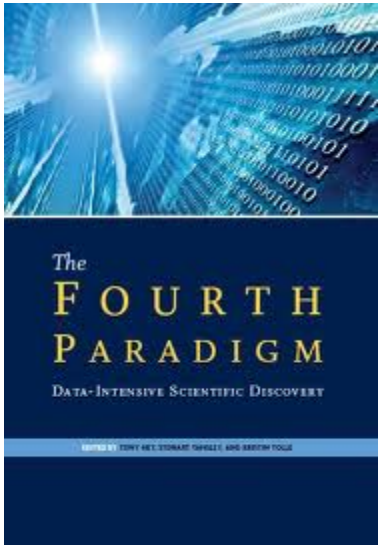
Automatic design of functional molecules and materials

Koji Tsuda
University of Tokyo / NIMS / RIKEN AIP

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
- スキル
  - 機械学習、データマイニング、統計学、離散アルゴリズム

第4のパラダイム：データ中心科学

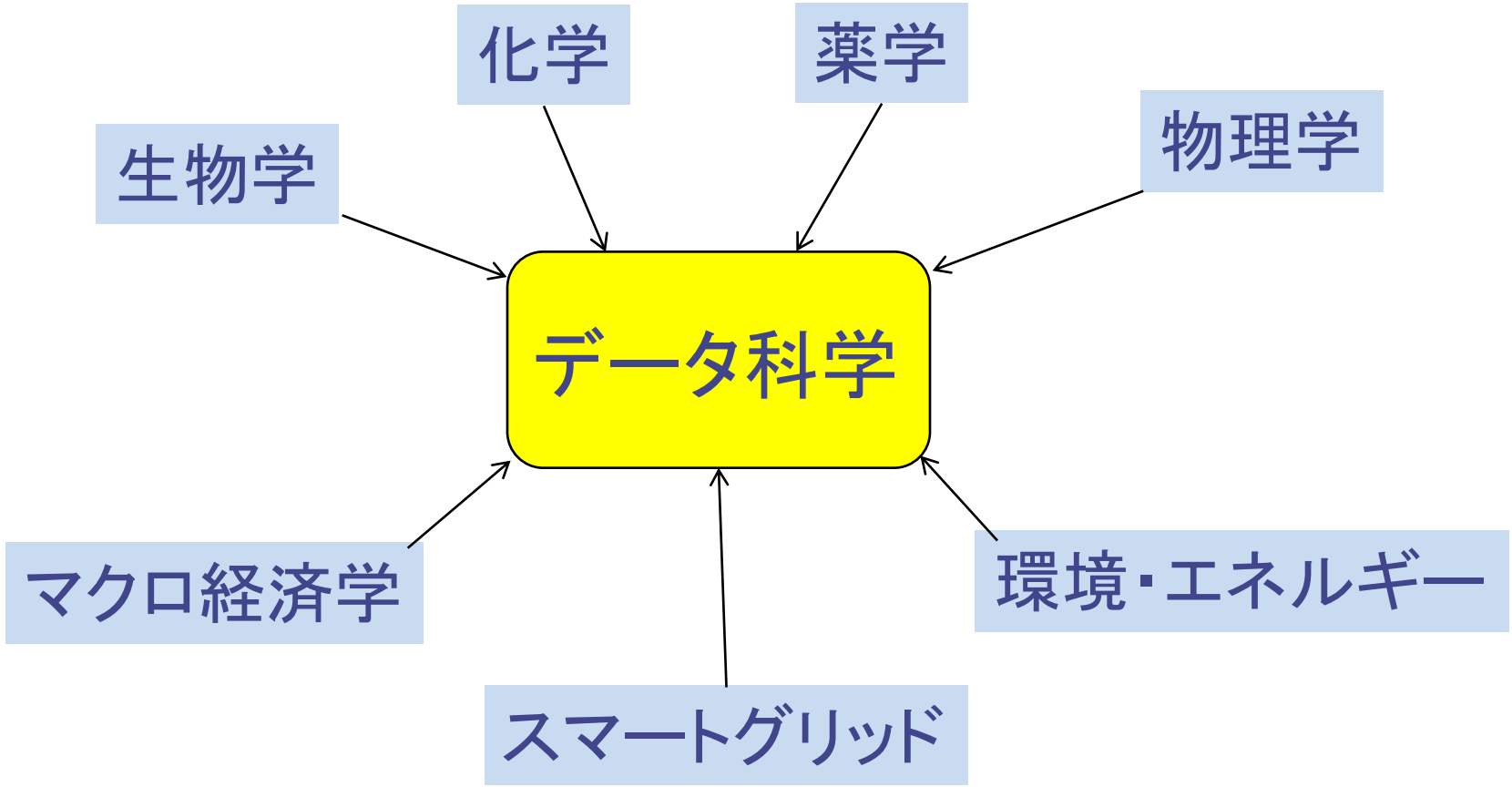


- 経験科学・理論科学・計算科学（シミュレーション）に代わる21世紀の科学

- 従来：　直感・経験に基づく仮説生成　⇒　データに基づく検証

- 第４：　データからの**知識発見**による仮説生成⇒データに基づく検証

様々な分野からの要望が、データ科学に集中： Computational Thinking



Materials Genome Initiative

Strategic Report by JST (2013)

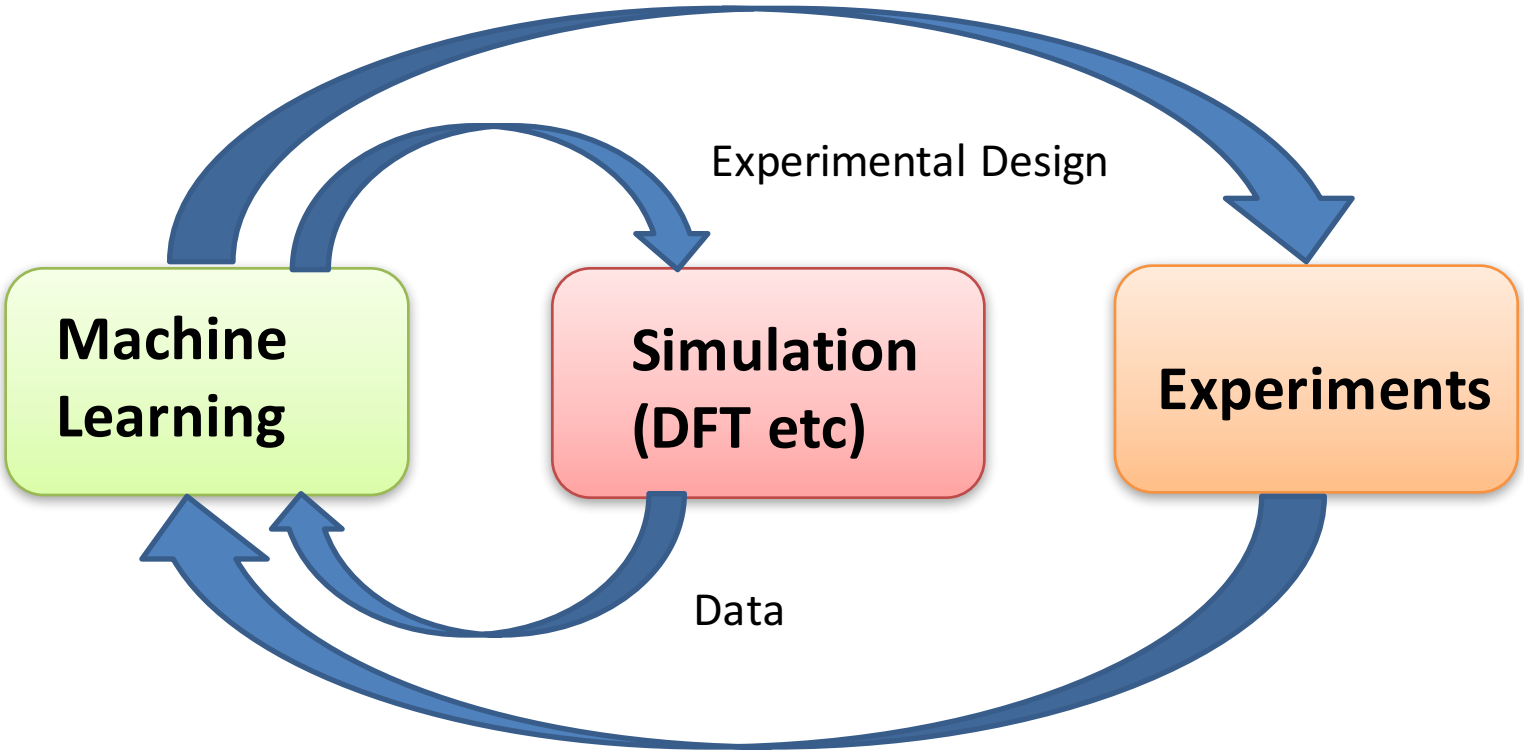
CRDS-FY2013-SP-01

戦略プロポーザル

データ科学との連携・融合による 新世代物質・材料設計研究の促進 （マテリアルズ・インフォマティクス）

～物質・材料研究を飛躍的に発展させるための新たなパラダイム～

Automatic Materials Design



Materials Design Examples

- Design of Si-Ge nanostructures (Phys Rev X 2017)
- De novo design of organic compounds with desired absorption wavelength
- Machine-learning-guided mutagenesis platform for accelerated discovery of novel functional proteins

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

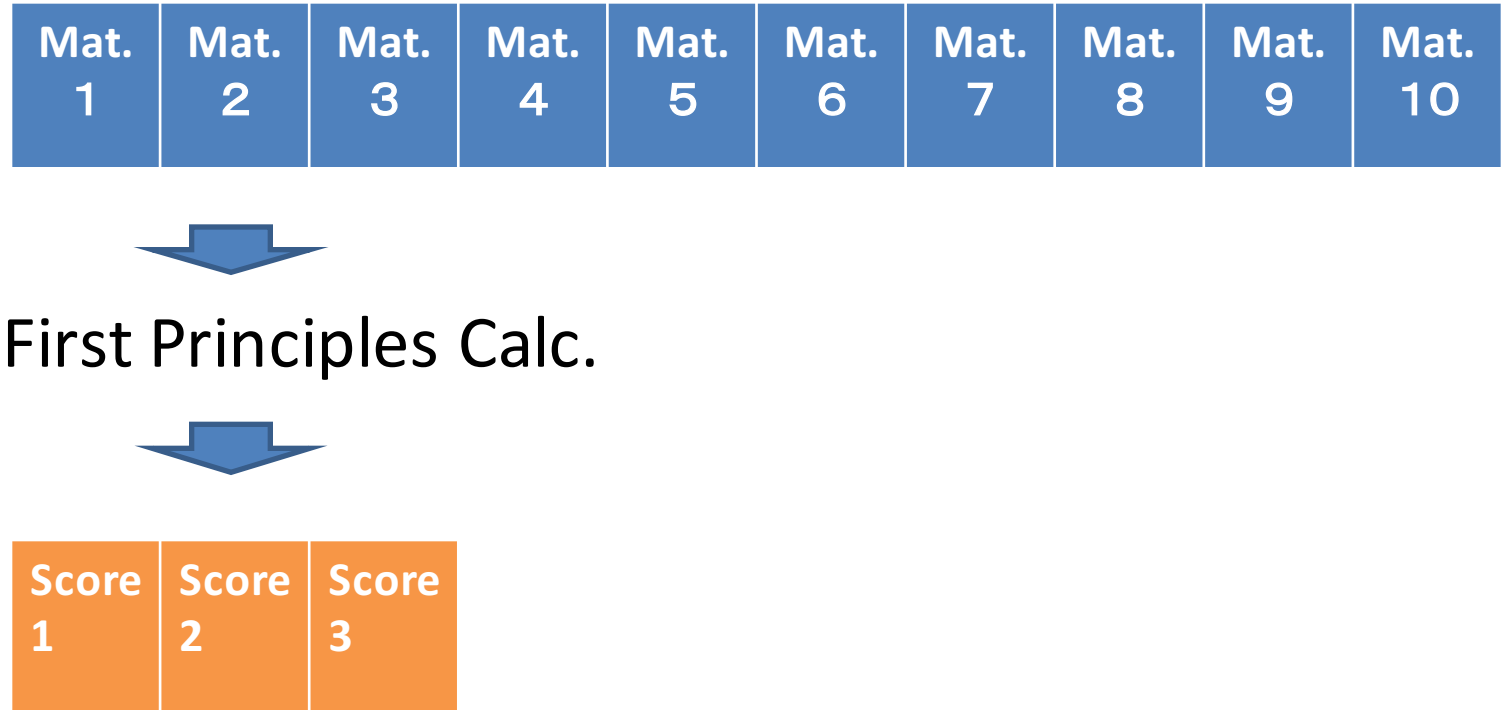
Bayesian Optimization

(Jones et al., 1998)

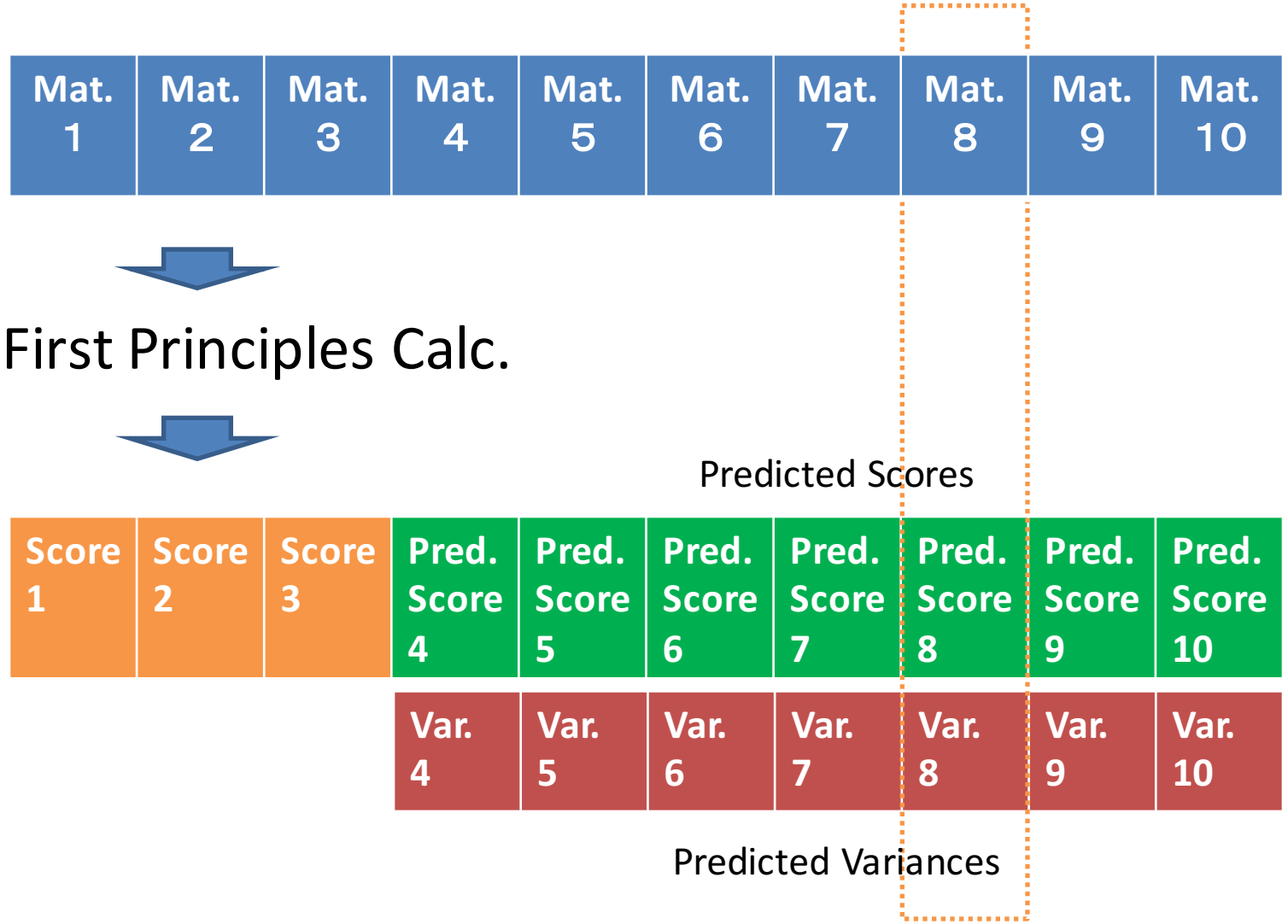
- Find best data points with minimum number of observations

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

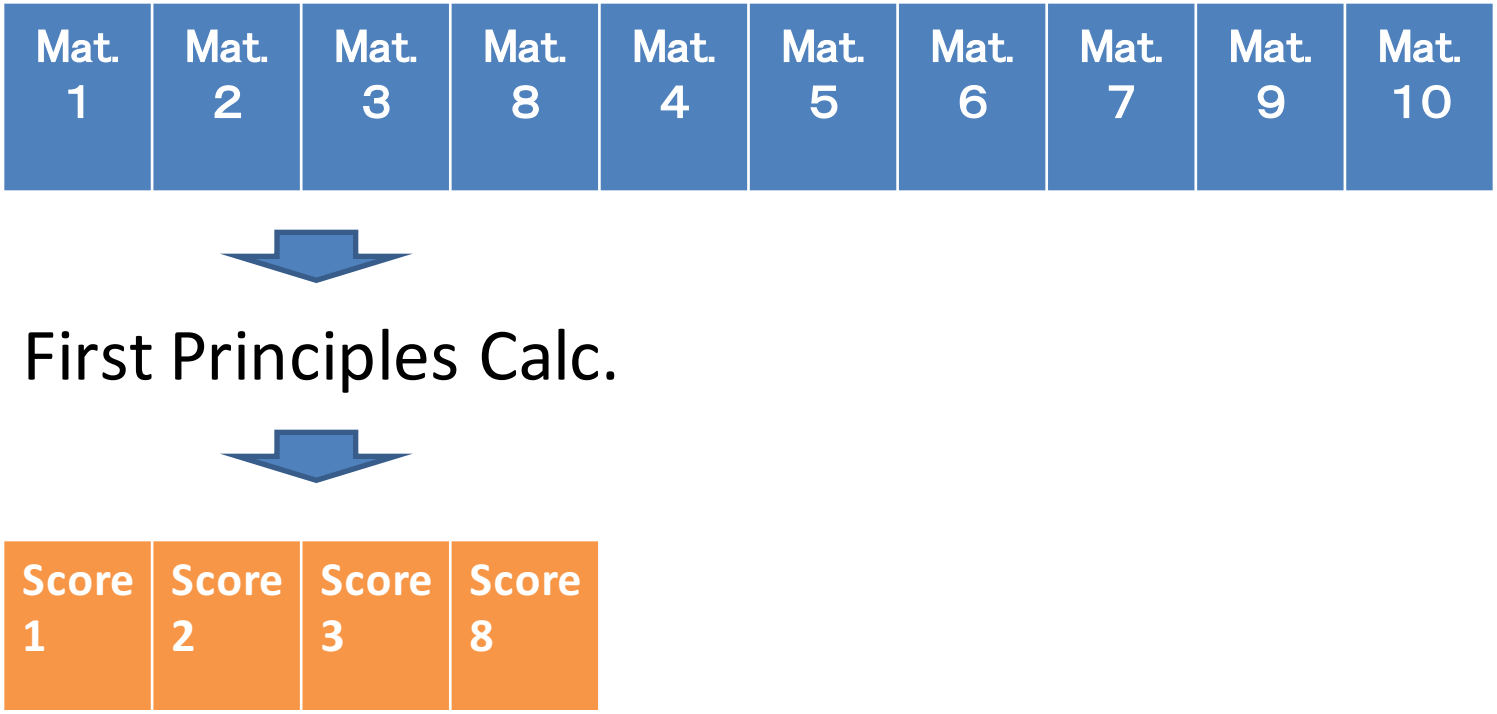
Bayesian Optimization (1)



