人工知能技術による機能分子・物質設計

津田　宏治

東京大学新領域
メディカル情報生命専攻

津田宏治（つだ こうじ）

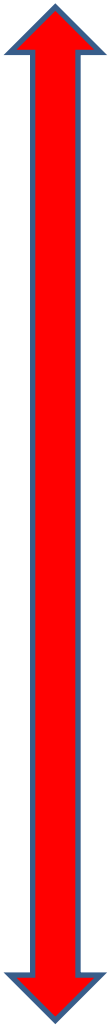
- 経歴
  - 1972 京都生まれ
  - 1998 京大で博士号取得、旧電子技術総合研究所入所
  - 2000 ドイツGMD FIRSTで在外研究
  - 2001 産総研CBRCに配属
  - 2003-2004, 2006-2008 ドイツ・マックスプランク研究所
  - 2014 東京大学大学院教授

  - さきがけ　マテリアルズインフォマティクス　領域アドバイザー
  - 理研革新知能統合研究センター　チームリーダ
  - NIMS MI2I グループリーダ

Discovery of new functional molecules and materials is of national importance

First Principles Calculations

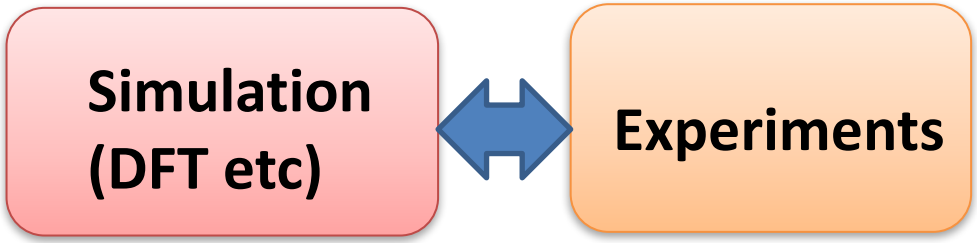
Accurate, Slow



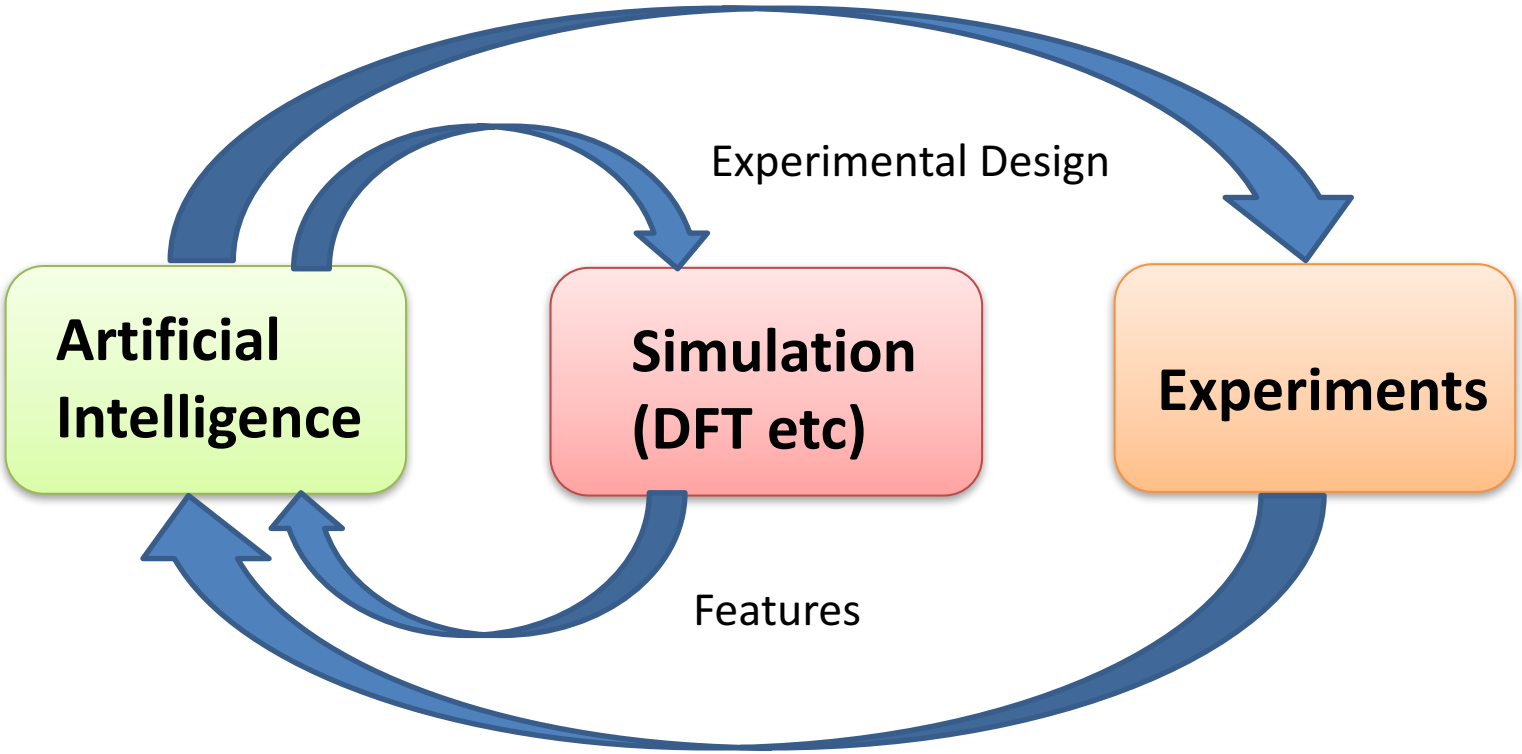
- Full configuration interaction
- Wave function based
- Density functional theory
- Semi-empirical
- Empirical potentials

Inaccurate, Fast

Old Picture



New Picture


Experimental Design
Artificial Intelligence
Simulation (DFT etc)
Experiments
Features

データ駆動科学とベイズ最適化

- データ駆動科学では、データに基づいて、新たな知見・事柄を発見することが求められる
- 単に予測を行うだけでなく、それに基づいて「行動」を設計することが必要
- これらは、「ベイズ最適化」の枠組みに乗ることが多い

講演の構成

- Part 1: ベイズ最適化の基礎
- Part 2: ベイズ最適化の応用

Part1概要

- ガウシアン分布
- ガウシアンプロセス
- ベイズ最適化
- ベイズ最適化による界面構造最適化
- COMBO

多次元ガウシアン分布

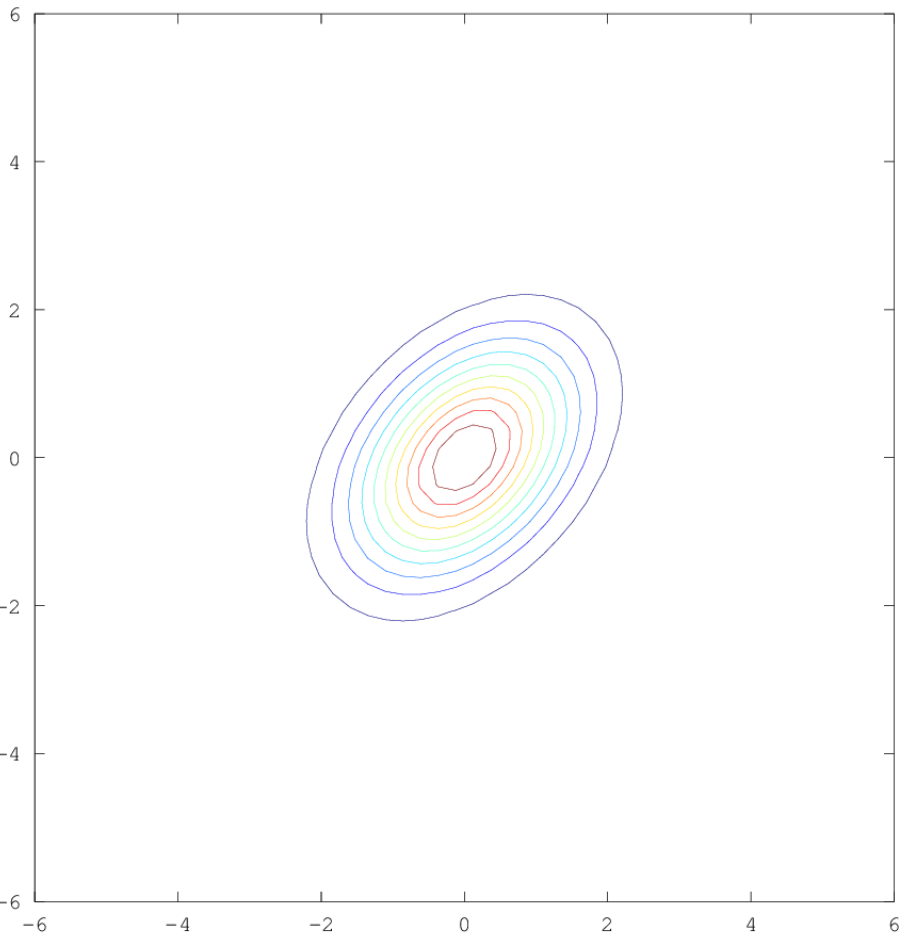
- 多次元ガウシアン分布の確率密度関数

$$p(\boldsymbol{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(\boldsymbol{x}-\mu)^{\top}\Sigma^{-1}(\boldsymbol{x}-\mu))$$

