Machine learning for materials discovery: **catalysts**, grain boundaries, superlattices and RNAs

Koji Tsuda

University of Tokyo

NIMS

RIKEN (from Oct 1)

Kashiwa Campus, University of Tokyo

www.tsudalab.org

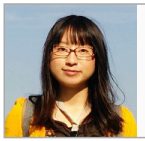
Tsuda Lab. | Home | **Members** | Publications | Projects | Lectures | Contact | 日本語

Koji Tsuda
Professor

Kaoru Shibutani
Secretary

David duVerle
Assistant Professor

Kazuki Yoshizoe
Post-Doc Researcher



Aika Terada
JST PRESTO Researcher

Thaer Dieb
Post-Doc Researcher

Kei Terayama
Post-Doc Researcher

Shogo Takeuchi
Post-Doc Researcher

Xiufeng Yang
Doctoral Student

Diptesh Das
Researcher

Koki Kitai
Master Student

Hiroaki Takada
Master Student

Kenichi Himeno
Master Student

Seiji Sasaoka
Undergraduate Student

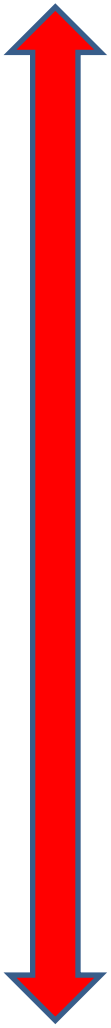
Kento Shin
Undergraduate Student

Open Postdoc Positions !

Discovery of new functional molecules and materials is of national importance

First Principles Calculations

Accurate, Slow



- Full configuration interaction
- Wave function based
- Density functional theory
- Semi-empirical
- Empirical potentials

Inaccurate, Fast

Talk Agenda:
From Prediction to Automatic Design

- Prediction of d-band center of bimetals (Takigawa et al., RSC Advances, 2016)

- Automatic Design by Bayesian Optimization
  - Optimization of grain boundary (Kiyohara et al., JJAP, 2016)
  - Design of Si-Ge nanostructures (Ju et al., in submission)
  - COMBO: Fast python library for Bayesian opt. (Ueno et al., Mater. Discov., 2016)

- Automatic Design by Monte Carlo Tree Search
  - Designing optimal RNA sequences (Yang et al., in submission)

Prediction of d-band center

- Catalytic bimetals: host and guest metals
  - Doped in the surface
  - Surface monolayer

- Binding between a metal surface and an adsorbate depends on the electronic structure of the metal
- D-band center = key indicator of catalytic activity

D-band center (Doped)

B. Hammer and J. K. Nørskov, Adv. Catal., 2000

M_g M_h	Fe	Co	Ni	Cu	Ru	Rh	Pd	Ag	Ir	Pt	Au
Fe	***-0.92***	-0.87	-1.12	-1.05	-1.21	-1.46	-2.16	-1.75	-1.28	-2.01	-2.34
Co	-1.16	***-1.17***	-1.45	-1.33	-1.41	-1.75	-2.54	-2.08	-1.53	-2.36	-2.73
Ni	-1.20	-1.10	***-1.29***	-1.10	-1.43	-1.60	-2.26	-1.82	-1.43	-2.09	-2.42
Cu	-2.11	-2.07	-2.40	***-2.67***	-2.09	-2.35	-3.31	-3.37	-2.09	-3.00	-3.76
Ru	-1.20	-1.15	-1.40	-1.29	***-1.41***	-1.58	-2.23	-1.68	-1.39	-2.03	-2.25
Rh	-1.49	-1.39	-1.57	-1.29	-1.69	***-1.73***	-2.27	-1.66	-1.56	-2.08	-2.22
Pd	-1.46	-1.29	-1.33	-0.89	-1.59	-1.47	***-1.83***	-1.24	-1.30	-1.64	-1.66
Ag	-3.58	-3.46	-3.63	-3.83	-3.46	-3.44	-4.16	***-4.30***	-3.16	-3.80	-4.45
Ir	-1.90	-1.84	-2.06	-1.90	-2.02	-2.26	-2.84	-2.24	***-2.11***	-2.67	-2.85
Pt	-1.92	-1.77	-1.85	-1.53	-2.11	-2.02	-2.42	-1.81	-1.87	***-2.25***	-2.30
Au	-2.93	-2.79	-2.93	-3.01	-2.86	-2.81	-3.39	-3.35	-2.58	-3.10	***-3.56***

D-band center (Monolayer)

B. Hammer and J. K. Nørskov, Adv. Catal., 2000

M_g \ M_h	Fe	Co	Ni	Cu	Ru	Rh	Pd	Ag	Ir	Pt	Au
Fe	***-0.92***	-0.78	-0.96	-0.97	-1.65	-1.64	-2.24	-2.17	-1.87	-2.40	-3.11
Co	-1.18	***-1.17***	-1.37	-1.23	-1.87	-2.12	-2.82	-2.53	-2.26	-3.06	-3.56
Ni	-0.33	-1.18	***-1.29***	-1.17	-1.92	-2.03	-2.61	-2.43	-2.15	-2.82	-3.39
Cu	-2.42	-2.29	-2.49	***-2.67***	-2.89	-2.94	-3.71	-3.88	-2.99	-3.82	-4.63
Ru	-1.11	-1.04	-1.12	-1.11	***-1.41***	-1.53	-1.88	-1.81	-1.54	-2.02	-2.27
Rh	-1.42	-1.32	-1.39	-1.51	-1.70	***-1.73***	-2.12	-1.81	-1.70	-2.18	-2.30
Pd	-1.47	-1.29	-1.29	-1.03	-1.94	-1.58	***-1.83***	-1.68	-1.52	-1.79	-1.97
Ag	-3.75	-3.56	-3.62	-3.68	-3.80	-3.63	-4.03	***-4.30***	-3.50	-3.93	-4.51
Ir	-1.78	-1.71	-1.78	-1.55	-2.12	-2.14	-2.53	-2.20	***-2.11***	-2.60	-2.70
Pt	-1.90	-1.72	-1.71	-1.47	-2.13	-2.01	-2.23	-2.06	-1.96	***-2.25***	-2.33
Au	-3.03	-2.82	-2.85	-2.86	-3.09	-2.89	-3.21	-3.44	-2.77	-3.13	***-3.56***

Machine Learning

- Randomly partition d-band centers into training / test sets

Guest metals

Host metals

	Fe	Co	Ni	Cu	Ru	Rh	Pd	Ag	Ir	Pt	Au
Fe	-0.92	-0.87	?	-1.05	?	?	-2.16	-1.75	-1.28	-2.01	?
Co	-1.16	?	-1.45	-1.33	?	?	?	?	-1.53	?	-2.73
Ni	?	-1.1	?	-1.1	-1.43	-1.6	-2.26	-1.82	-1.43	-2.09	-2.42
Cu	?	-2.07	-2.4	?	?	?	-3.31	-3.37	?	-3	?
Ru	?	?	?	-1.29	?	-1.58	?	?	-1.39	-2.03	-2.25
Rh	?	?	?	?	?	-1.73	-2.27	-1.66	?	-2.08	-2.22
Pd	?	-1.29	?	-0.89	-1.59	-1.47	?	-1.24	?	?	?
Ag	-3.58	?	-3.63	-3.83	?	?	?	?	?	?	?
Ir	-1.9	?	-2.06	-1.9	?	?	?	-2.24	?	-2.67	?
Pt	-1.92	-1.77	?	?	-2.11	?	-2.42	-1.81	-1.87	-2.25	?
Au	?	?	?	?	?	-2.81	?	-3.35	-2.58	?	-3.56

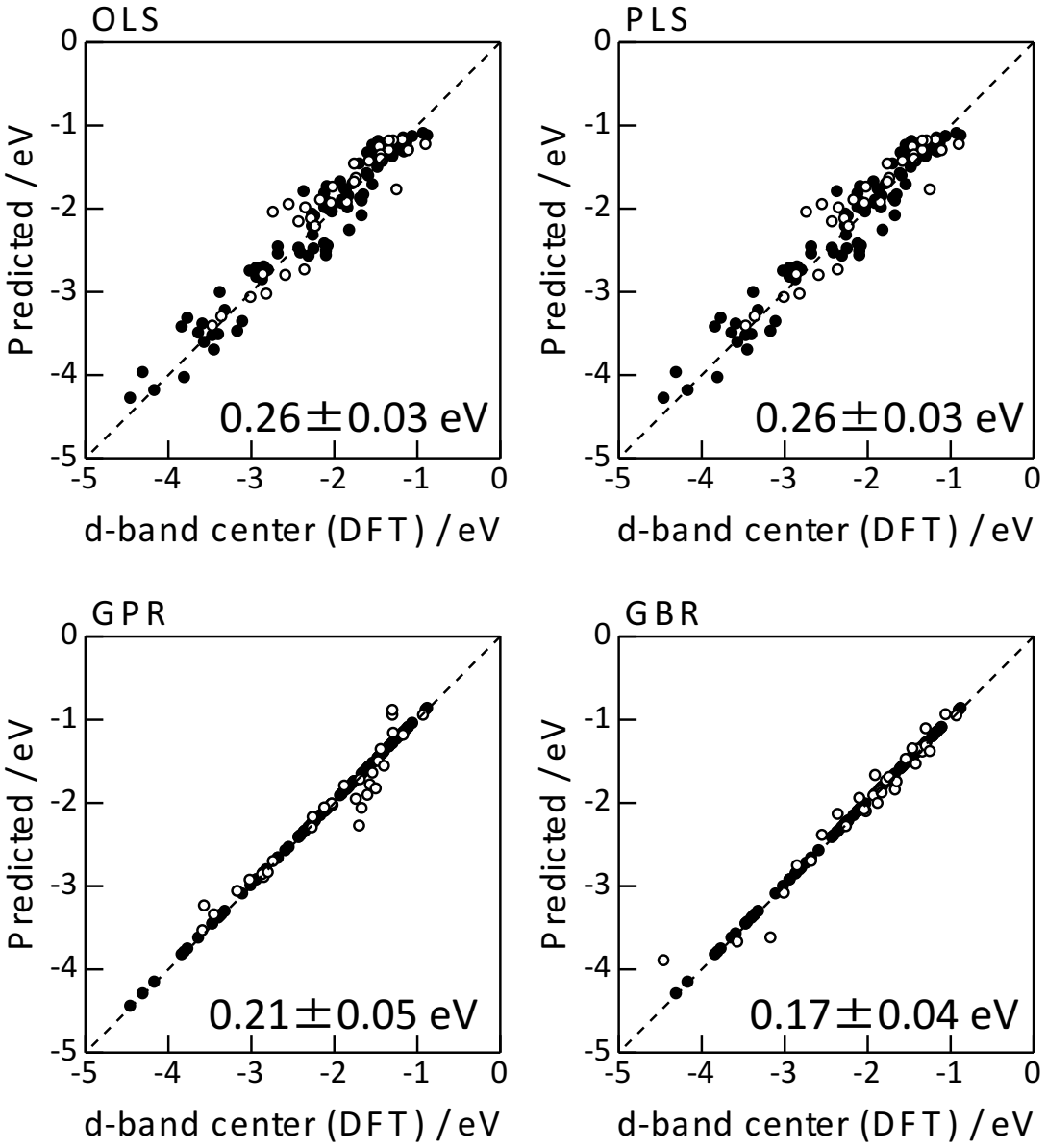
Descriptors (Host & Guest)

- Group (*G*)
- Bulk Wigner-Seitz radius (*R*) in Å
- Atomic number (*AN*)
- Atomic mass (*AM*) in g mol^{-1}
- Period (*P*)
- Electronegativity (*EN*)
- Ionization energy (*IE*) in eV
- Enthalpy of fusion ($\Delta_{fus}H$) in J g^{-1}
- Density at 25 C (ρ) in g cm^{-3}

