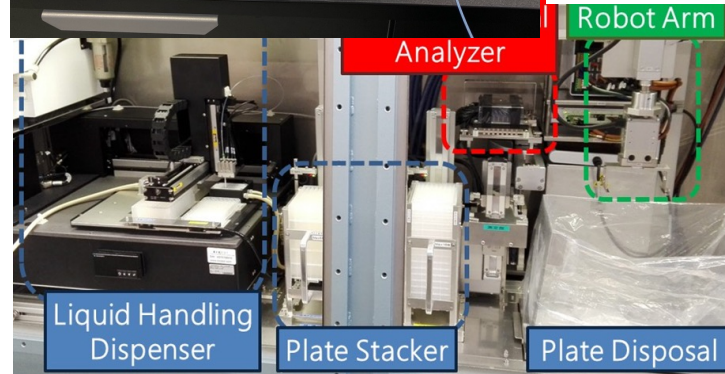
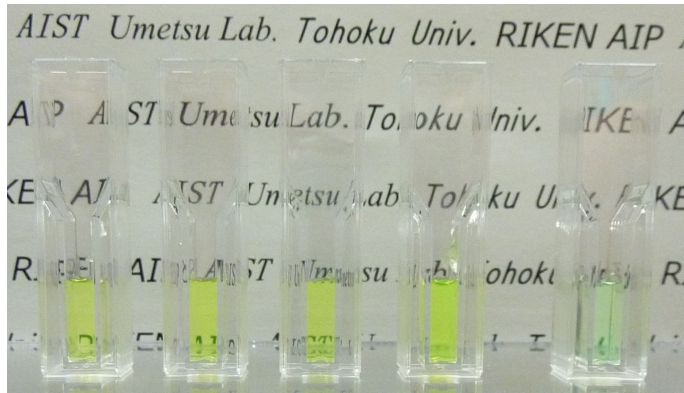
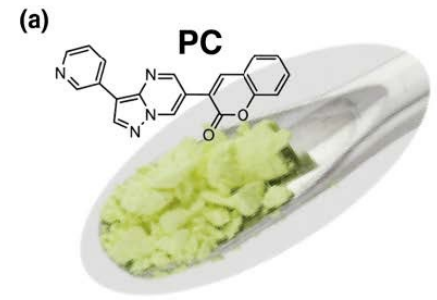
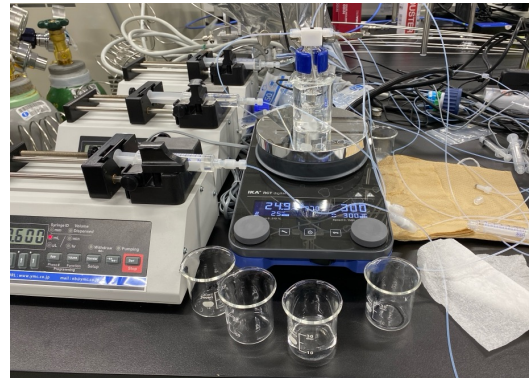
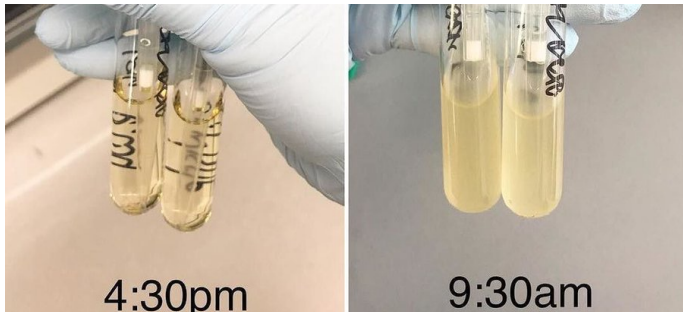


# AI4Scienceにおける量子CAEの可能性

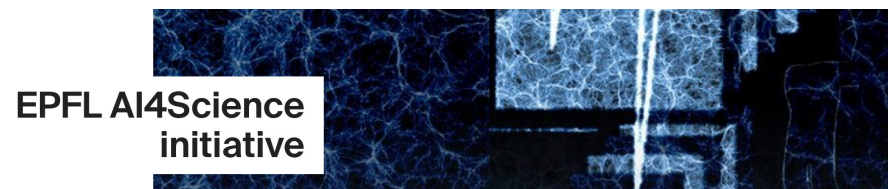
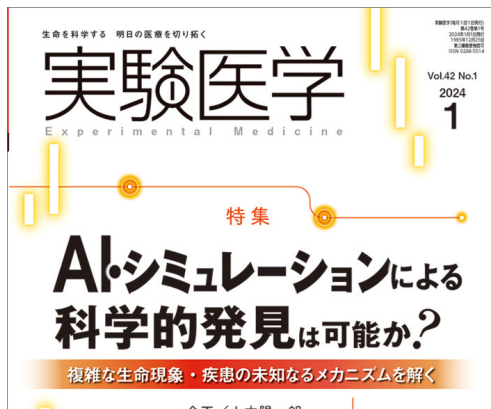
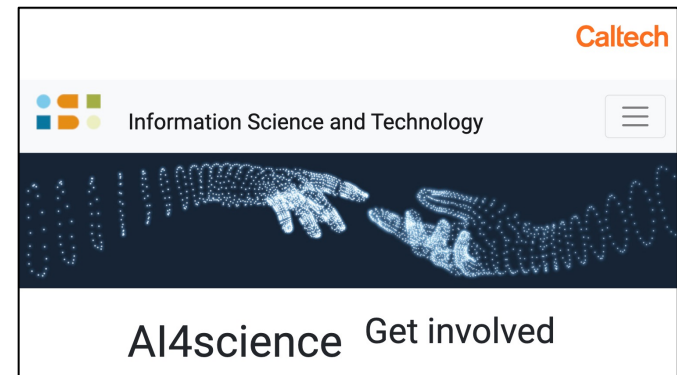
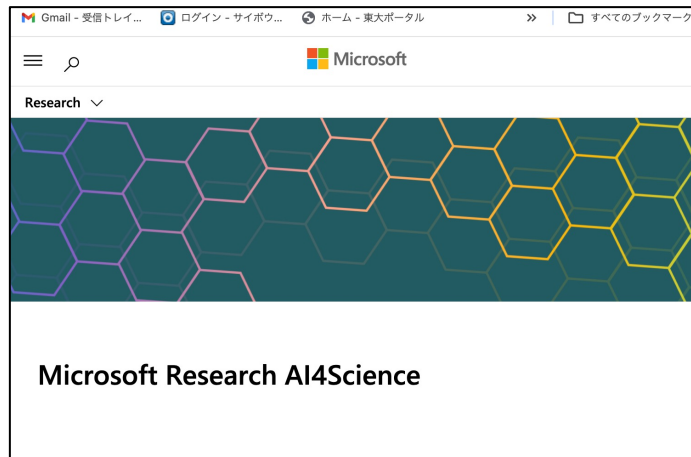
津田宏治

