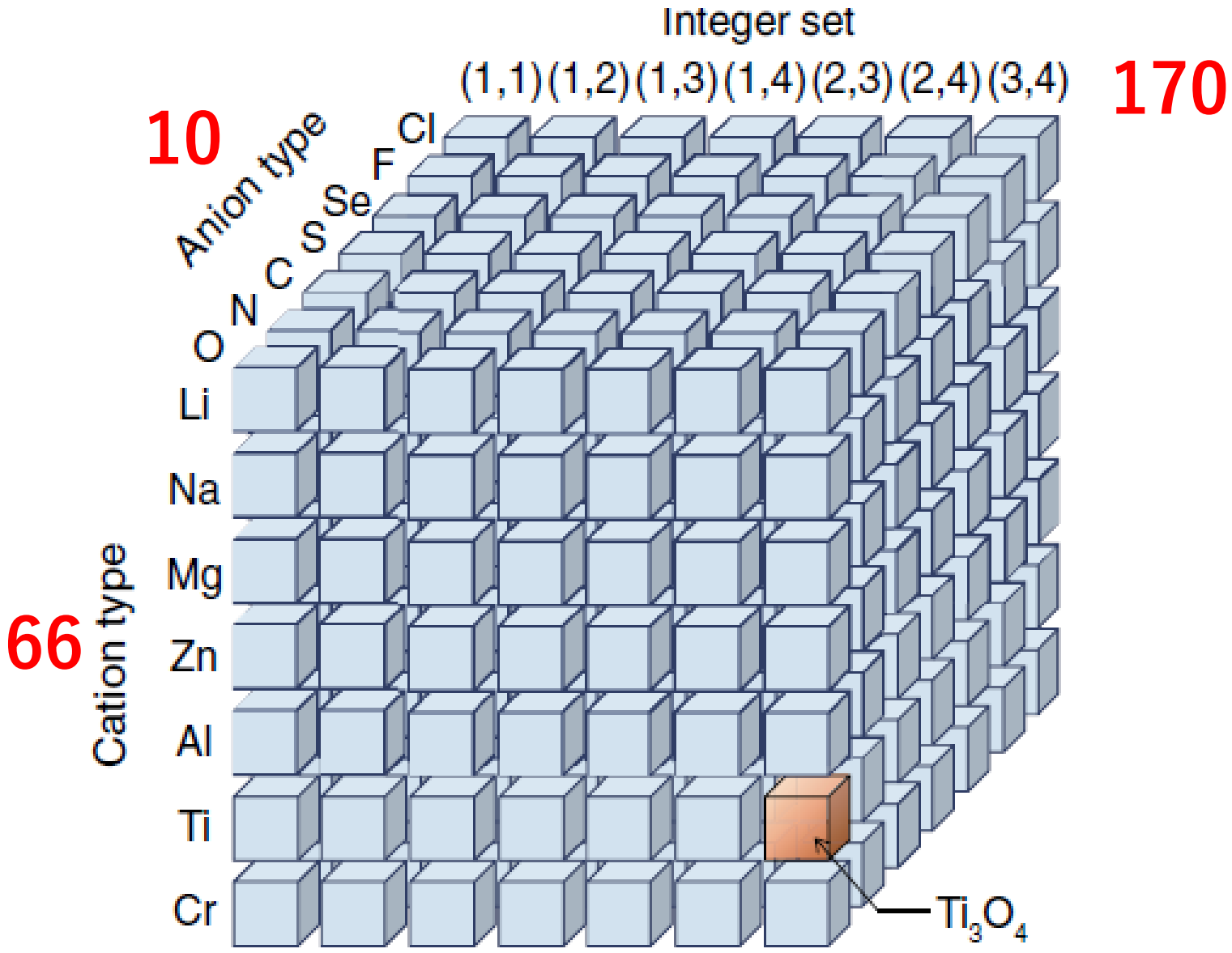


TOP3,000 compositions with high predicted rating.
Discovery rate
> 21% !!

TOP100 compositions with high predicted rating.
Discovery rate
> 45% !!