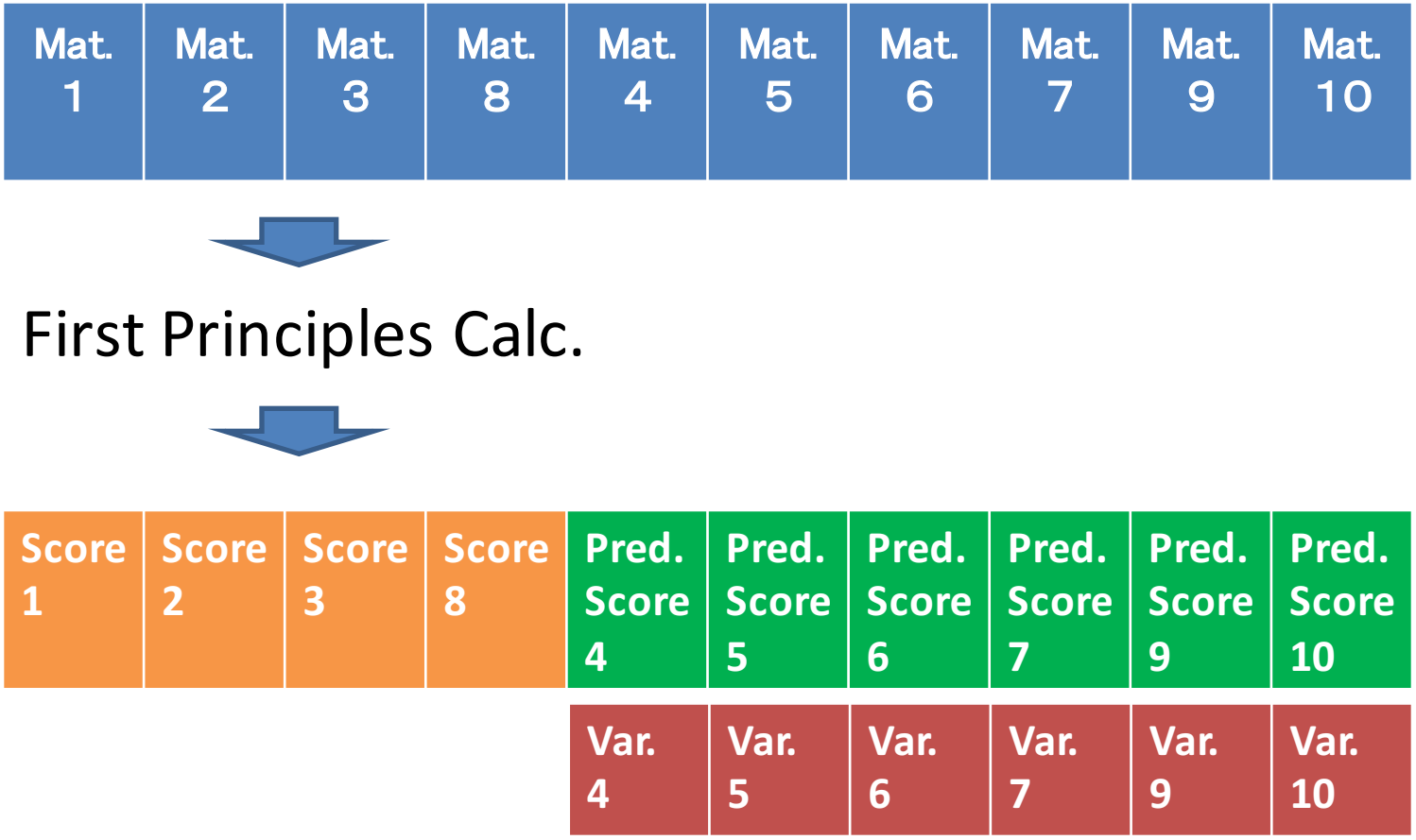
Bayesian Optimization (2)



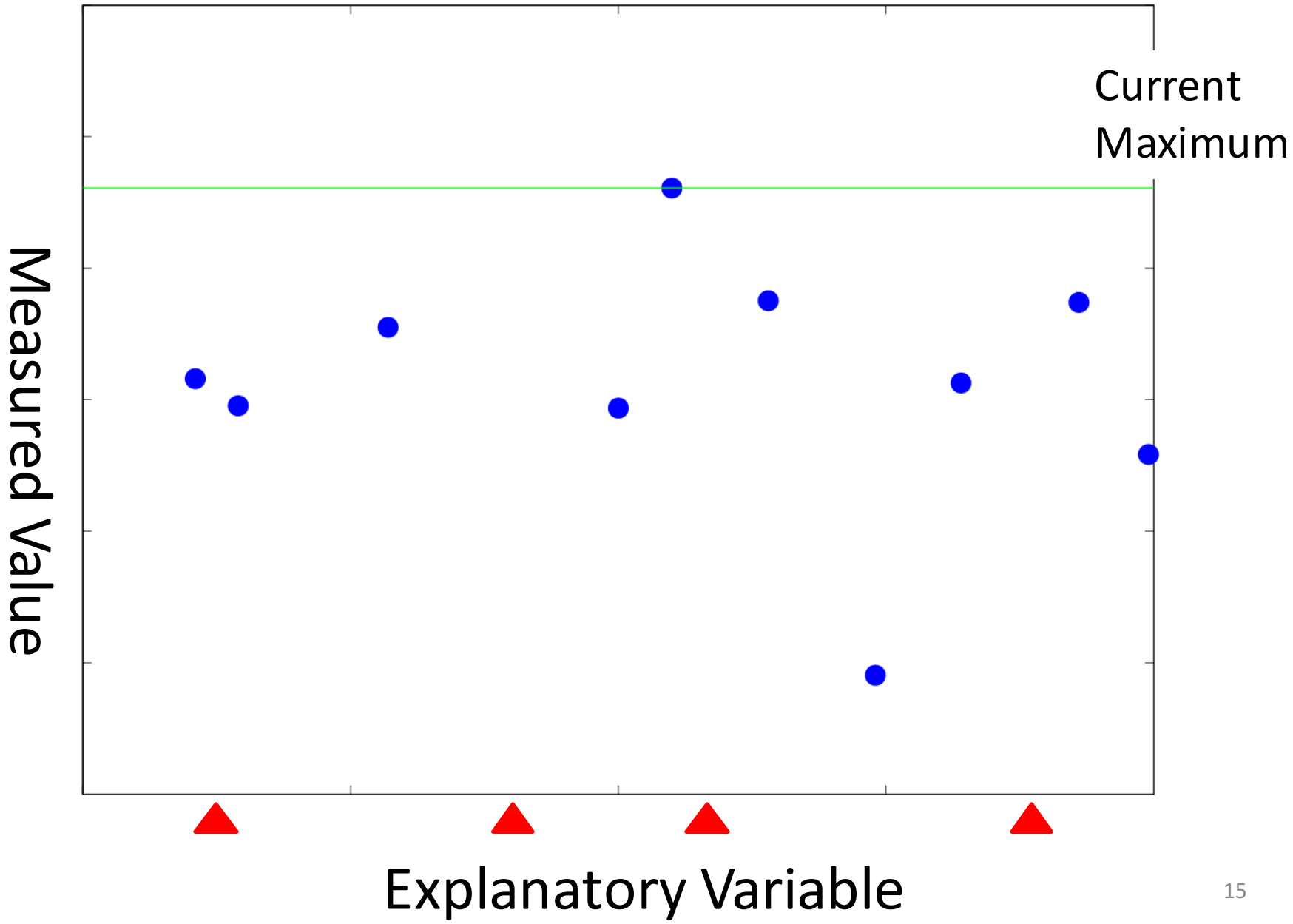
Bayesian Optimization (3)



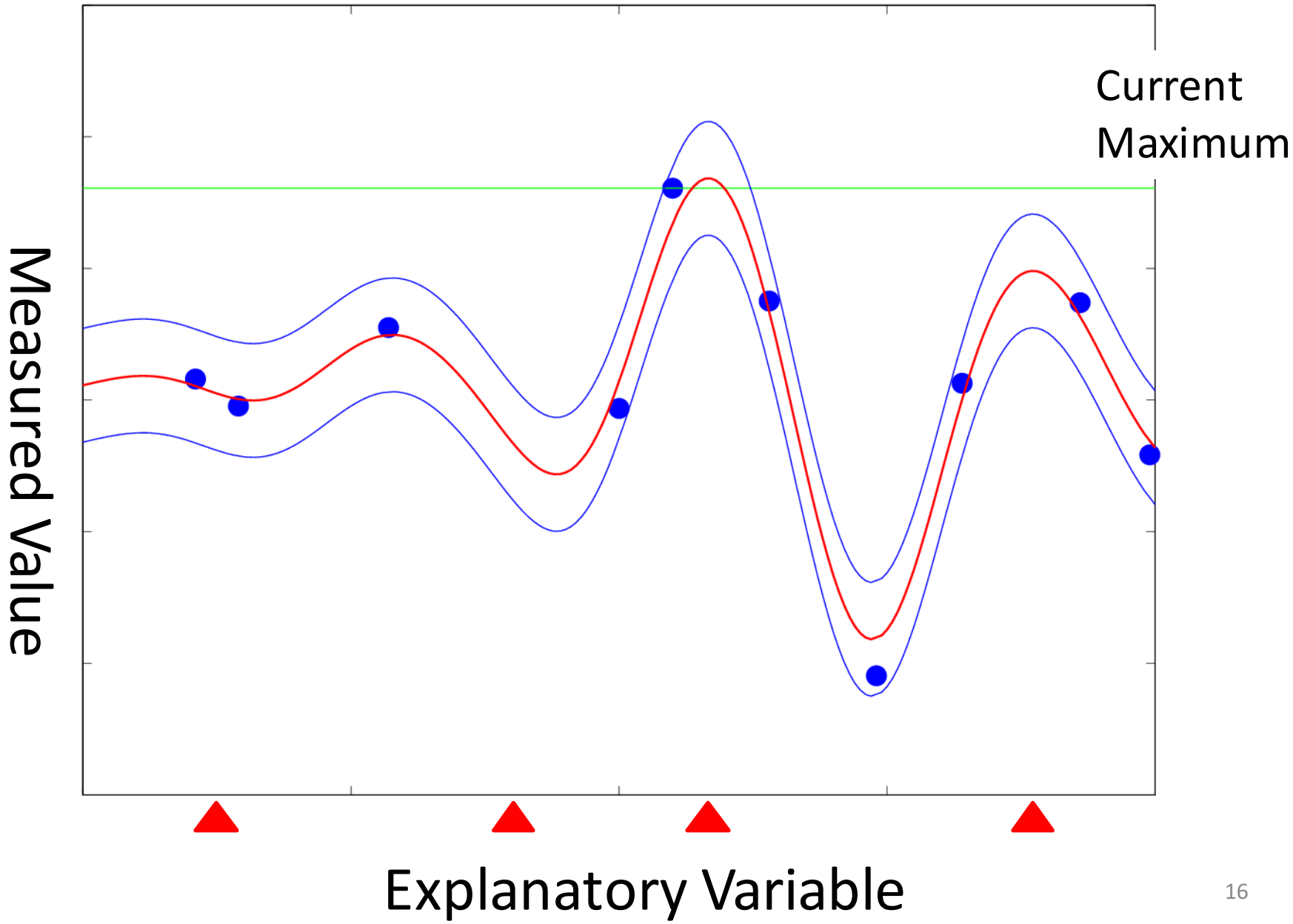
Bayesian Optimization (4)



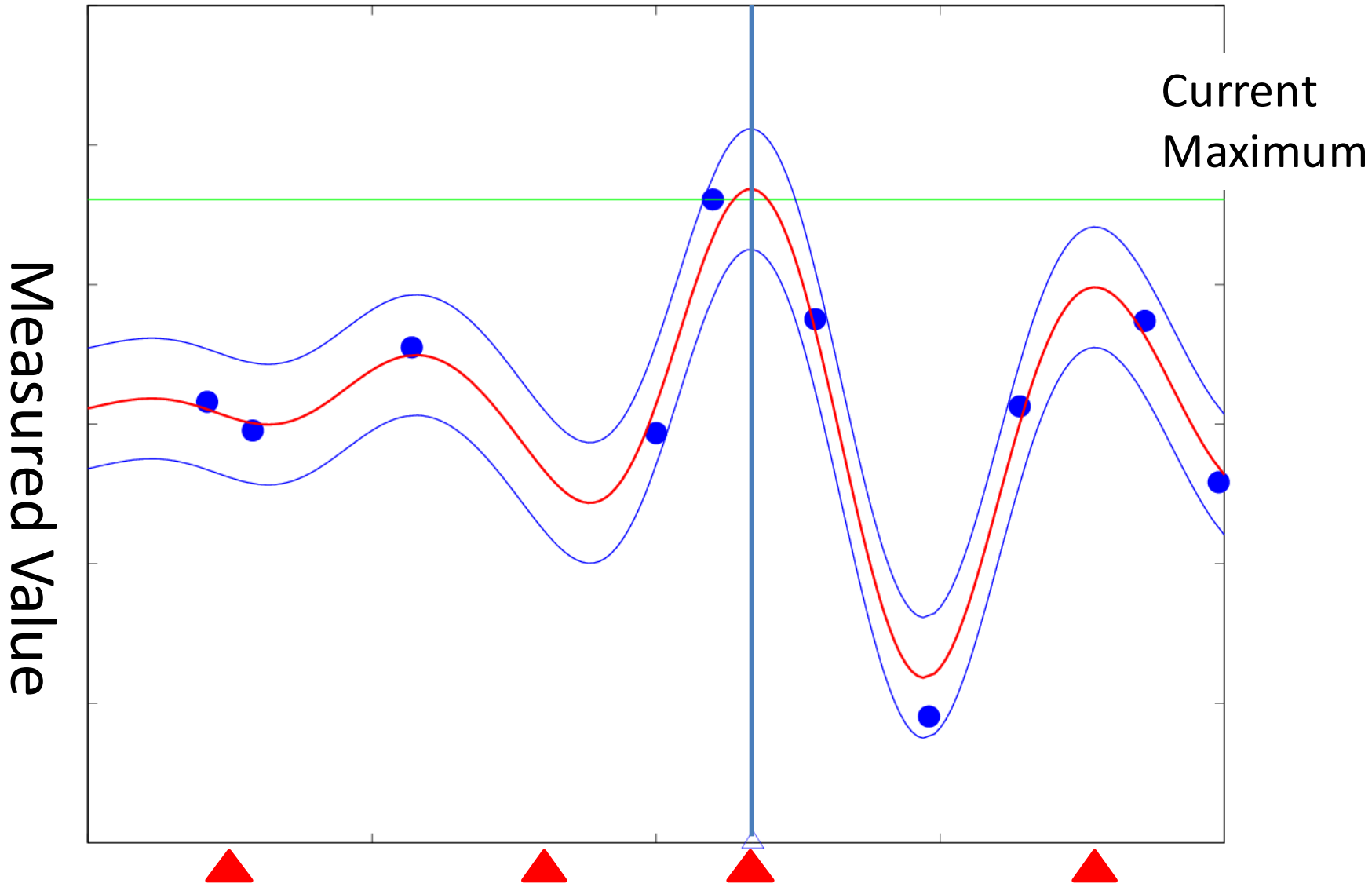
Where to observe next?



Gaussian Process

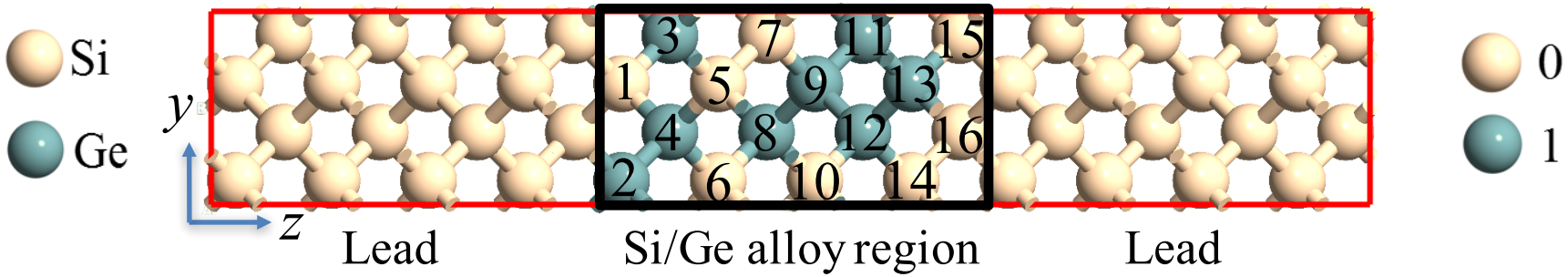


Maximum probability of improvement



Alloy Structure Optimization (Phys Rev X, 2017)

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



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

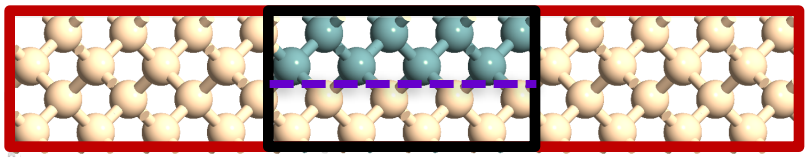
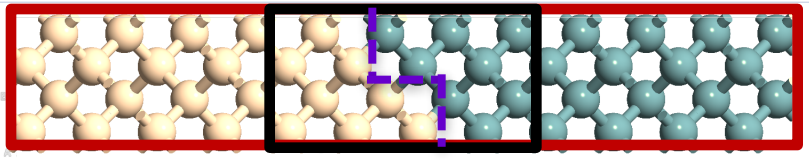
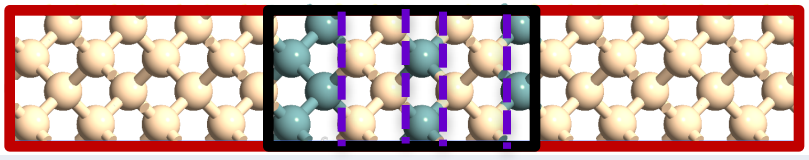
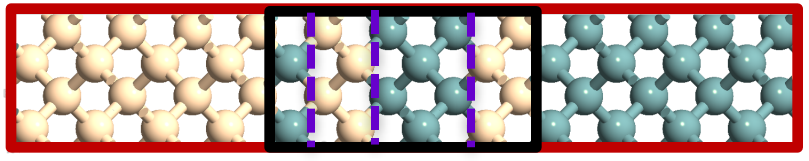
Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

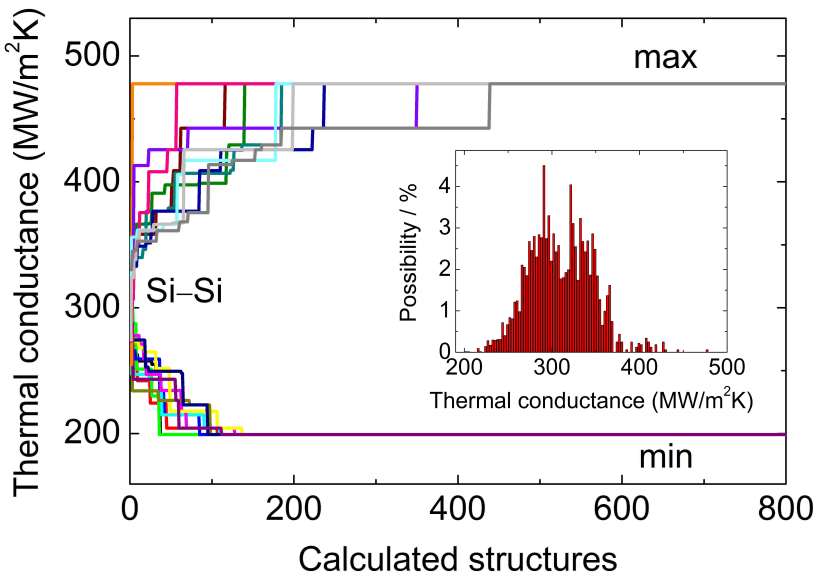
Calculator: Atomistic Green's Function (AGF): Phonon transmission

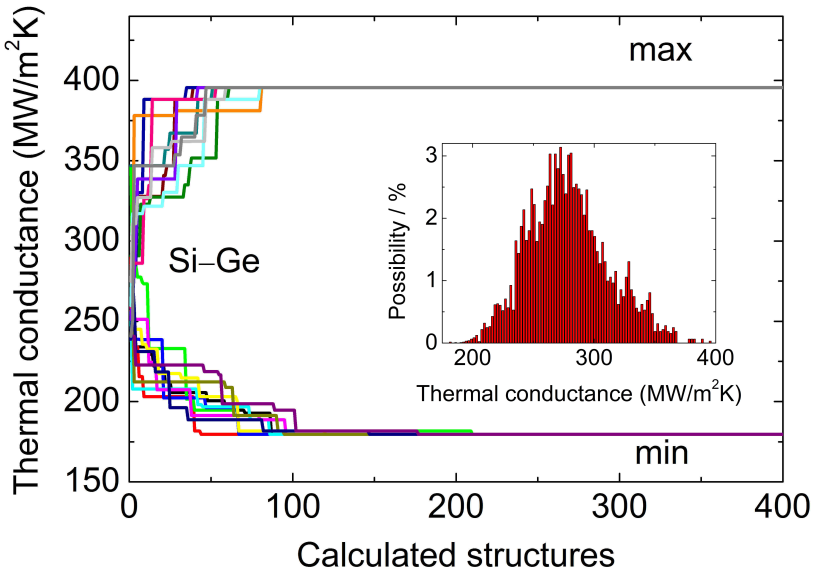
Evaluator: Interfacial Thermal Conductance (ITC)

Optimization method: Thompson Sampling (Bayesian Optimization)

Alloy Structure Optimization

ITC	Si-Si	Si-Ge
Max		
Min		

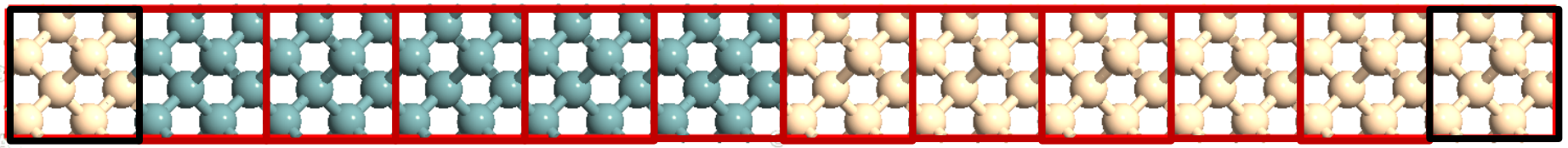




Optimal structures were obtained by calculating only 3.4% of all candidates.