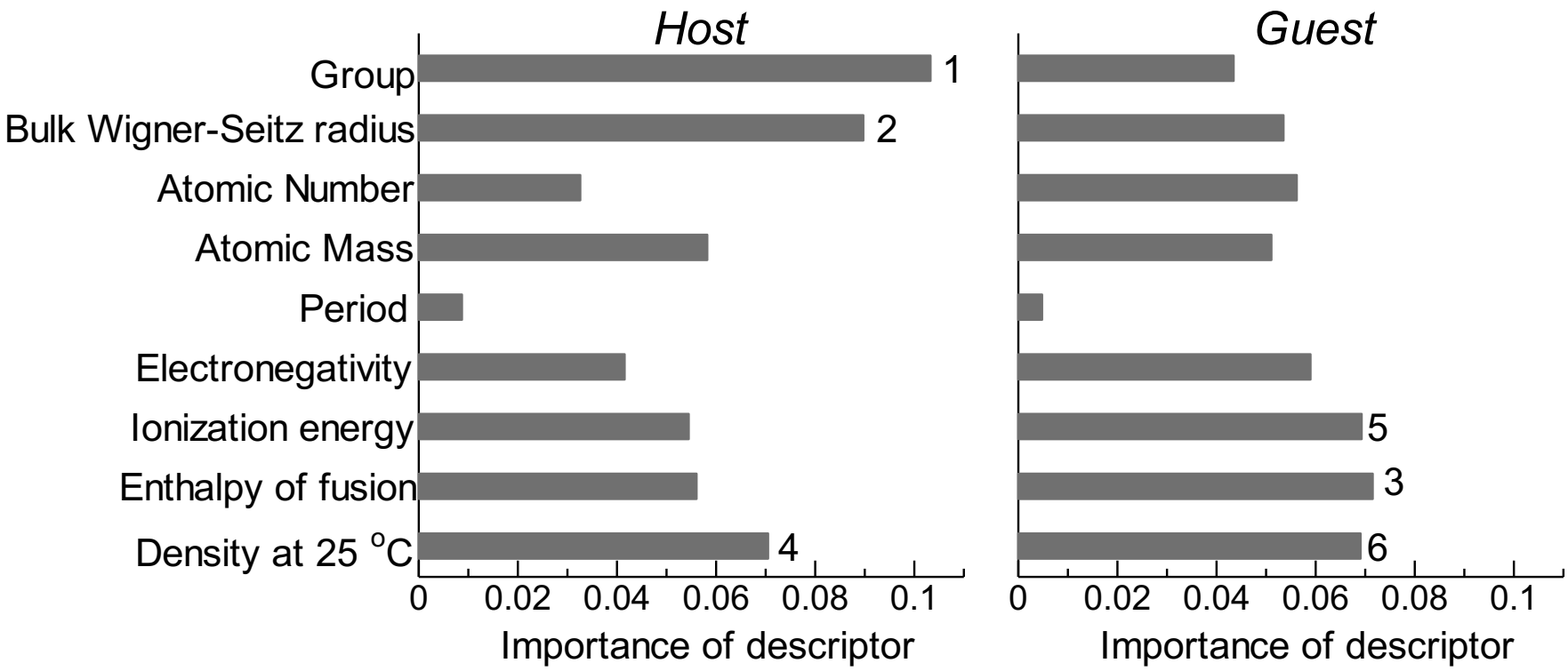
Methods Used

Abbreviation	Method	Tuning parameters [tested range]
Linear Methods		
OLS	Ordinary least squares regression	(No tuning parameters)
PLS	Partial least squares regression	n_components $\in$ [1,2,…,# of vars]
Nonlinear Methods		
GPR	Gaussian process regression	theta0 $\in$ [1.0,10^{-1},10^{-2},10^{-3},10^{-4},10^{-5}]
GBR	Gradient boosting regression	learning_rate $\in$ [1.0,10^{-1},10^{-2},10^{-3},10^{-4},10^{-5}] max_depth $\in$ [4,6,8,10] n_estimators $\in$ [100,250,500]

‘Doped’ Averaged 100 Times (25% hidden)

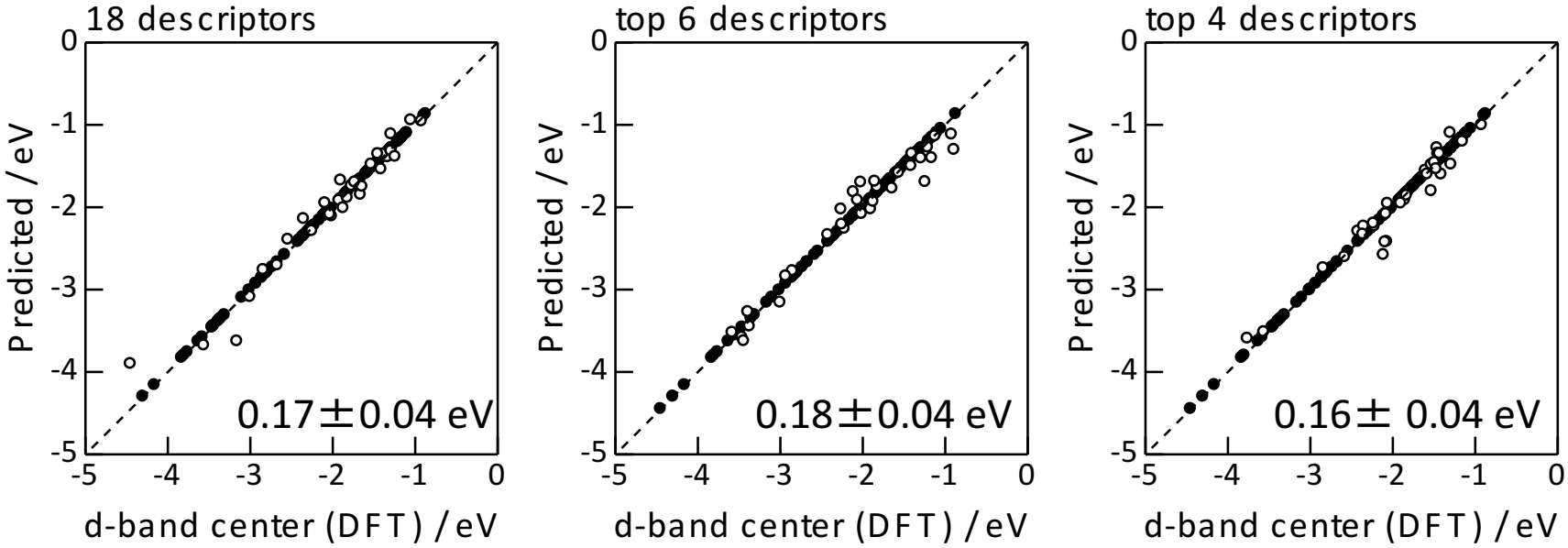


Variable Importance in Gradient Boosting Regression



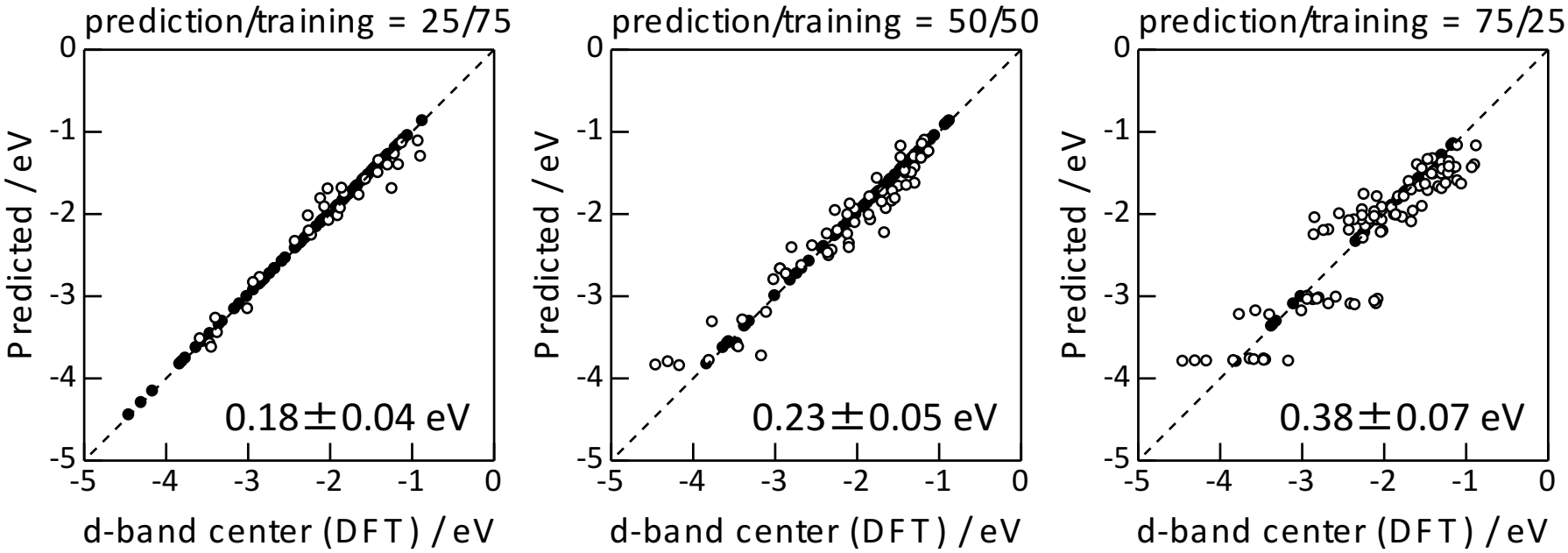
Using Top Descriptors Only

‘Doped’ Averaged 100 Times (25% hidden)



Changing Train/Test Fraction

‘Doped’ averaged 100 Times w/ selected 6 descriptors



NEWS

Machine-learning accelerates catalytic trend spotting

BY ANNA MEEHAN | 9 JUNE 2016

Example of what you can gain when 'people from different disciplines start looking at the same problems'

Researchers in Japan have used a machine-learning method to cut the time it takes to predict the catalytic potential of different metals.

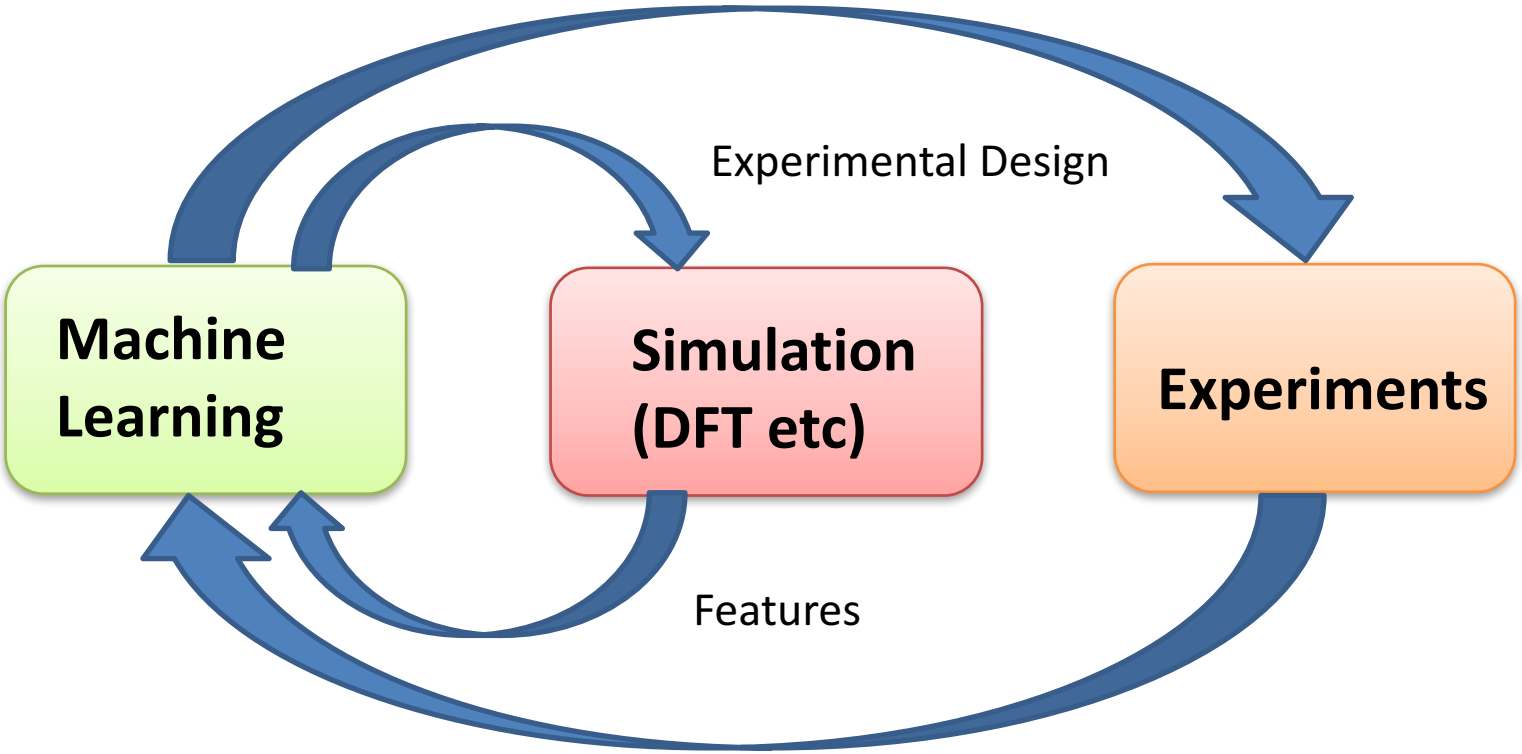
Binding between a metal surface and an adsorbate mainly depends on the electronic structure of the metal. More energy at centre of the metal's d-band creates a stronger bond between its surface and the adsorbate. Based on this theory, scientists have long regarded a value called the d-band centre as a key indicator of a metal's catalytic activity.