- ✓ Dependence on rank is weak.
- ✓ SVD performs slightly better than NMF.
- ✓ Type 2 representation works best.

Tensor representation of binary composition

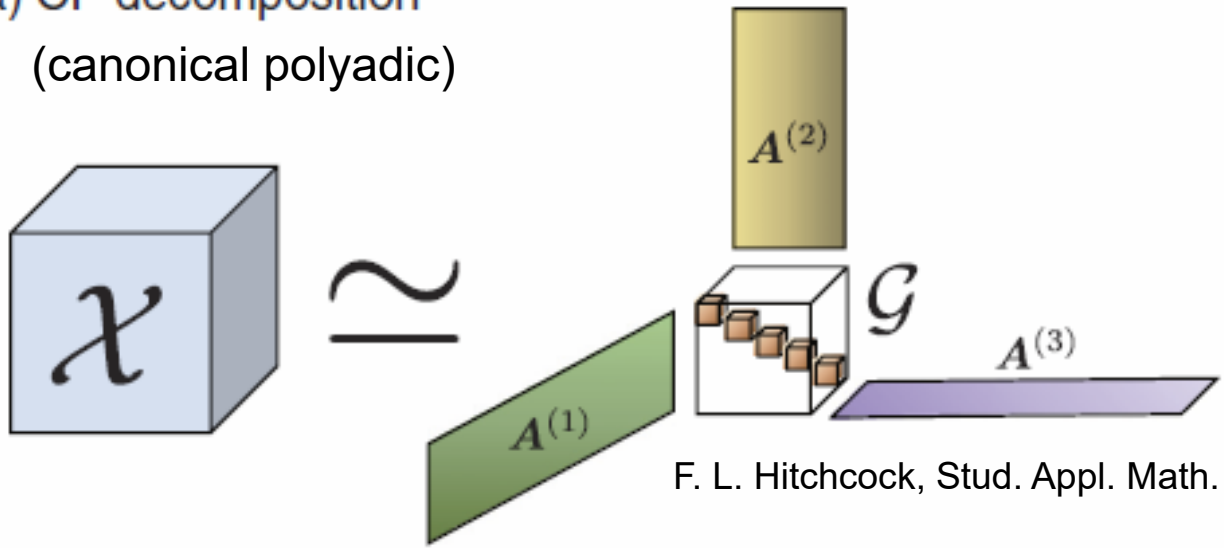


Binary: # Elements：66x10x170=112,200

Tensor factorization

(a) CP decomposition

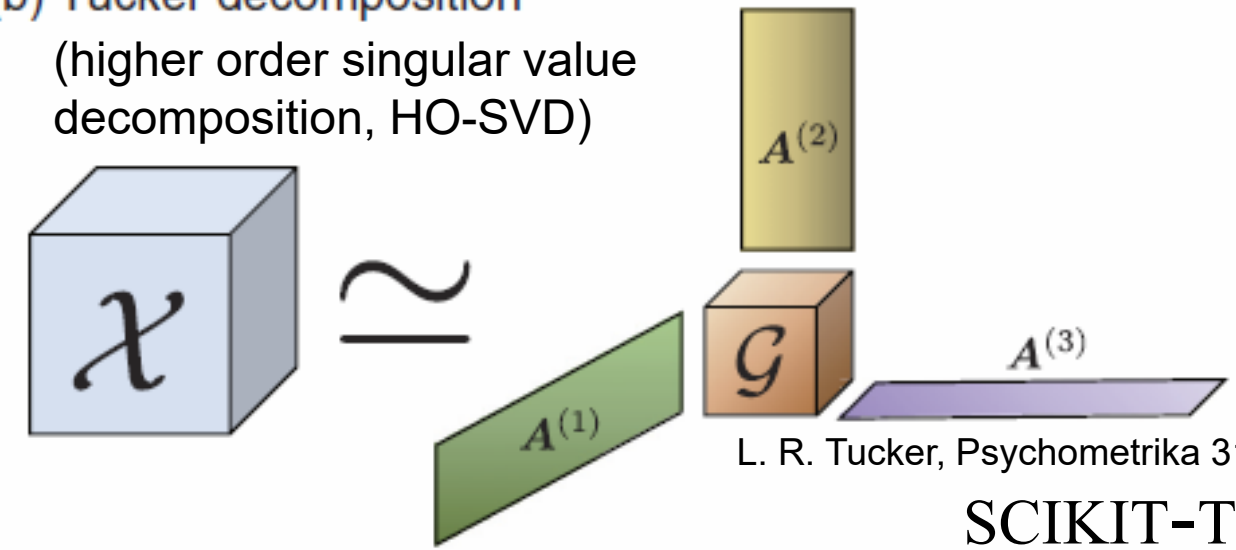
(canonical polyadic)