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)

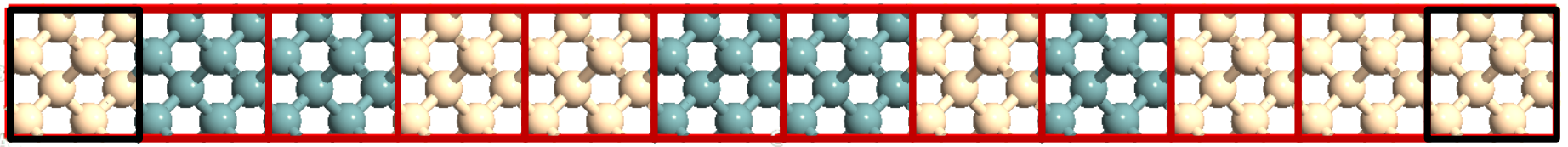


Lead Si Ge Si/Ge superlattice region Lead

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
…	…	…	…	…	…	…	…	…	…	…

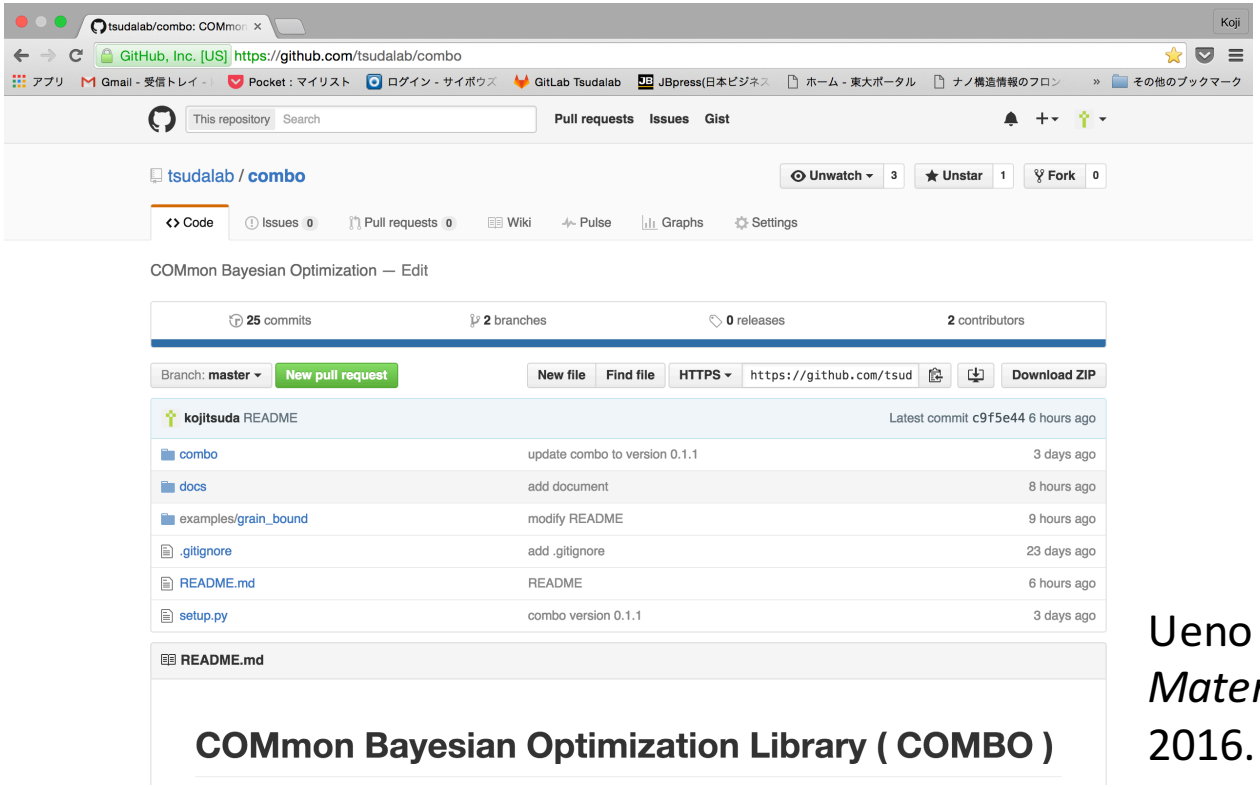
Best Structure: (1101010001)



COMBO: COMmon Bayesian Optimization Python Library

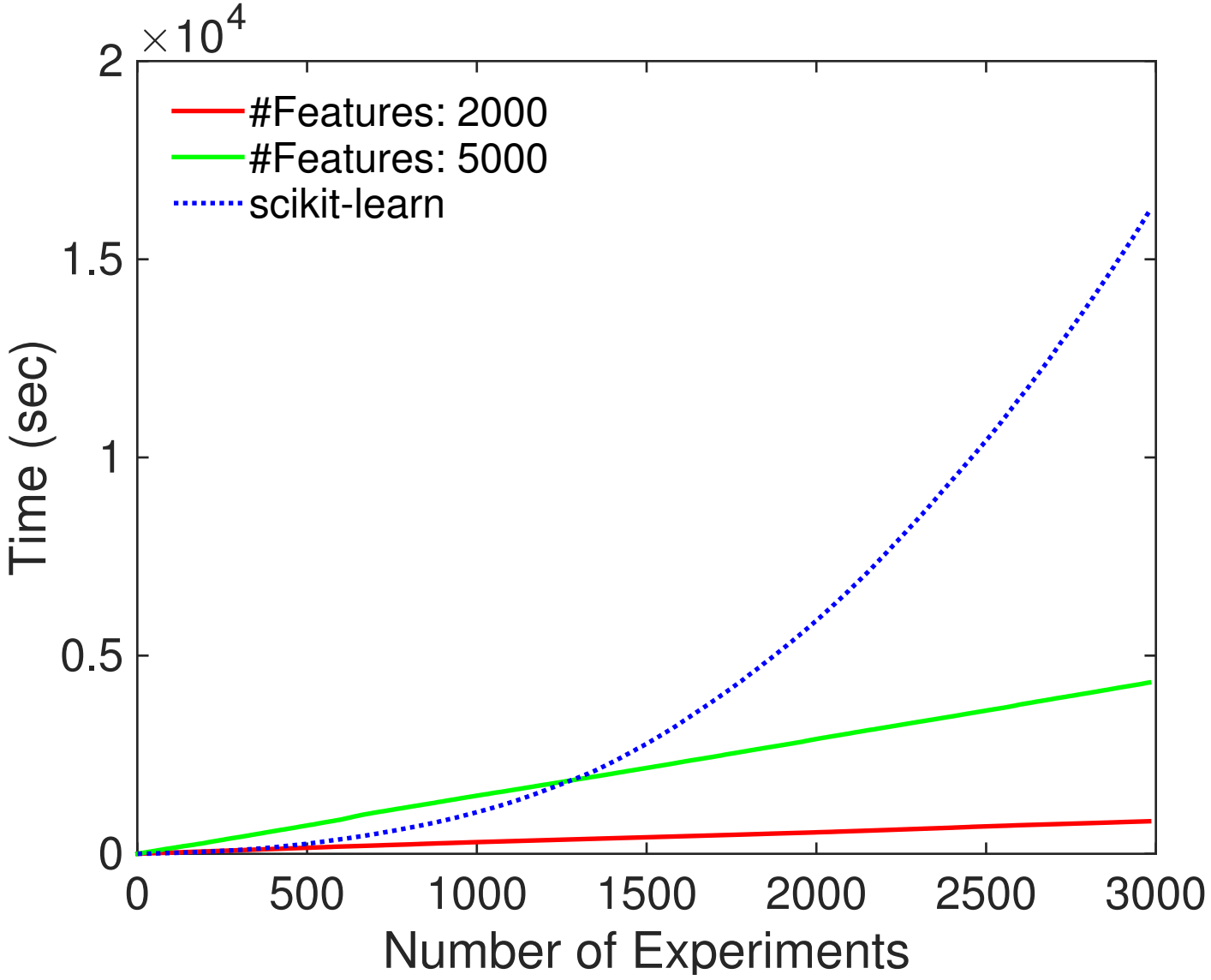
https://github.com/tsudalab/combo

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



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

Computational Time of COMBO



Monte Carlo Tree Search

ARTICLE

doi:10.1038/nature16961

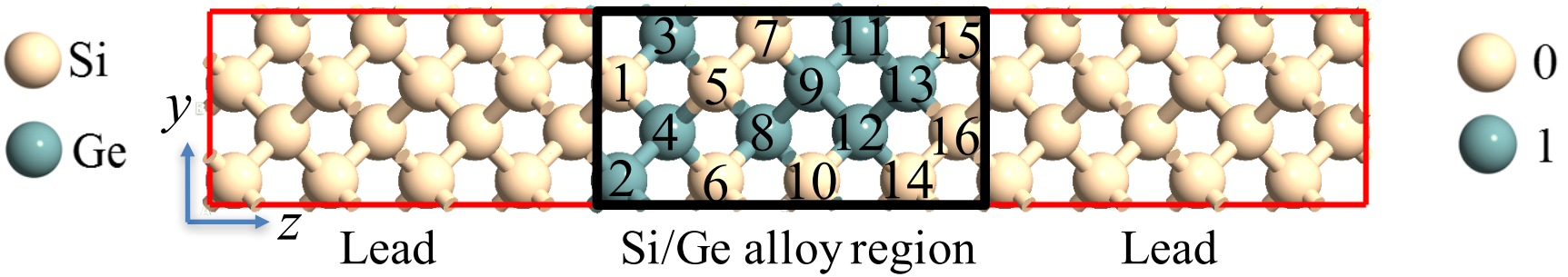
Mastering the game of Go with deep neural networks and tree search

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.

Atom assignment problem

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



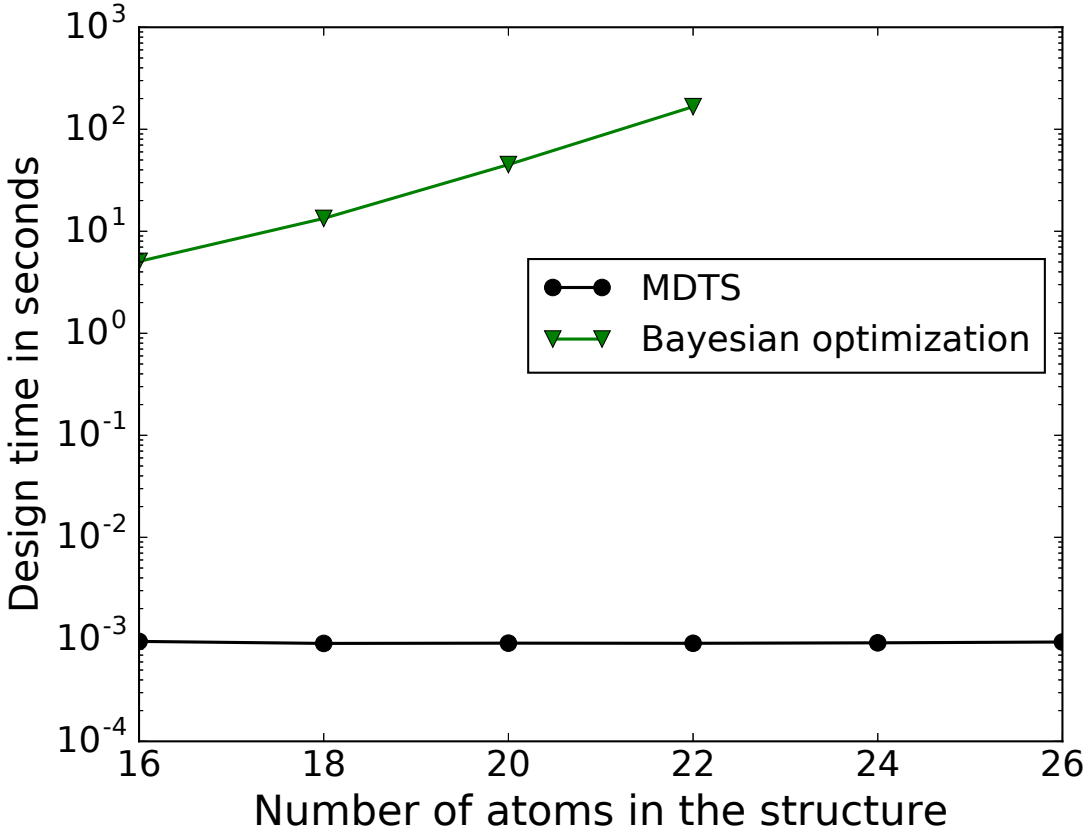
- Discrete black-box optimization:

$$\text{Maximize } f(\mathbf{x}),\ \mathbf{x} \in \{0, 1, k-1\}^N$$
$$\text{s.t. } \sum_{\ell=1}^{N} I(x_\ell = j) = n_j, \quad j = 0, \ldots, k-1$$

Search space is too large for Bayesian Optimization

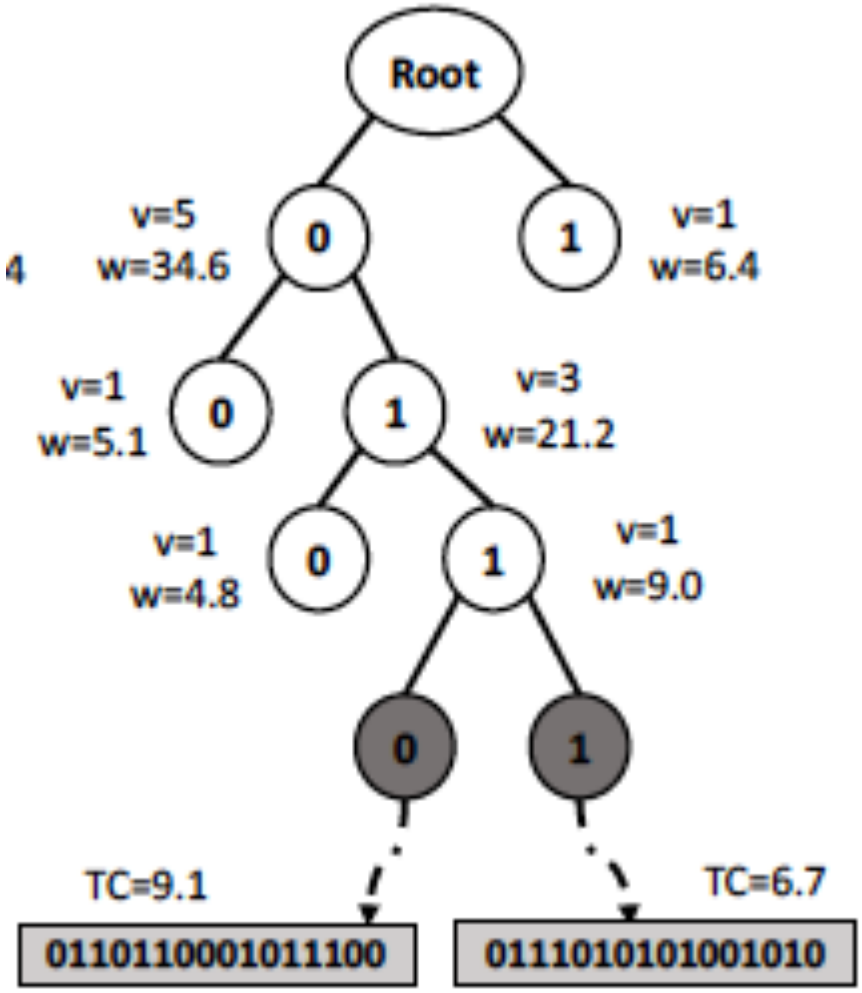
(Jones et al., 1998)

- Need to compute scores for all candidates

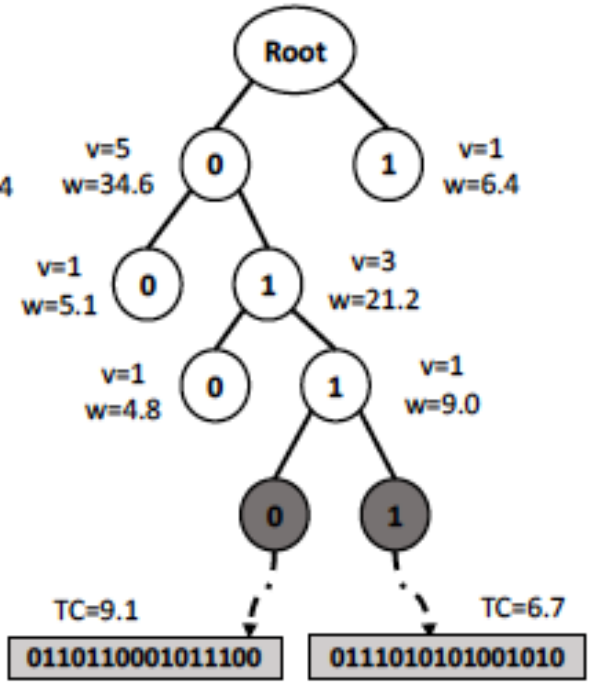


Monte Carlo Tree Search

- Candidates at leafs of search tree
- Merit at leafs: Thermal Conductivity
- Merit at intermediate node by Rollout
  - Random traverse to a leaf (engineered)
  - UCB score: average merit + Premium