東京大学新領域/理研AIP/NIMS



# AI for Science

- 科学的発見を加速するアルゴリズム
- 実験・シミュレーションも含め、科学的探索をアルゴリズムと捉える
- 計算機の中だけで完結しない研究



Values



## Methods-Driven ML

algorithms that perform well on benchmarks or admit theoretical guarantees.

stereotyped benchmarks



narrow set of evaluation metrics



massive datasets



problem-agnostic methods

## Application-Driven ML

algorithms and systems that address challenges in real-world applications.

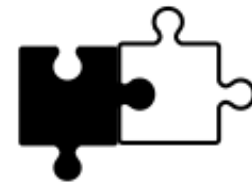
real-world tasks



application-specific evaluation metrics



auxiliary domain knowledge

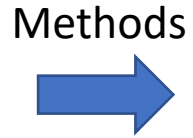


use-inspired methods

Traditional AI Startups

Deep Tech Startups

Research  
“of AI”



Research  
“by AI”



Domain

Neurips, ICML, JMLR etc

Machine Learning: Science and Technology (IOP, 2021-), Digital Discovery (RSC, 2021-), Nature Machine Intelligence (NS, 2019-), STAM Methods (TF, 2021-), Patterns (Cell, 2020-)

Physics, Chemistry, Biology, etc.

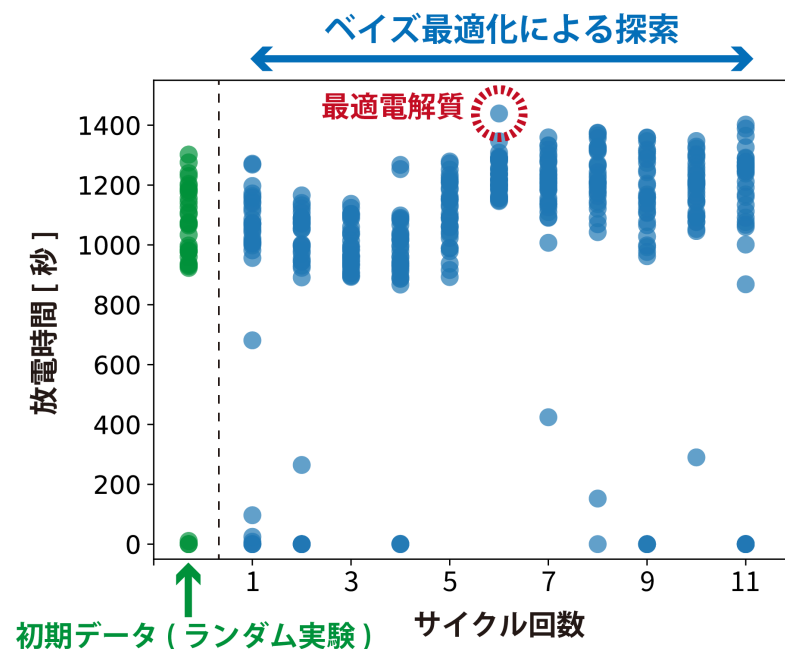
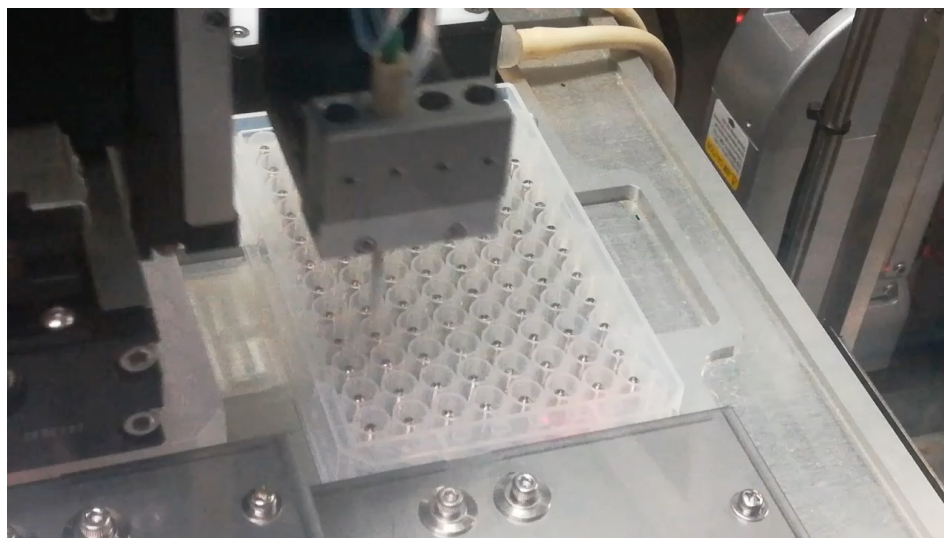