F. L. Hitchcock, Stud. Appl. Math. 6, 164 (1927).

(b) Tucker decomposition

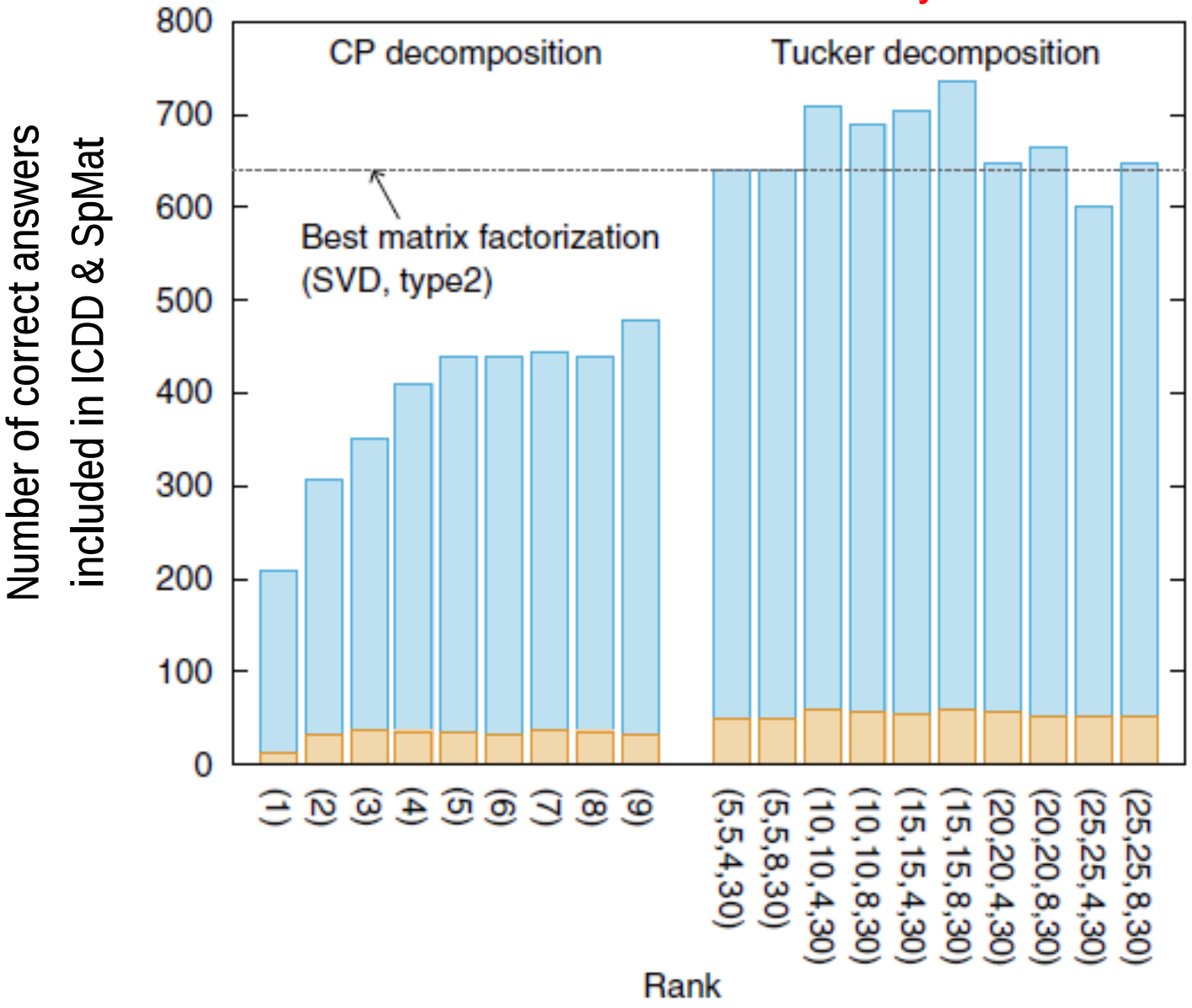
(higher order singular value decomposition, HO-SVD)