Variables at each node i



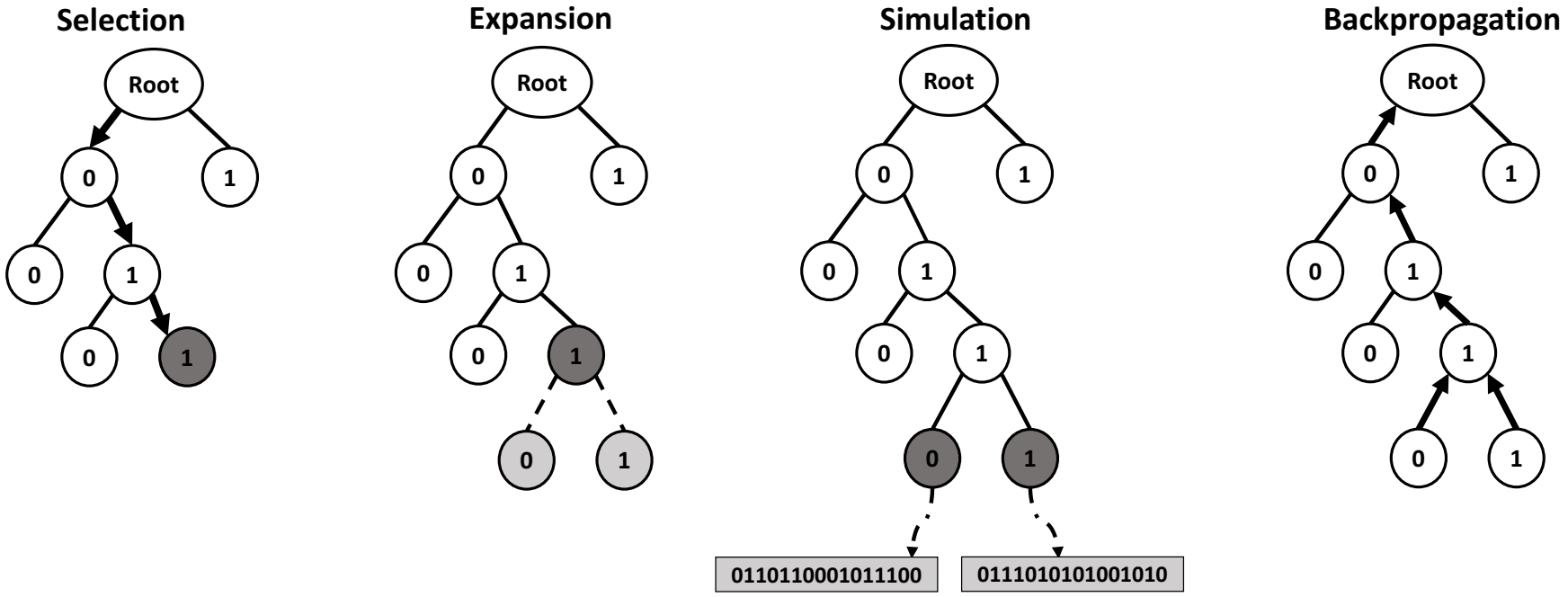
- f_i : Immediate merit by rollout
- w_i : Cumulative merit (downward sum of f_i)
- v_i : Visit count

- UCB score

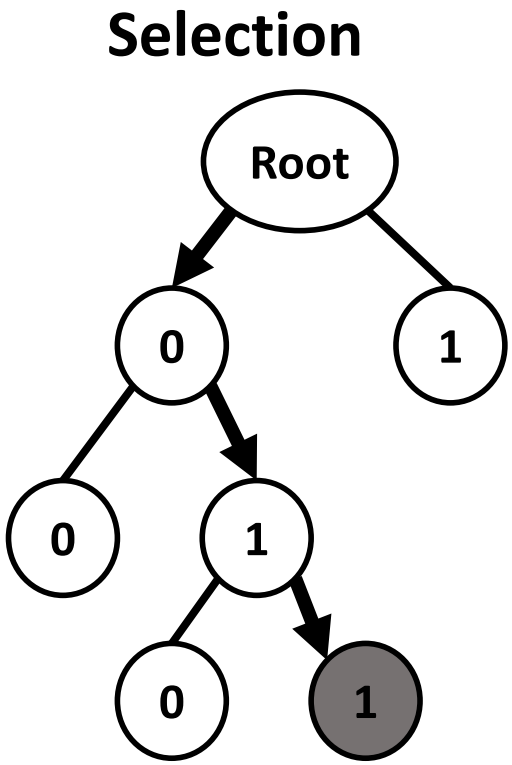
$$u_i = \frac{w_i}{v_i} + C\sqrt{\frac{2\ln v_{parent}}{v_i}}$$

MCTS steps

- Initially, only root node exists
- j-th level corresponds to x_j
- Tree grows by iterating the following 4 steps

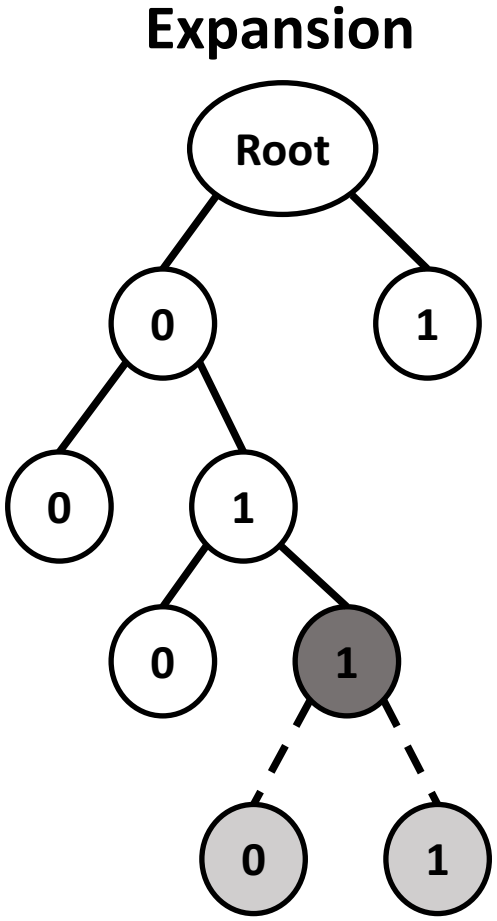


1. Selection



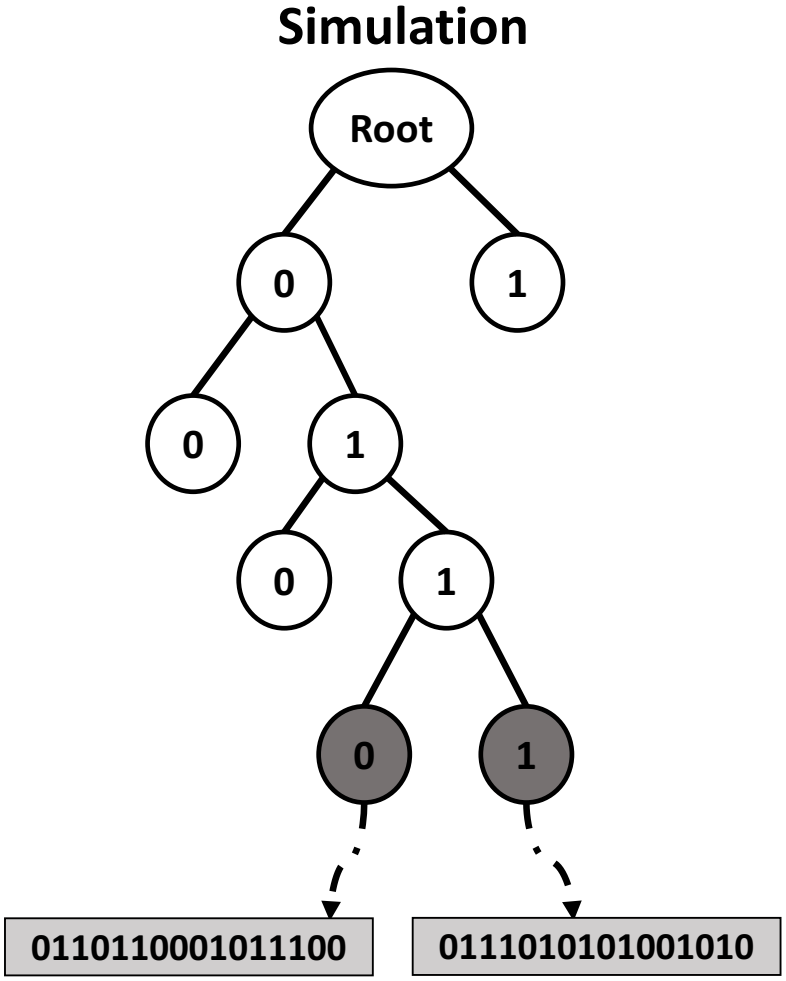
- Reach a leaf by following the child with largest UCB score
- In case of tie, choose randomly

2. Expansion



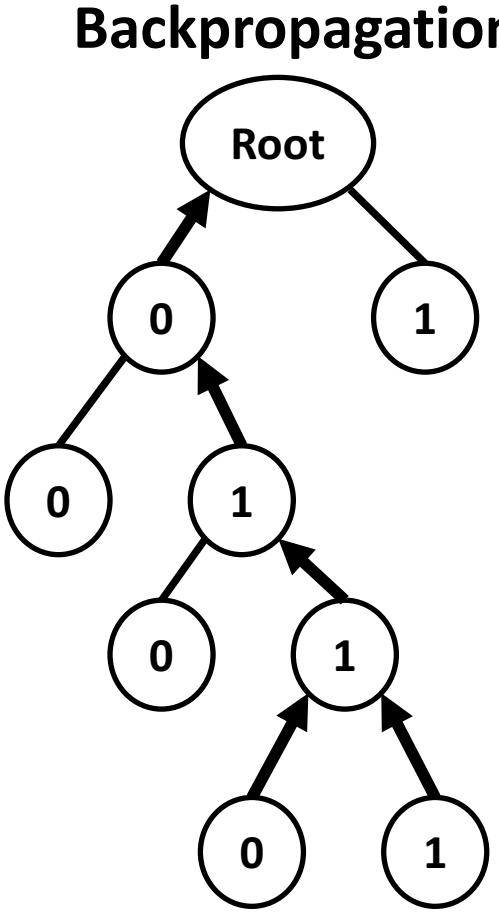
- Generate k children under the selected node
- If atom j is exhausted, no child is made
- Initialize the variables

3. Simulation



- Traverse randomly to a leaf
- Respect composition constraints
- Call first principle calculation and obtain thermal conductivity

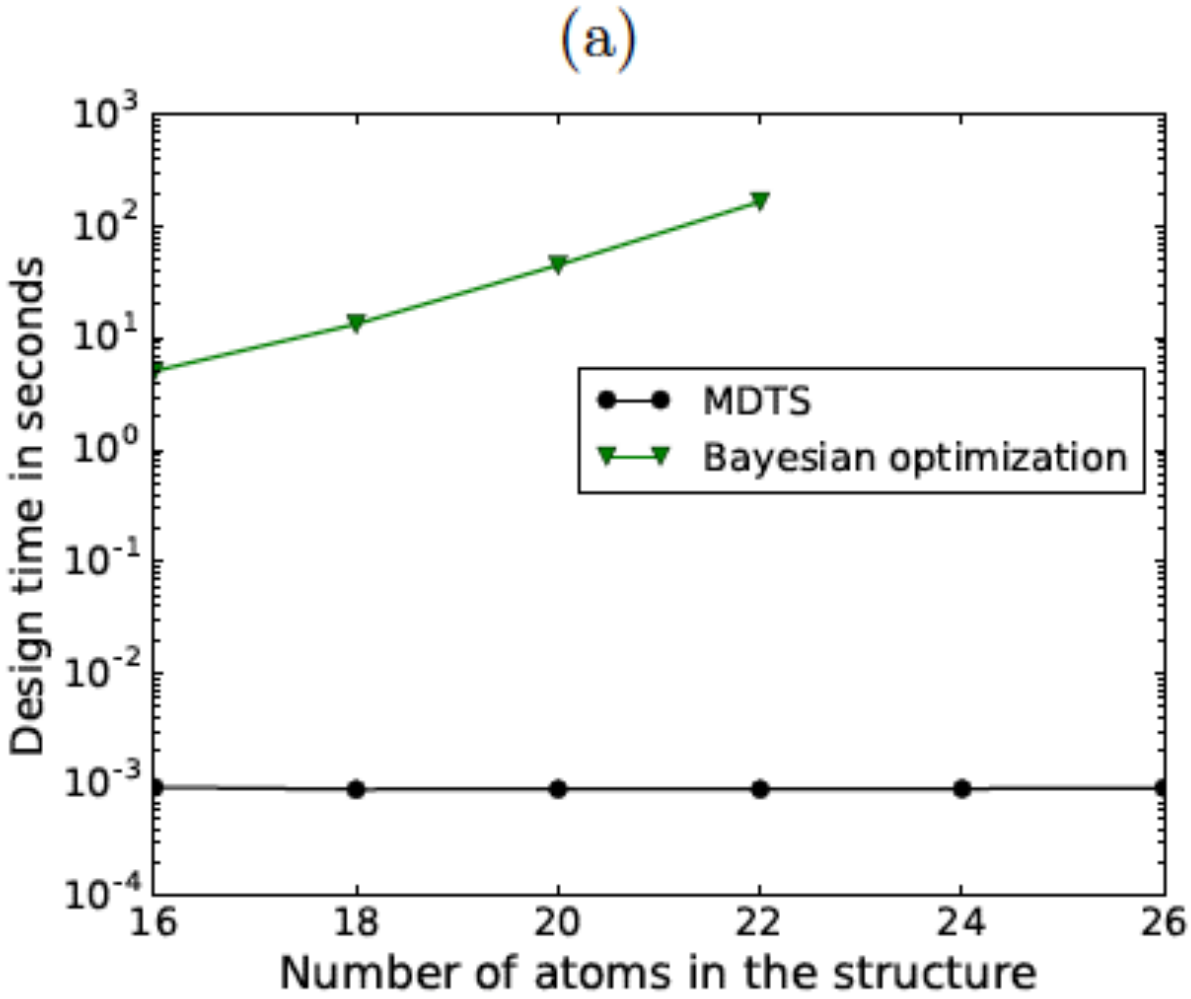
4. Backpropagation



- Visit count v_i of all ancestor nodes is incremented
- Cumulative merit w_i is also updated