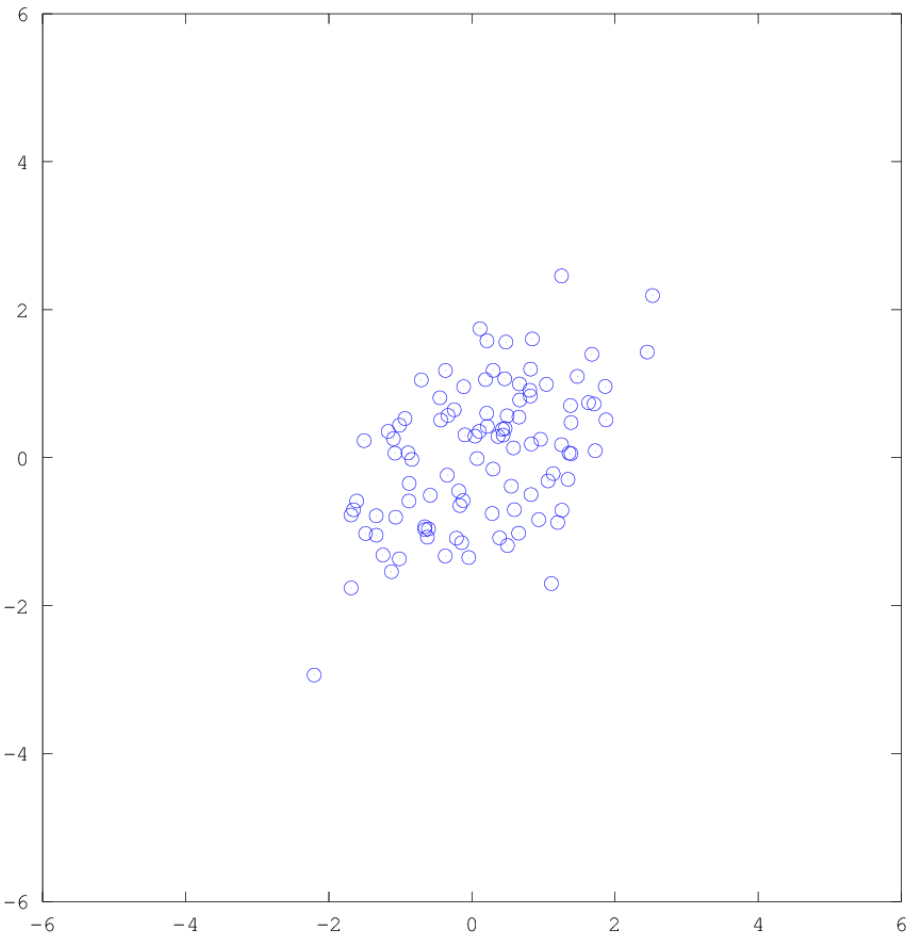
$\boldsymbol{\mu}$　中心点

$\sum$　分散共分散行列

確率密度関数



サンプル (100個）



$$\boldsymbol{\mu} = (0, 0)^\top$$

$$\Sigma = \begin{pmatrix} 0.4 & 1 \\ 1 & 0.4 \end{pmatrix}$$

条件つき分布

平均　　分散共分散行列

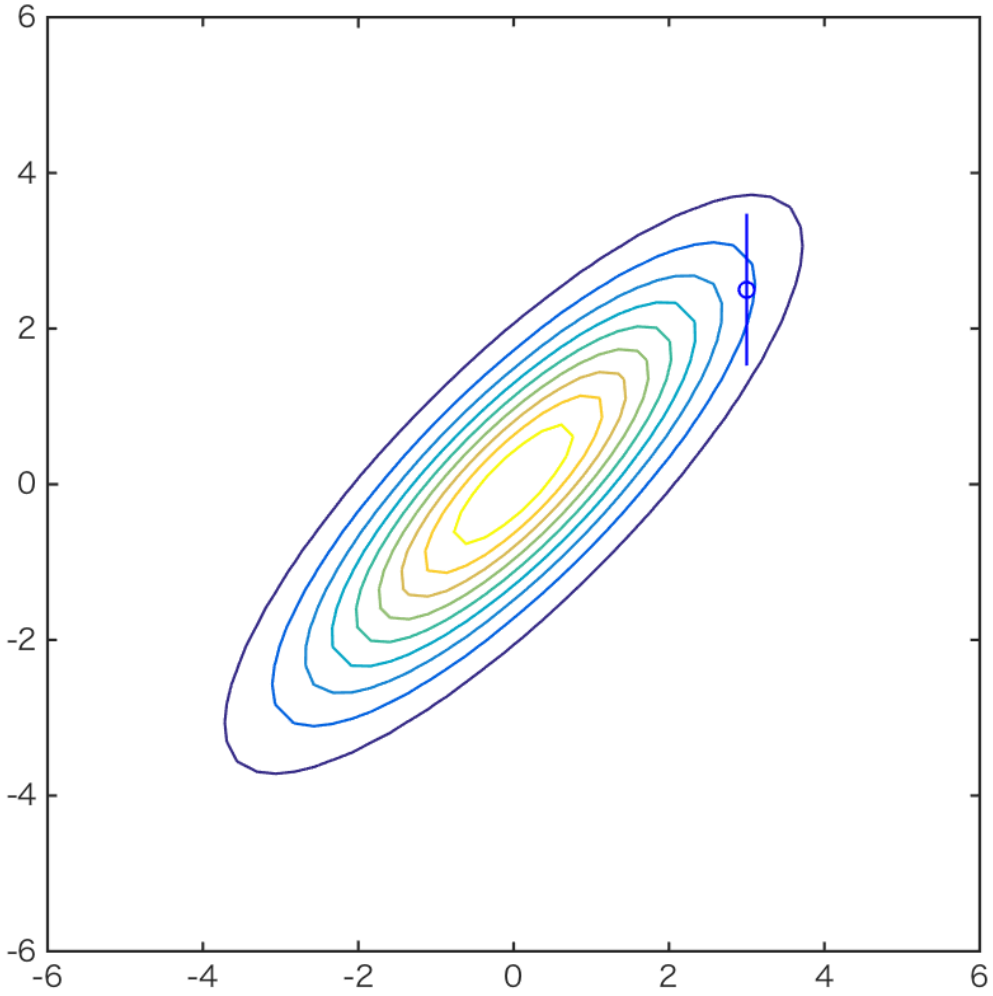
$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$$P(x_1 \mid x_2 = a) = \mathcal{N}(\mu_c, \Sigma_c)$$

$$\mu_c = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

$$\Sigma_c = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

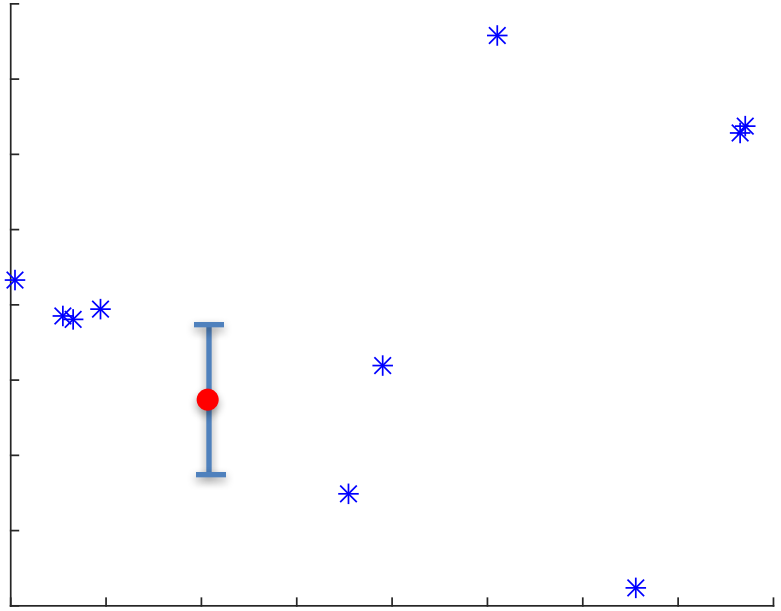
x＝3における条件付き分布



$$\Sigma = \begin{pmatrix} 3 & 2.5 \\ 2.5 & 3 \end{pmatrix}$$

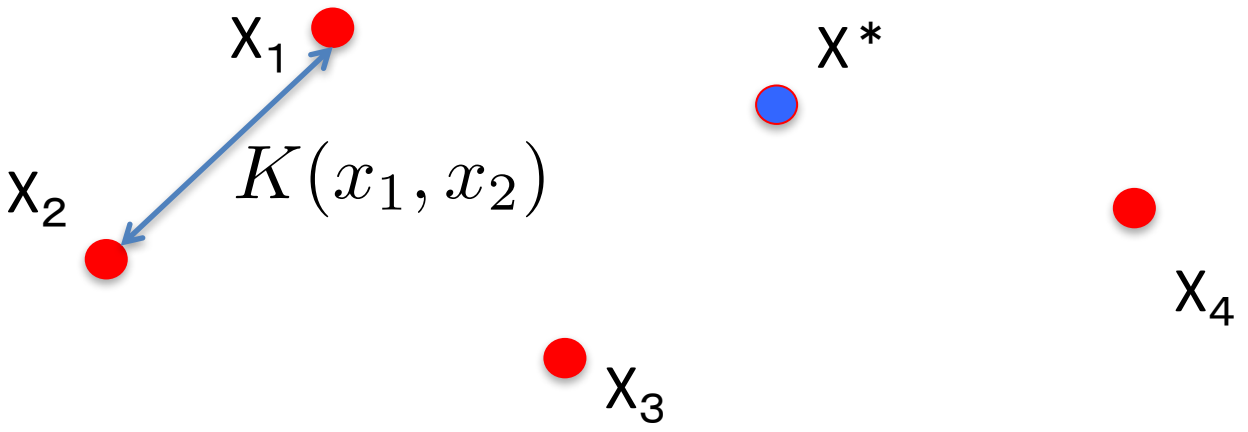
ガウシアンプロセス

- 回帰分析のためのカーネル法
- テストサンプルに対して、予測値だけでなく、予測分散も与えることができる。



ガウシアンプロセス（観測ノイズなし）

- 訓練サンプル点 $\{x_i\}_{i=1,\ldots,n}$、テスト点x^*
- それらにおける観測値y_i, y^*は、n+1次元の多次元ガウシアン分布に従う
- y_iの平均は0であり、分散・共分散は、カーネル関数$K(x_i,x_j)$で与えられる。



ガウシアンカーネルによる
分散共分散行列

$$\begin{pmatrix} k(\mathbf{x}^*, \mathbf{x}^*) & \mathbf{k}^{*\top} \\ \mathbf{k} & K \end{pmatrix}$$

$$K(x, x') = \exp(-\|x - x'\|^2/\eta)$$

Gaussian Process （ノイズなし）

- K: 訓練サンプルに関するカーネル行列
- **y**: 訓練サンプルに関する観測値
- テスト点 ***x****に**おける予測値

$$E[y^*] = \mathbf{k}^{*\top} K^{-1} \mathbf{y}$$

- 予測分散

$$V[y^*] = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^{*\top} K^{-1} \mathbf{k}^*$$

観測ノイズがある場合

- 観測値に、平均0、分散σ^2のノイズが含まれているとき
- 分散共分散行列

$$\begin{pmatrix} k(\mathbf{x}^*, \mathbf{x}^*) + \sigma^2 & \mathbf{k}^{*\top} \\ \mathbf{k} & K + \sigma^2 I \end{pmatrix}$$

Gaussian Process （ノイズあり）

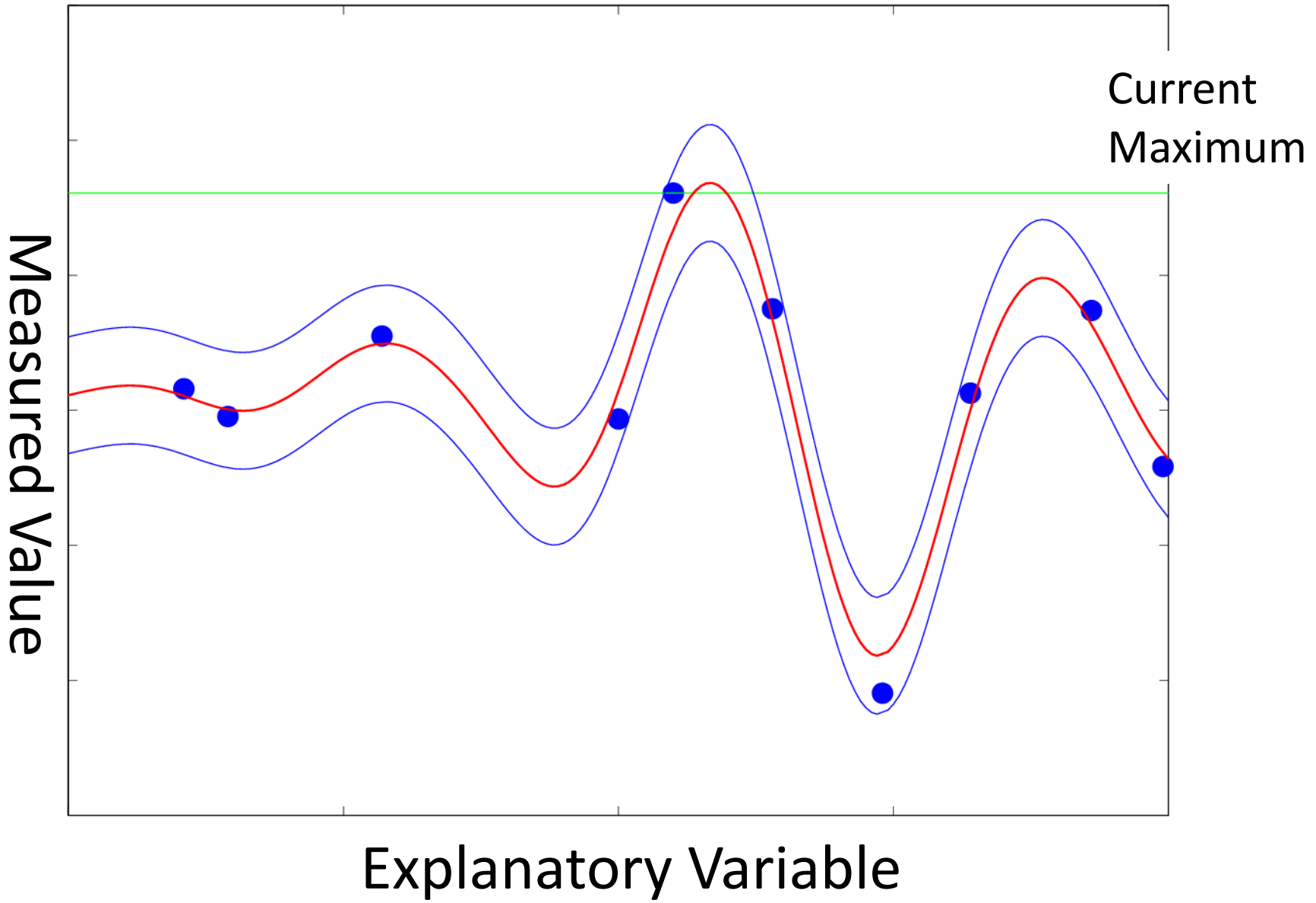
- K: 訓練サンプルに関するカーネル行列
- **y**: 訓練サンプルに関する観測値
- テスト点 ***x****に**おける予測値

$$E[y^*] = \mathbf{k}^{*\top}(K + \sigma^2 I)^{-1}\mathbf{y}$$

- 予測分散

$$V[y^*] = k(\mathbf{x}^*, \mathbf{x}^*) + \sigma^2 - \mathbf{k}^{*\top}(K + \sigma^2 I)^{-1}\mathbf{k}^*$$

Gaussian Process: グリッド点での予測値・分散



ベイズ最適化

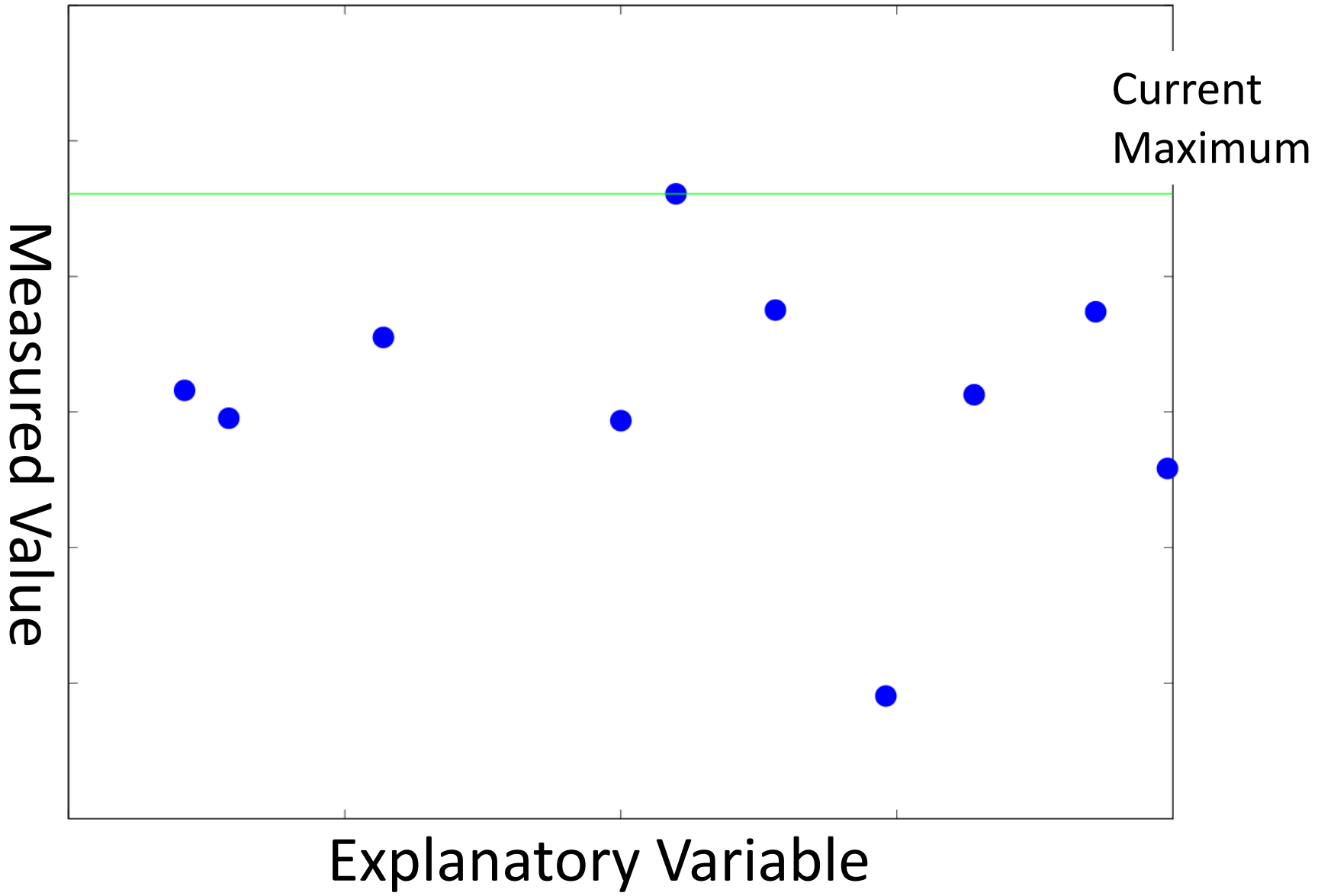
- M個の候補点があり、この中から最大の観測値を持つものを探したい
- できるだけ実験数を少なくしたい