MOST POPULAR MOST COMMENTED


World's first commercial MOF keeps fruit fresh


Nucleic acid instability challenges RNA world hypothesis

Automatic Design by Machine Learning



Screening by first principles calculations alone

Mat. 1	Mat. 2	Mat. 3	Mat. 4	Mat. 5	Mat. 6	Mat. 7	Mat. 8	Mat. 9	Mat. 10

First Principles Calc.

Score 1	Score 2	Score 3	Score 4	Score 5	Score 6	Score 7	Score 8	Score 9	Score 10

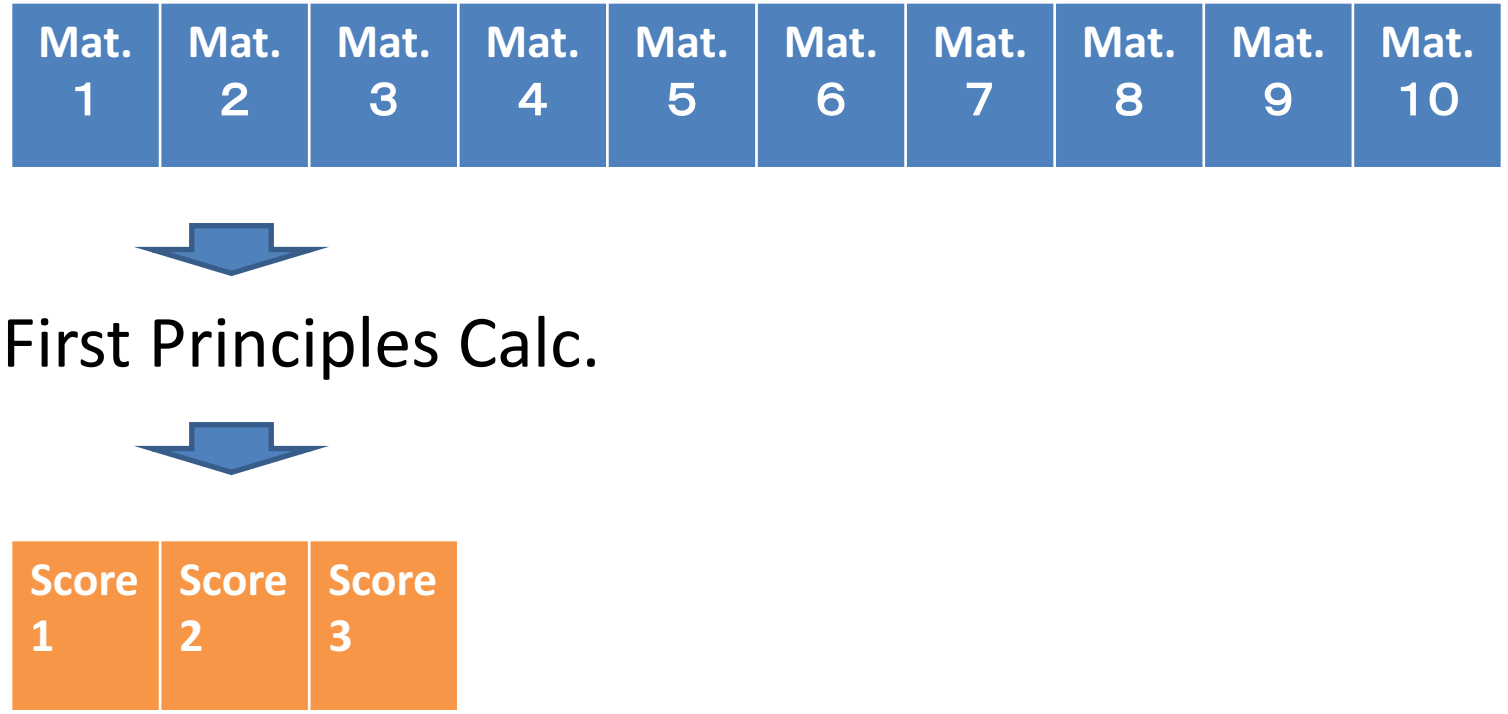
Bayesian Optimization

(Jones et al., 1998)

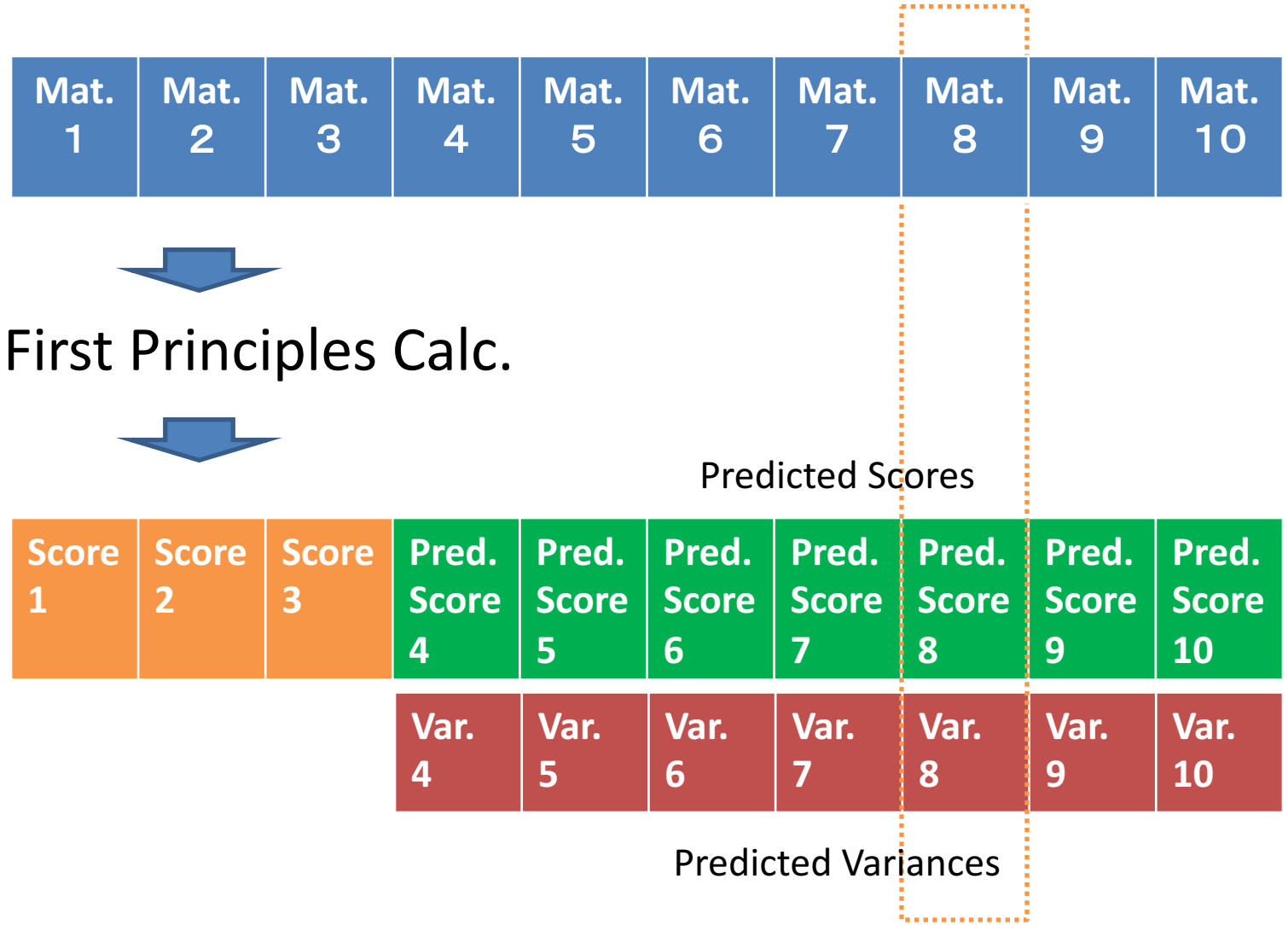
- Find best data points with minimum number of observations

- Choose next point to observe to discover the best ones as early as possible

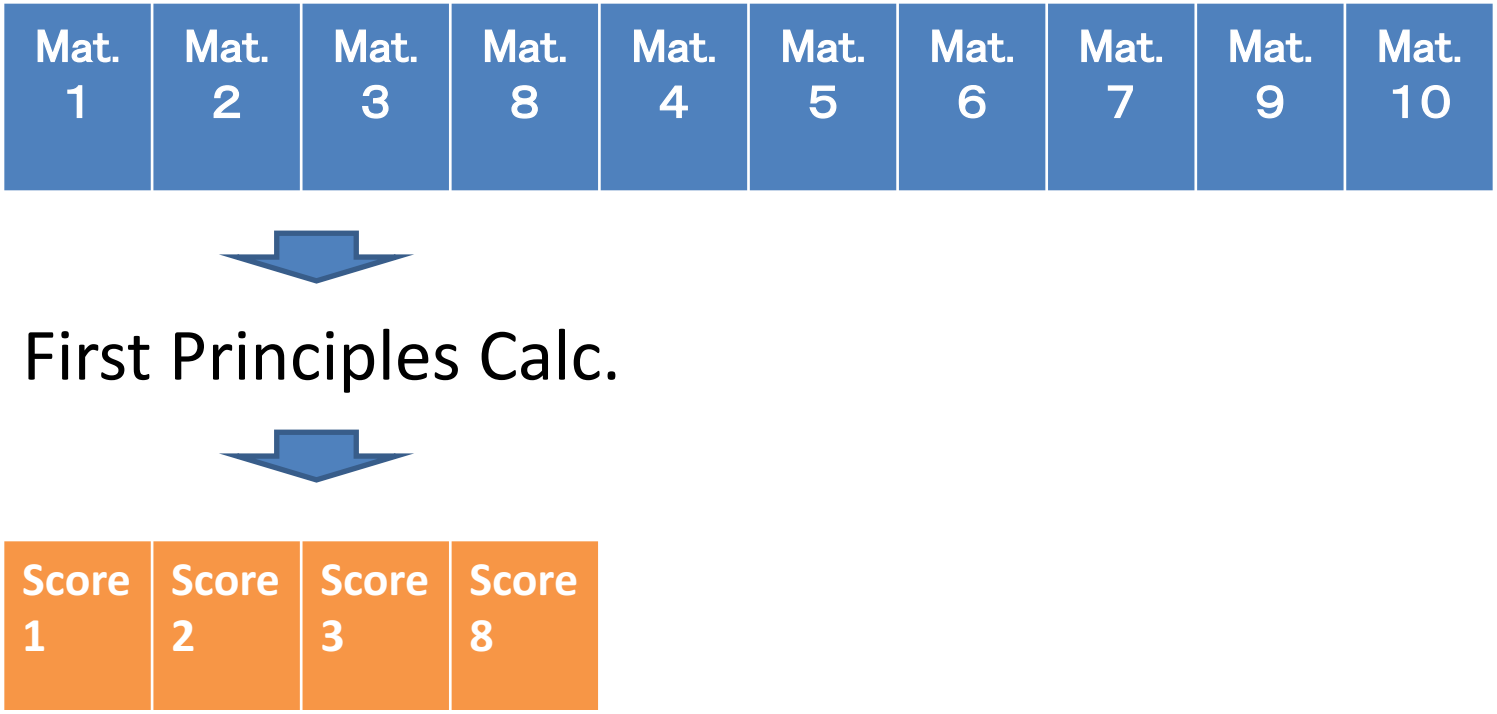
Bayesian Optimization (1)