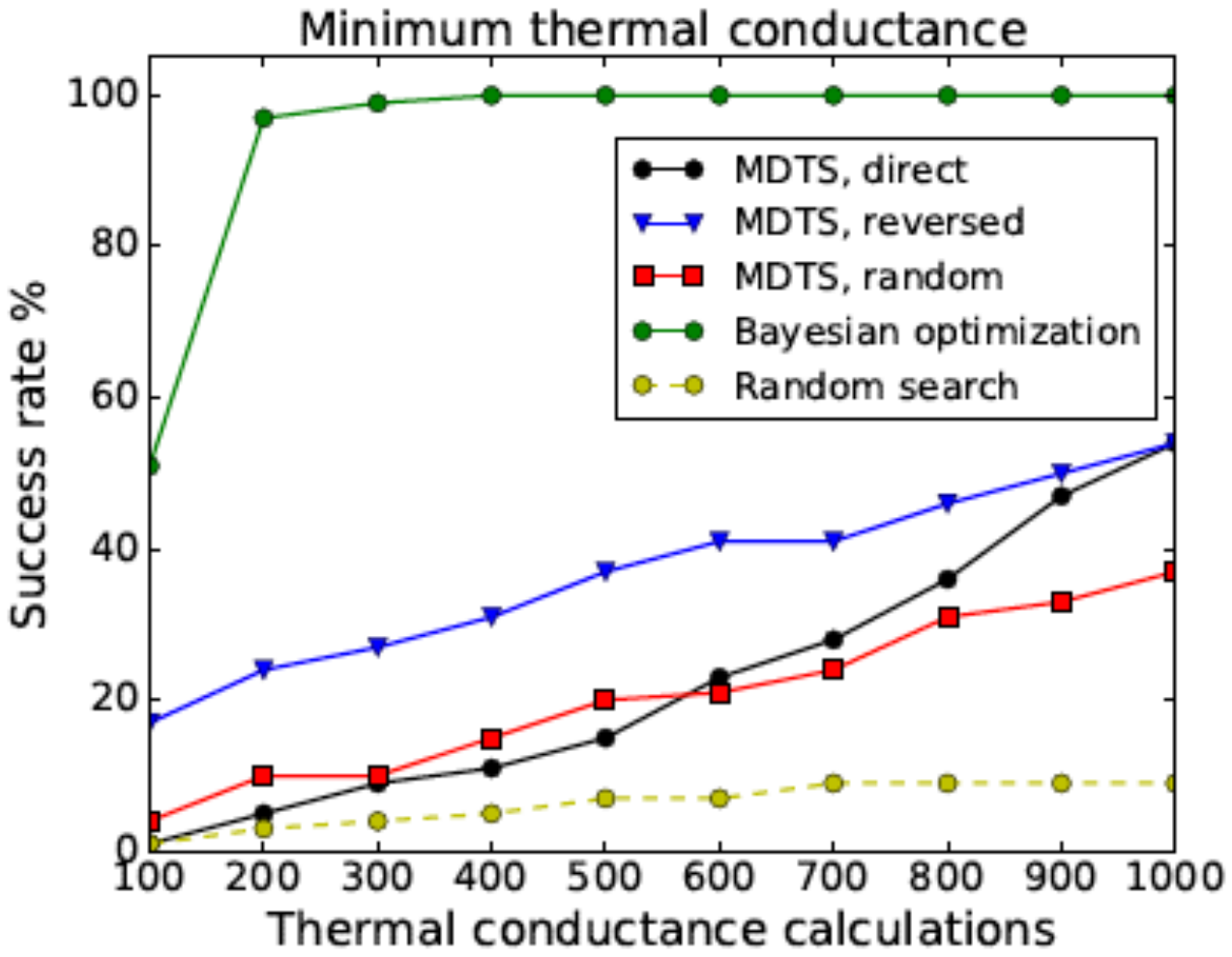
Experiments: Design time

- Total time = Design time + Simulation time



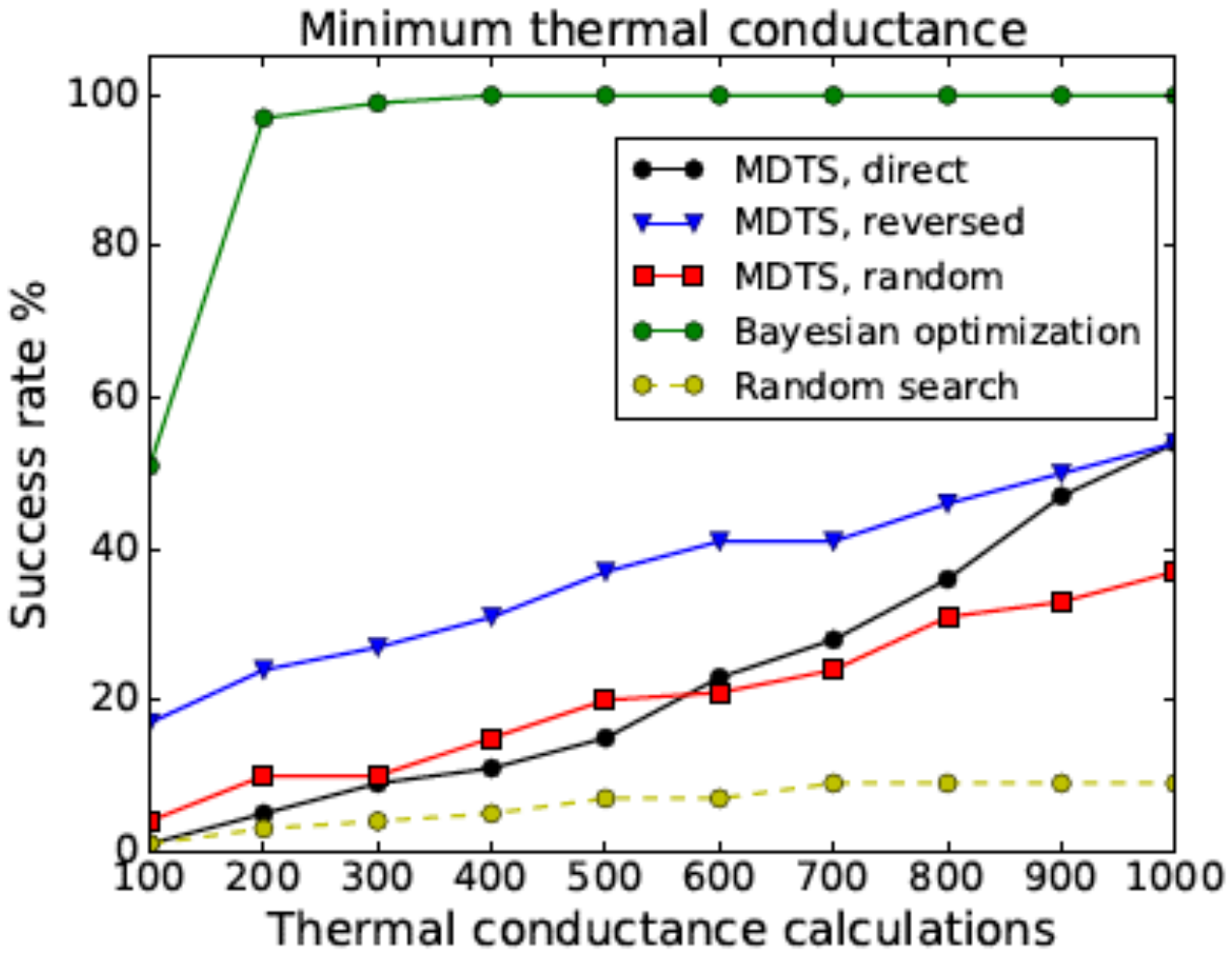
calculations vs success rate

- Success rate = Fraction of discovering the best in 100 runs

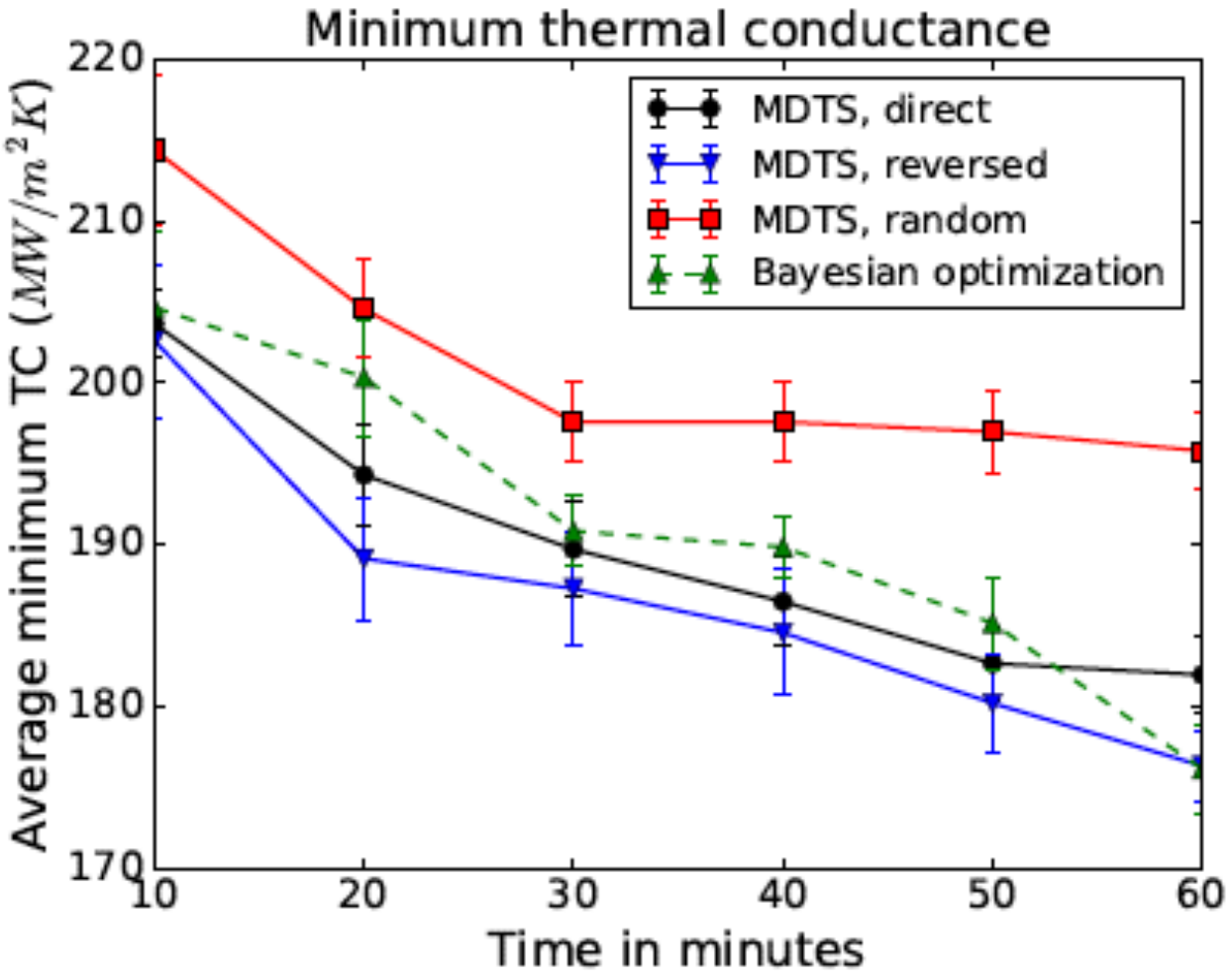


calculations vs success rate

- Success rate = Fraction of discovering the best in 100 runs



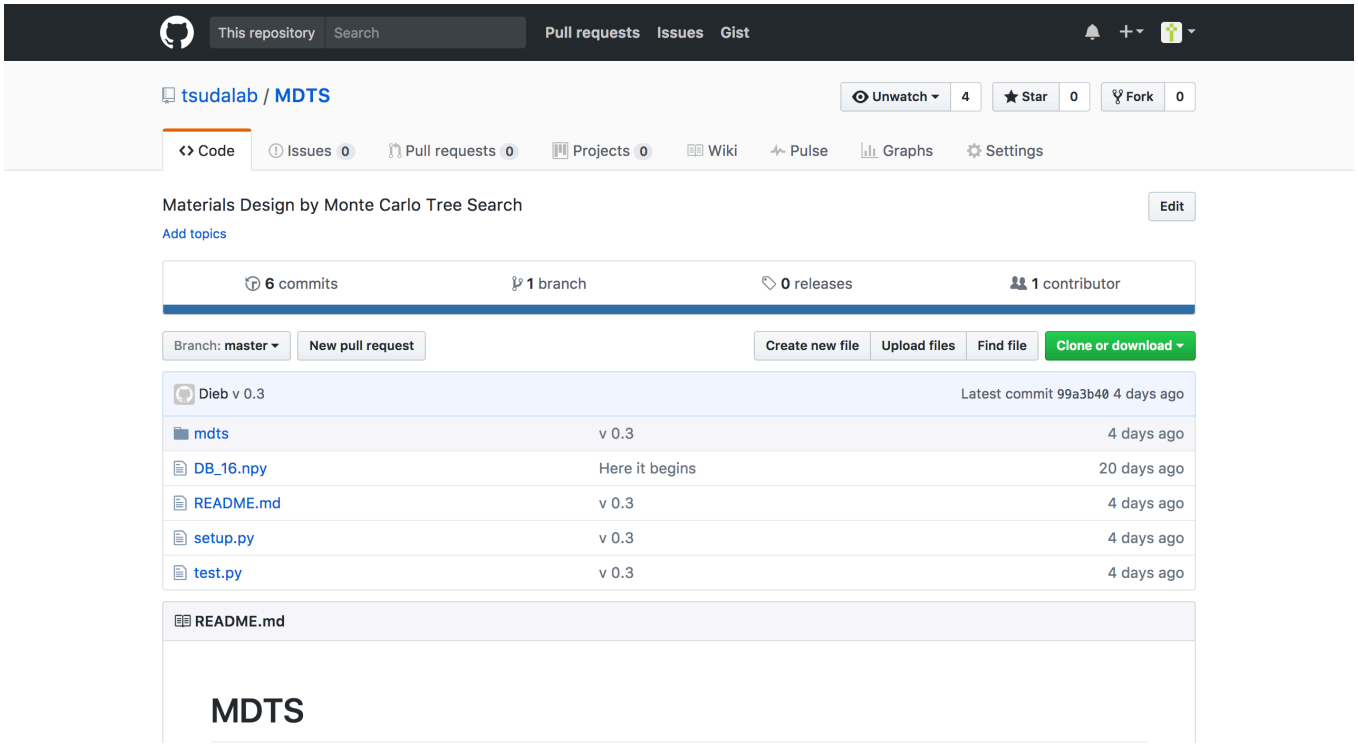
Total time vs achieved TC (N=22): fast design = more calculations



MDTS: Materials Design with Tree Search

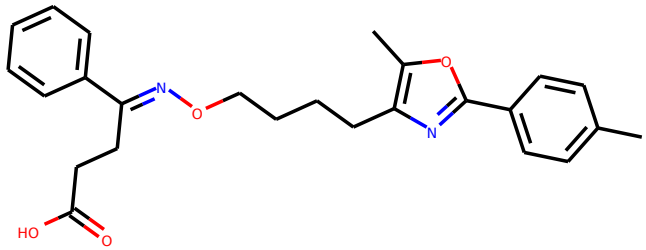
https://github.com/tsudalab/MDTS

- MCTS Implementation for atom assignment problems with composition constraints
- Parameter free



de novo Molecular Generation

Machine learning (prediction)

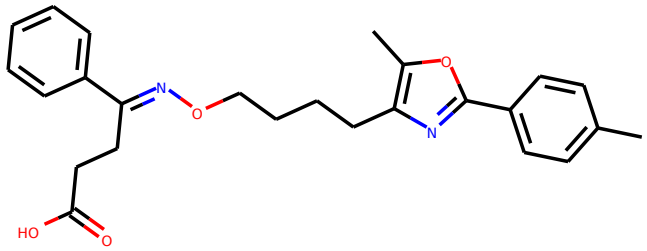


Properties:

Binding Score
Bio Activity
ADMET
…

Generation of molecules with desired properties

SMILES



Cc3ccc(c2nc(CCCCO/N=C(CCC(O)=O)c1ccccc1)c(C)o2)cc3

Atom: {C, c, o, O, N, F, [C@@H], n, -, S,Cl, [O-],[C@H], [NH+],[C@], s, Br, [nH], [NH3+], [NH2+], [C@@], [N+], [nH+], [S@], [N-], [n+],[S@@], [S-], I, [n-], P, [OH+],[NH-], [P@@H], [P@@], [PH2], [P@], [P+], [S+],[o+], [CH2-], [CH-], [SH+], [O+], [s+], [PH+], [PH], [S@@+] }
Bonds: {/,=, ¥# }
Ring: {1,2,3,4,5,6,7,8,9}
Branch: {(,)}

De novo Molecular Generation

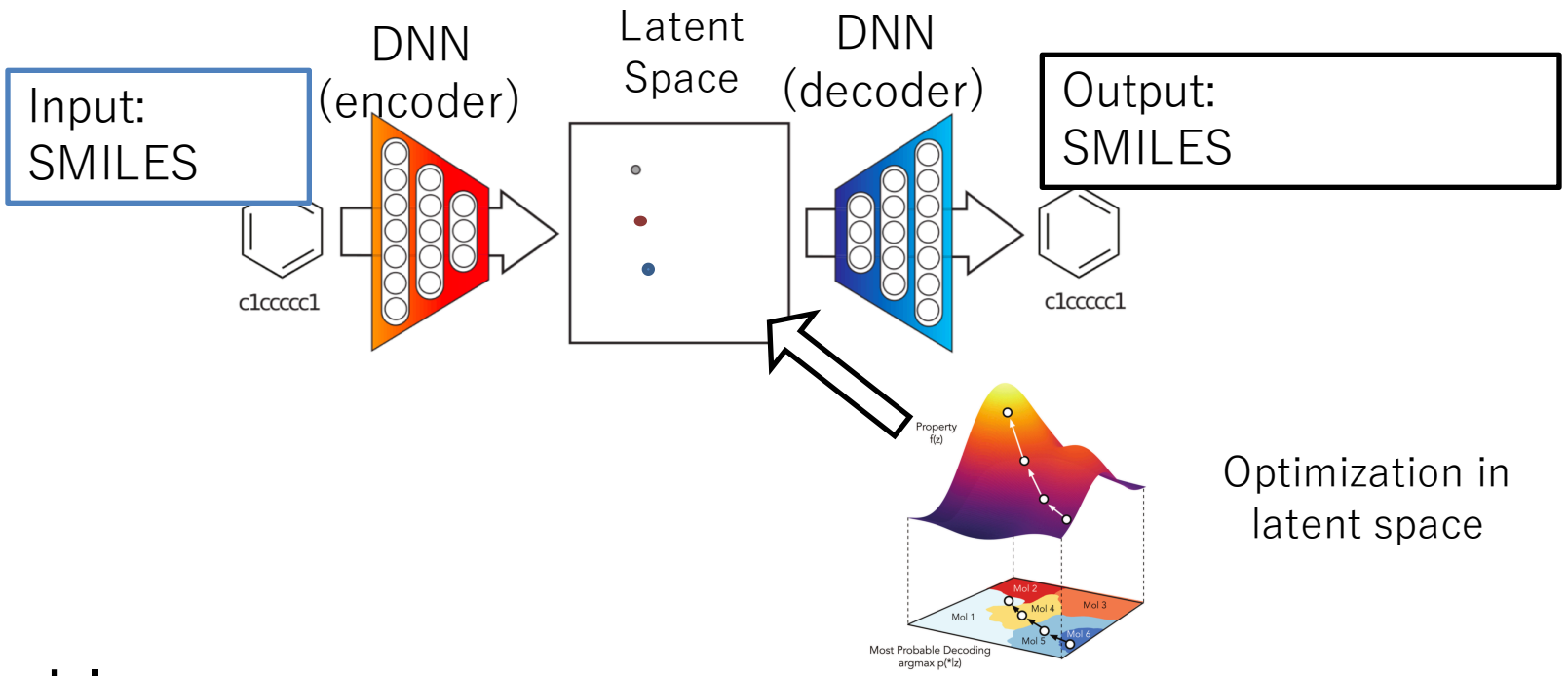
- Most existing methods make molecules by combining predetermined fragments

- De novo generation by deep neural networks
  - Variational autoencoder (Kusner et al., ICML 2017)
  - Recurrent neural network (Segler et al., Arxiv, 2017)

- ChemTS (https://github.com/tsudalab/ChemTS)
  - MCTS + Recurrent Neural Network

Generation with VAE

- Variational Auto Encoder (VAE).

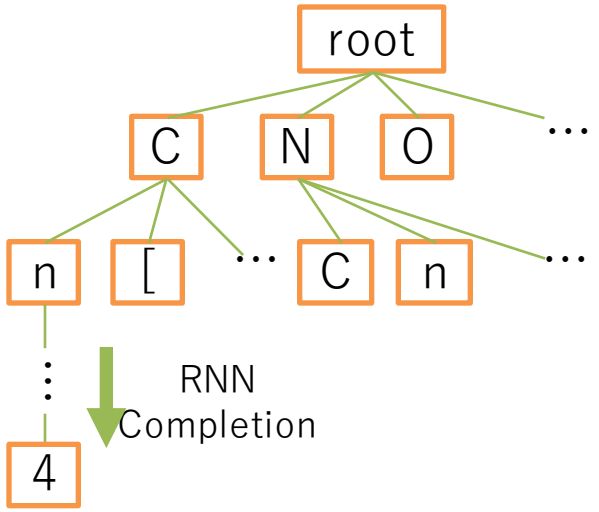


- Problem
  - Low chance of creating valid SMILES (0.4%～5%)
  - Slow

Gómez-Bombarelli et al., “Automatic chemical design using a data-driven continuous representation of molecules,” Arxiv, 2016.
Kusner et al., "Grammar Variational Autoencoder," ICML, 2017.

ChemTS: MCTS and RNN

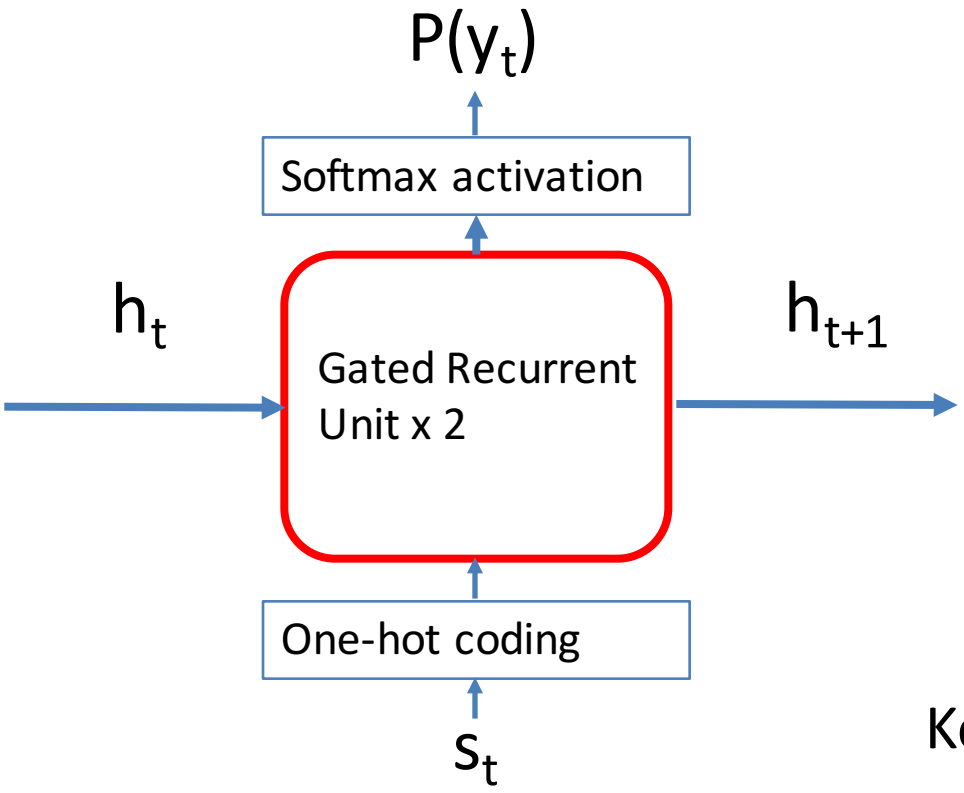
- Search tree of SMILES
- Grow tree by MCTS
- Intermediate node = prefix of SMILES
- Rollout: String completion by RNN



High chance of valid SMILES creation (>40%)
Fast generation

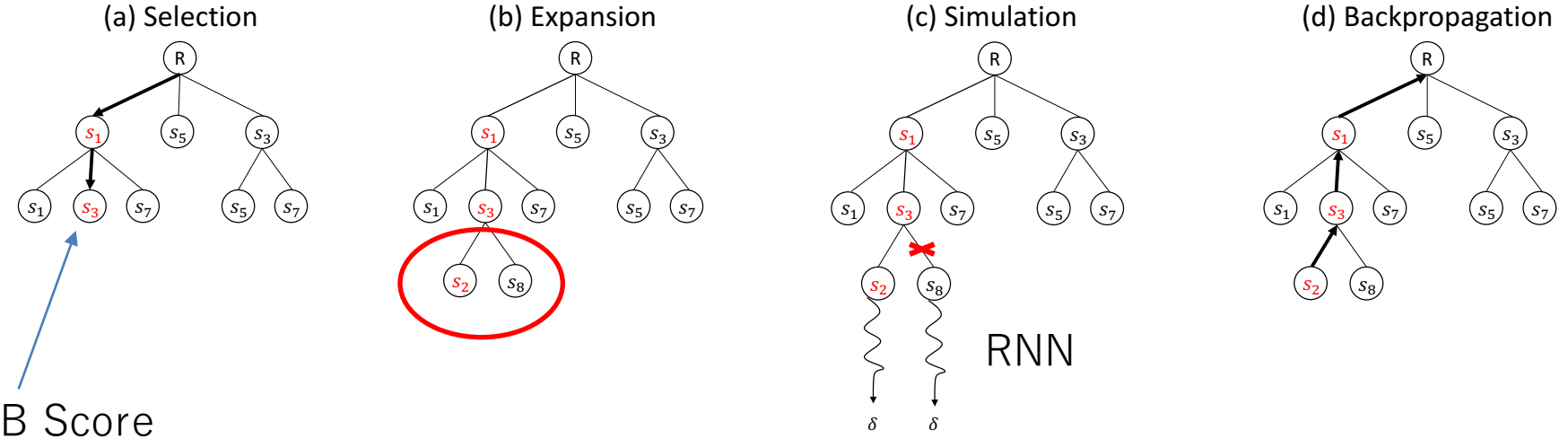
RNN : trained by 250,000 compounds

- Input : String $s_1,\ldots,s_T$
- Output : Distribution of string $P(y_1),\ldots,P(y_T)$
- Output string is right-shift of input string



Keras & Tensorflow

MCTS



UCB Score

$$u_i = \frac{w_i}{v_i} + C\frac{\sqrt{2\ln V_{parent}}}{v_i}$$

Average reward

Premium

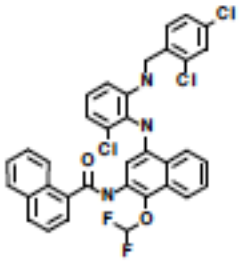
Evaluate drug-likeliness score

$$J(m) = logP(m) - SA(m) - RingPenalty(m)$$

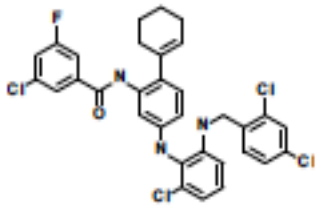
Table 1. Maximum score J at time points 2,4,6 and 8 hours achieved by different molecular generation methods. The rightmost column shows the number of generated molecules per minute. The average values and standard deviations over 10 trials are shown.