# How to explore the property space of materials

(beyond black-box optimization)

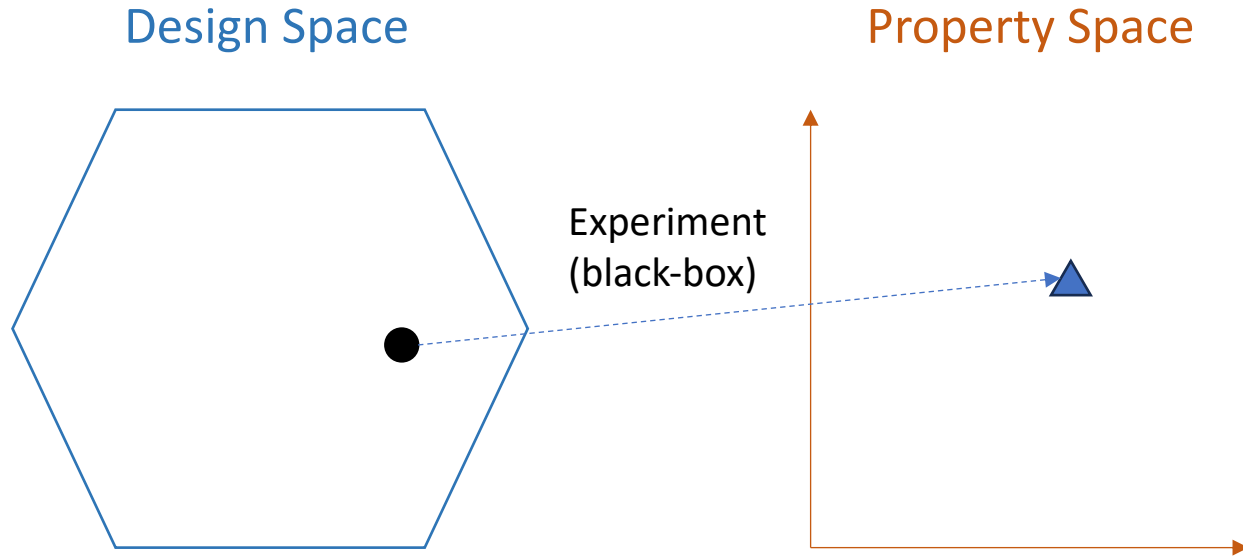
# Golden recipe for materials discovery

- Bayesian optimization + Automated experiments
- Optimization is not everything



# Exploration in materials science

- Sampling at **design space** to gain knowledge about **property space**



Example 1: Set of organic molecules

Binding affinity, Toxicity

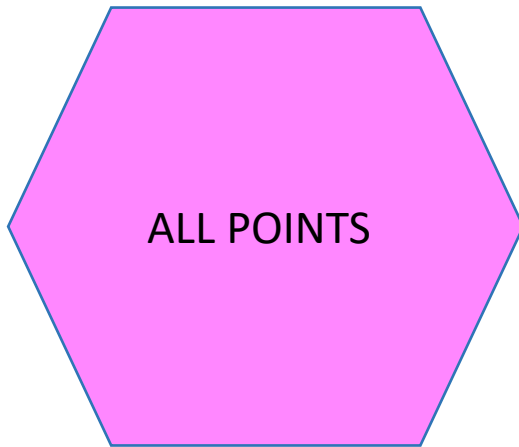
Example 2: Set of polymers

Thermal conductivity, Melting temp.

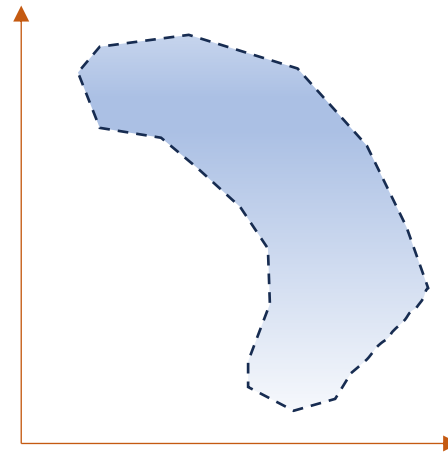
# Two problems to address

- Boundary Exploration
- Density Estimation

Design Space



Property Space



# Part 1: Boundary Exploration

Chemical  
Science



EDGE ARTICLE

[View Article Online](#)

[View Journal](#) | [View Issue](#)



Cite this: *Chem. Sci.*, 2020, **11**, 5959

All publication charges for this article have been paid for by the Royal Society of Chemistry

## Pushing property limits in materials discovery *via* boundless objective-free exploration†

Kei Terayama,<sup>id</sup>\*abcd Masato Sumita,<sup>id</sup>ae Ryo Tamura,<sup>id</sup>efg Daniel T. Payne,<sup>id</sup>h  
Mandeep K. Chahal,<sup>id</sup>e Shinsuke Ishihara,<sup>id</sup>e and Koji Tsuda\*afg

Materials chemists develop chemical compounds to meet often conflicting demands of industrial applications. This process may not be properly modeled by black-box optimization because the target property is not well defined in some cases. Herein, we propose a new algorithm for automated materials discovery called BoundLess Objective-free eXploration (BLOX) that uses a novel criterion based on kernel-based Stein discrepancy in the property space. Unlike other objective-free exploration methods, a boundary for the materials properties is not needed; hence, BLOX is suitable for open-ended scientific endeavors. We demonstrate the effectiveness of BLOX by finding light-absorbing molecules from a drug database. Our goal is to minimize the number of density functional theory calculations required to discover out-of-trend compounds in the intensity–wavelength property space. Using absorption spectroscopy, we experimentally verified that eight compounds identified as outstanding exhibit the expected optical properties. Our results show that BLOX is useful for chemical repurposing, and we expect this search method to have numerous applications in various scientific disciplines.