- N個の候補点に対する実験が終わった。M-N個の候補点が残っている
- 次のN+1個目の候補点を最適に選びたい
- N個の化合物から予測モデルを学習し、それを用いて、残りの候補点をスコアリングし決定

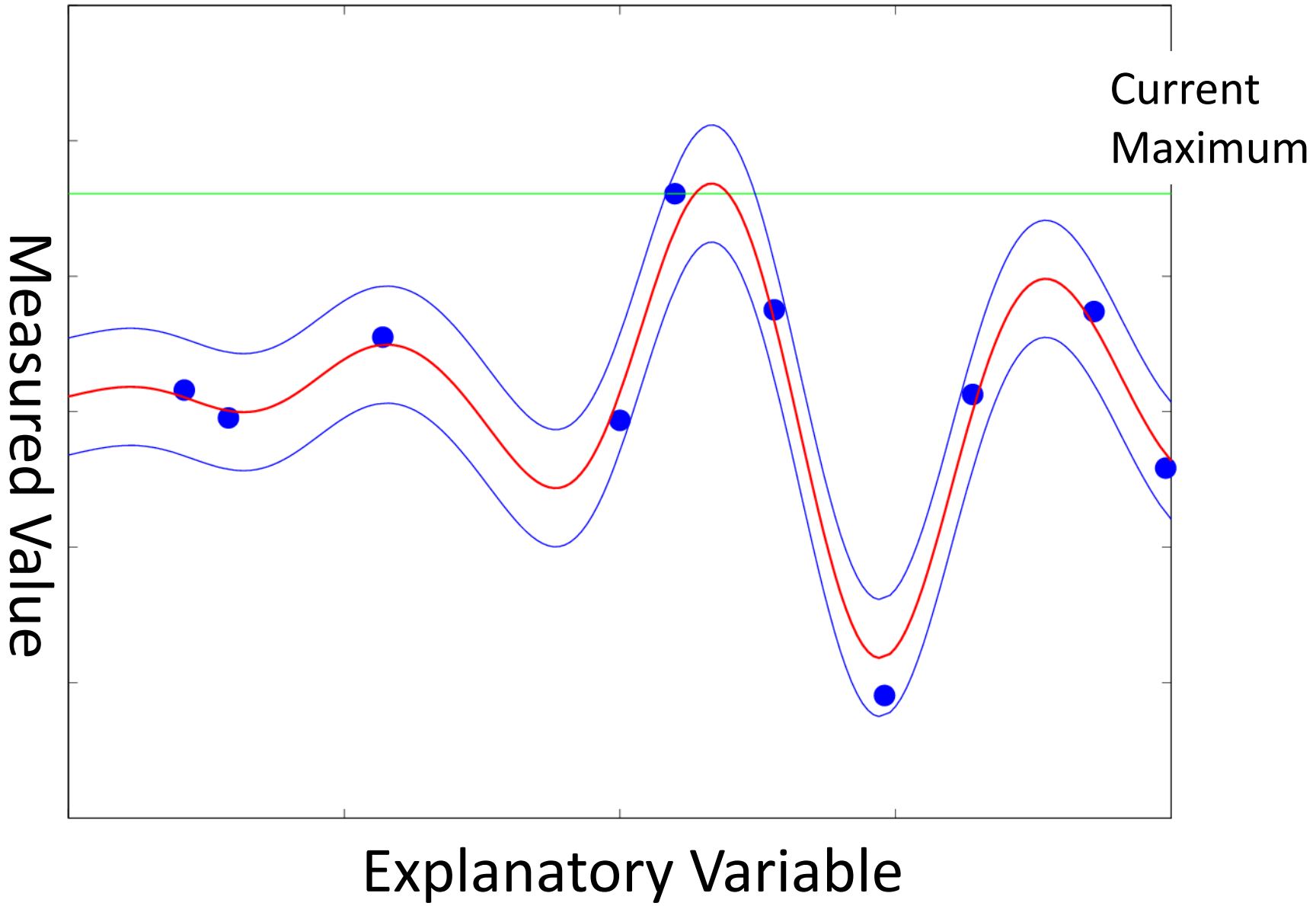
三種類のスコア

- Maximum Probability of Improvement
  - Current Maxを超える確率
- Maximum Expected Improvement
  - (観測値-Current Max)の期待値
- Thompson Sampling
  - 残りのM-N個候補点に対して、条件つき結合確率からサンプリングを行う
  - そのサンプリング値をスコアとする

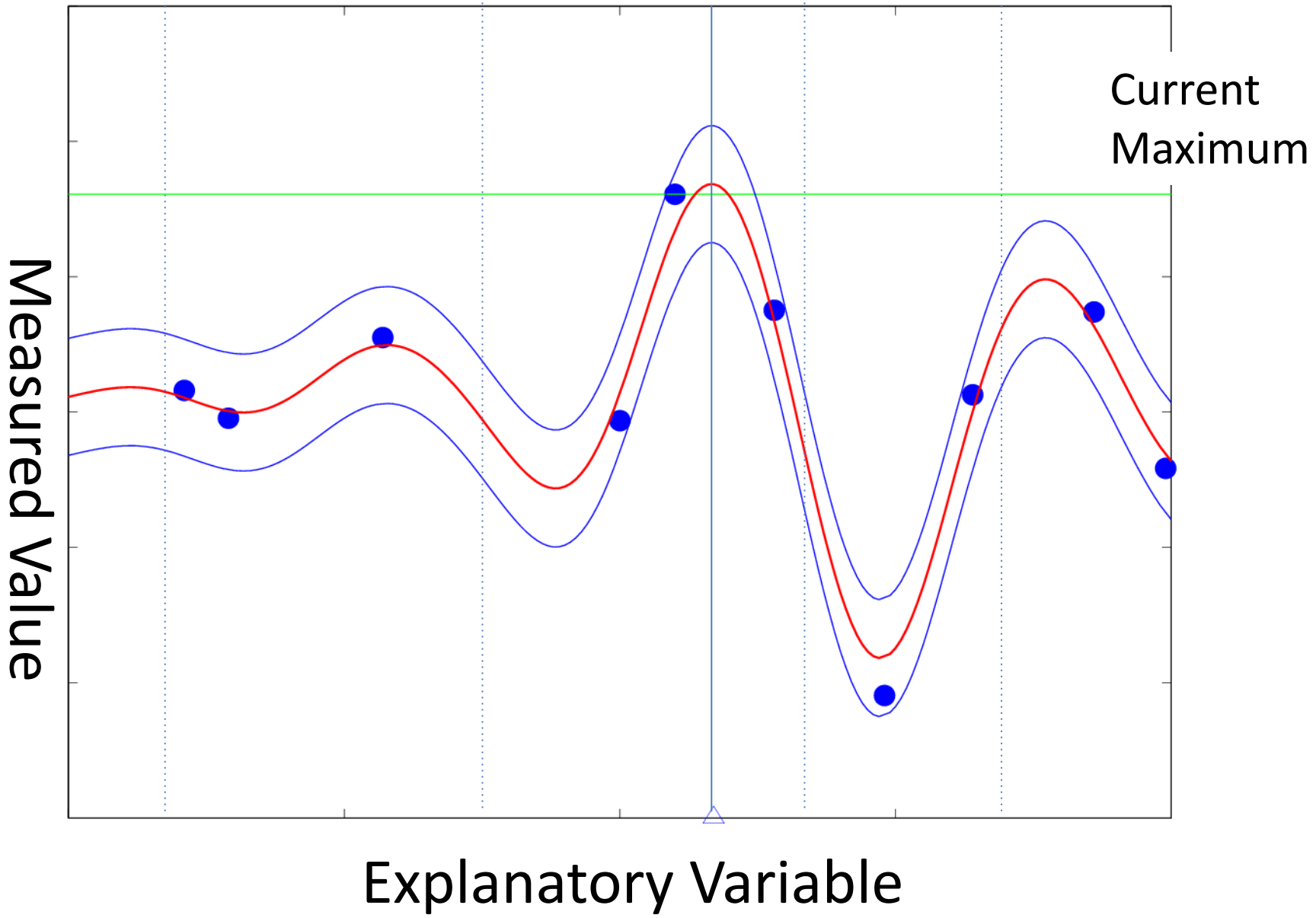
Where to observe next?



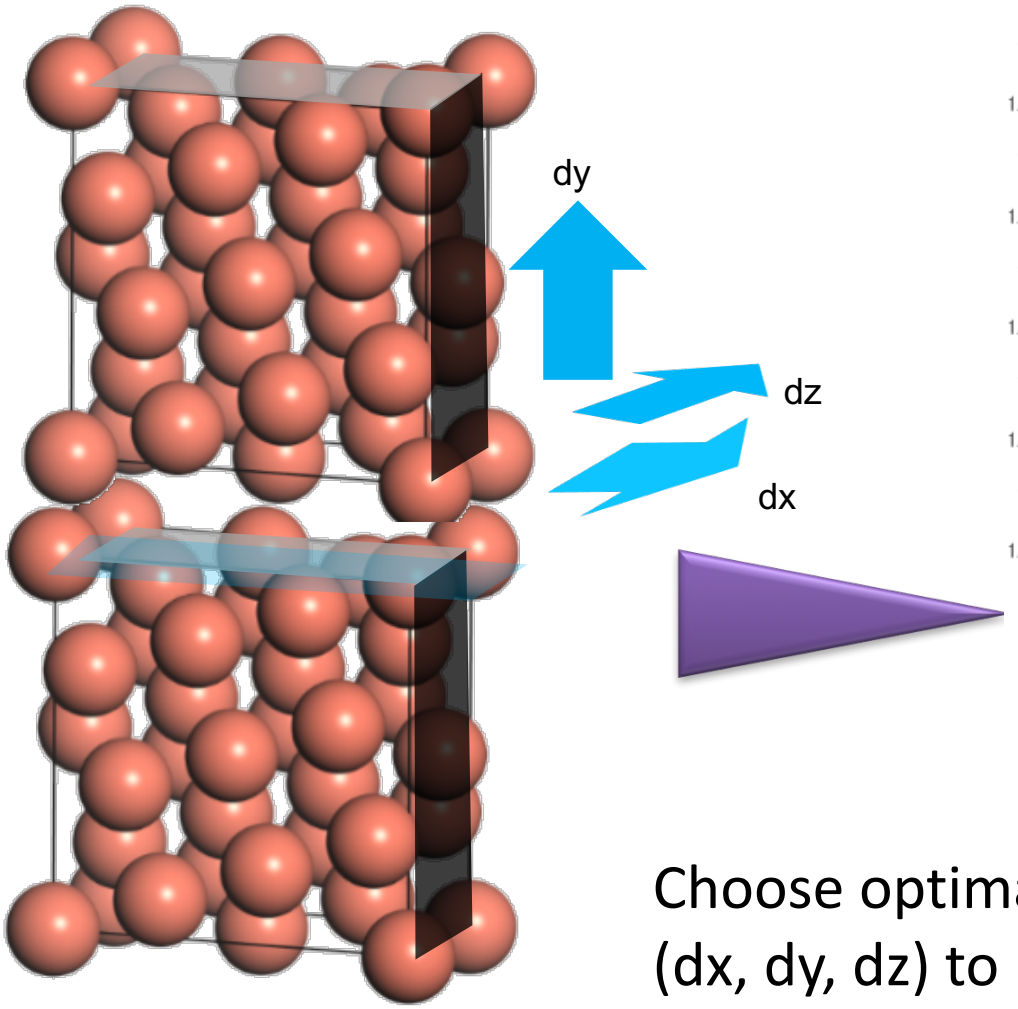
Gaussian Process: Obtain Posterior Distribution