Method	2h	4h	6h	8h	Molecules/Min
ChemTS	4.91 ± 0.38	5.41 ± 0.51	5.49 ± 0.44	5.58 ± 0.50	40.89 ± 1.57
RNN+BO	3.54 ± 0.27	4.46 ± 0.24	4.46 ± 0.24	4.46 ± 0.24	8.33 ± 0.00
Only RNN	4.51 ± 0.27	4.62 ± 0.26	4.79 ± 0.25	4.79 ± 0.25	41.33 ± 1.42
CVAE+BO	-30.18 ± 26.91	-1.39 ± 2.24	-0.61 ± 1.08	-0.006 ± 0.92	0.14 ± 0.08
GVAE+BO	-4.34 ± 3.14	-1.29 ± 1.67	-0.17 ± 0.96	0.25 ± 1.31	1.38 ± 0.91

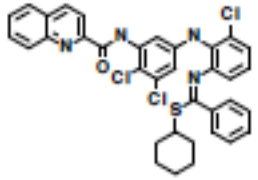
SMILES representation	J
O=C(Nc1cc(Nc2c(Cl)cccc2NCc2ccc(Cl)cc2Cl)c2ccccc2c1OC(F)F)c1cccc2ccccc12	6.56
O=C(Nc1cc(Nc2c(Cl)cccc2NCc2ccc(Cl)cc2Cl)ccc1C1=CCCCC1)c1cc(F)cc(Cl)c1	6.43
O=C(Nc1cc(Nc2c(Cl)cccc2N=C(SC2CCCCC2)c2ccccc2)cc(Cl)c1Cl)c1ccc2ccccc2n1	6.34
O=C(Nc1cc(Oc2ccc(Cl)cc2Cl)ccc1Nc1cc(Cl)ccc1Cl)c1ccc(Cl)cc1	6.33
O=C(Nc1cc(Nc2c(Cl)cccc2Cl)c(Cl)cc1Br)N(c1ccccc1)c1ccc(Cl)cc1	6.26
O=C(Nc1cc(Oc2c(Cl)cccc2Oc2ccc(-c3ccccc3)cc2)ccc1Cl)c1ccccc1	6.19
O=C(Nc1cc(Nc2c(Cl)cccc2Cl)c(Cl)c(C(=O)N(Cc2ccccc2)c2ccccc2)c1Cl)c1ccccc1F	6.08
O=C(Nc1cc(Oc2ccc(Cl)cc2Cl)cc(Cl)c1Cl)c1ncoc1-c1ccc(Sc2ccccc2)cc1	6.007
O=C(Nc1cc(Nc2c(Cl)cccc2NCc2ccc(Cl)cc2Cl)c2nccccc2c1Cl)c1ccc(Cl)cc1	6.0067
O=C(Nc1cc(Nc2c(Cl)cccc2NCc2ccc(Cl)cc2)c(Cl)cc1Cl)c1cc(F)ccc1Cl	6.0062



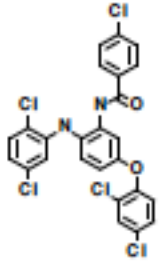
$J = 6.56$



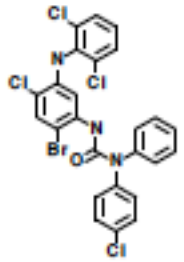
$J = 6.43$



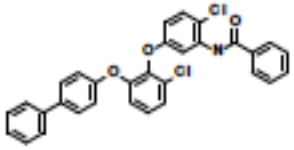
$J = 6.34$



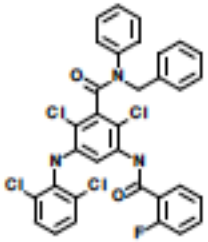
$J = 6.33$



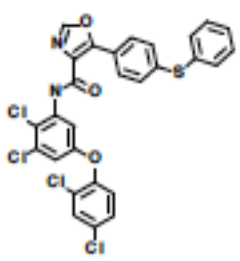
$J = 6.26$



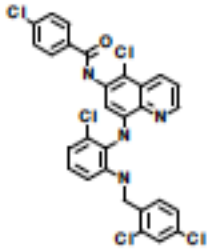
$J = 6.19$



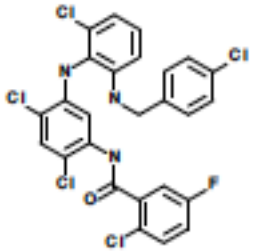
$J = 6.08$



$J = 6.007$



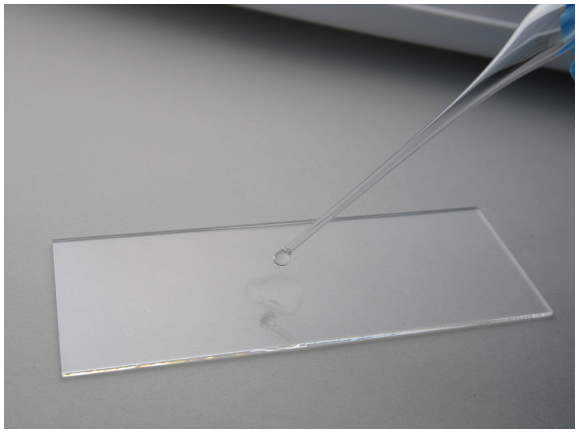
$J = 6.0067$

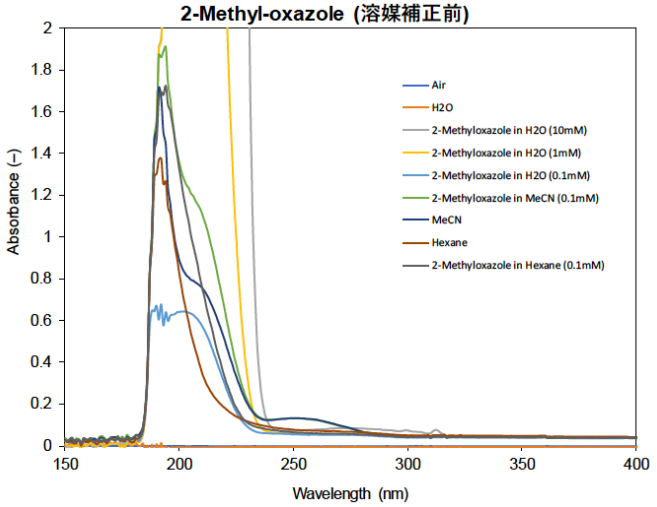


$J = 6.0062$

Generating molecules with desired absorption wavelength

- Question: Can ChemTS generate synthesizable, functional compounds within realistic computational time?





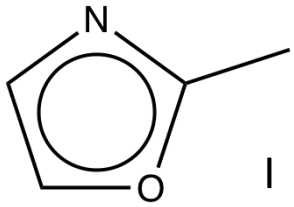
Sumita et al., in preparation

- ChemTS + TD-DFT (Gaussian)
- 2 days computation on a 12-core server for each of five target wavelengths
- 6 molecules synthesized, 5 experimentally validated

Target wavelength (nm)	200	300	400	500	600
Generated	646	757	629	607	638
DFT wavelength error < 20nm	34	26	13	12	1
Synthesized	2	2	1	1	0
Confirmed in experiments	1	2	1	1	0

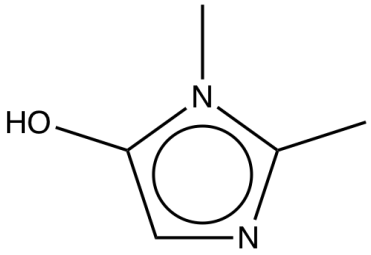
- Six compounds with known synthetic routes and high oscillation strengths are synthesized

200nm



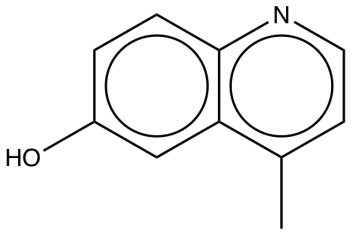
I

200nm



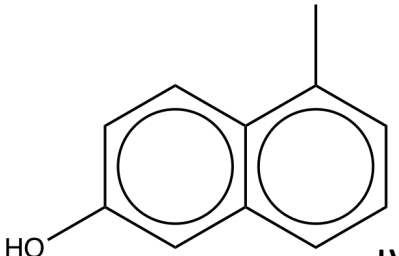
II

300nm



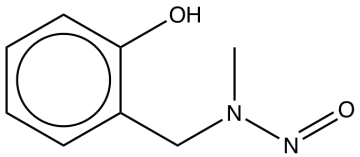
III

300nm



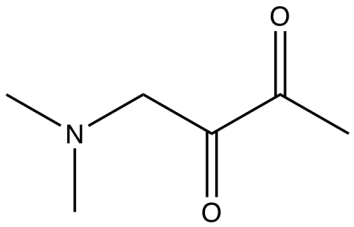
IV

400nm



V

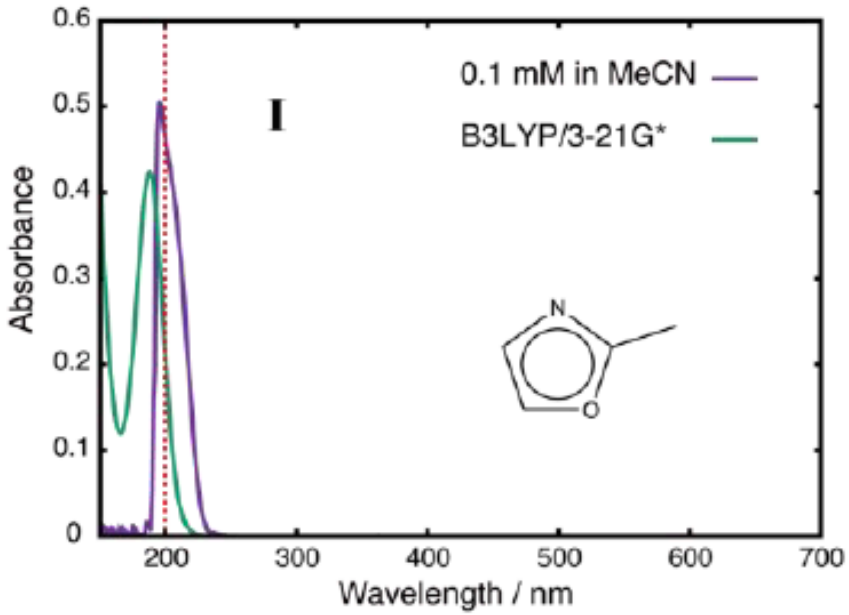
500nm

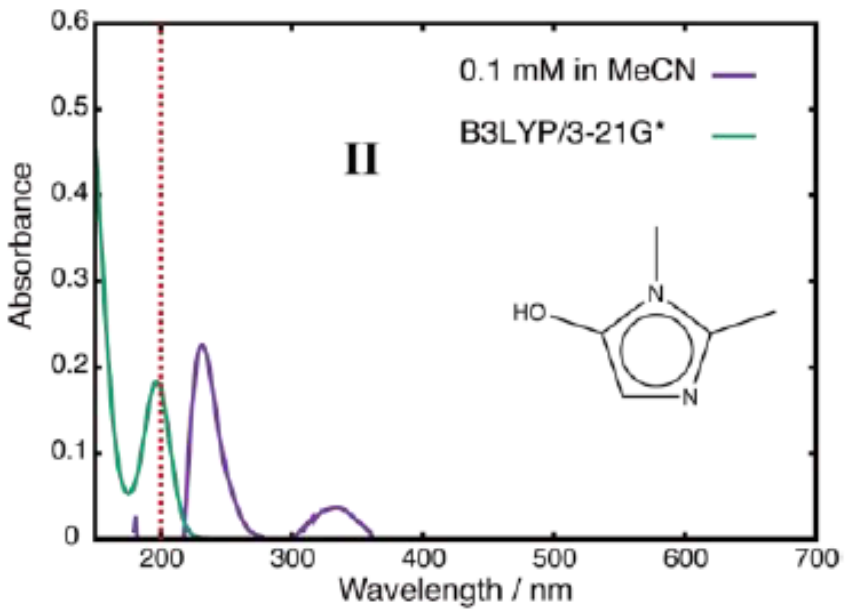


VI

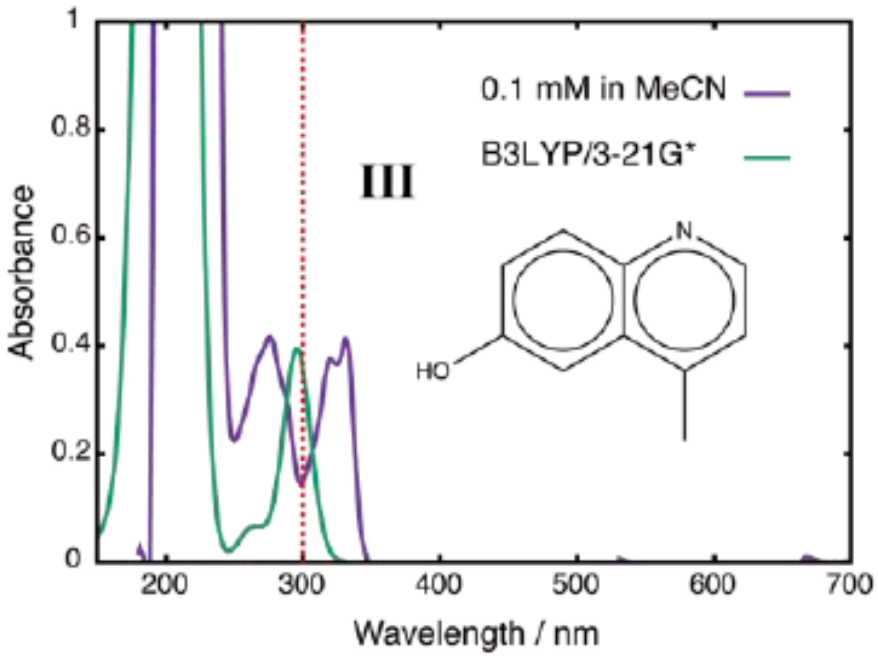
UV spectroscopy

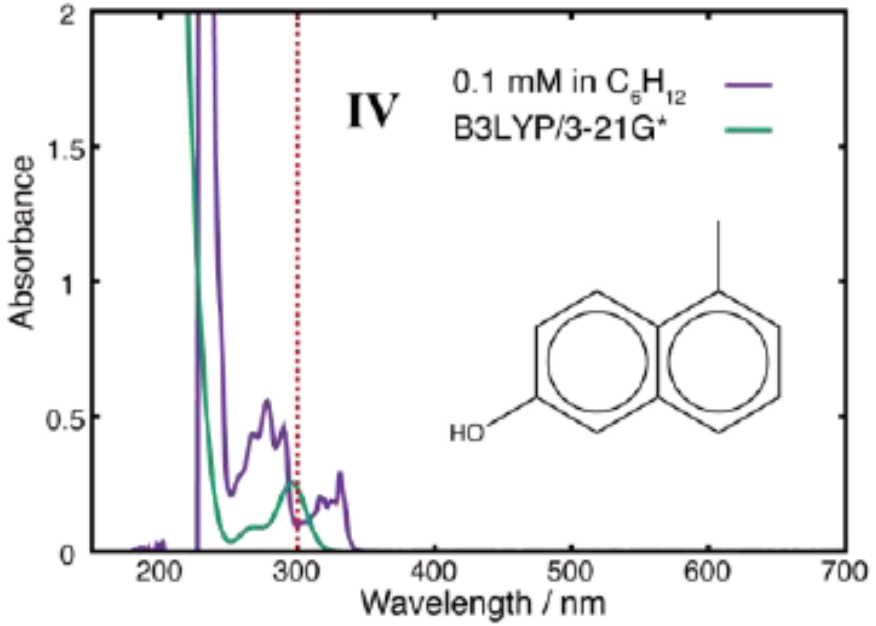
- Five out of six reproduced TD-DFT wavelength