Received 19th February 2020  
Accepted 4th May 2020

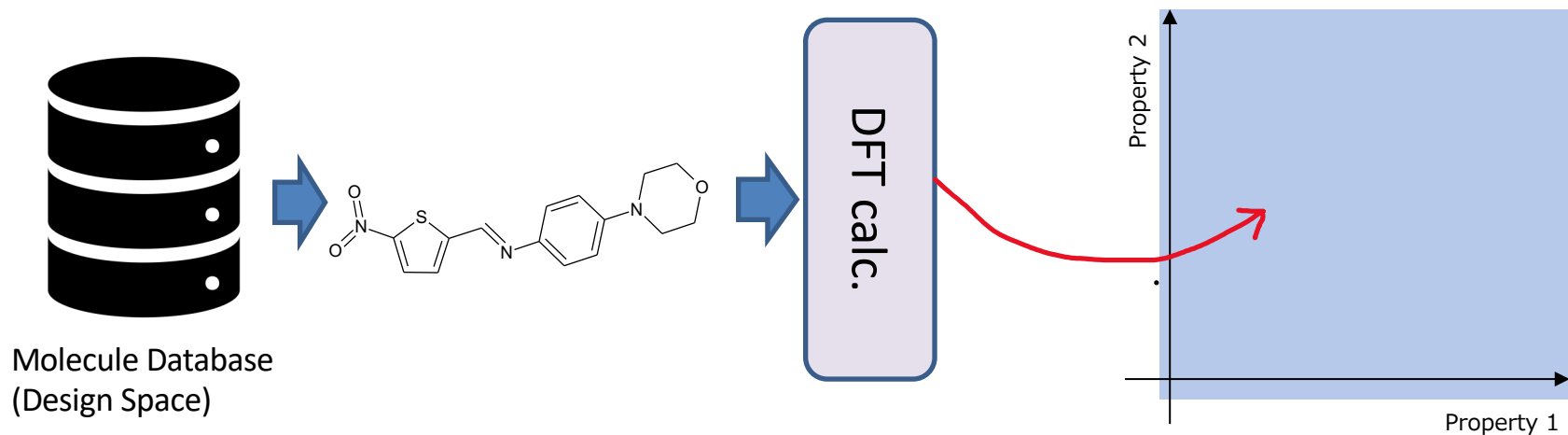
DOI: 10.1039/d0sc00982b

[rsc.li/chemical-science](http://rsc.li/chemical-science)



# Boundless objective-free exploration (BLOX)

- **Try to sample molecules uniformly in the property space**
- Not to be confused: Uniform sampling from the database



# Non-uniformity Measure

- Given a set of samples  $\mathbf{V}$ , measure deviation from the uniform distribution
- Kernel-based Stein discrepancy between  $p$  and  $q$

$$S(p, q) = E_{\mathbf{x}, \mathbf{x}' \sim p} [\delta_{p, q}(\mathbf{x})^T k(\mathbf{x}, \mathbf{x}') \delta_{p, q}(\mathbf{x}')^T],$$

$$\delta_{p, q}(\mathbf{x}) = s_q(\mathbf{x}) - s_p(\mathbf{x}) \quad s_p = \nabla_{\mathbf{x}} \log p(\mathbf{x})$$

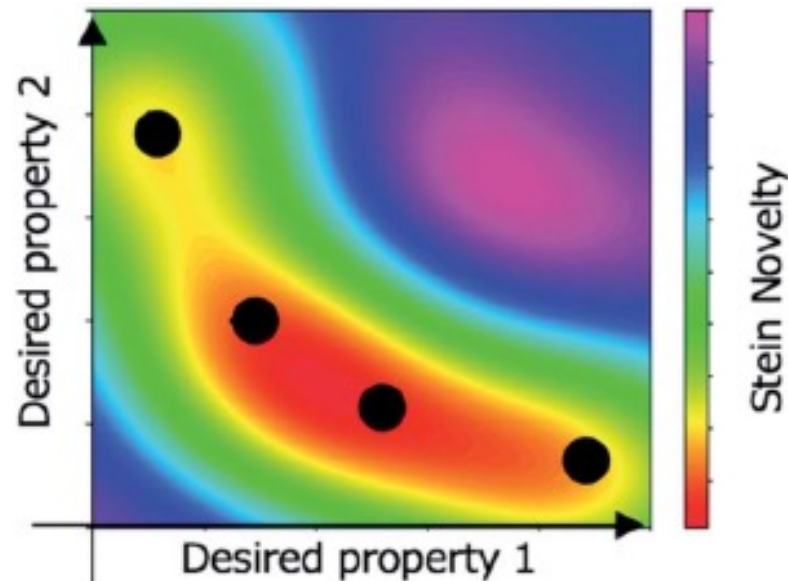
- Set  $q$  to uniform, Sample-average approximation

$$\hat{S}(V) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \sum_{t=1}^d \frac{\partial^2}{\partial v_t \partial v'_t} k(\mathbf{v}_i, \mathbf{v}_j).$$