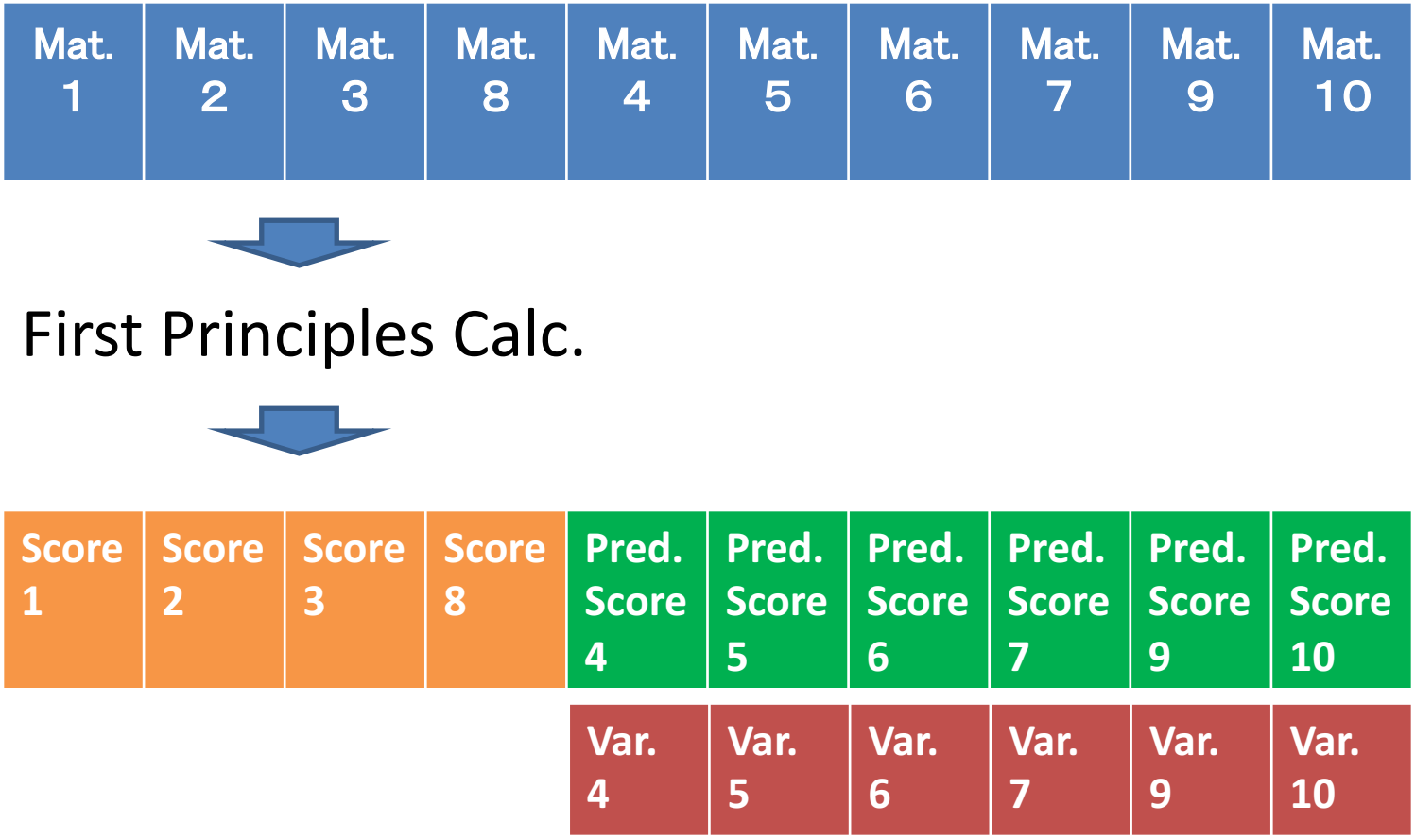
Bayesian Optimization (2)



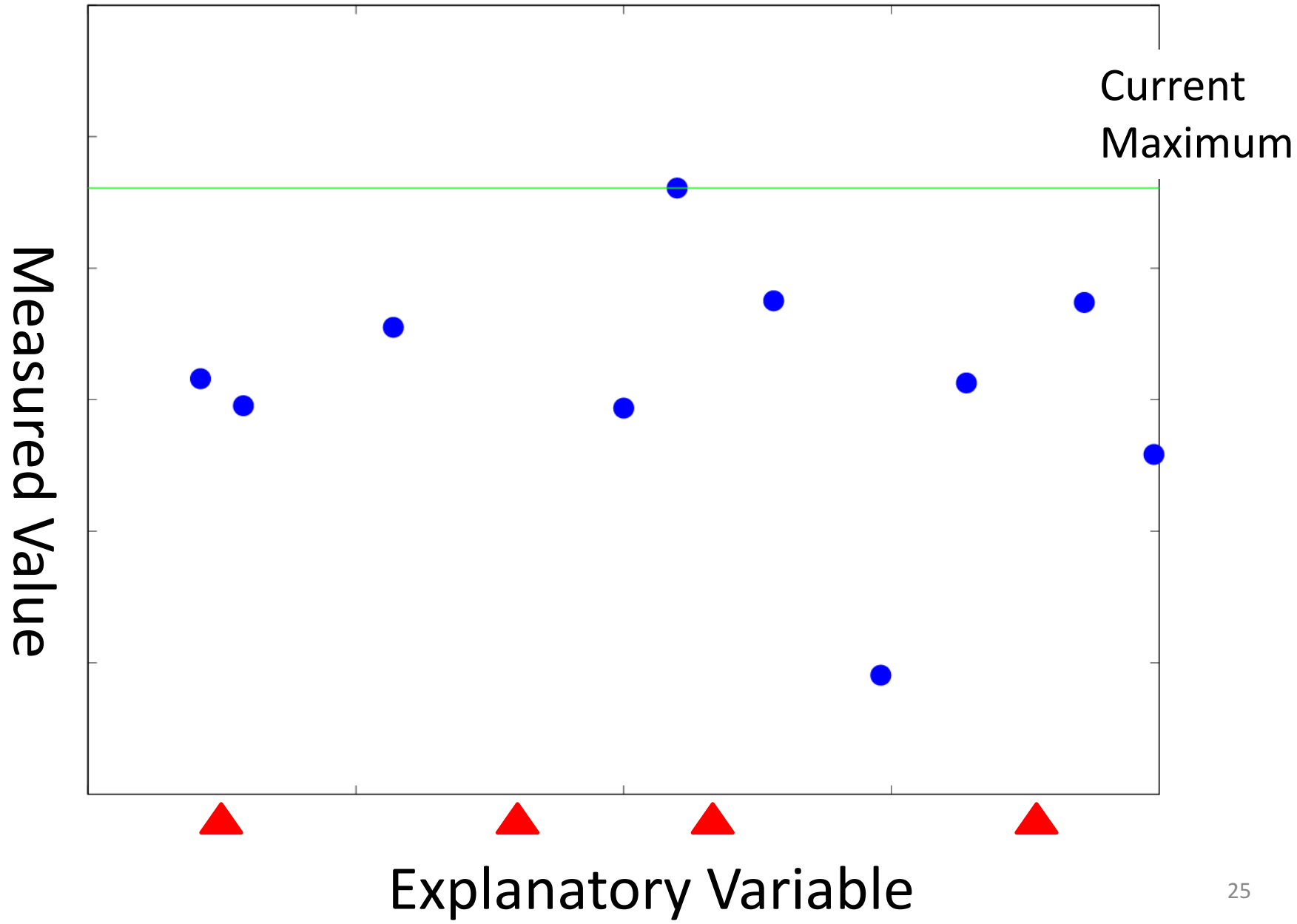
Bayesian Optimization (3)



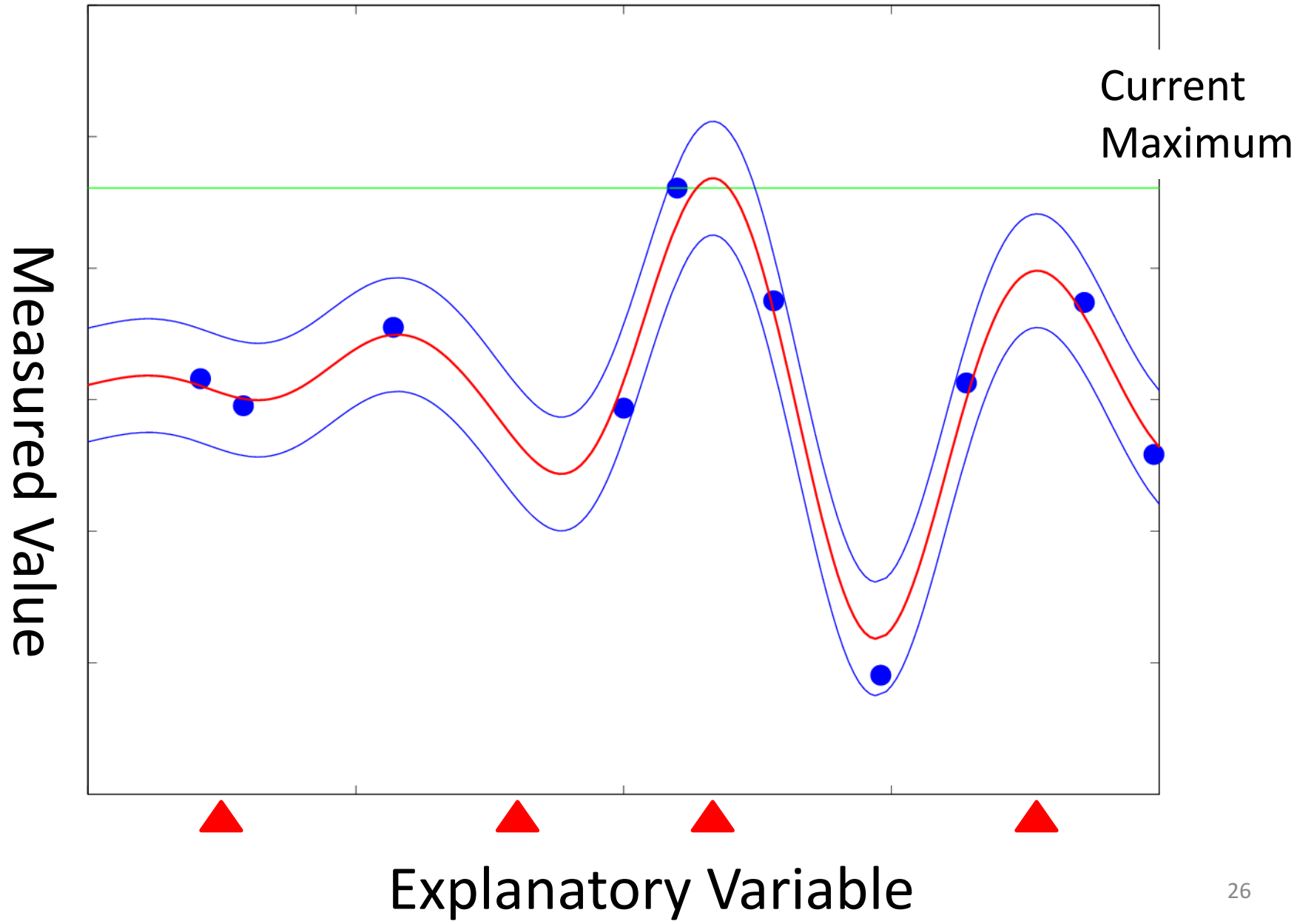
Bayesian Optimization (4)



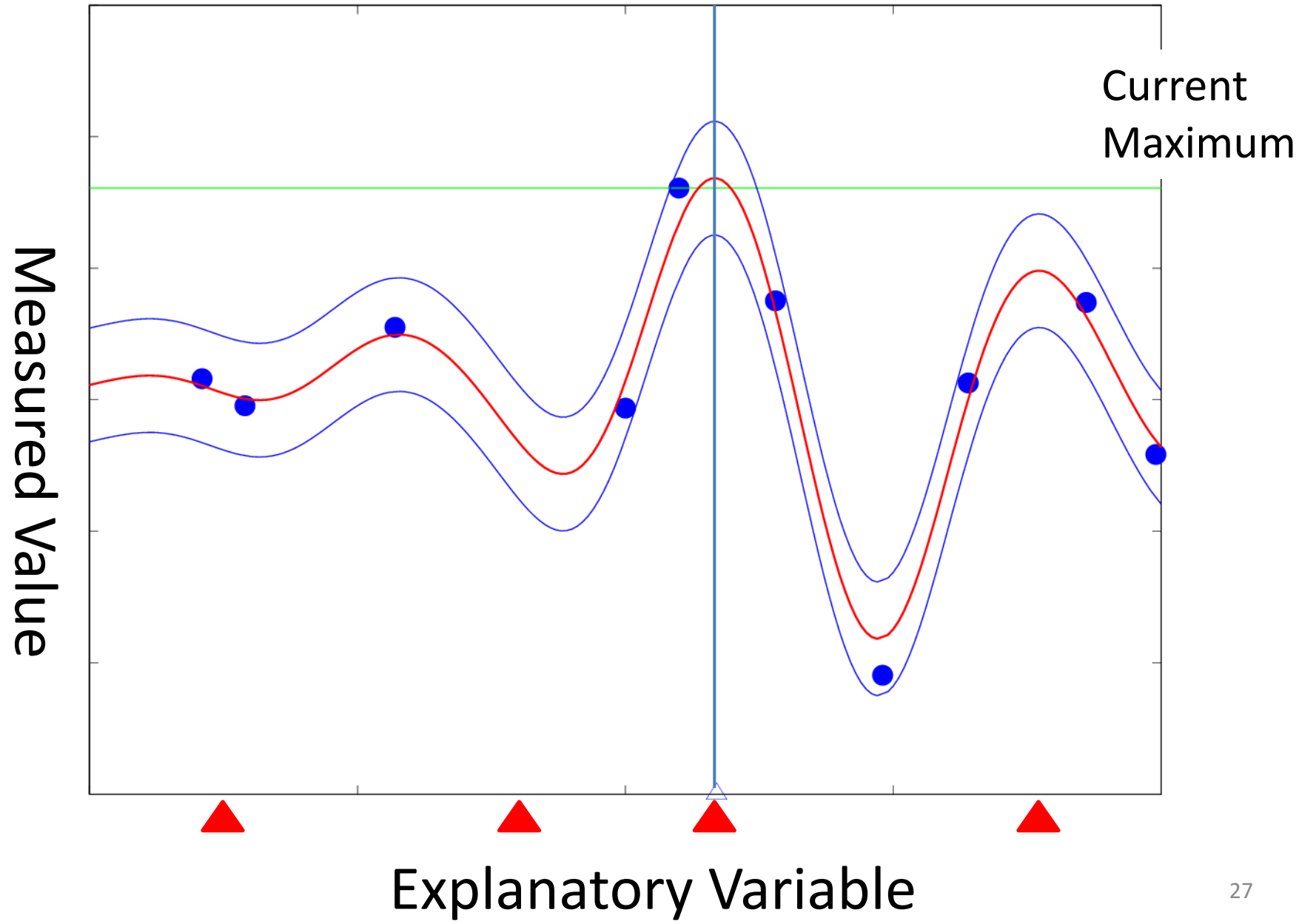
Where to observe next?