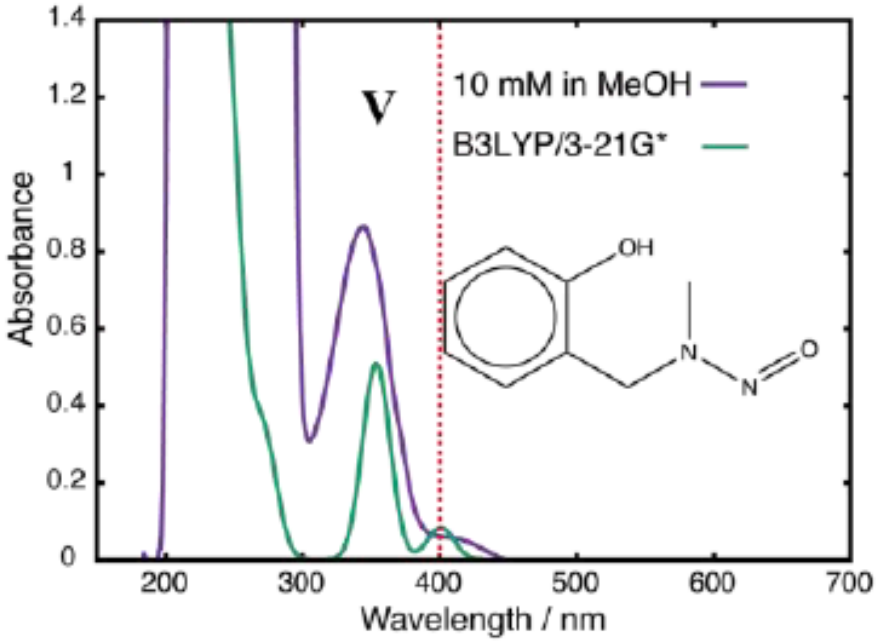


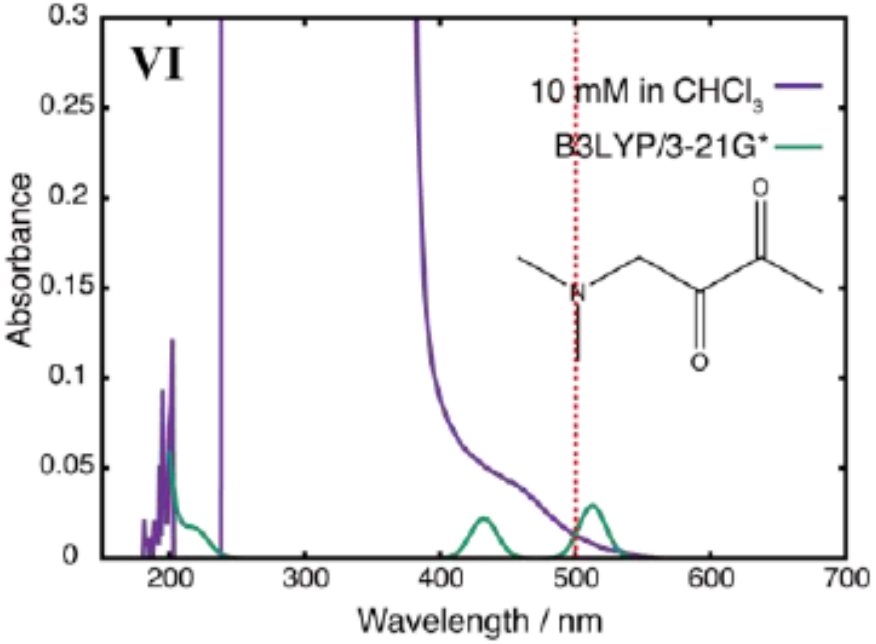


Tautomer


III
0.1 mM in MeCN
B3LYP/3-21G*
Absorbance
Wavelength / nm
HO
N

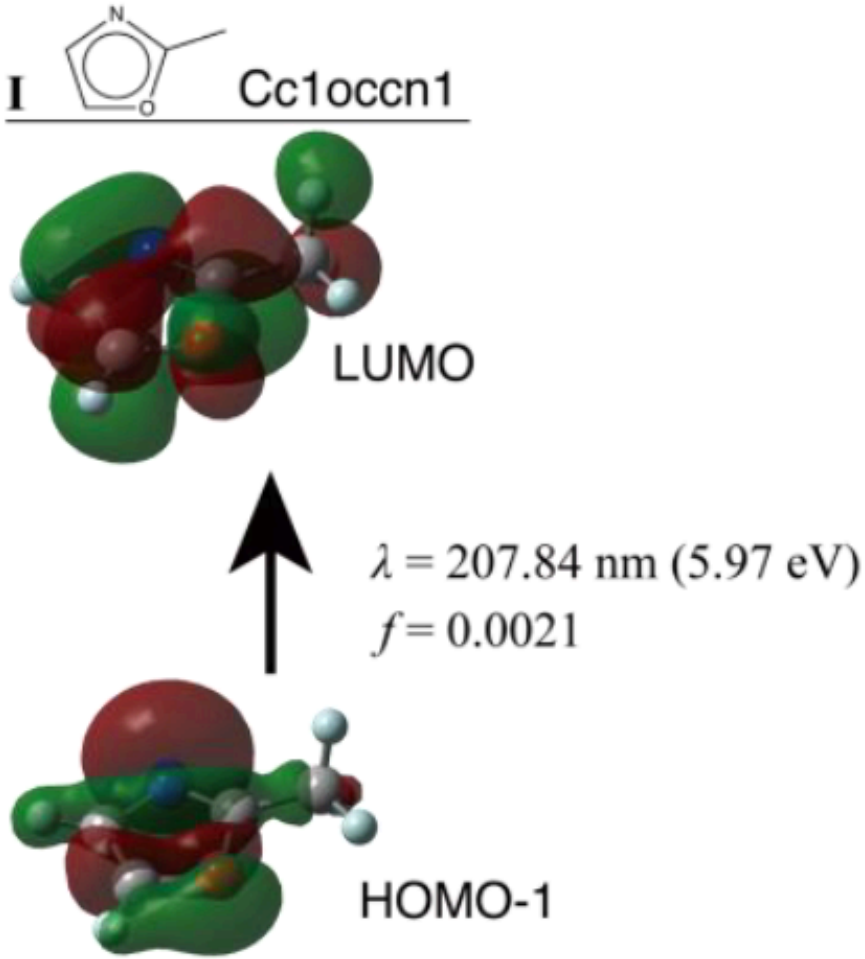

IV
0.1 mM in C_6H_{12}
B3LYP/3-21G*
Absorbance
Wavelength / nm
HO


V
10 mM in MeOH
B3LYP/3-21G*
Absorbance
Wavelength / nm
OH
N
N
O


VI
10 mM in $CHCl_3$
B3LYP/3-21G*
Absorbance
Wavelength / nm
N
O
O

Electronic origin of molecule I

- n-π* excitation, not HOMO-LUMO excitation
- Typical n-π* excitations appear at lower level (300-400 nm)

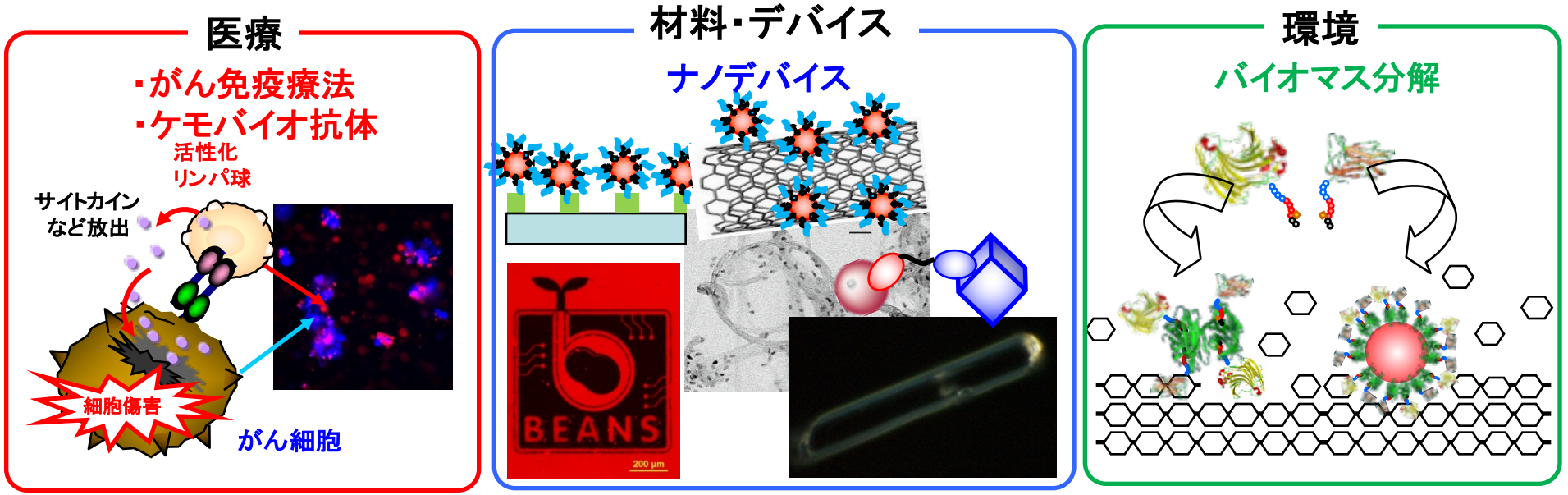


ChemTS Conclusion

- ChemTS worked well in a proof-of-concept study
- Five functional compounds in 10 days with an inexpensive server !
- Future work
  - Automated detection of quantum simulation flaws
  - Synthetic route planning
  - High-throughput experiments (robots?)

理研AIP分子情報科学チーム（PI: 津田宏治）

- 人工知能を用いて、分子・物質を設計する手法を確立
- 試行錯誤による設計よりも時間的・金銭的コストダウン
- 医療・材料・環境に関連する物質設計への応用



タンパク質工学
(protein engineering)

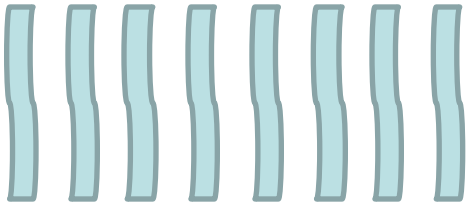
- 抗体医薬(オブジーボ)、酵素など、タンパク質は広範な産業応用を持つ
- タンパク質は、20種のアミノ酸の鎖
- 変異を導入することで機能強化・改変
- K箇所に変異導入： 20のK乗の可能性

```
1          10          20          30          40          50          60
MSKGEELFTG VVPILVELDG DVNGHKFSVS GEGEGDATYG KLTLKFICTT GKLPVPWPTL

61         70          80          90          100         110         120
VTTFSYGVQC FSRYPDHMKR HDFFKSAMPE GYVQERTISF KDDGNYKTRA EVKFEGDTLV
```

通常の手法

既存タンパク質

ランダム変異の入った
タンパク質



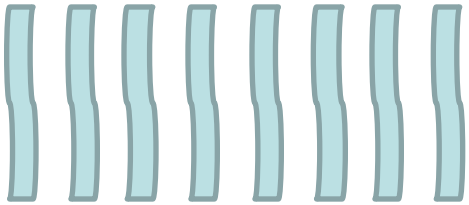
特性データ取得

優れたタンパク質を選択

AI支援タンパク質工学プラットフォーム

既存タンパク質

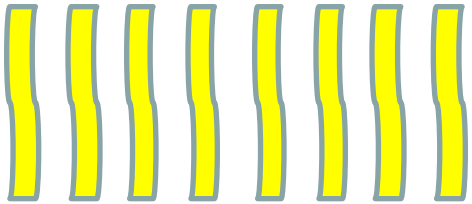
ランダム変異の入った
タンパク質



特性データ取得

機械学習モデルを学習

機械学習によって設計された
変異の入ったタンパク質



特性データ取得

優れたタンパク質を選択

機械学習による実験計画

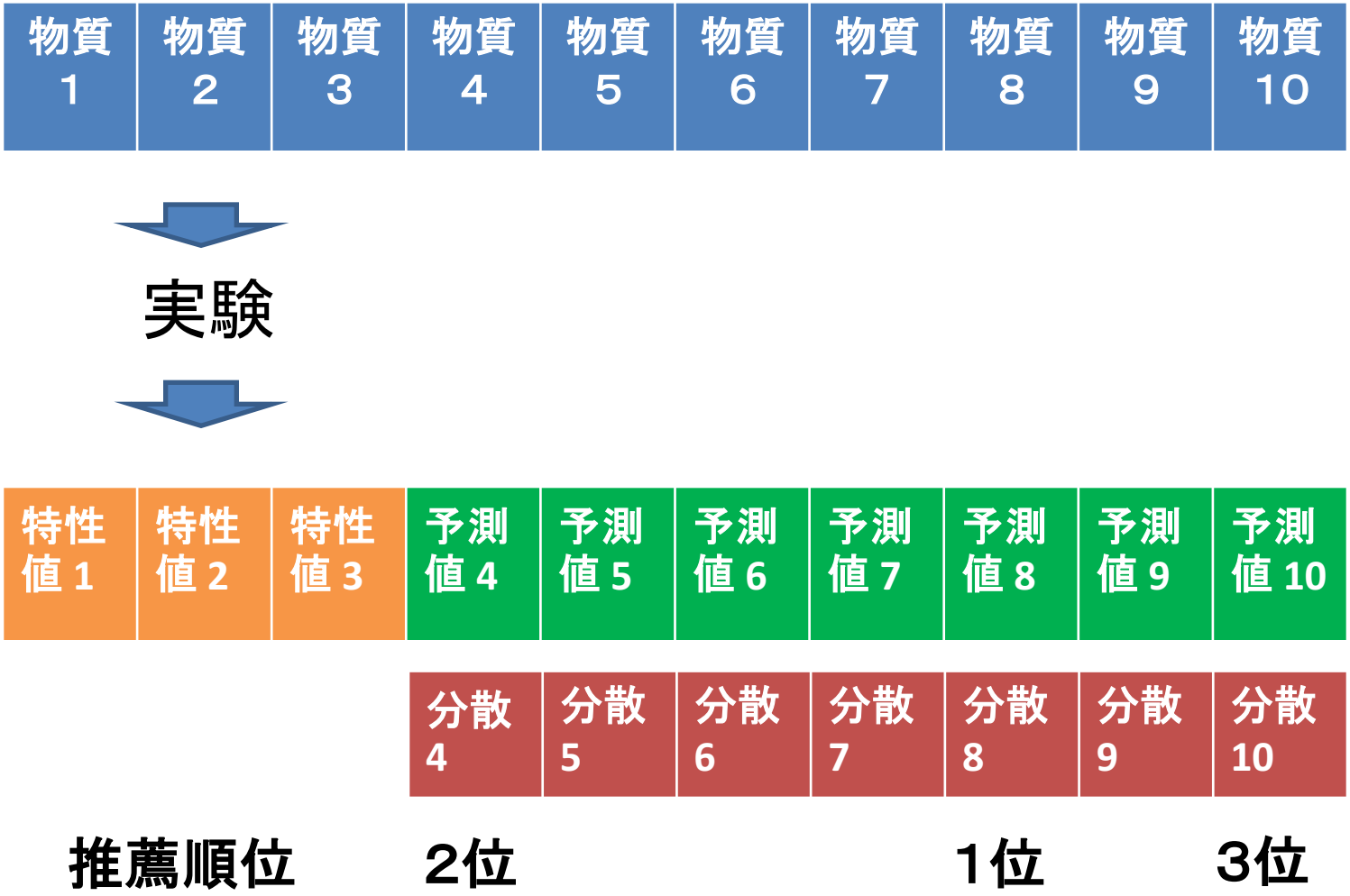
物質1	物質2	物質3	物質4	物質5	物質6	物質7	物質8	物質9	物質10