# Stein novelty of a new point $\mathbf{v}_p$

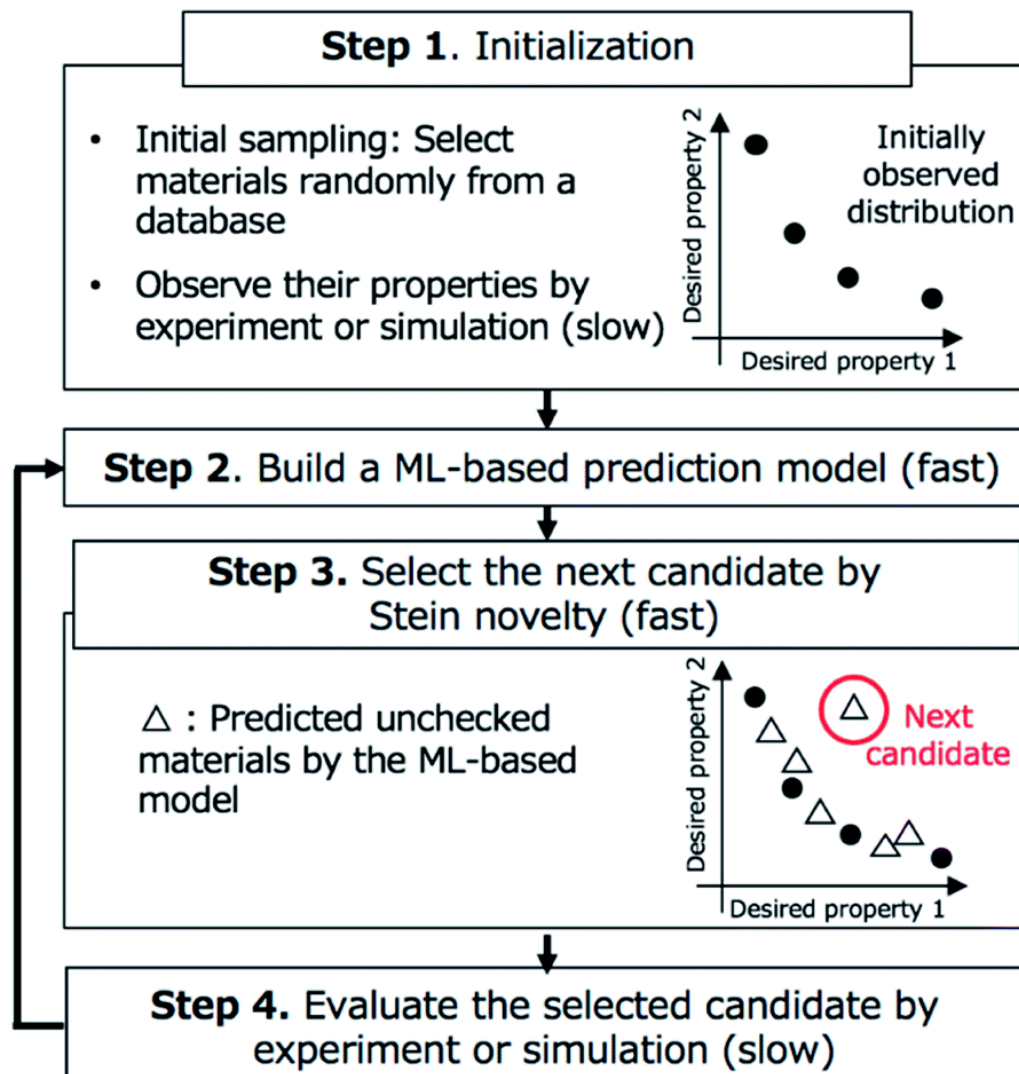
- Novelty is measured by the decrement of non-uniformity

$$N(V, \mathbf{v}_p) = \hat{S}(V) - \hat{S}(V \cup \{\mathbf{v}_p\}).$$



# Drawing a sample from the database

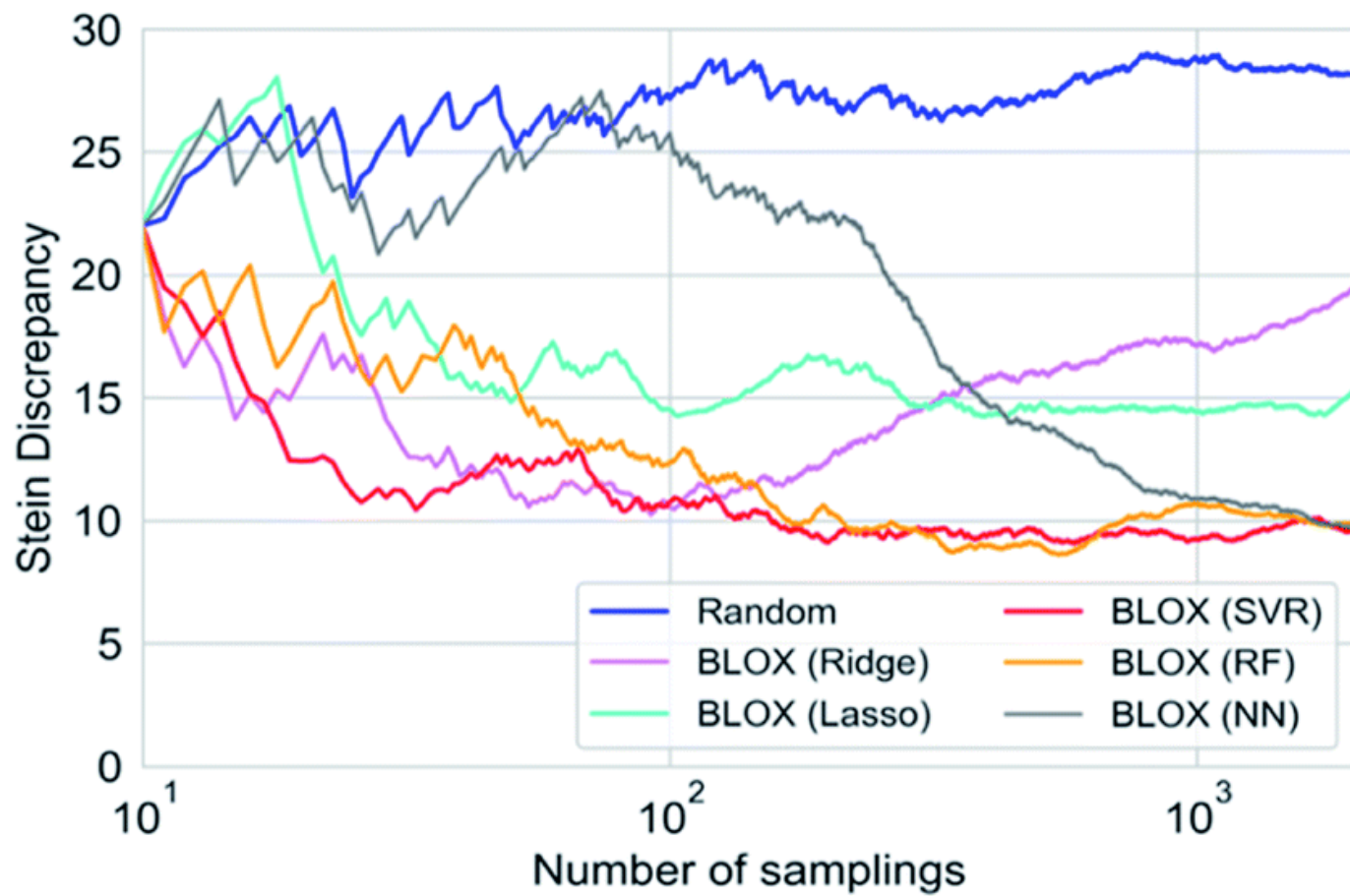
- Random forest (RF) predicts the properties
- The molecule with highest predicted Stein novelty is selected