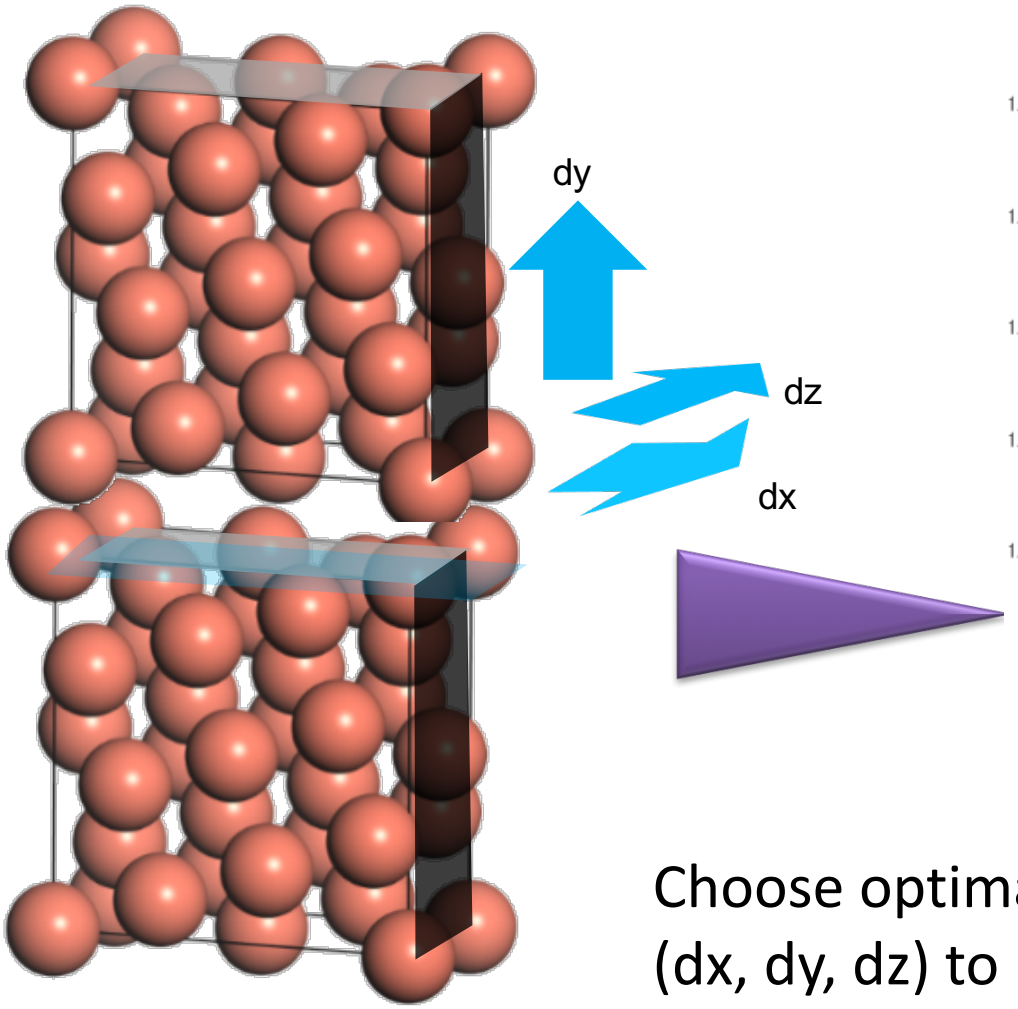
Gaussian Process

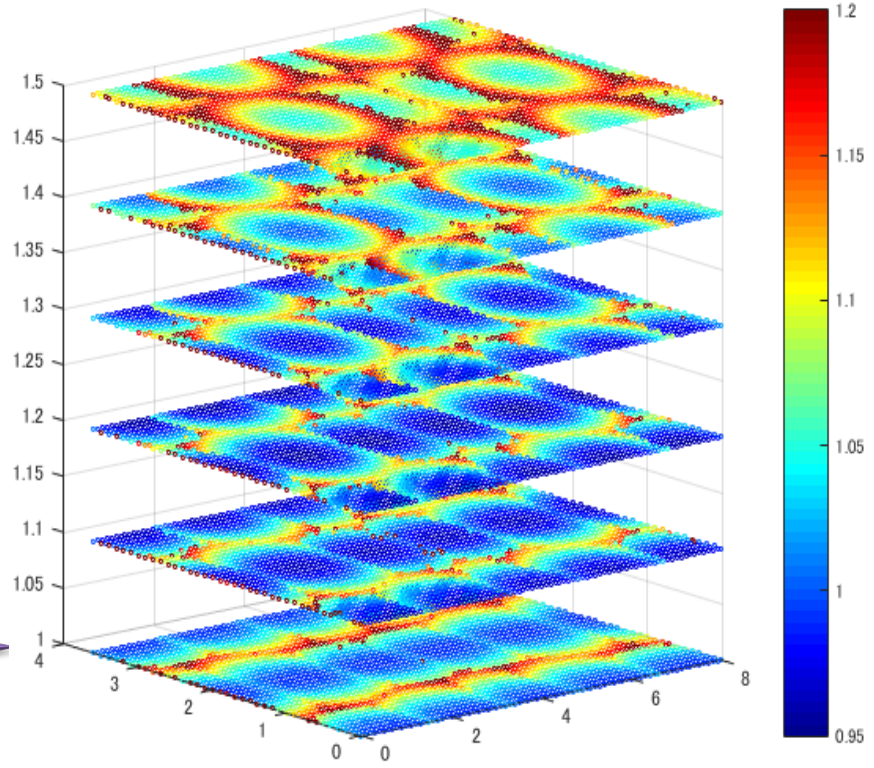


Maximum probability of improvement



Grain boundary structure determination



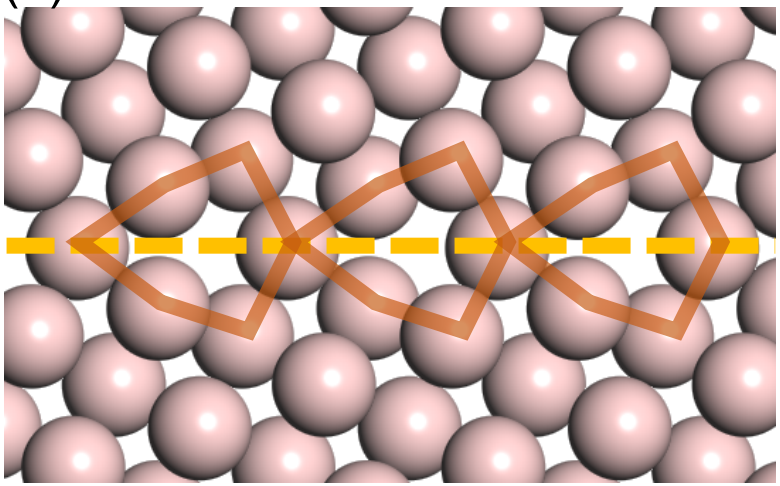


Choose optimal translation parameters (dx, dy, dz) to minimize the grain boundary energy

Acceleration of Discovery

Cu [001] (210) Σ5 grain boundary

(a)

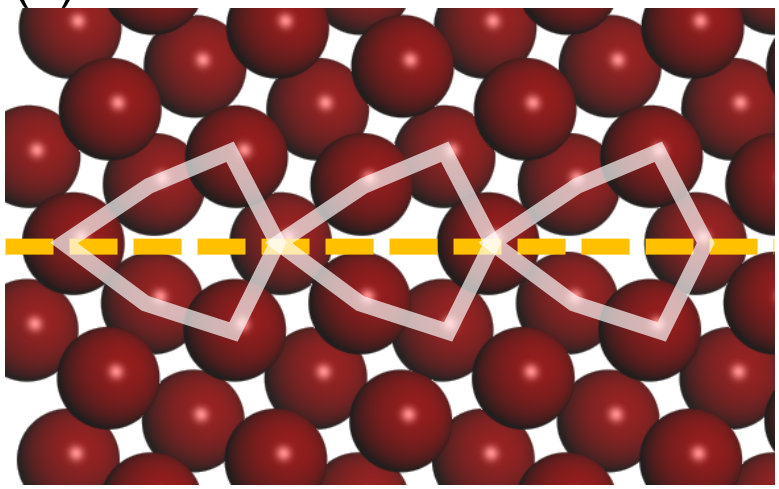


Exhaustive calculations

GB energy=0.96J/m^2

Number of energy calculations
＝16,983

(b)