実験

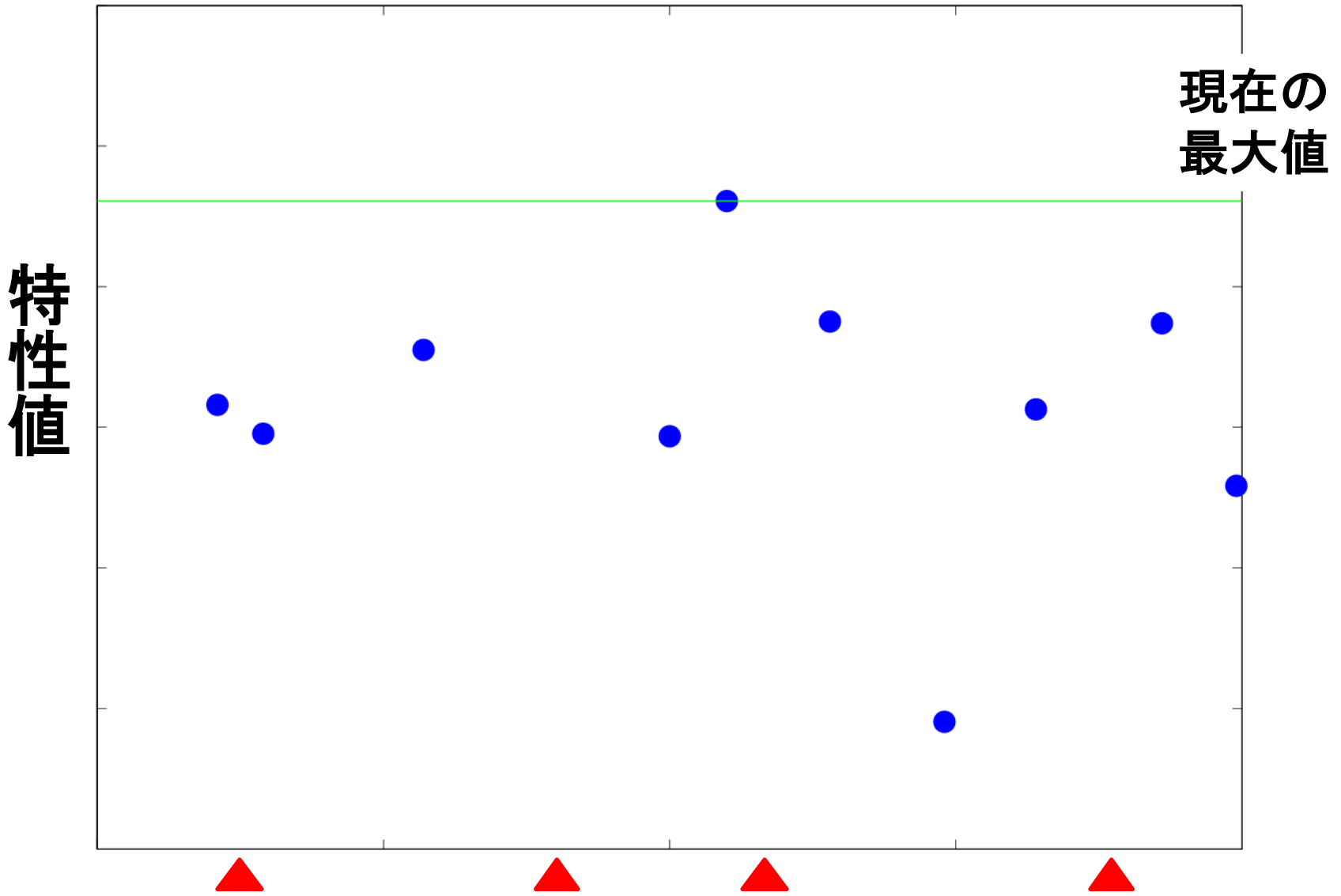
特性値1	特性値2	特性値3

機械学習による実験計画

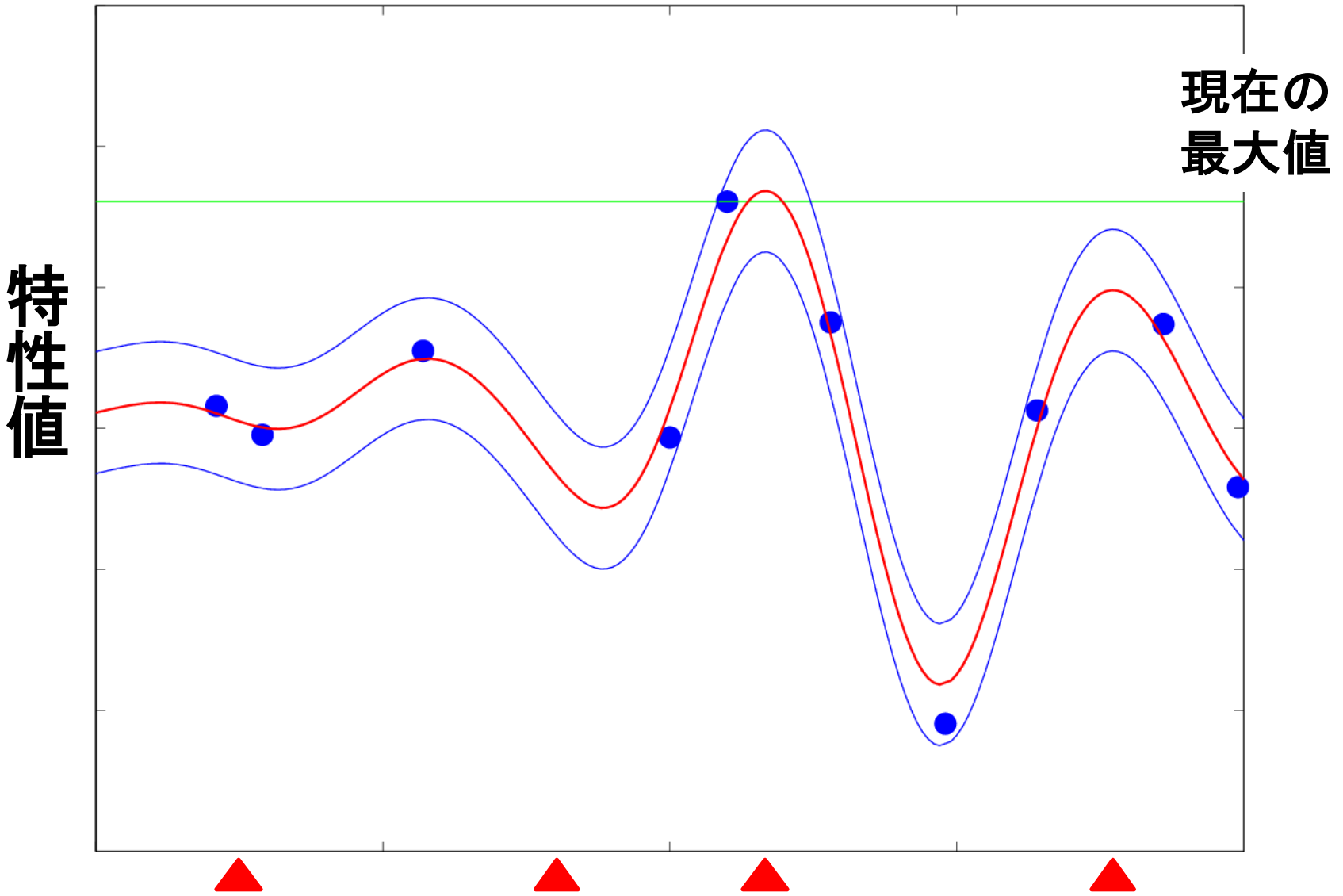
機械学習による実験計画



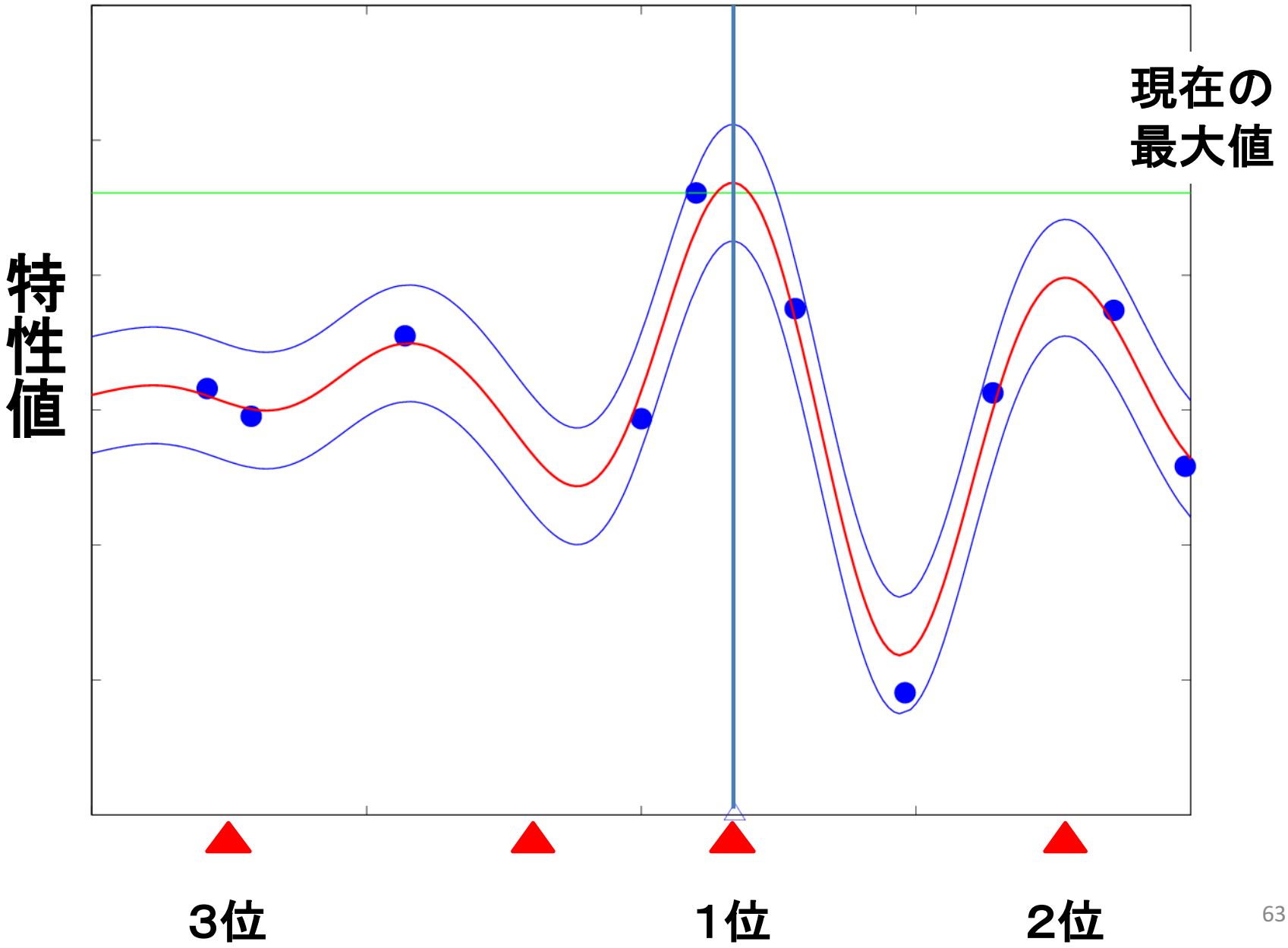
どれを観測すべきか？　推薦順位をつける



ガウス過程による機械学習

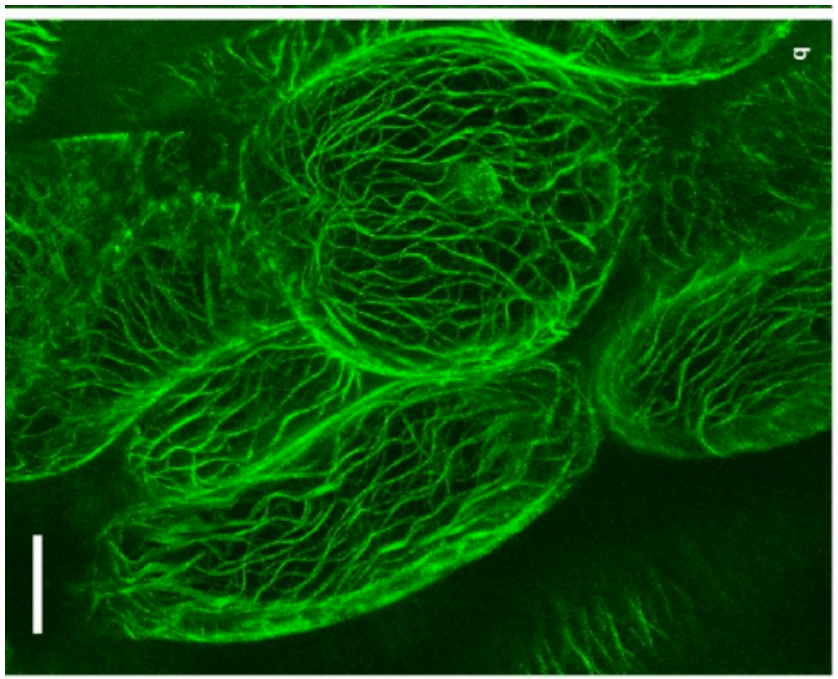


「改善確率」に基づく推薦



緑色蛍光タンパク質 (GFP)

- 標的遺伝子に結合して、標的遺伝子の生体組織における発現を標識するために用いられる
- 下村脩により発見 (2008年ノーベル化学賞)



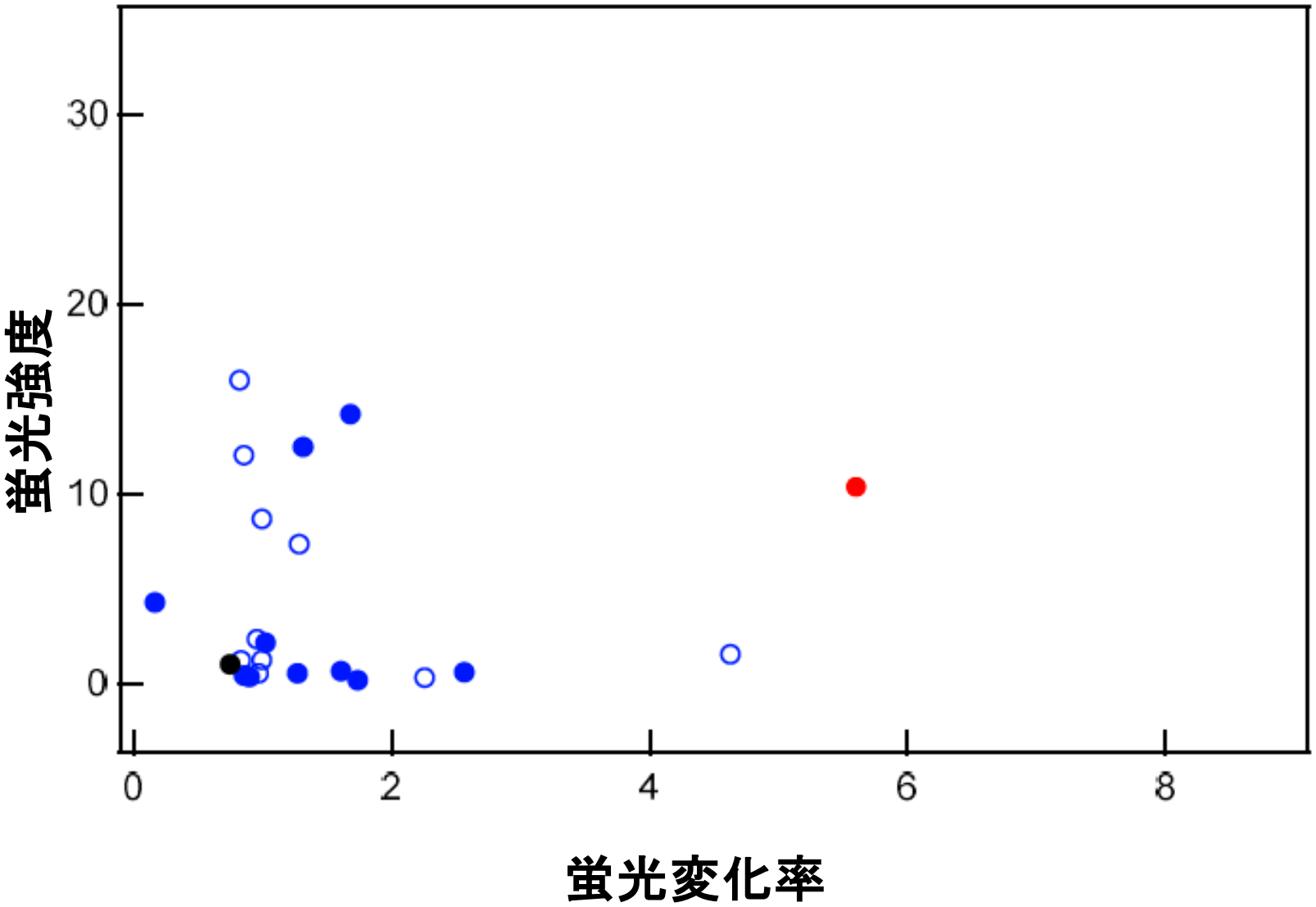
実証実験：AI支援プラットフォームを用いてGFPの機能を改変する

- GFPの４箇所を改変 (65,72,77,203)

```
1          10          20          30          40          50          60
MSKGEELFTG VVPILVELDG DVNGHKFSVS GEGEGDATYG KLTLKFICTT GKLPVPWPTL

61         70          80          90          100         110         120
VTTFSYGVQC FSRYPDHMKR HDFFKSAMPE GYVQERTISF KDDGNYKTRA EVKFEGDTLV

121      130         140         150         160         170         180
NRIELKGIDF KEDGNILGHK LEYNYNSHNV YITADKQKNG IKANFKIRHN IEDGSVQLAD

181      190         200         210         220         230        238
HYQQNTPIGD GPVLLPDNHY LSTQSALSKD PNEKRDHMVL LEFVTAAGIT HGMDELYK
```

- 候補数は16万 ： 網羅的実験は不可能
- より波長が長く、蛍光が強いほど、大きくなるように統合特性値を定義
- アミノ酸の説明変数：T-scale (Tian et al., 2007)

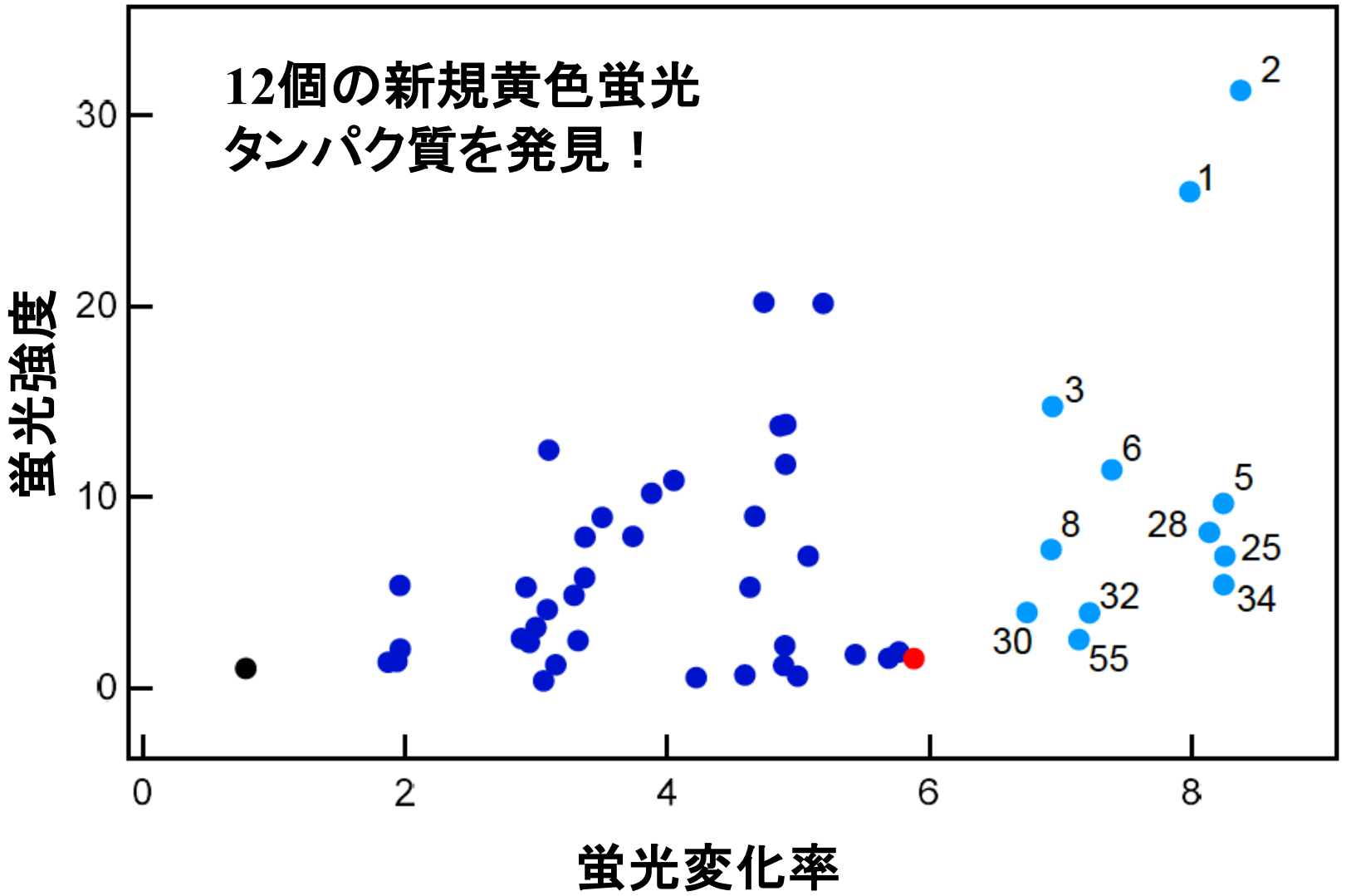
ランダム変異

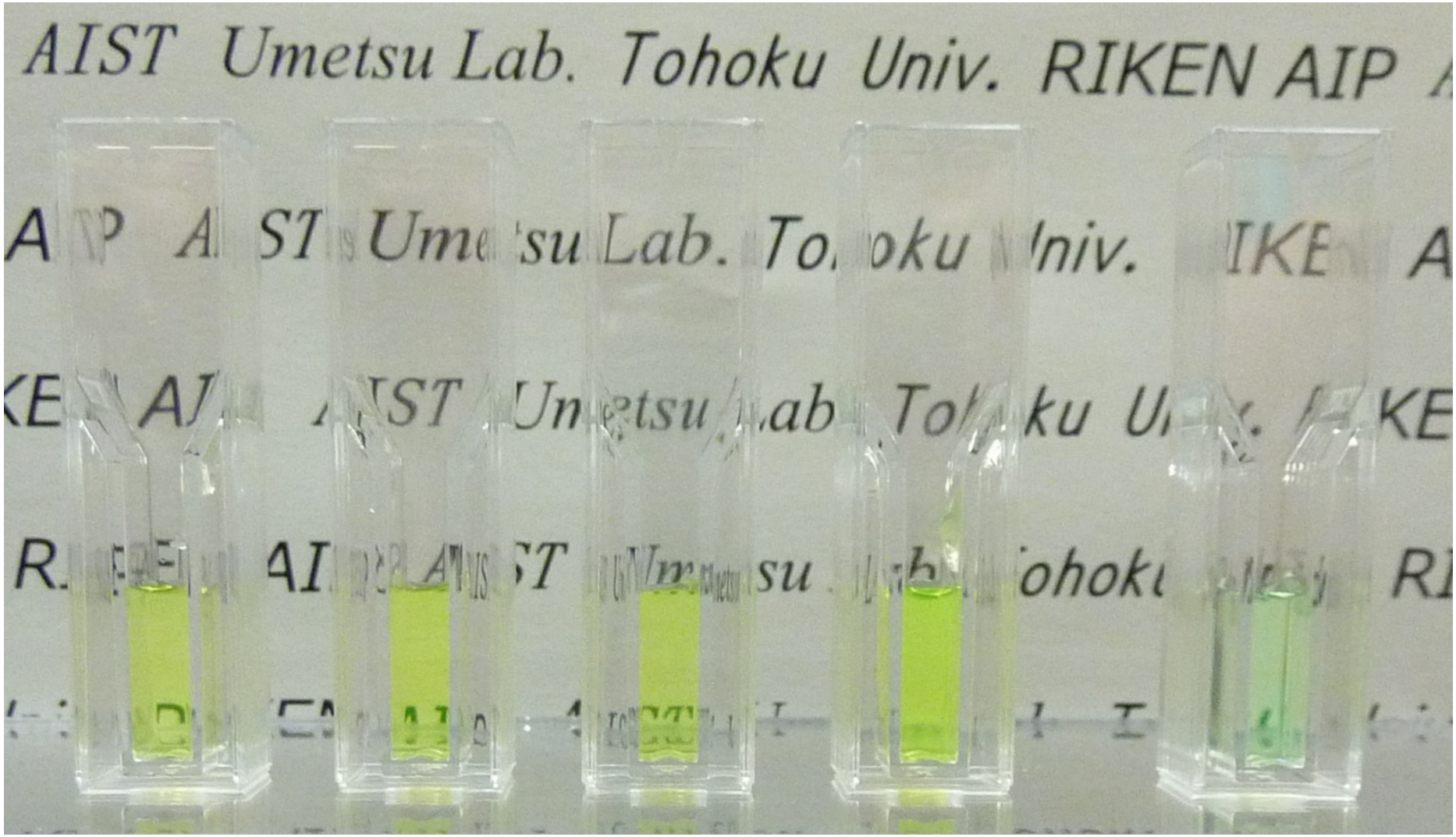


機械学習による推薦リスト(100個）

Ranking	Amino acid residue			
	65	72	77	203
1	G	A	Y	Y
2	G	A	H	Y
3	G	A	F	Y
4	G	A	H	H
5	G	G	H	Y
6	G	G	Y	Y
7	G	A	Y	H
8	G	G	F	Y
9	G	A	F	H
10	G	G	H	H

機械学習で設計した変異を導入したタンパク質





今回発見した黄色蛍光タンパク質　　GFP

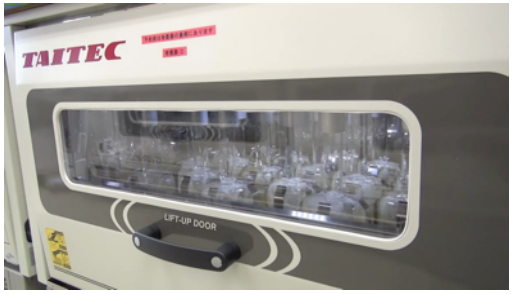
まとめ：人工知能技術により物質設計は加速できる

AIによる配列設計

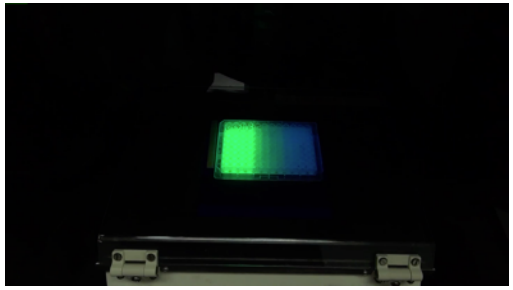
…MVSKGEELFT…

…MTSKVEGLFT…

大腸菌培養



蛍光観察



定量評価

共同研究：
梅津光央 (東北大)
斎藤裕 (産総研)
亀田倫史（産総研）