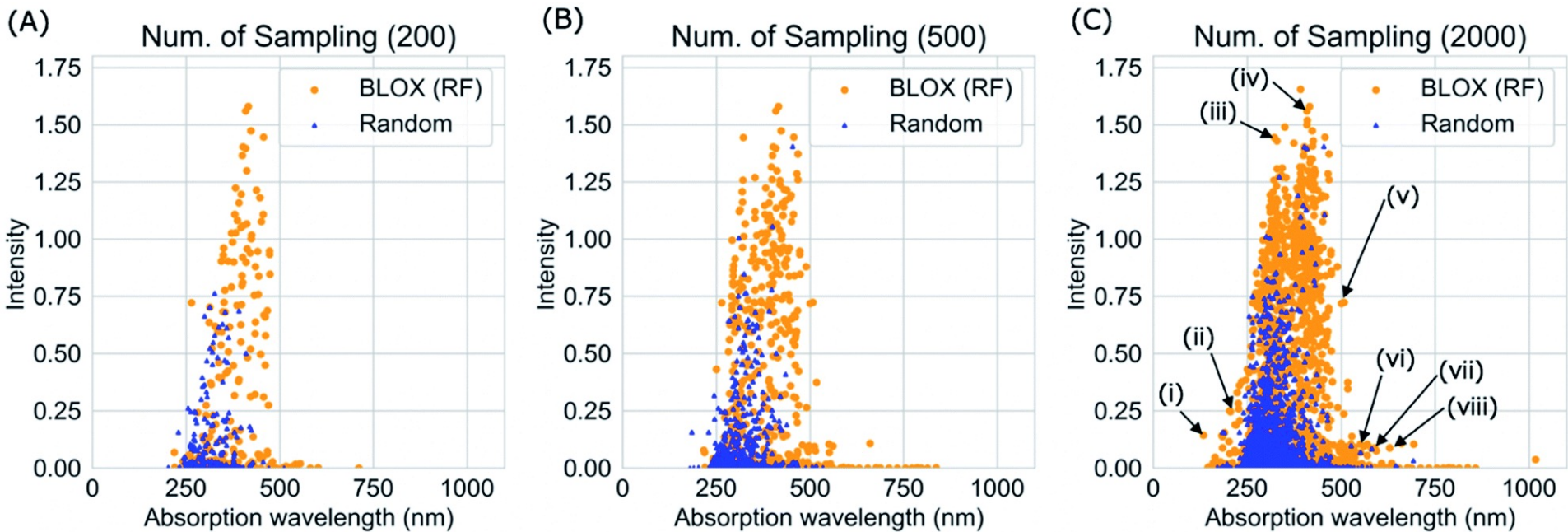
# Application: Finding dyes in a drug database

- 100,000 molecules from ZINC database
- Property space: Absorption wavelength and Oscillator strength
- TD-DFT at B3LYP/6-31G\* level
- Picked up 8 BLOX-chosen molecules, purchased them, experimentally confirmed their absorption spectra
- Efficiently discovered “colored” drugs !