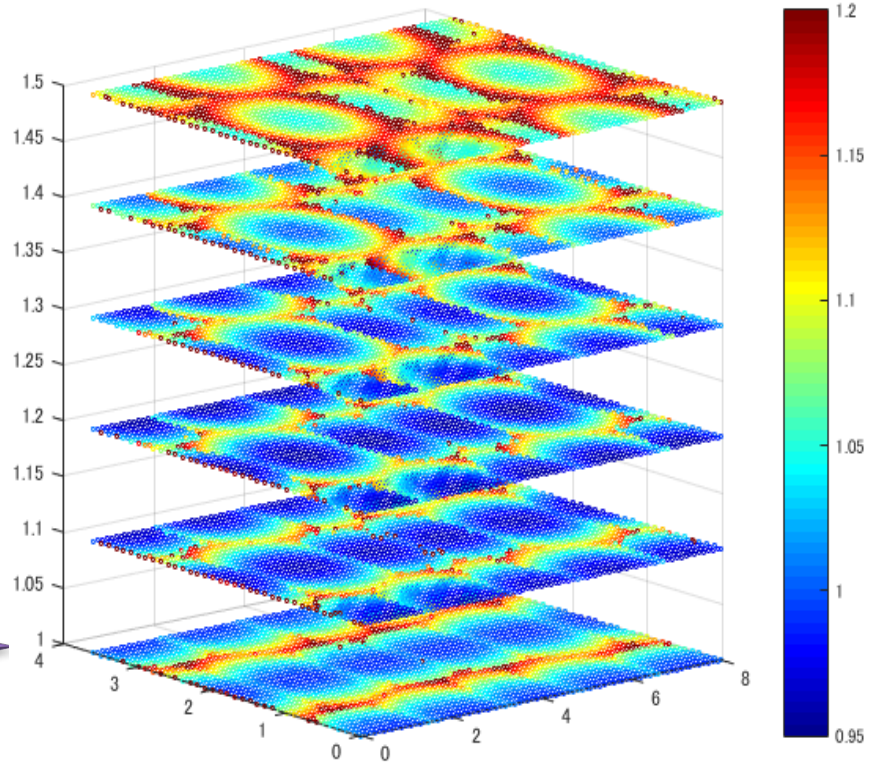


Maximum Probability of Improvement



Grain boundary structure determination



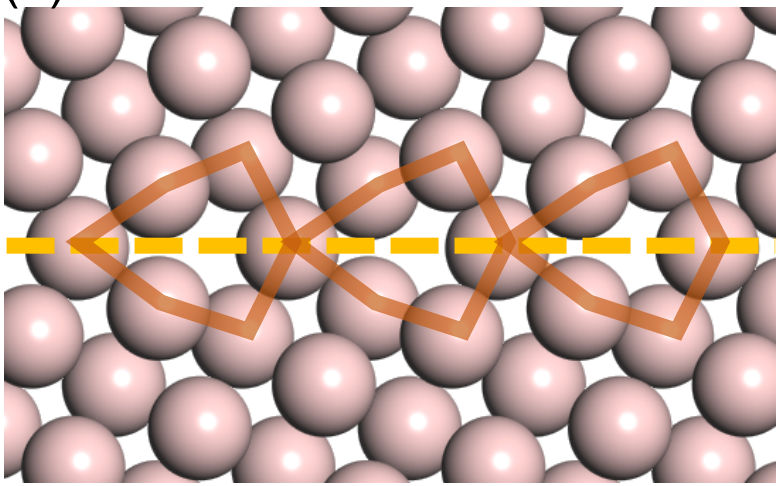


Choose optimal translation parameters (dx, dy, dz) to minimize the grain boundary energy

Acceleration of Discovery

Cu [001] (210) Σ5 grain boundary

(a)

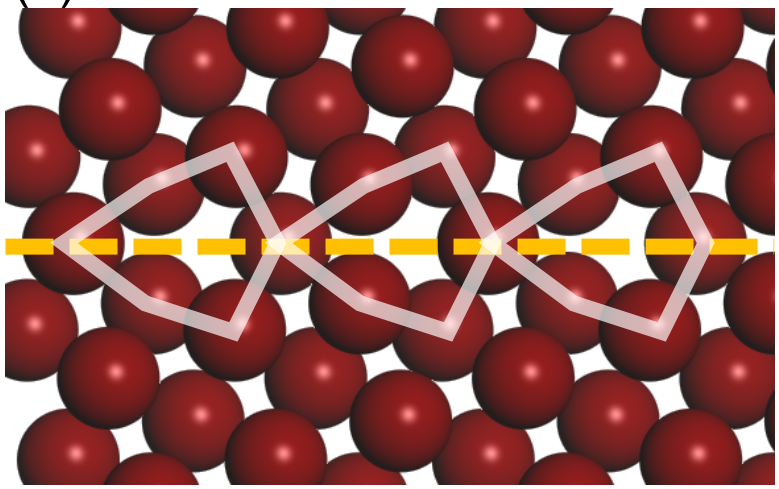


Exhaustive calculations

GB energy=0.96J/m^2

Number of energy calculations
＝16,983

(b)

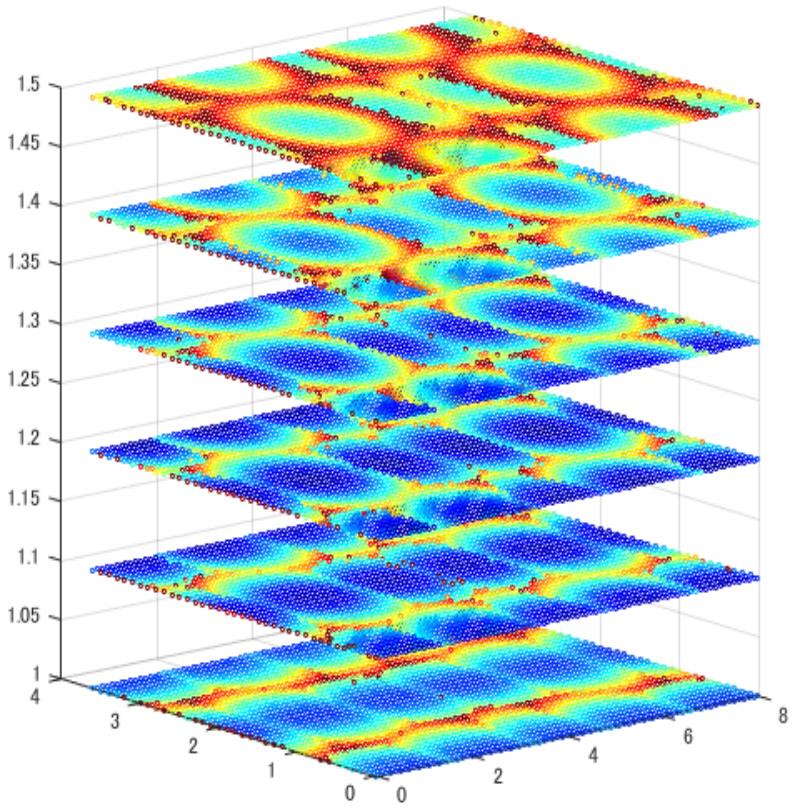


Bayesian optimization

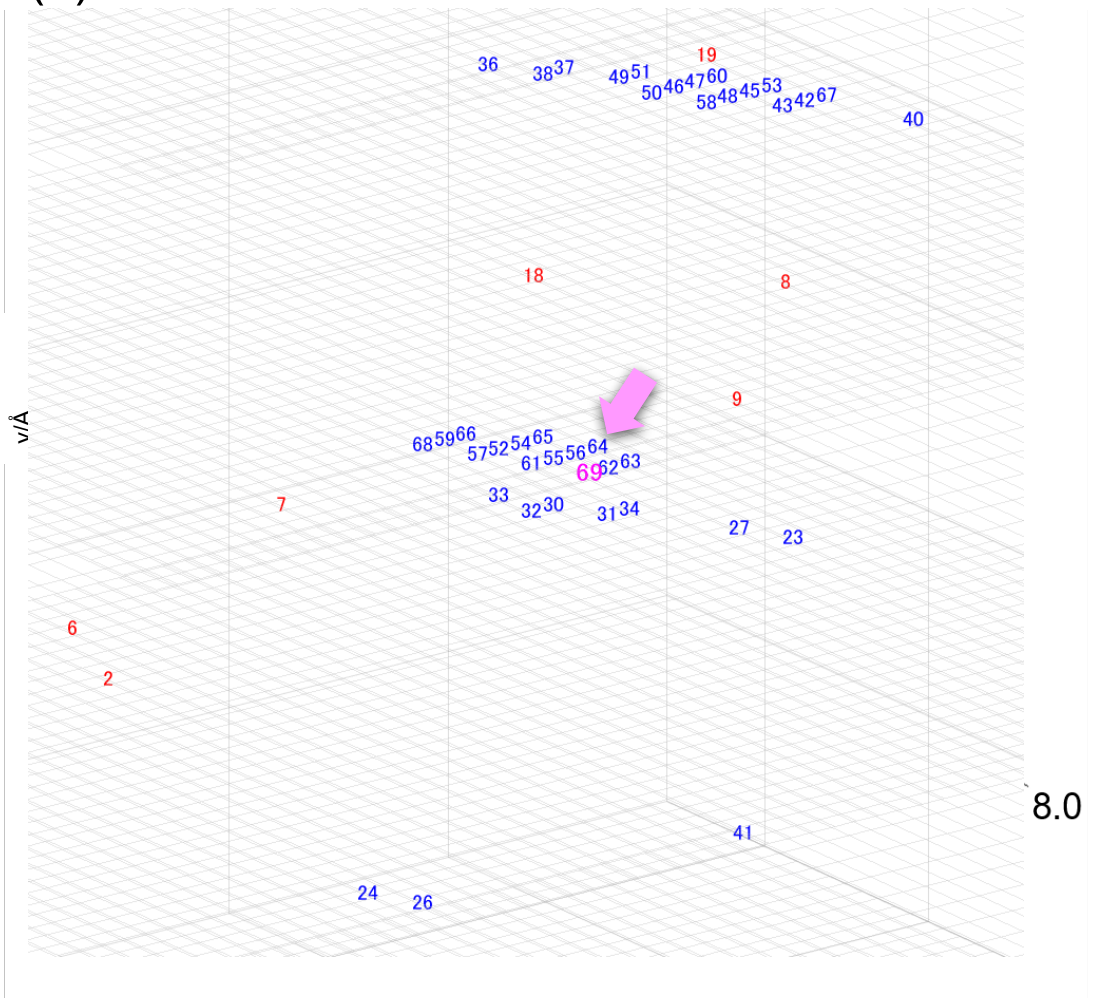
GB energy=0.96J/m^2

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

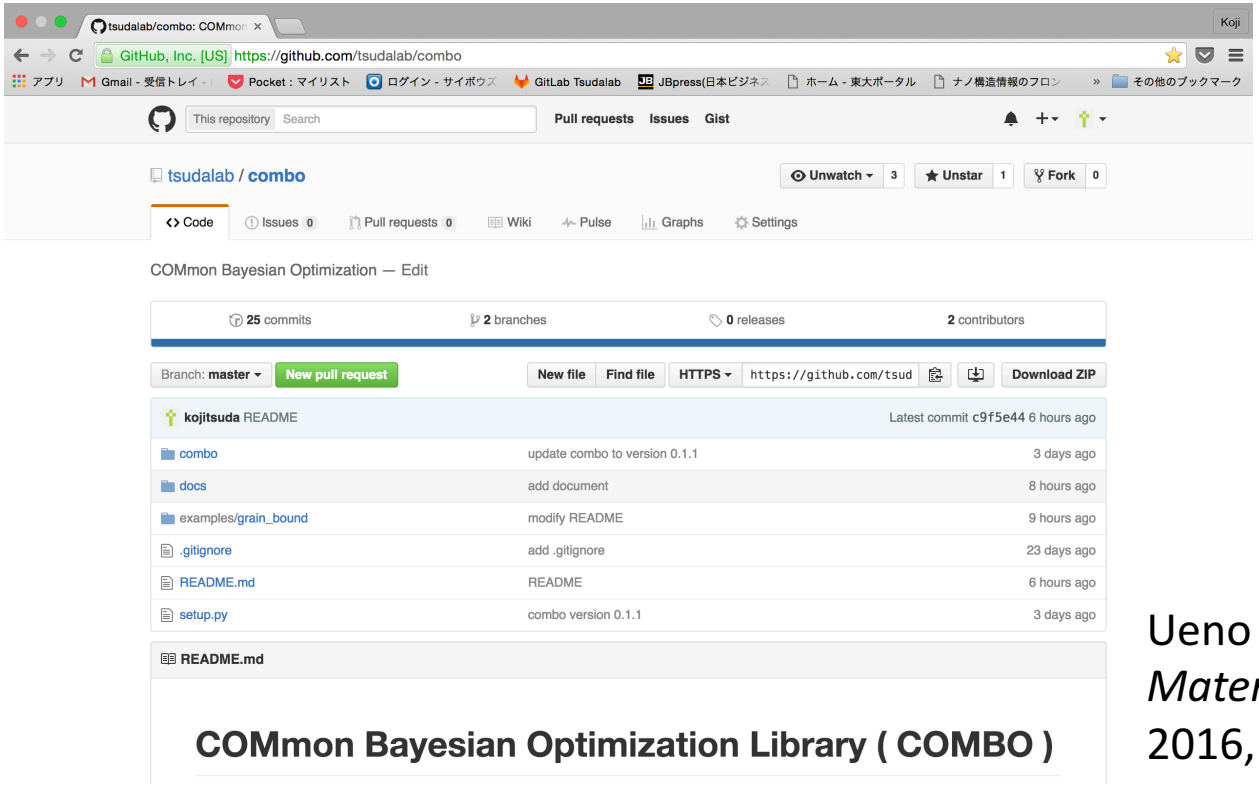
(b)


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

COMBO: COMmon Bayesian Optimization Python Library

https://github.com/tsudalab/combo

- Fast learning by random feature maps
- Automatic hyperparameter initialization & update



Ueno et al., *Materials Discovery*, 2016, accepted.