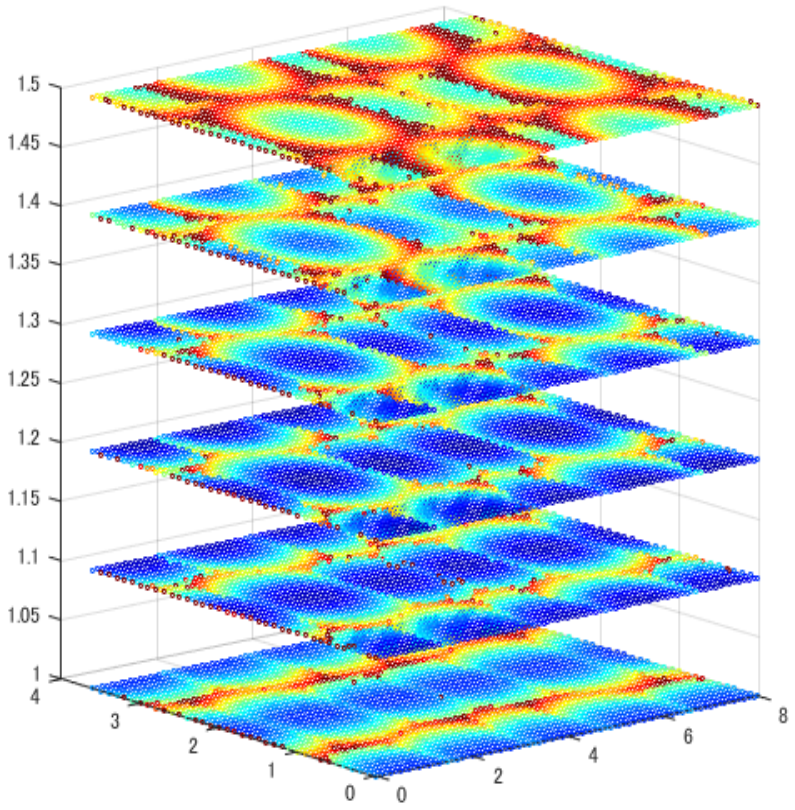


Bayesian optimization

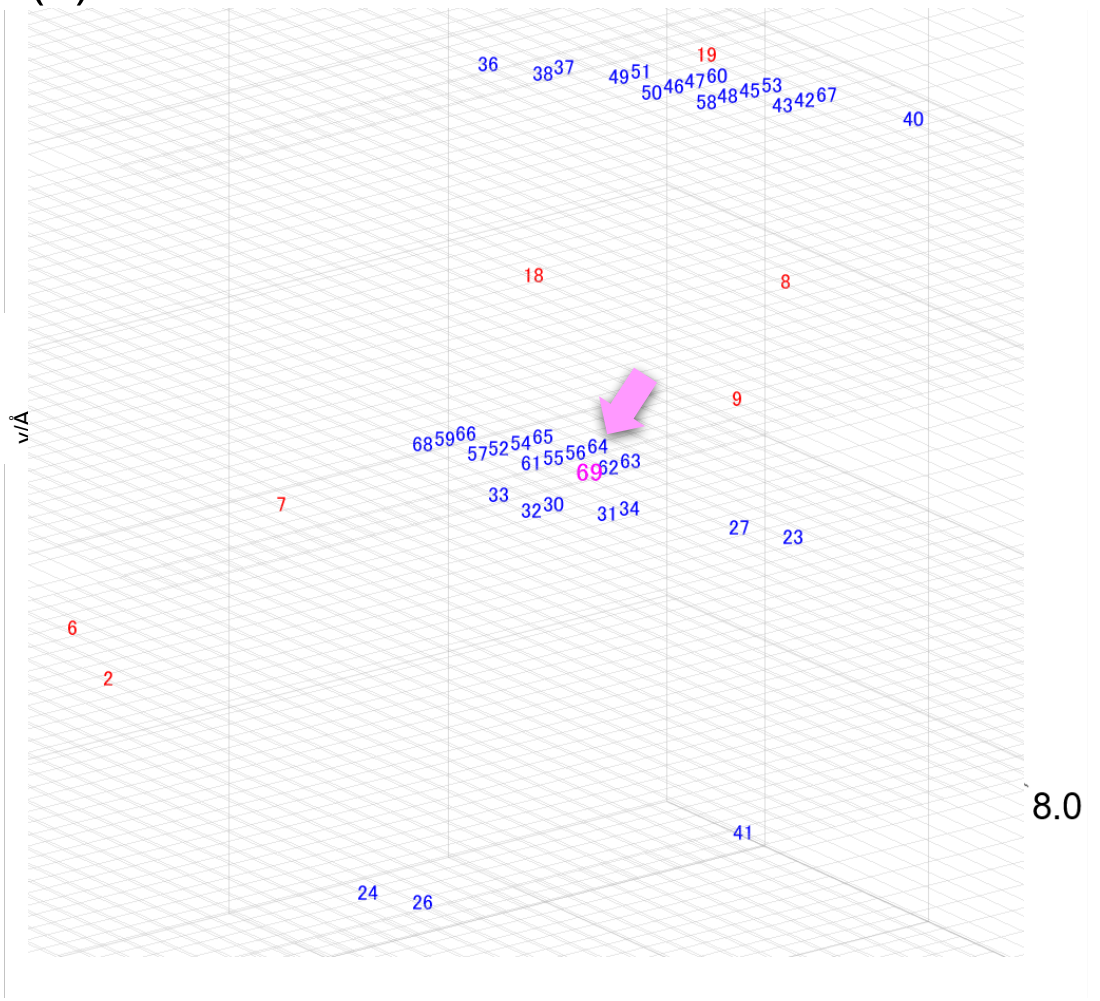
GB energy=0.96J/m^2

Number of energy calculations
＝69

S. Kiyohara et al., Jpn. J. Appl. Phys., 2016.


1.5
1.45
1.4
1.35
1.3
1.25
1.2
1.15
1.1
1.05
1
4
3
2
1
0
0
2
4
6
8

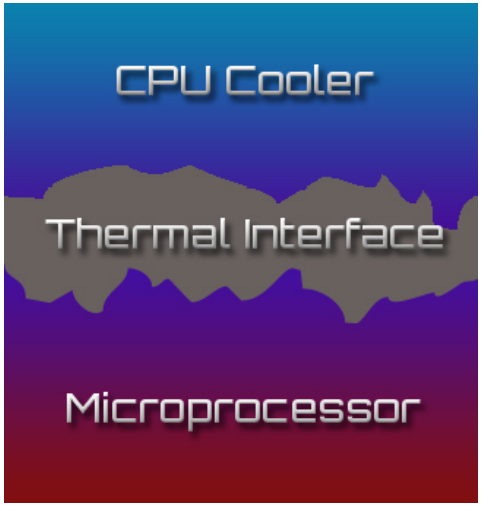
(b)

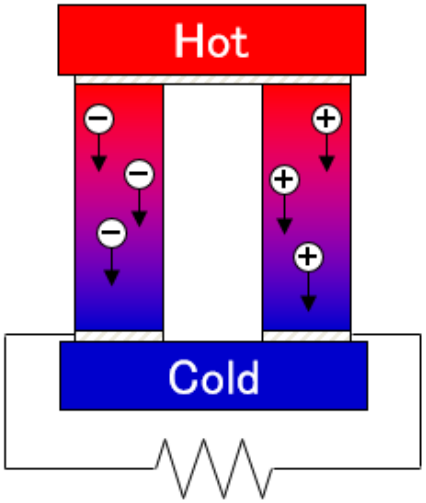

v/Å
36
38
37
49
51
19
50
46
47
60
58
48
45
53
43
42
67
40
18
8
9
68
59
66
57
52
54
65
61
55
56
64
69
62
63
33
32
30
31
34
7
27
23
6
2
41
24
26
8.0

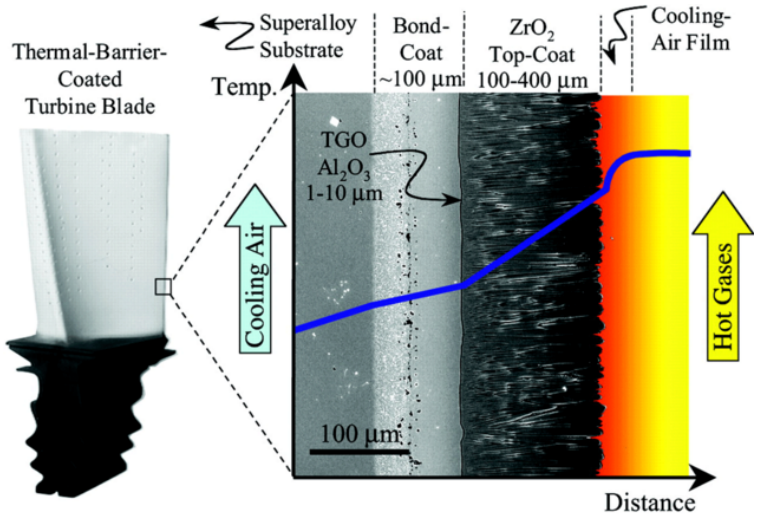
Design nanostructures for phonon transport via material informatics

Interface structure design has wide application in thermal devices.

High Conductance ⟷ **Low Conductance**







Interface materials | **Thermoelectric** | **Thermal barrier coating**

- ❑ various parameters
- ❑ parameters effect coupled
- ❑ transport vs local atomic configurations
- ❑ interference/resonance effects

⇒

- ➢ low developing efficiency
- ➢ high experimental cost
- ➢ long calculation time

Alloy Structure Optimization

Question: How to organize 16 alloy atoms (Si: 8, Ge: 8) to obtain the largest and smallest interfacial thermal conductance?

Descriptors: $C_{16}^{8} = 12{,}870$

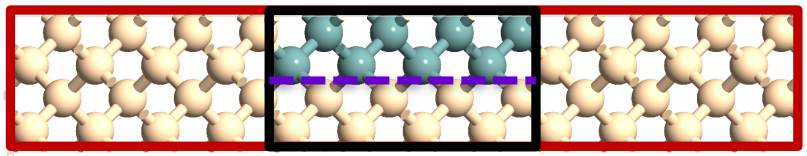
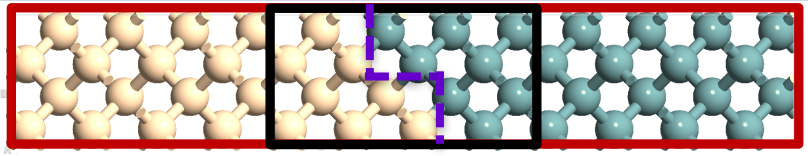
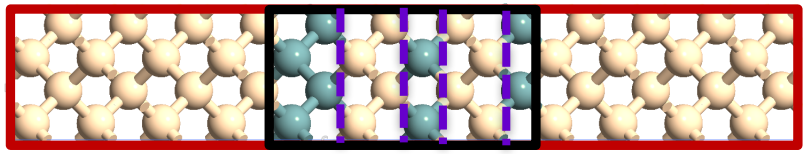
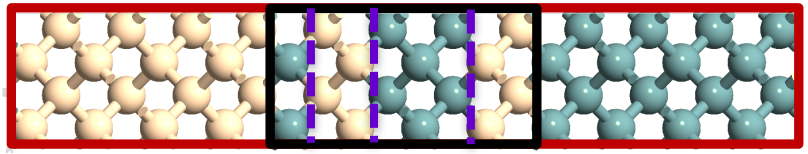
Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

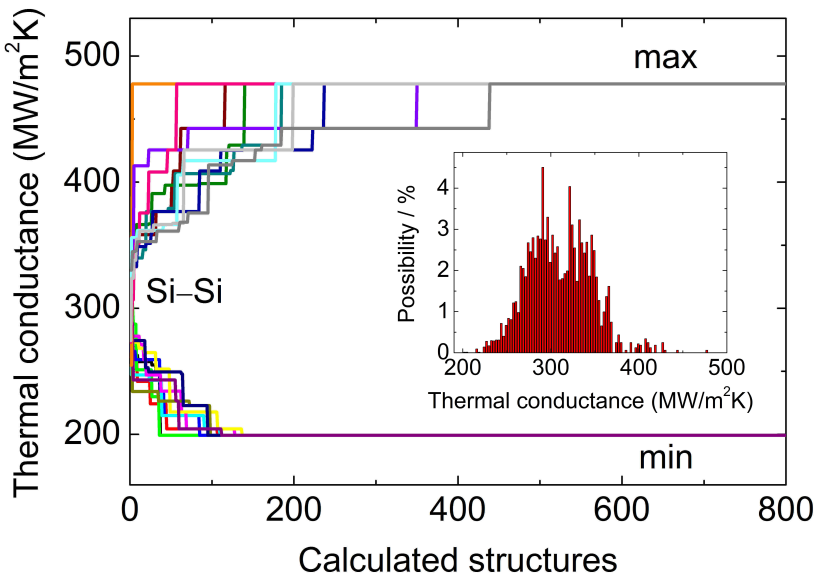
Calculator: Atomistic Green's Function (AGF): Phonon transmission

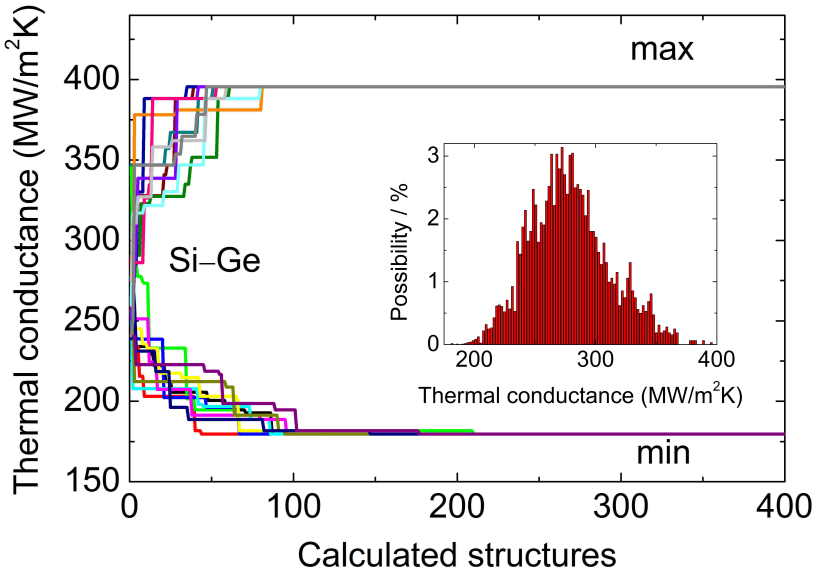
Evaluator: Interfacial Thermal Conductance (ITC)

Optimization method: Thompson Sampling (Bayesian Optimization)

Alloy Structure Optimization

ITC	Si-Si	Si-Ge
Max		
Min		

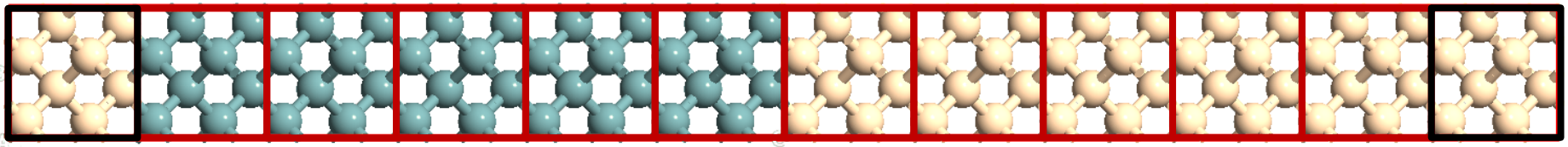




Optimal structures were obtained by calculating only 3.4% of all candidates.