L. R. Tucker, Psychometrika 31, 279 (1966).

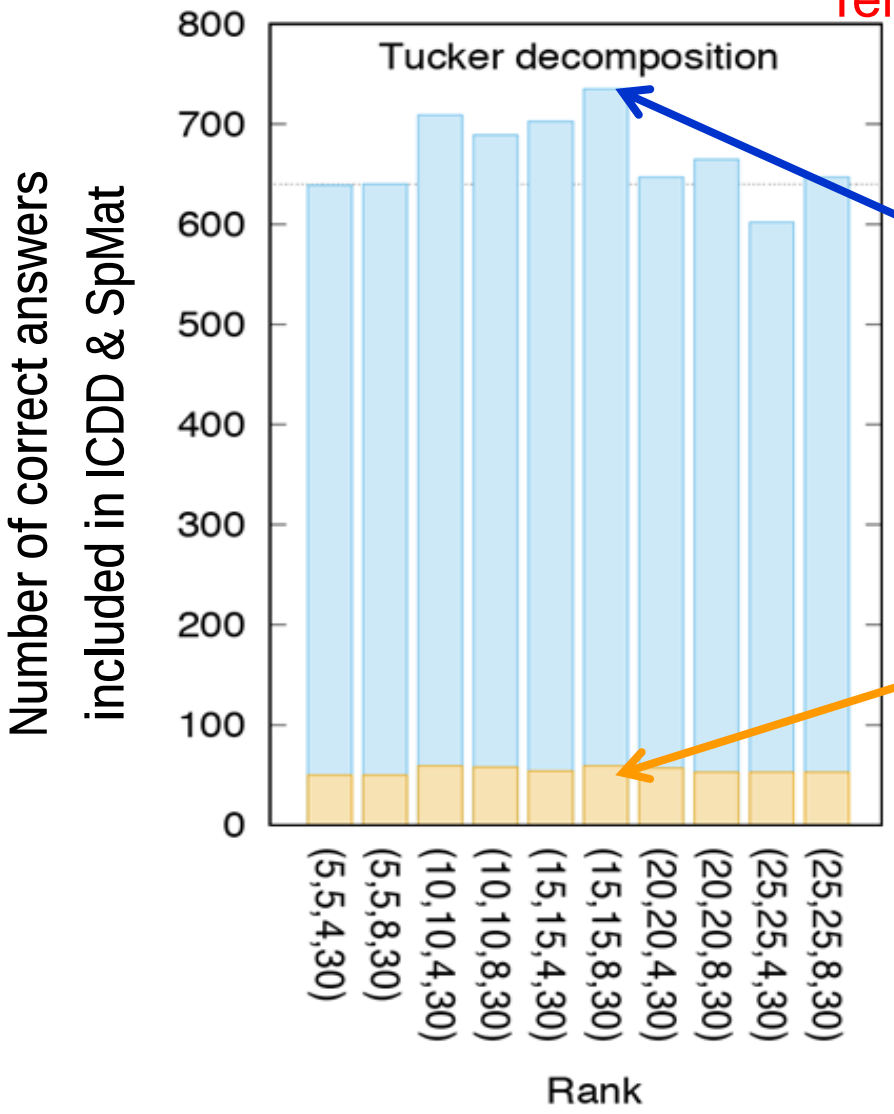
Validation of CRC prediction by a recommender system for ternary compounds using Tucker decomposition

Ternary # Elements： 7,405,200



Validation of CRC prediction by a recommender system for ternary compounds using Tucker decomposition

Ternary # Elements： 7,405,200



TOP3,000 compositions with high predicted rating.