GP = Random Feature Map + Bayesian Linear Regression

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

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

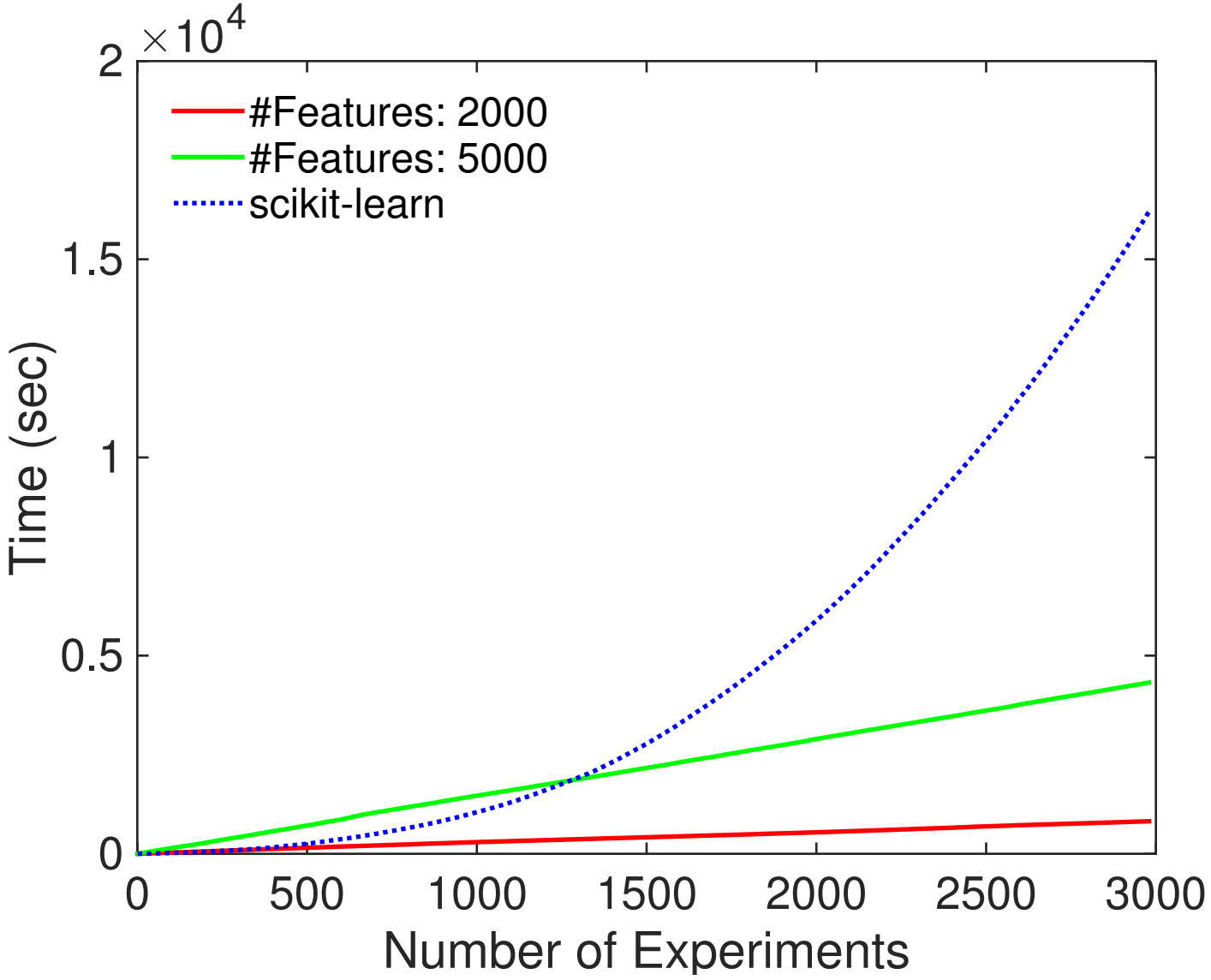
- Approximation by random feature maps (Rahimi and Recht, NIPS 2007)

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

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

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

Computational Time of COMBO



Part 1 まとめ

- データ駆動科学では、データに基づいて、新たな知見・事柄を発見することが求められる
- 単に予測を行うだけでなく、それに基づいて次の「行動」を設計することが必要
- これらは、ベイズ最適化の枠組みに乗ることが多い

Part 2 概要

- Virtual Screening
  - Discovering low-LTC compounds
- Bayesian Optimization
  - Optimization of melting temperature
  - Design of Si-Ge nanostructures

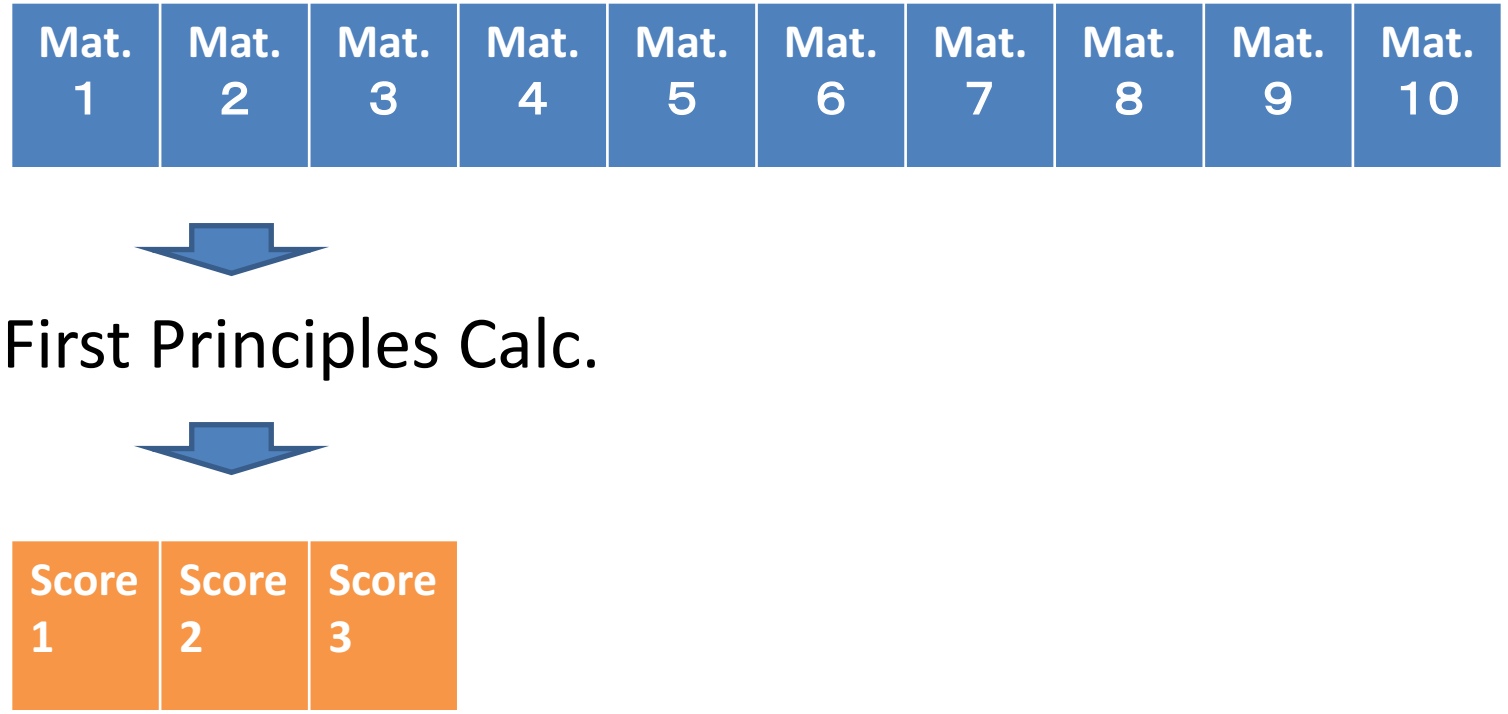
Screening by first principles calculations alone

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

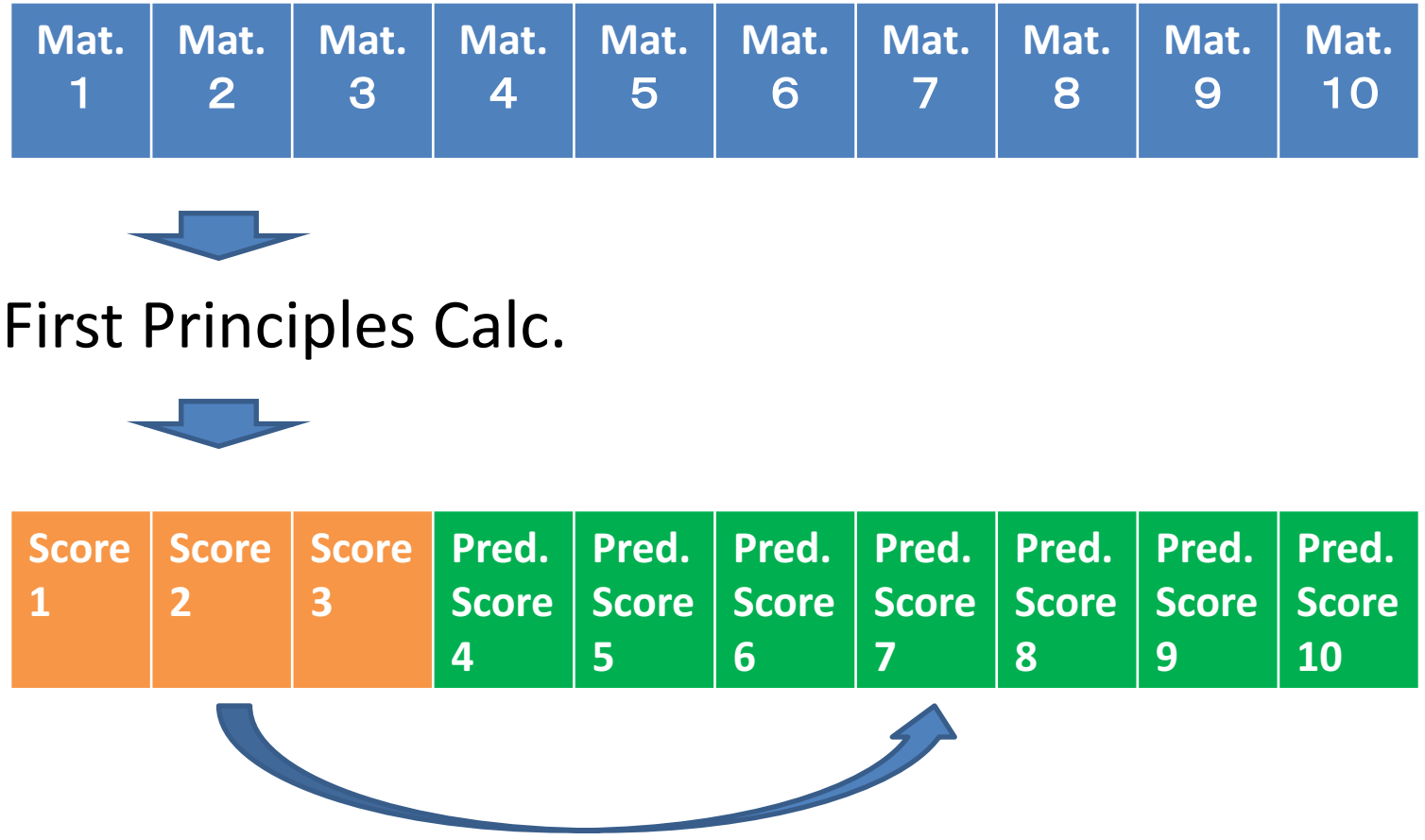
First Principles Calc.