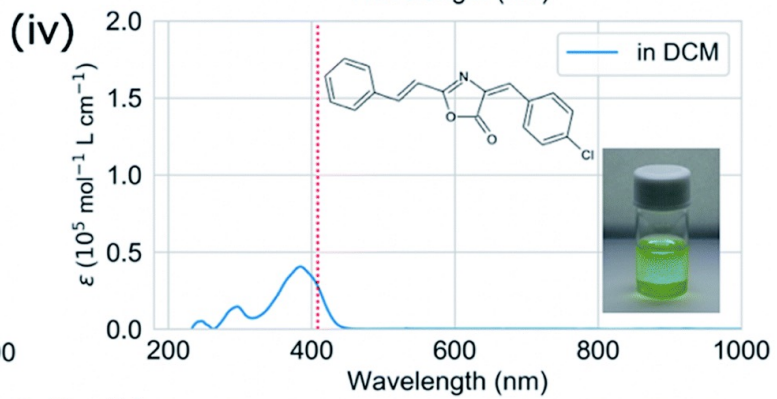
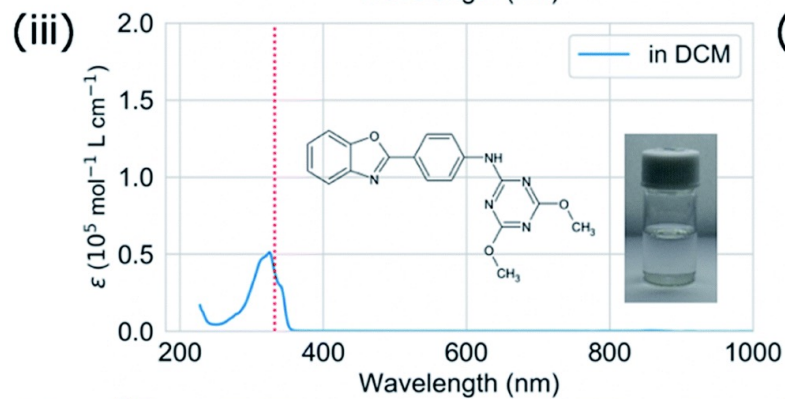
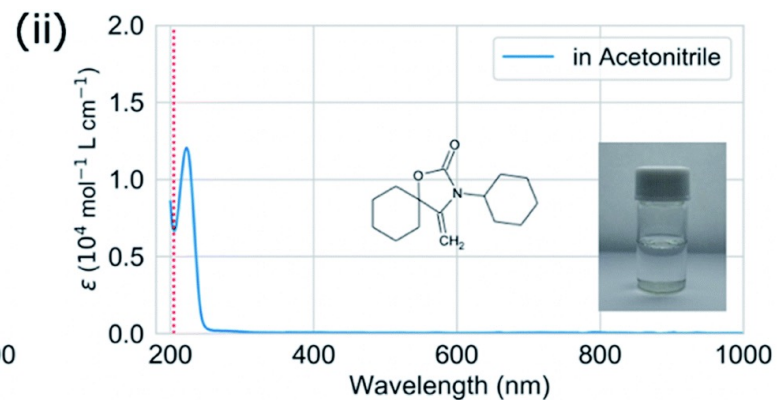
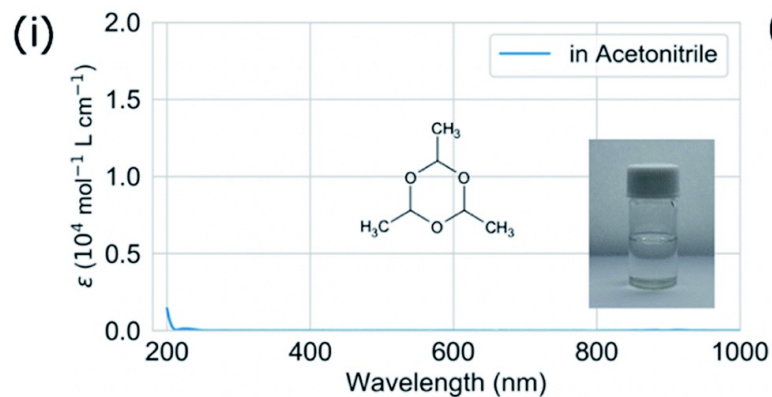




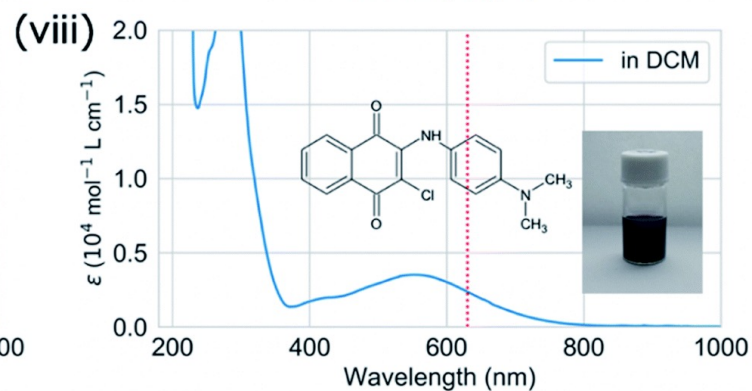
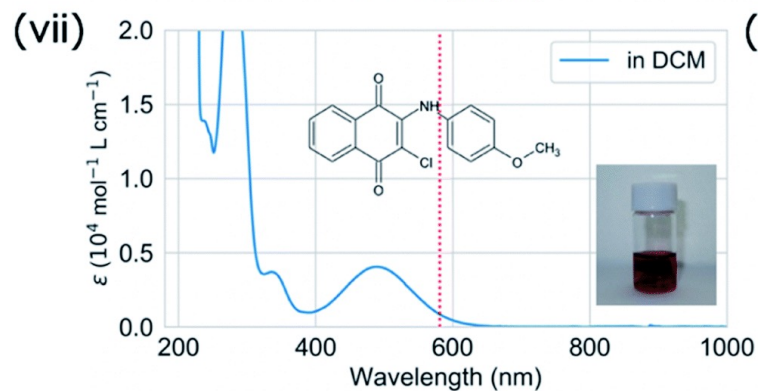
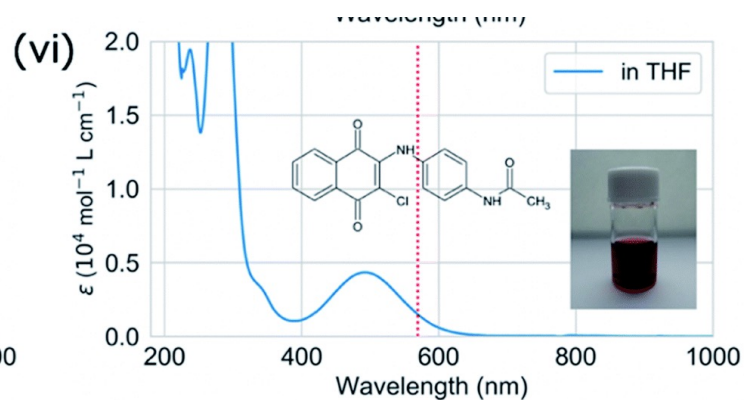
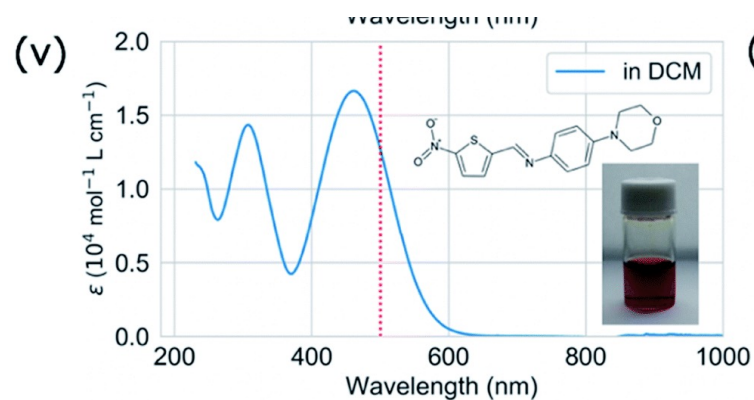
# Distribution in property space