Superlattices Structure Optimization

Topic: Arrange 10-layer superlattices structure (5 layers Si + 5 layers of Ge) between Si and Si to obtain minimal thermal conductance (1 layer thickness = 5.43 A)

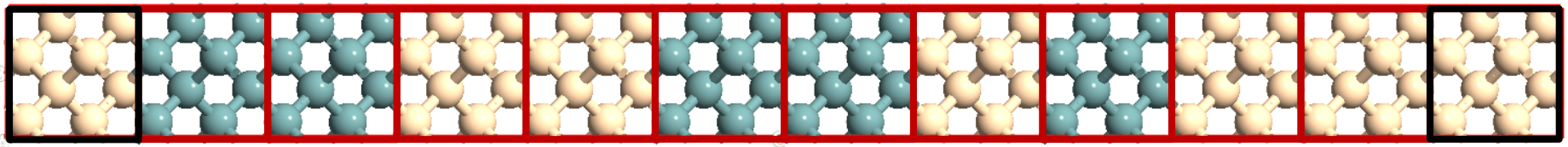


Lead Si Ge Si/Ge superlattice region Lead

Descriptors: 0 1 $C_{10}^{5} = 252$

Case	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	1	0	0	0	0
3	1	1	1	1	0	0	1	0	0	0
…	…	…	…	…	…	…	…	…	…	…

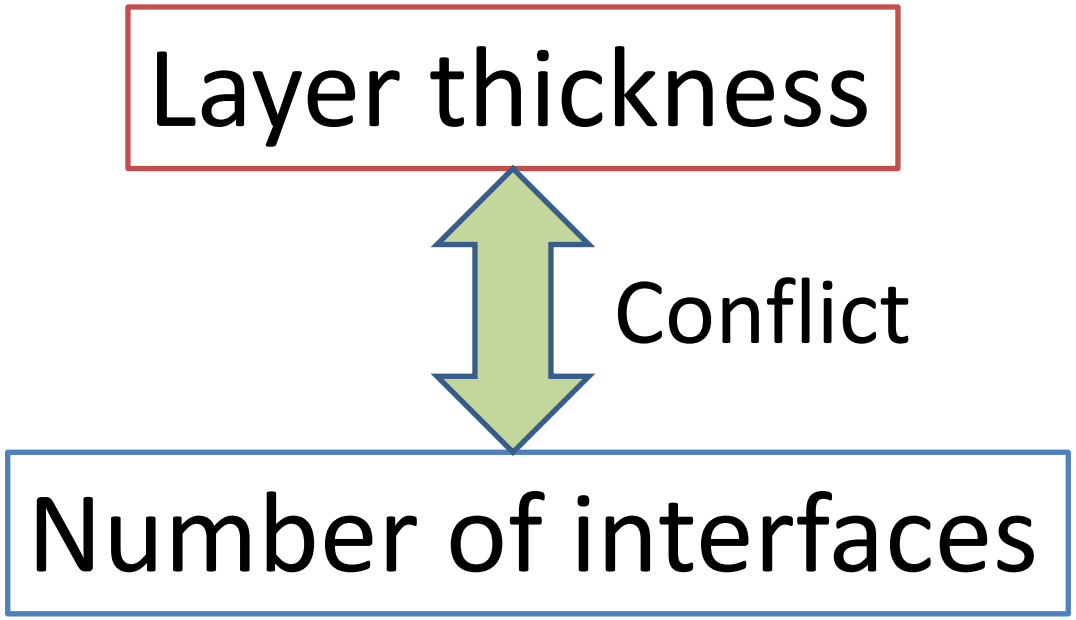
Best Structure:



Superlattices Structure Optimization

Layers	Si-Si Si:Ge=1:1	Si-Si Si:Ge=no limit	Si-Ge Si:Ge=1:1	Si-Ge Si:Ge= no limit
8	11000101 (70)	11101101 (256)	10100110 (70)	10100110 (256)
10	1101010001 (252)	1110110101 (1024)	1000101 10 (252)	1000101 10 (1024)
12	101100100101 (924)	110110101001 (4096)	100101010110 (924)	100101010110 (4096)
14	11011000101001 (3432)	11001010110111 (16384)	10011001010110 (3432)	10010101101110 (16384)
16	1100010010110101 (12870)	1100101110110101 (65536)	1010110110010010 (12870)	1001010101101110 (65536)

Why aperiodic patterns?

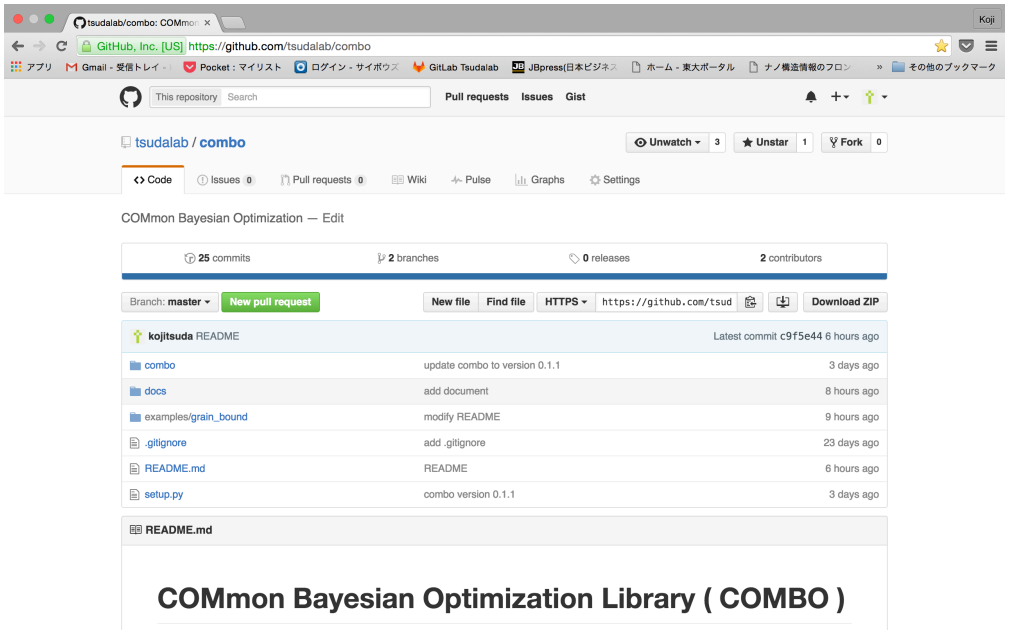


Ju et al., Arxiv 1609.04972, 2016

COMBO: COMmon Bayesian Optimization Python Library

https://github.com/tsudalab/combo

- Fast learning by Thompson sampling, random feature maps, one-rank Cholesky update
- Automatic hyperparameter initialization & update
- Multi-probe design



Ueno et al.,
Materials Discovery,
2016, published online.

GP = Random Feature Map + Bayesian Linear Regression

- Gaussian process (GP) is slow $O(n^3)$ due to the use of kernel function

$$k(\Delta) = \exp(-\|\Delta\|^2/2)$$

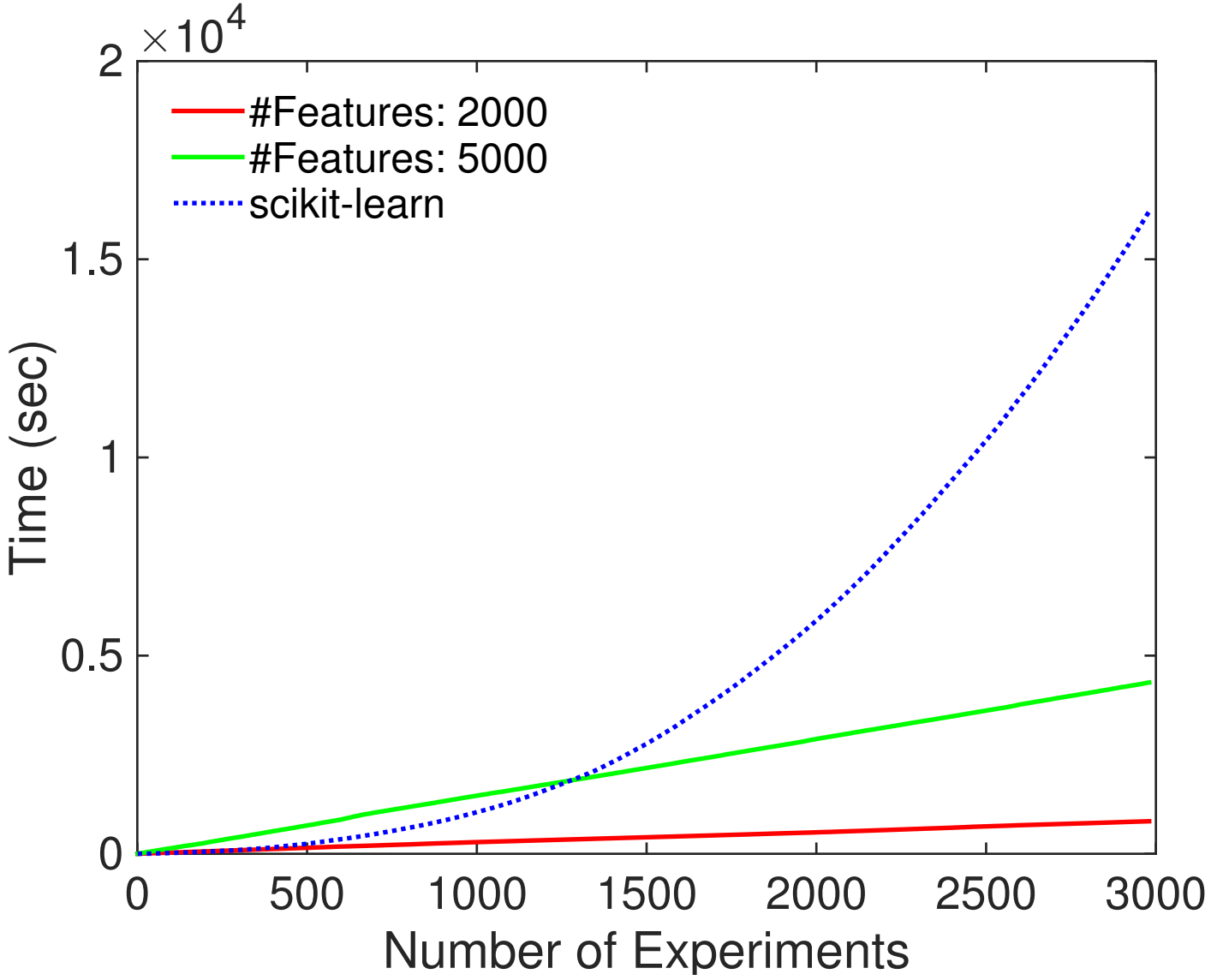
- Approximation by random feature maps (Rahimi and Recht, NIPS 2007)

$$E[z_{\boldsymbol{\omega},b}(\boldsymbol{x}) z_{\boldsymbol{\omega},b}(\boldsymbol{x}')] = k(\boldsymbol{x} - \boldsymbol{x}')$$

$$z_{\omega,b}(\boldsymbol{x}) = \sqrt{2}\cos(\boldsymbol{\omega}^\top \boldsymbol{x} + b)$$

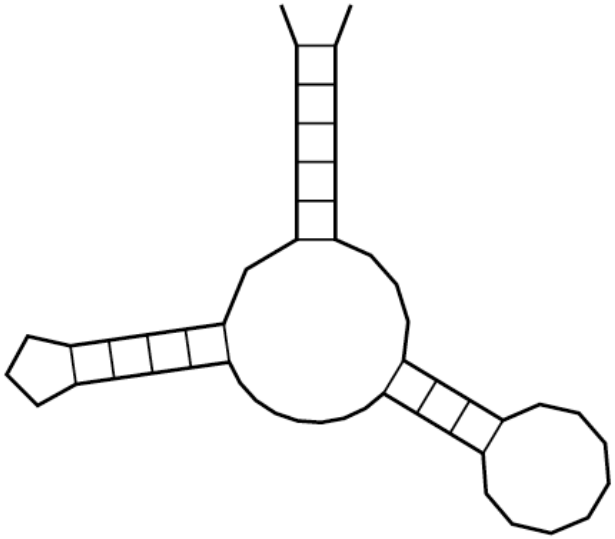
ω is a vector of random samples from unit Gaussian distribution
b is drawn uniformly from $[0,2\pi]$