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

Virtual Screening



Virtual Screening



Thermoelectric materials

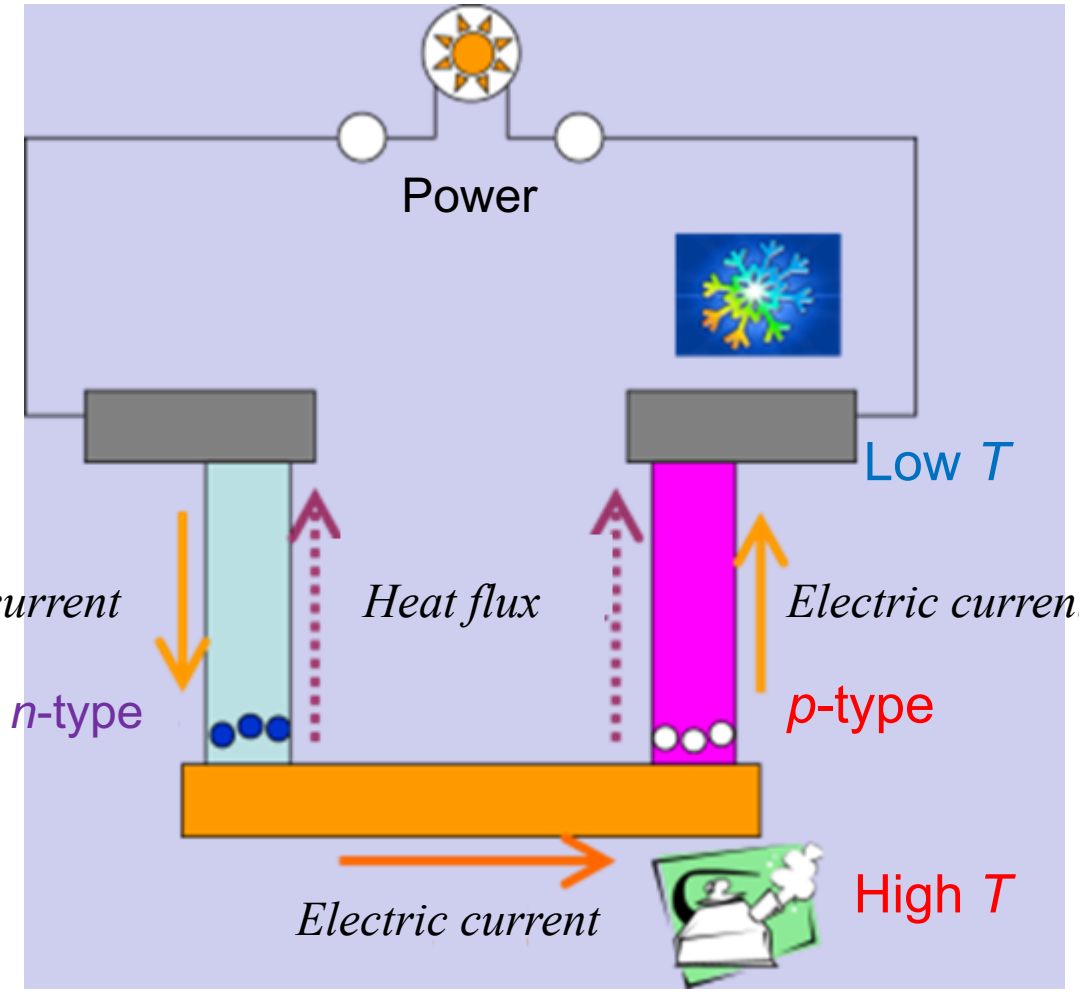


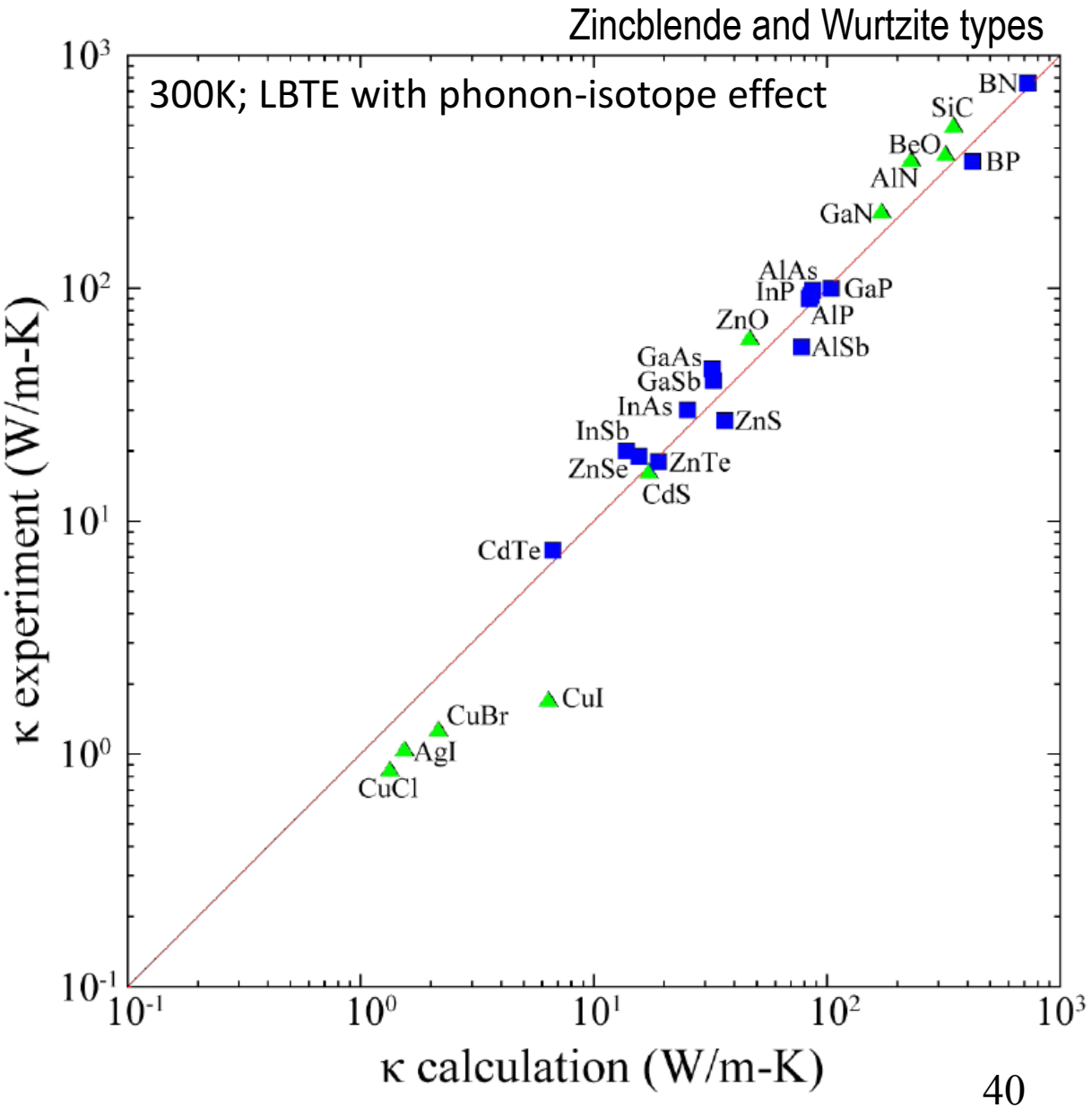
Figure of Merit $(S^2 \sigma/\kappa)T$

S: Seebeck coefficient
σ: electrical conductivity
κ: thermal conductivity

Discovering Compounds of Low Thermal Conductivity: Motivation

- Isao Tanaka's Lab developed a system capable of calculating lattice thermal conductivity (LTC)
  - First-principles anharmonic lattice dynamics calculations
  - Solving Boltzmann transport equation with the single-mode relaxation-time approximation
- Too slow for screening in a large database
  - One LTC can take one WEEK with hundreds of cores

Lattice thermal conductivity : calc. vs. exp.



40

Discovering Compounds of Low Thermal Conductivity from Database

(Seko et al., PRL 2015)

- Compute LTC of 101 prototypical compounds
  - Rocksalt, Zincblende, Wurtzite-type
  - Best LTC: 0.9 W/m·K
- Predict LTC of 54779 compounds in Materials Project Database
  - Additional LTC calculations for best 8 compounds
  - Five had impressive LTC of < 0.2 W/m·K (@300K)

TABLE I. First principles LTCs and Z-scores for highly ranked compounds by the virtual screening. Band gaps by DFT-PBE are taken from MPD library[29, 33].

Ranking	Z-score	Formula	Space group	LTC (W/mK)	Band gap (eV)
1	1.90	$PbRbI_3$	$Pnma$	0.10	2.46
2	1.76	PbIBr	$Pnma$	0.13	2.56
3	1.56	$PbRb_4Br_6$	$R\bar{3}c$	0.08	3.90
4	1.56	PbICl	$Pnma$	0.18	2.72
5	1.56	PbClBr	$Pnma$	0.09	3.44
7	1.44	PbI_2	$R\bar{3}m$	0.29	2.42
8	1.43	PbI_2	$P6_3mc$	0.29	2.45
121	0.39	K_2CdPb	$Ama2$	0.45	0.18
144	0.29	$Cs_2[PdCl_4]I_2$	$I4/mmm$	0.31	0.88

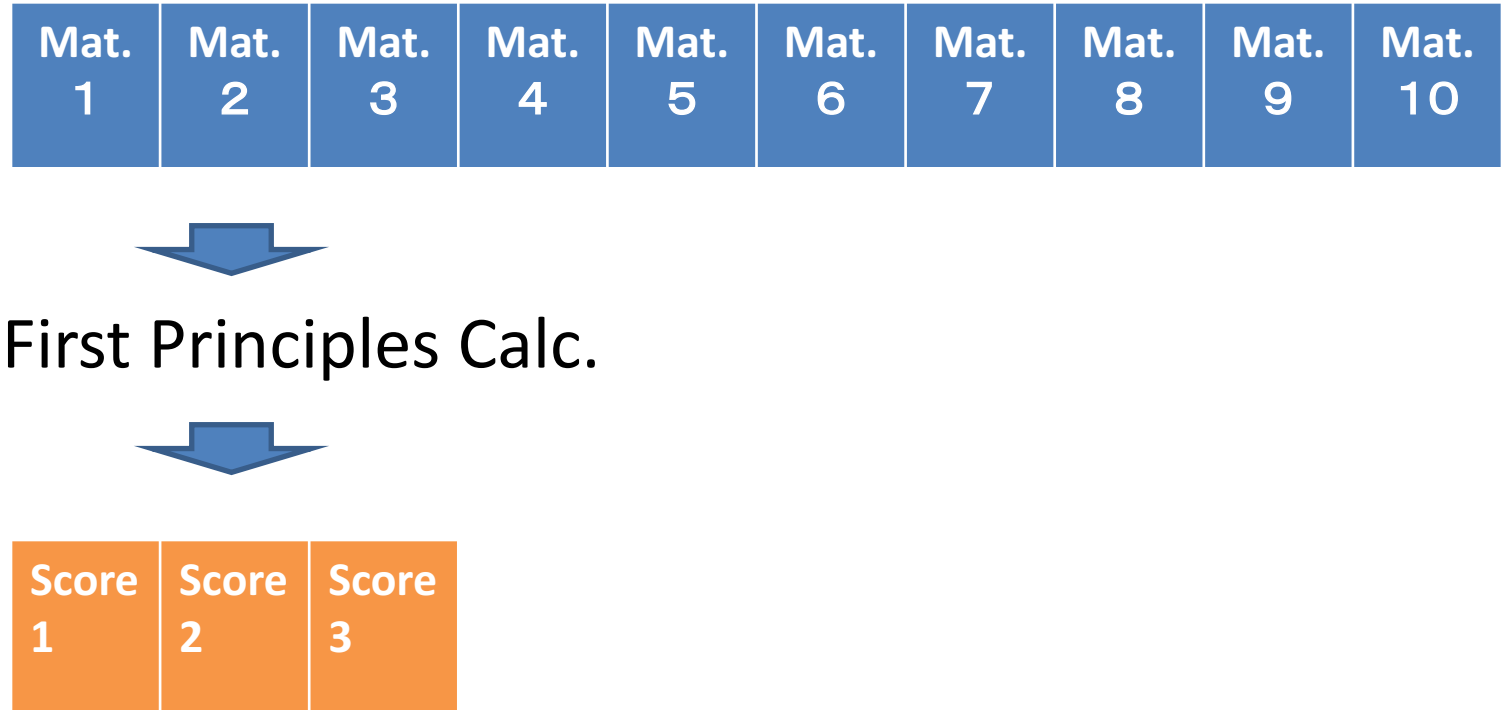
Bayesian Optimization

(Jones et al., 1998)

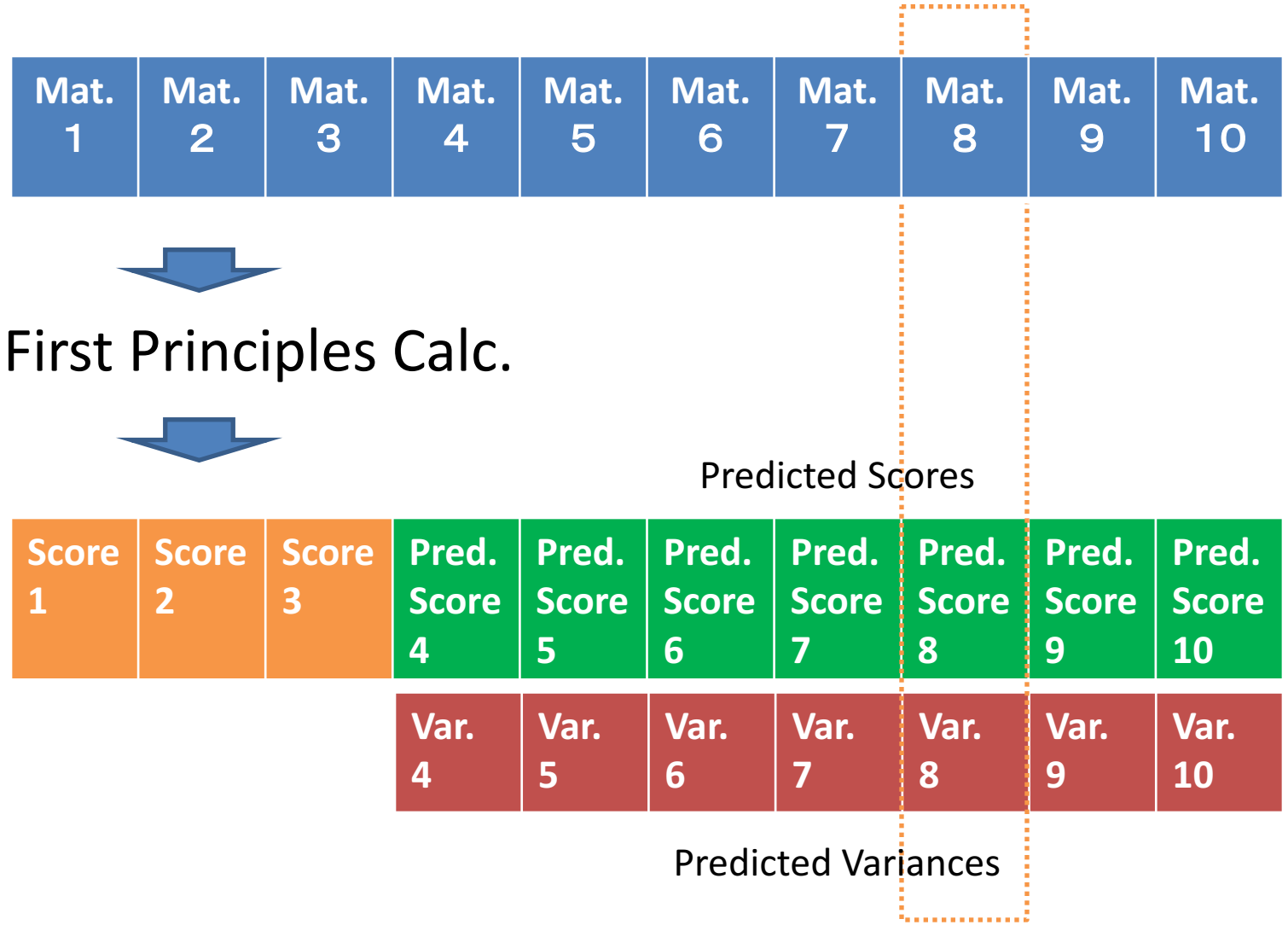
- Find best data points with minimum number of observations