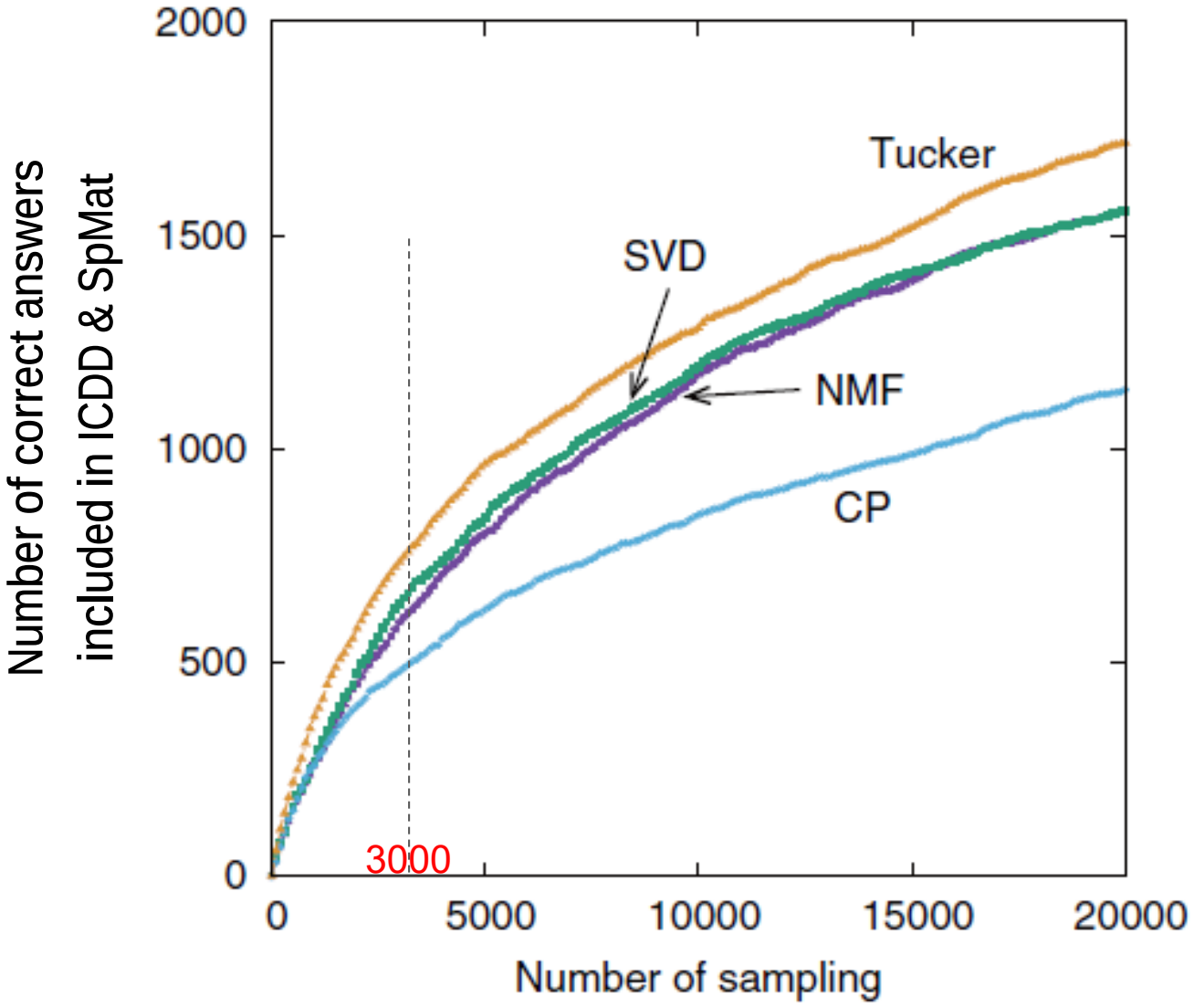
Discovery rate > 25% !!