Computational Time of COMBO

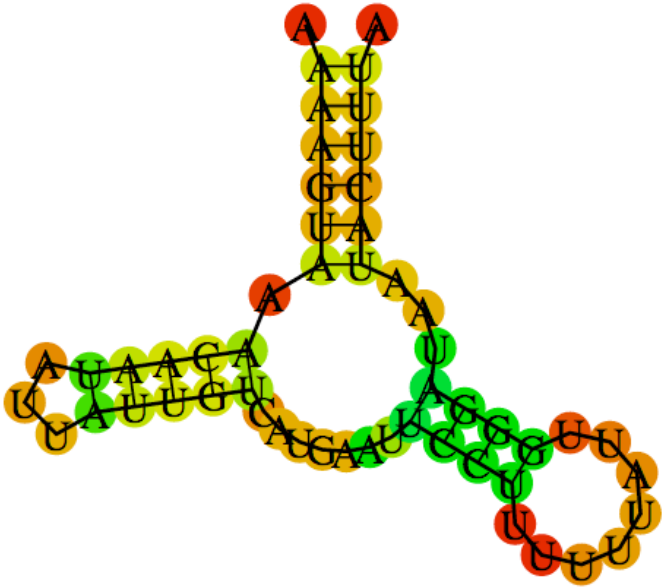


RNA Inverse Folding

- Design RNA whose structure matches target
- 4^n candidates: Too many for Bayesian Opt



Target Structure



AAAAGUAAACAAUAUUAUUGUCAUGAAUUCC
UUUUUUAUUGGGAUAAUACUUUA

Monte Carlo Tree Search

ARTICLE

doi:10.1038/nature16961

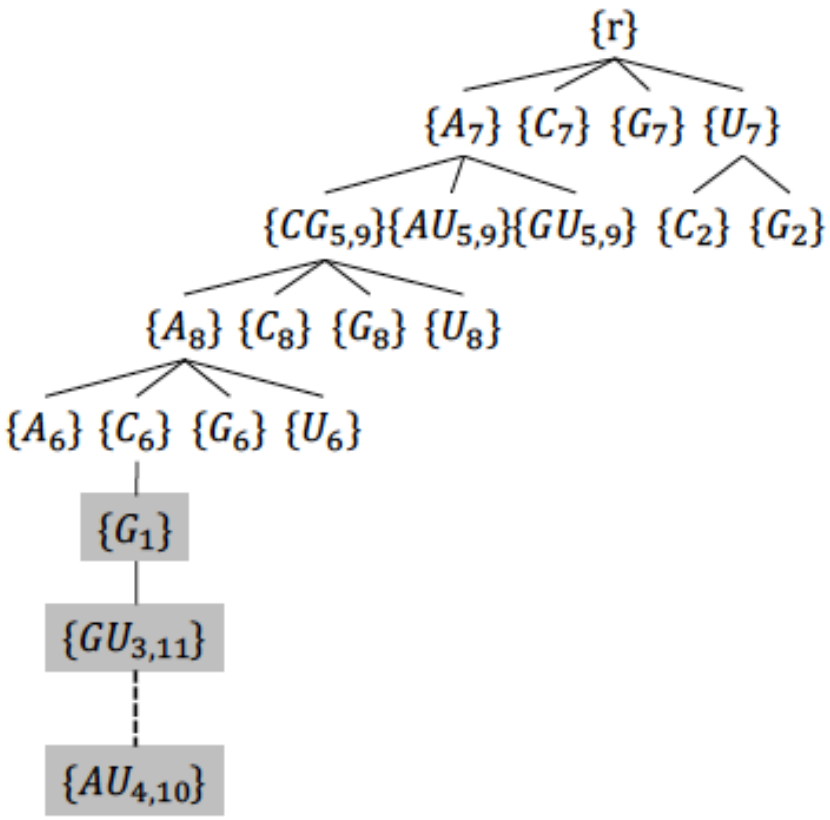
Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

The game of Go has long been viewed as the most challenging of classic games for artificial intelligence owing to its enormous search space and the difficulty of evaluating board positions and moves. Here we introduce a new approach to computer Go that uses 'value networks' to evaluate board positions and 'policy networks' to select moves. These deep neural networks are trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play. Without any lookahead search, the neural networks play Go at the level of state-of-the-art Monte Carlo tree search programs that simulate thousands of random games of self-play. We also introduce a new search algorithm that combines Monte Carlo simulation with value and policy networks. Using this search algorithm, our program AlphaGo achieved a 99.8% winning rate against other Go programs, and defeated the human European Go champion by 5 games to 0. This is the first time that a computer program has defeated a human professional player in the full-sized game of Go, a feat previously thought to be at least a decade away.

Monte Carlo Tree Search

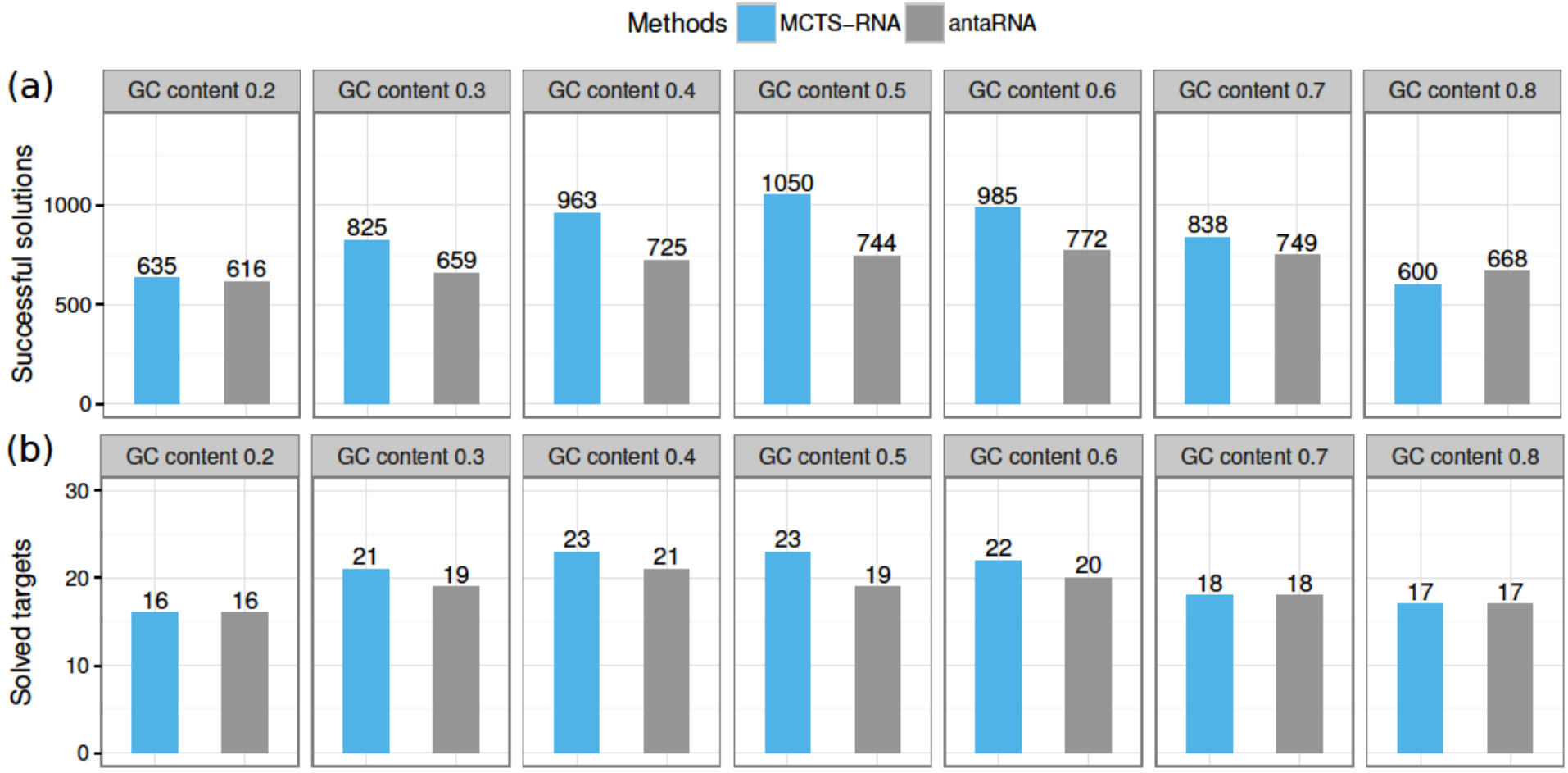
- Candidates at leafs of search tree
- Reward at leafs: Energy
- Score at intermediate node by Playout
  - K random traverses to leafs
  - UCB score: Average reward + Penalty



MCTS-RNA and Existing Tools

Tools	Algorithm	GC content
RNAinverse (Hofacker, 2003)	Local search	No
RNA-SSD (Andronescu *et al.*, 2004)	Stochastic local search+structure decomposition	No
INFO-RNA (Busch and Backofen, 2006)	Dynamic programming+local search	No
NUPACK (Zadeh *et al.*, 2011)	Minimization of ensemble defect and structure decomposition	No
MODENA (Taneda, 2011)	Multi-objective genetic algorithm	No
Frnakenstein (Lyngsø *et al.*, 2012)	Genetic algorithm	No
ERD (Esmaili-Taheri *et al.*, 2014)	Evolutionary algorithm and structure decomposition	No
RNAifold (Garcia-Martin *et al.*, 2013)	Constraint programming and structure decomposition	Yes
IncaRNAtion (Reinharz *et al.*, 2013)	Weighted sampling algorithm	Yes
antaRNA (Kleinkauf *et al.*, 2015)	Ant colony optimization	Yes
MCTS-RNA	Monte Carlo tree search	Yes

Comparison with antaRNA



Data				MCTS-RNA		ERD		MODENA	
Rfam	RfamID	N	ℓ	Sc	time	Sc	time	Sc	time
RF00001	5S_rRNA	117	83	**50/50**	8.38	10/50	3.1	38/50	34.26
RF00002	5_8S_rRNA	151	127	**32/50**	88.32	12/50	3.86	41/50	35.29
RF00003	U1	161	121	**48/50**	83.02	0/50	–	14/50	53.86
RF00004	U2	193	149	**50/50**	1.35	21/50	2.62	38/50	13.16
RF00005	tRNA	74	53	**50/50**	0.3	31/50	1.35	48/50	14.3
RF00006	Vault	89	69	**50/50**	0.167	38/50	0.88	48/50	14.3
RF00007	U12	154	112	**50/50**	0.18	30/50	1.52	46/50	27.37
RF00008	Hammerhead_3	54	39	**50/50**	0.026	33/50	0.67	39/50	12.01
RF00009	RNaseP_nuc	348	293	23/50	61.7	32/50	19.25	**40/50**	118.85
RF00010	RNaseP_bact_a	357	255	0/50	–	0/50	–	0/50	–
RF00011	RNaseP_bact_b	382	286	0/50	–	0/50	–	0/50	–
RF00012	U3	215	176	**50/50**	4.08	8/50	15.3	47/50	72.14
RF00013	6S	185	137	**50/50**	0.6	28/50	2.43	39/50	52.12
RF00014	DsrA	87	58	**50/50**	0.03	32/50	0.77	45/50	15.88
RF00015	U4	140	109	**50/50**	0.73	25/50	1.74	45/50	24.91
RF00016	SNORD14	129	112	0/50	–	0/50	–	0/50	–
RF00017	SRP_euk_arch4	301	200	**50/50**	3.15	2/50	2.48	47/50	175.89
RF00018	CsrB	360	311	0/50	–	0/50	–	0/50	–
RF00019	Y_RNA	83	60	**50/50**	0.1	18/50	0.86	43/50	17.18
RF00020	U5	119	89	0/50	–	0/50	–	0/50	–
RF00021	Spot_42	118	81	**50/50**	0.06	38/50	0.83	31/50	35.07
RF00022	GcvB	148	115	**50/50**	1.05	31/50	1.94	34/50	29.27
RF00024	Telomerase-vert	451	346	0/50	–	0/50	–	0/50	–
RF00025	Telomerase-cil	210	173	**50/50**	20.23	6/50	4.59	44/50	57.59
RF00026	U6	102	97	**50/50**	2	**50/50**	0.73	44/50	14.6
RF00027	let-7	79	48	**50/50**	0.08	46/50	0.76	48/50	16.84
RF00028	Intron_gp	344	291	**19/50**	91.32	**19/50**	46.16	0/50	–
RF00029	Intron_gpI	73	54	**50/50**	1.2	25/50	0.79	46/50	10.82
RF00030	RNase_MRP	340	276	**48/50**	71.4	0/50	–	44/50	122.07
total				**1070/1450**		532/1450		909/1450	

Conclusion

- Machine learning techniques combined with first principles calculation have enormous power