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

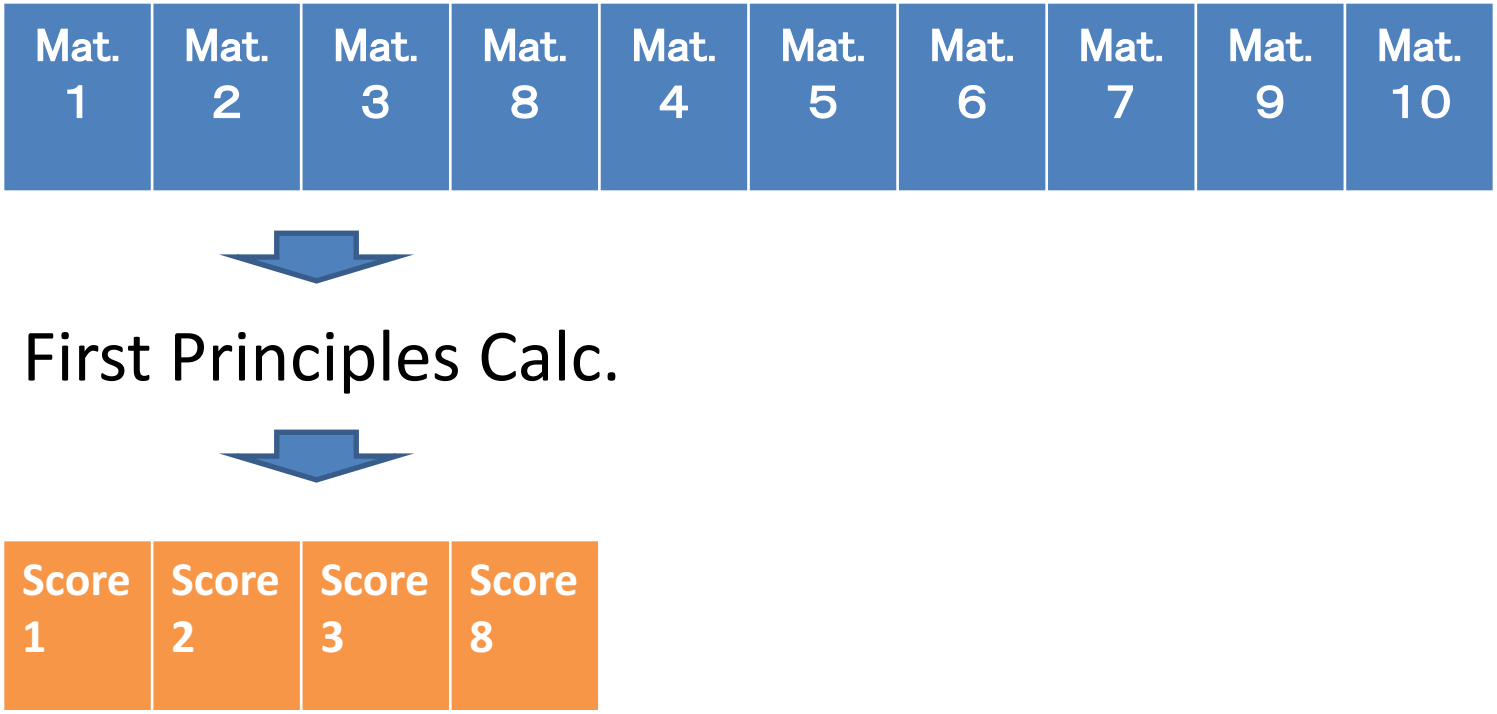
Bayesian Optimization (1)



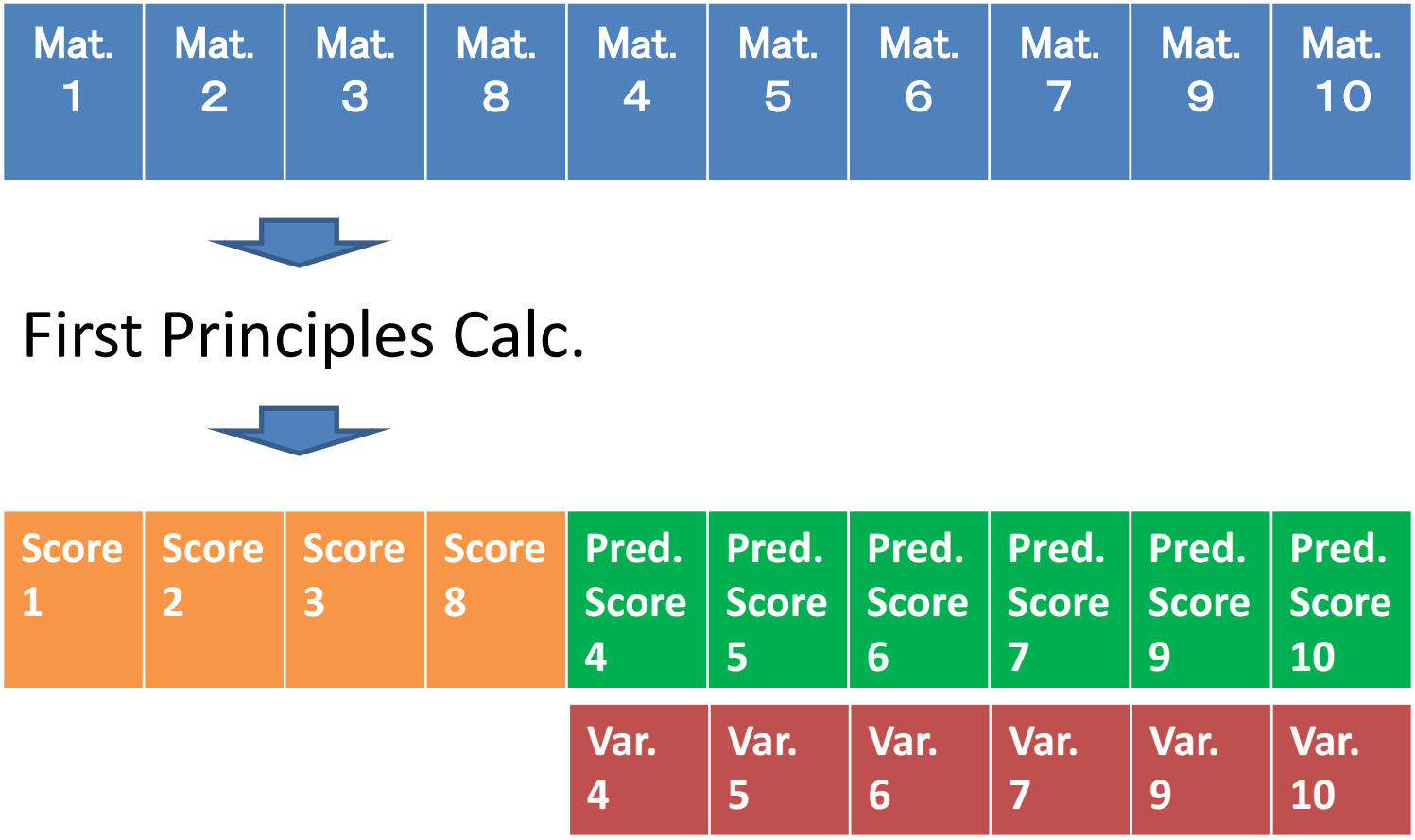
Bayesian Optimization (2)



Bayesian Optimization (3)



Bayesian Optimization (4)



二元化合物の融点データを用いた計算実験

- 226個ある材料のなかから、融点が最高のものを発見する
- 5%をランダムに選んで融点を観測する
- その後、ベイズ最適化を用いて、観測順を自動的に決定していく

１７個の説明変数

#Ecoh:一原子あたりの凝集エネルギー(計算値)

#bm:体積弾性率(計算値)

#V:一原子あたりの格子体積(計算値)

#NN:最近接原子間距離(計算値)

#c:組成

#Z1:構成元素の原子番号の二乗和

#Z2:構成元素の原子番号の積

#Z3:構成元素の原子番号の和

#M1:構成元素の原子量の二乗和

#M2:構成元素の原子量の積

#M3:構成元素の原子量の和

#n1:構成元素の価電子数の二乗和

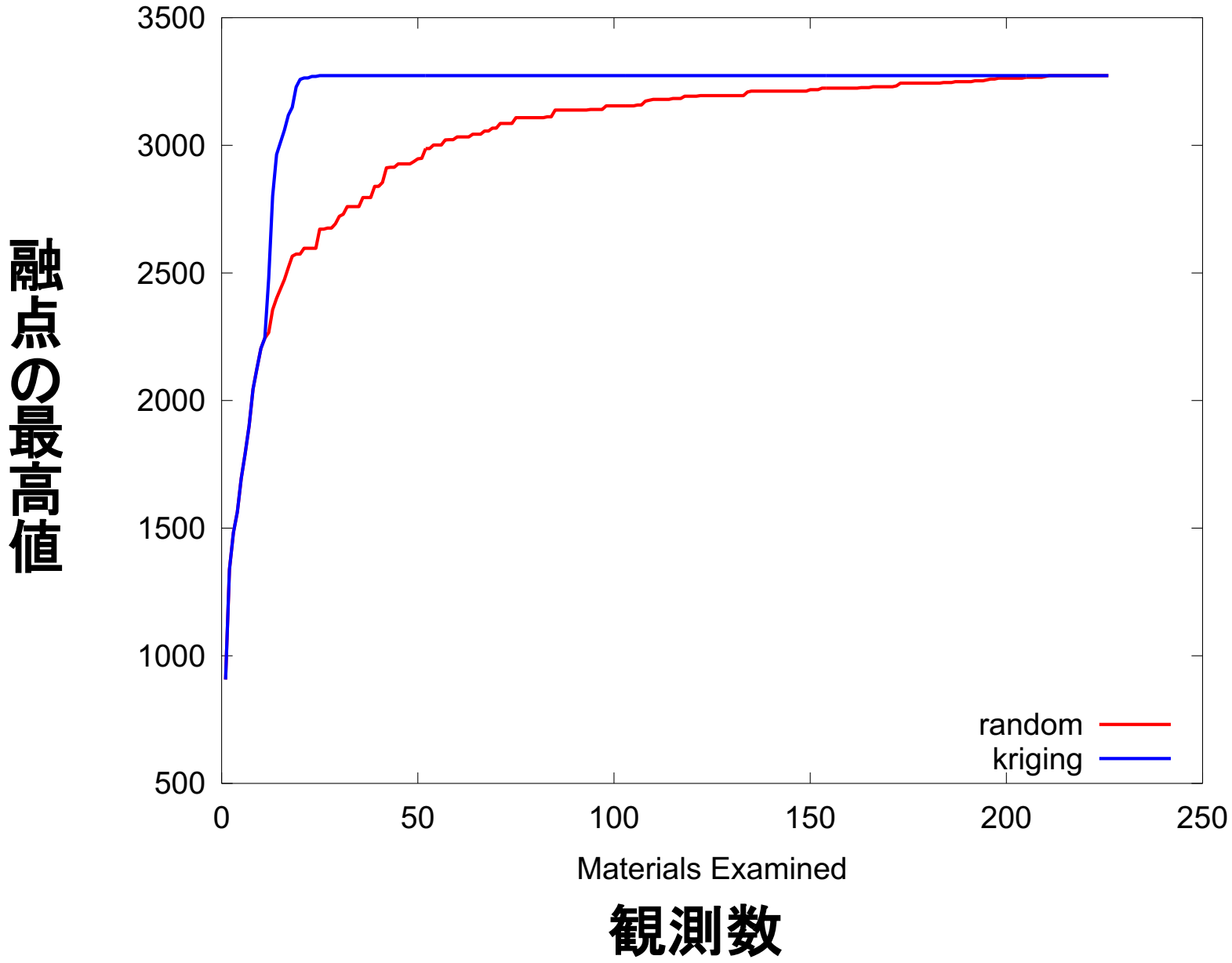
#n2:構成元素の価電子数の積

#n3:構成元素の価電子数の和

#p1:構成元素の周期の二乗和

#p2:構成元素の周期の積

#p3:構成元素の周期の和


融点の最高値
3500
3000
2500
2000
1500
1000
500
0
50
100
150
200
250
Materials Examined
観測数
random
kriging