TOP100 compositions with high predicted rating.

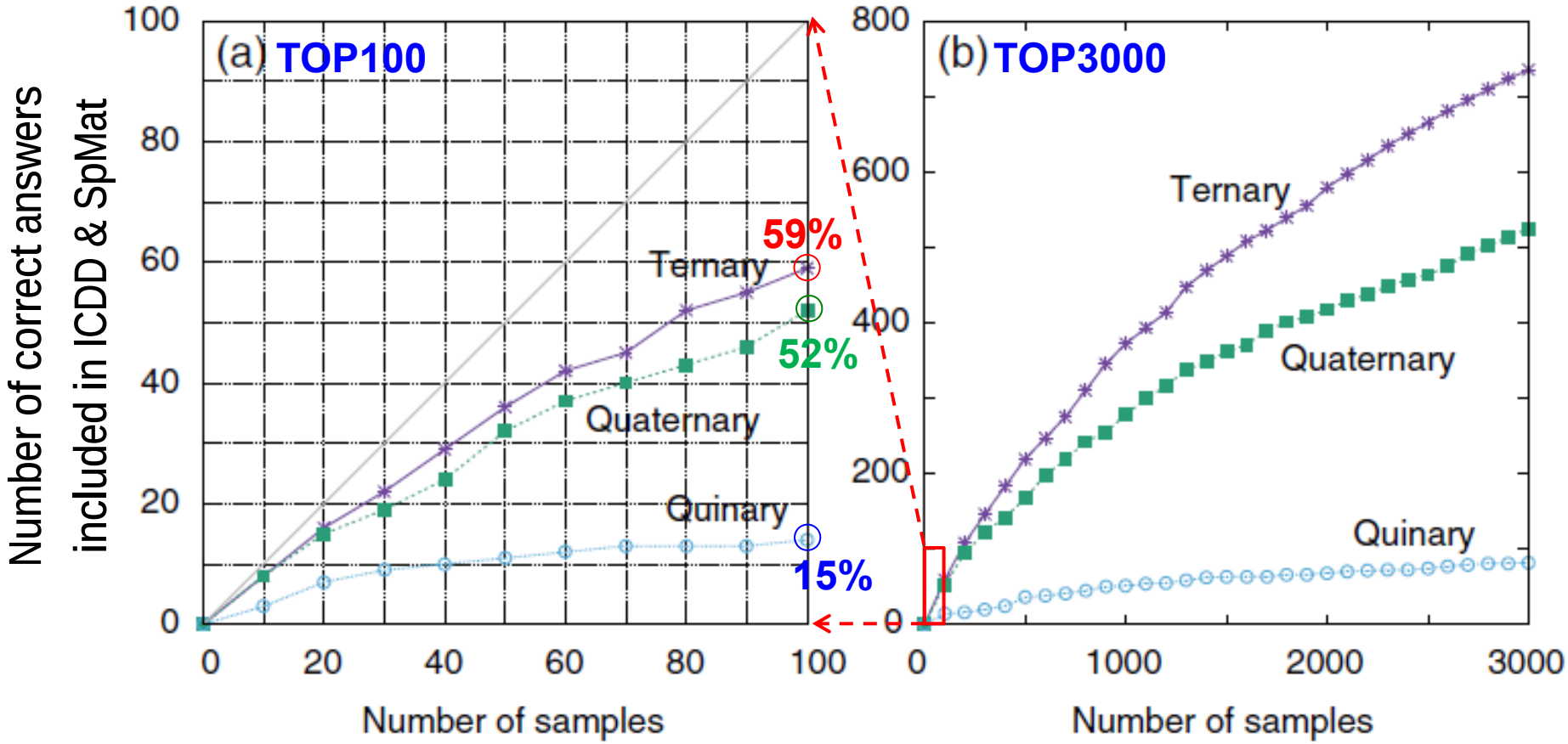
Discovery rate > 59% !!