最高融点の材料を見つけ出すまでの平均観測数

ベイズ最適化
16.1回

ランダム
133.4回

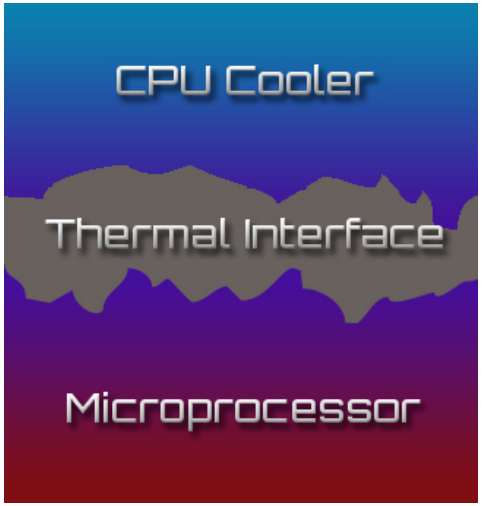
ベイズ最適化による実験順

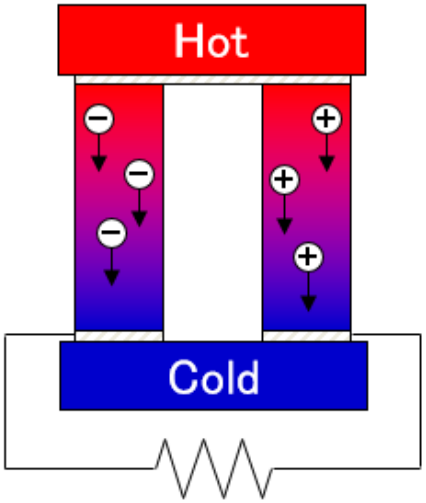
- AlBr3 - As4S4 - GeSe - Se - BaSe - SnO2 - Sb2S3 - Sb2Te3 - Pb - SnF2 - GeBr2 - SnSe2 - BaO - BaS - SrSe - SiC - BeO - **[AlN]** - Be3N2 - Al2O3 - Si3N4 - Al4C3 - MgO - CaO - CaC2 - LiH - Cs - Be - BaH2 - Bi2O4 - K - BeF2 - Tl - RbN3 - LiF - PbTe - CsI - Li - P2O5 - Tl2O3 - BaF2 - Bi - Ba - CaS - SrO - CaSi - PbO - CaF2 - Rb - MgH2 - Si - BaSi2 - IBr - Bi2O3 - SrS - NaF - Ga2O3 - Al - TlI - CsO2 - KCl - In - I2 - BiF3 - SrF2 - LiCl - InN - CsBr - ICl - SrH2 - Pb3O4 - Na - Na2O2 - In2O3 - RbI - S - PbF2 - Bi2Te3 - Sn - CaH2 - KF - InSb - Ca - BiI3 - CsCl - K2O2 - MgF2 - Ge - PbS - SrSi2 - TeO2 - TlSe - Sr - BaI2 - AlP - Li2O - RbO2 - CsF - P4S3 - BiF5 - Mg - GeO2 - NaCl - CaSi2 - BaCl2 - Te - PbSe - TeF4 - PbI2 - TlF - KI - P - MgS - SnTe - NaO2 - GaAs - RbCl - Tl2O - SiS2 - KO2 - InAs - BaBr2 - P2S3 - Sb - KBr - TeI4 - Li3N - TeO3 - RbBr - SiI4 - LiBr - GaSb - TlCl - SeO3 - GaP - RbF - SnI4 - Cs2O - As2O3 - SrCl2 - Mg2Si - TlBr - AlAs - LiI - P4S7 - Bi2S3 - Mg2Sn - CaCl2 - AlI3 - As2O5 - SnSe - Ca3N2 - Li2S - NaBr - InI3 - BeCl2 - Sb2O3 - NaI - Mg2Ge - InI - BiBr3 - GeS - BeI2 - SeBr4 - Tl2S - InP - GaTe - P2S5 - SbF3 - K2S - BiCl3 - SrBr2 - InF3 - GeTe - SbI3 - AlSb - In2Te3 - GeF2 - Mg3Sb2 - SrI2 - PbCl2 - GaS - PI3 - Na2S - SnS - Al2S3 - GaI3 - Rb2S - GaSe - MgCl2 - TeCl4 - Rb2Se - PbBr2 - GeI4 - K2Se - CaI2 - BeBr2 - P2I4 - Sb2Se3 - CaBr2 - As2Te3 - In2Se3 - AlCl3 - InS - GeBr4 - As2S3 - Ga2Se3 - SnBr4 - InCl - As2Se3 - AsBr3 - AsI3 - GaBr3 - Al2Te3 - In2S3 - SbBr3 - MgI2 - InBr3 - GeS2 - MgBr2 - Ga2S3 - GaCl3 - SbCl3 - SnBr2 - GaCl2 - SnCl2 -

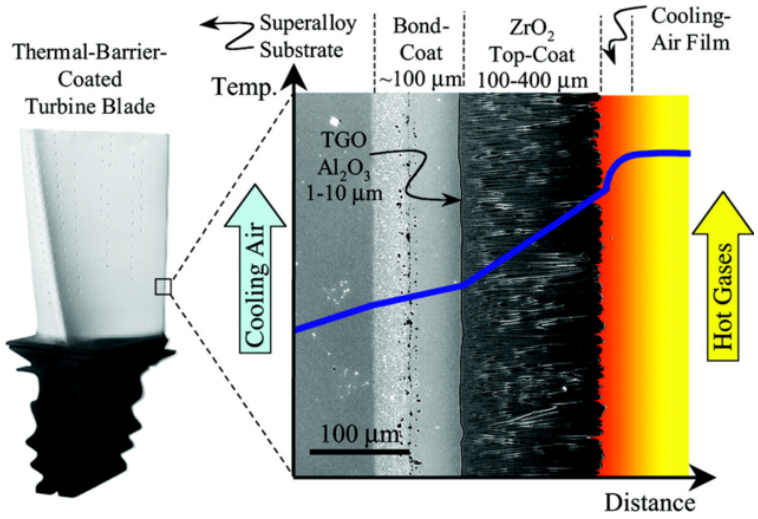
Design nanostructures for phonon transport via material informatics (Phys Rev X, in press)

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

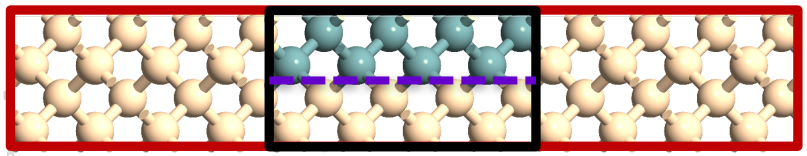
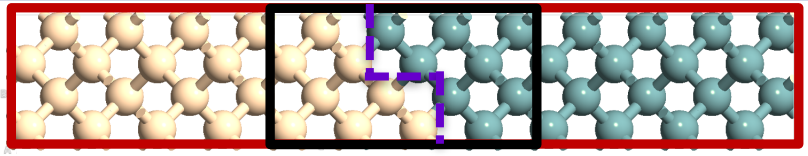
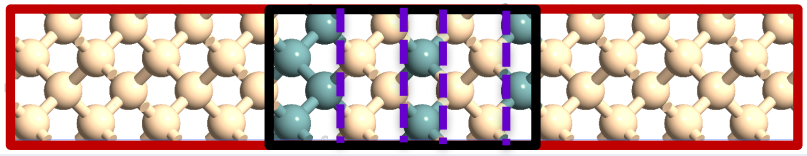
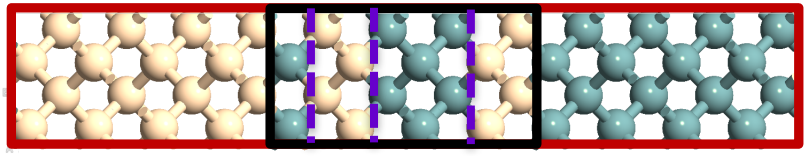
Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

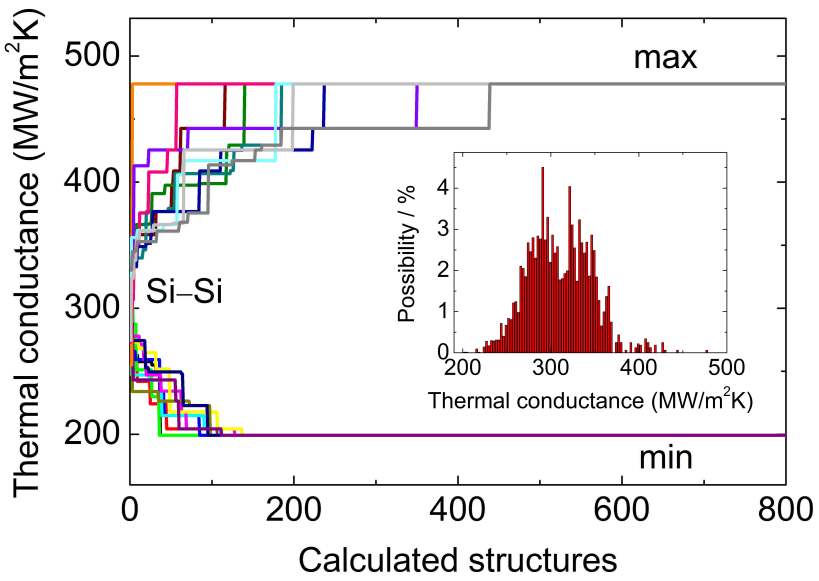
Calculator: Atomistic Green's Function (AGF): Phonon transmission

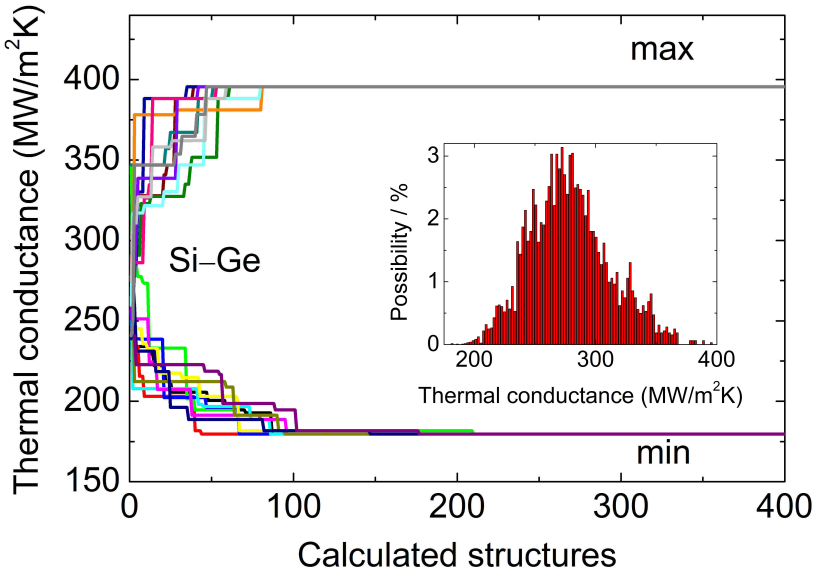
Evaluator: Interfacial Thermal Conductance (ITC)

Optimization method: Thompson Sampling (Bayesian Optimization)

Alloy Structure Optimization

ITC	Si-Si	Si-Ge
Max		
Min		

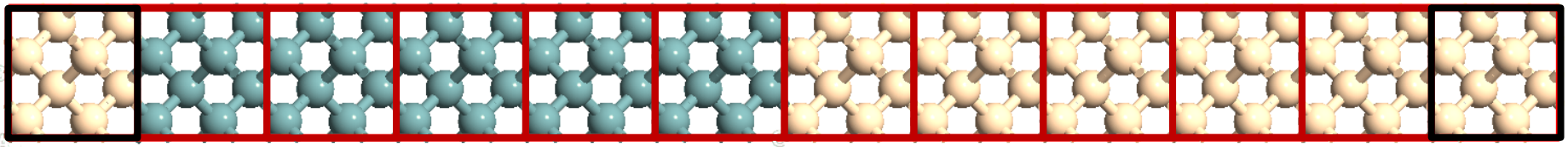




Optimal structures were obtained by calculating only 3.4% of all candidates.

Superlattices Structure Optimization

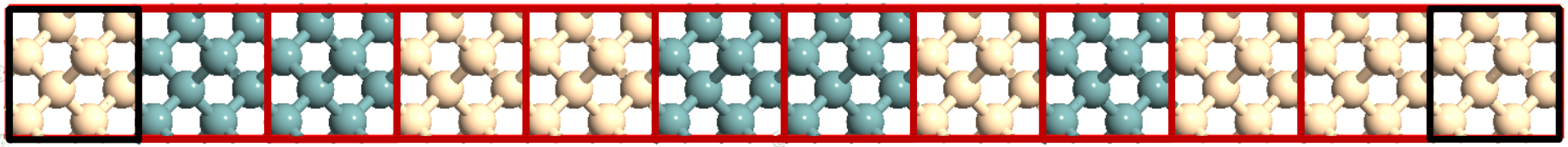
Topic: Arrange 10-layer superlattices structure (5 layers Si + 5 layers of Ge) between Si and Si to obtain minimal thermal conductance (1 layer thickness = 5.43 A)



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: (1101010001)



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)

Conclusion

- Artificial intelligence techniques combined with first principles calculation have enormous power