Validation of CRC prediction by a recommender system for ternary compounds using Tucker decomposition

Ternary # Elements： 7,405,200



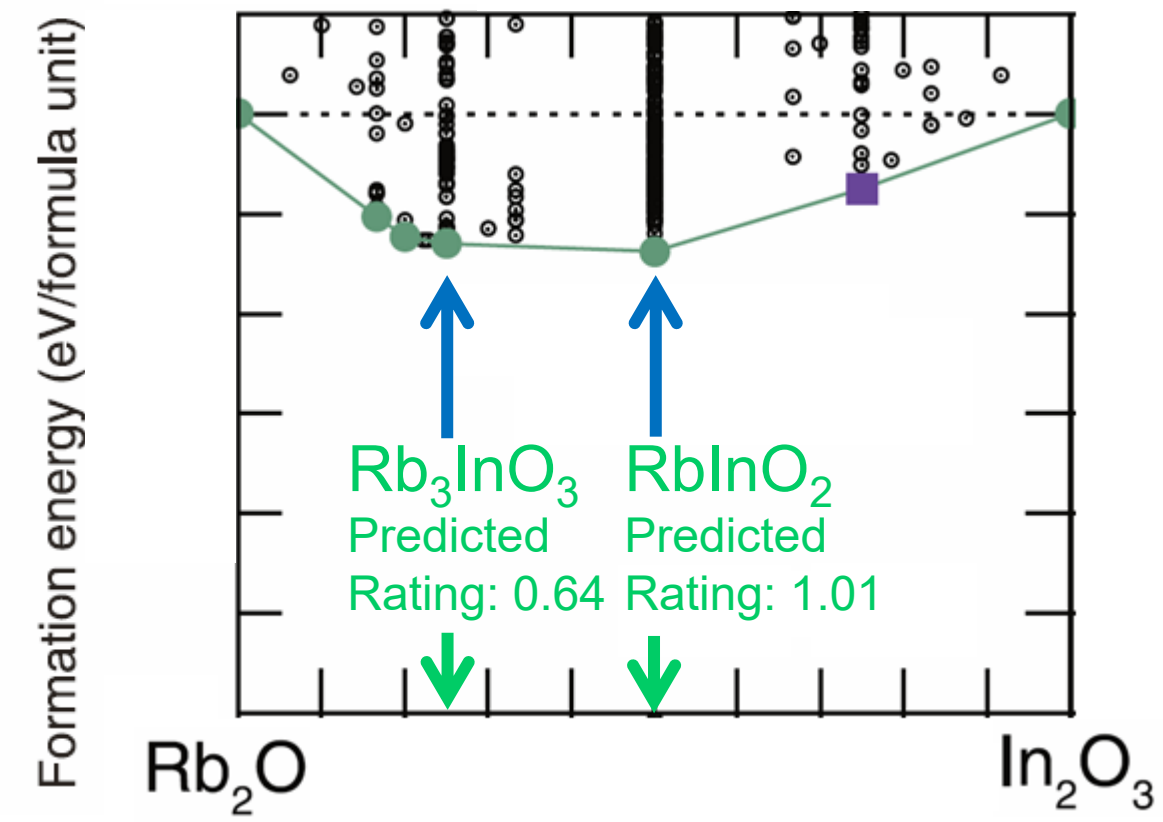
Results for quarternary/quinary systems



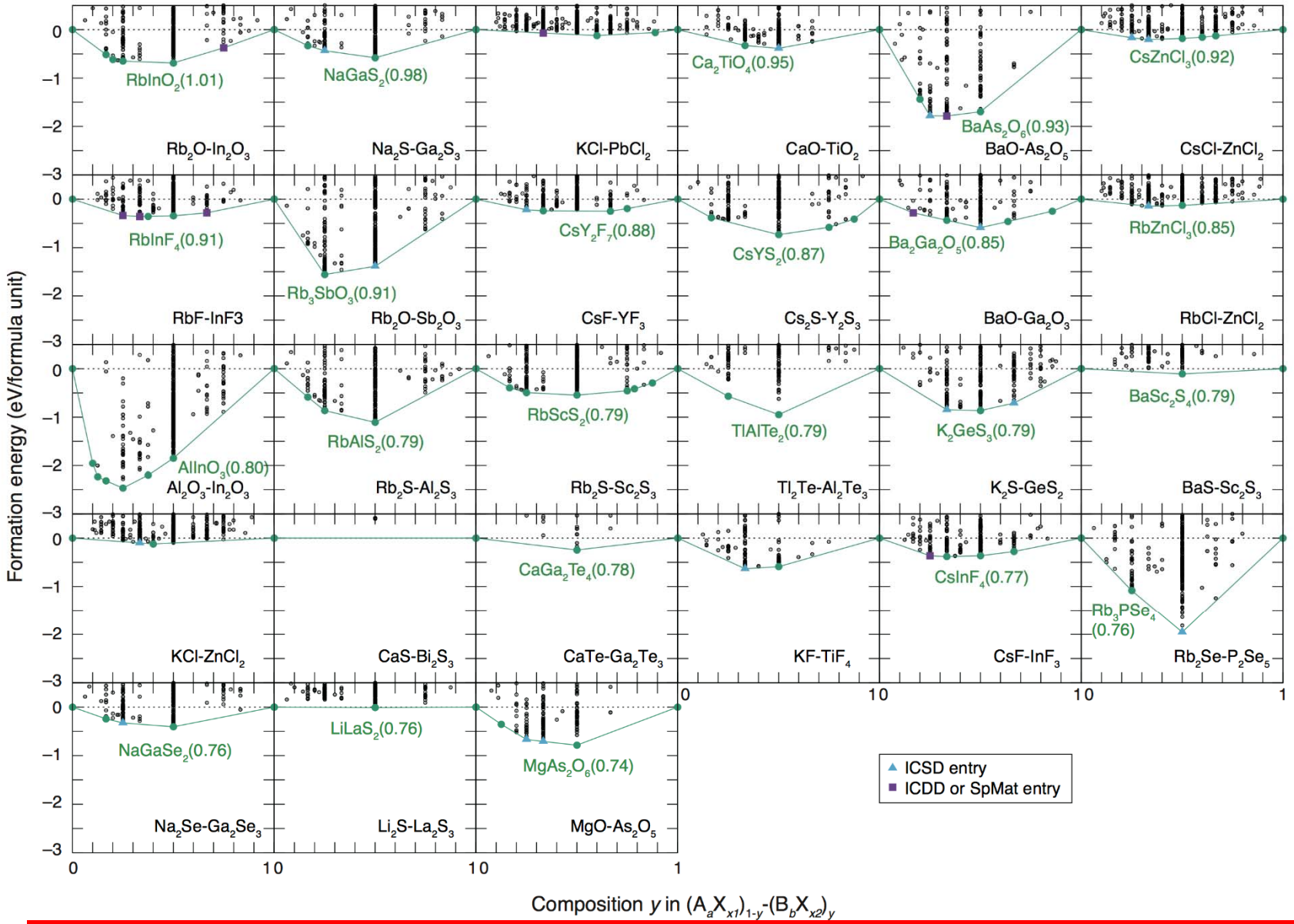
Number of entry compounds

	Ternary	Quaternary	Quinary
FIZ Karlsruhe ICSD	9,313	7,742	1,321

Further validation by first principles calculations for pseudo-binary compounds with high predicted rating



Further validation by first principles calculations for TOP 27 pseudo-binary compounds with high predicted rating



23 among 27 compositions(85%) are thermodynamically stable by DFT !

Systematic discovery of as-yet-unknown CRC

CRC (chemically relevant composition)

- ✓ Use of tensor-based recommender system ONLY with inorganic crystal database, ICSD.

  Rating prediction with neither descriptors, nor DFT results.

- ✓ Validation by two other databases, ICDD-PDF & Springer Materials. Discovery rate is 59/52/15% for TOP 100 ternary/quarternary/quinary CRC.

- ✓ Validation by DFT calculations. Among TOP 27 ternary (pseudo-binary oxides), 85% are thermodynamically stable.