# UV-vis absorption spectra



# UV-vis absorption spectra



# Part 2: Density Estimation

ORGANIC PROCESS RESEARCH & DEVELOPMENT

# OPR&D



pubs.acs.org/OPRD

Article

## Understanding Chemical Processes with Entropic Sampling

Yuji Kaiya, Ryo Tamura,\* and Koji Tsuda\*



Cite This: *Org. Process Res. Dev.* 2022, 26, 3276–3282



Read Online

ACCESS |



Metrics & More



Article Recommendations



Supporting Information

**ABSTRACT:** Kinetic models are widely used in simulating the relationship between the input space and the outcome space of a chemical process. Ignoring the computational cost, complete profiling, i.e., performing simulations at all grid points in the input space, would be the best way to understand the model because it provides us with a complete picture of intervariable relationships. Optimization methods that sample favorable input points can only provide narrower views. In this paper, we employ entropic sampling, a statistical physics method, to approximate complete profiling. It is cost-effective and provides a holistic picture of the model, where one can perform post hoc exploratory analyses across any region of the outcome space. Using a kinetic model of the nucleophilic aromatic substitution reaction, we analyze how the failure rate is related to process parameters and elucidate different ways to achieve low failure rates.

**KEYWORDS:** *density of states, entropic sampling, kinetic model*

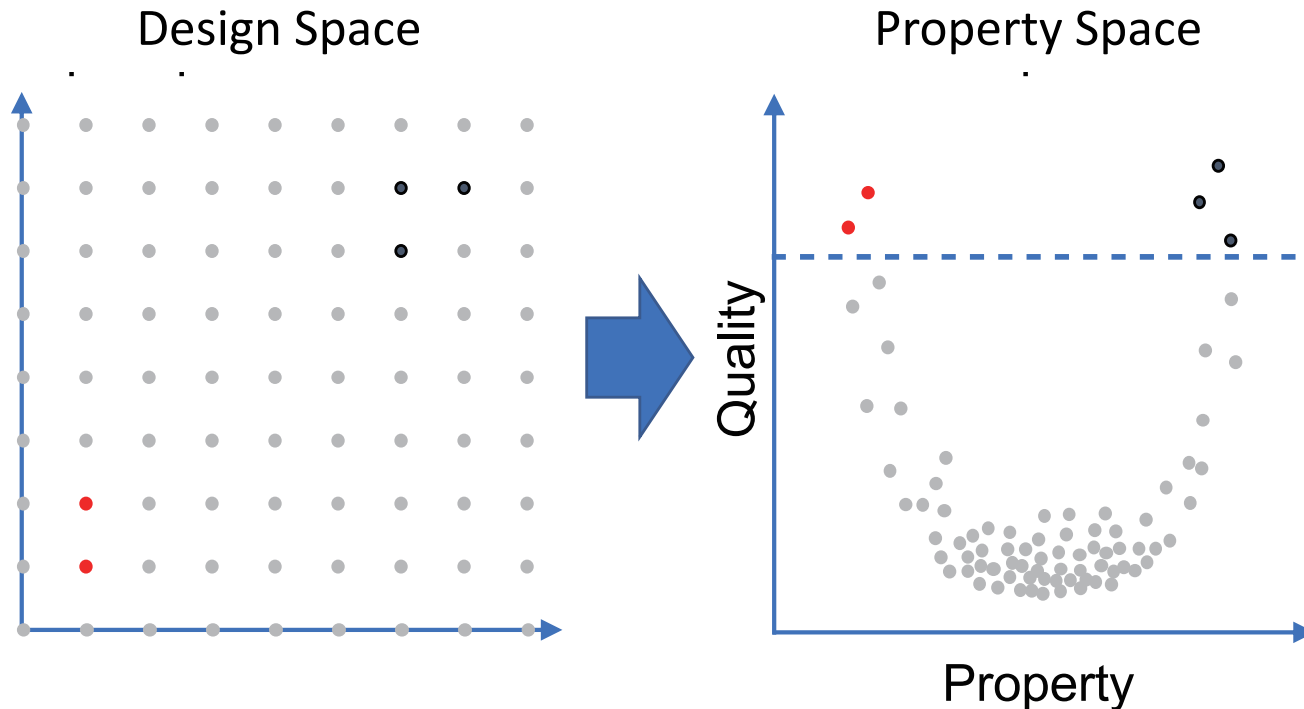


# What is entropic sampling?

- Method for computing *density of states* developed in statistical physics
  - Entropic population annealing
  - Wang-Landau sampling
  - Nested sampling
- Optimize and understand a black-box function *at the same time !*

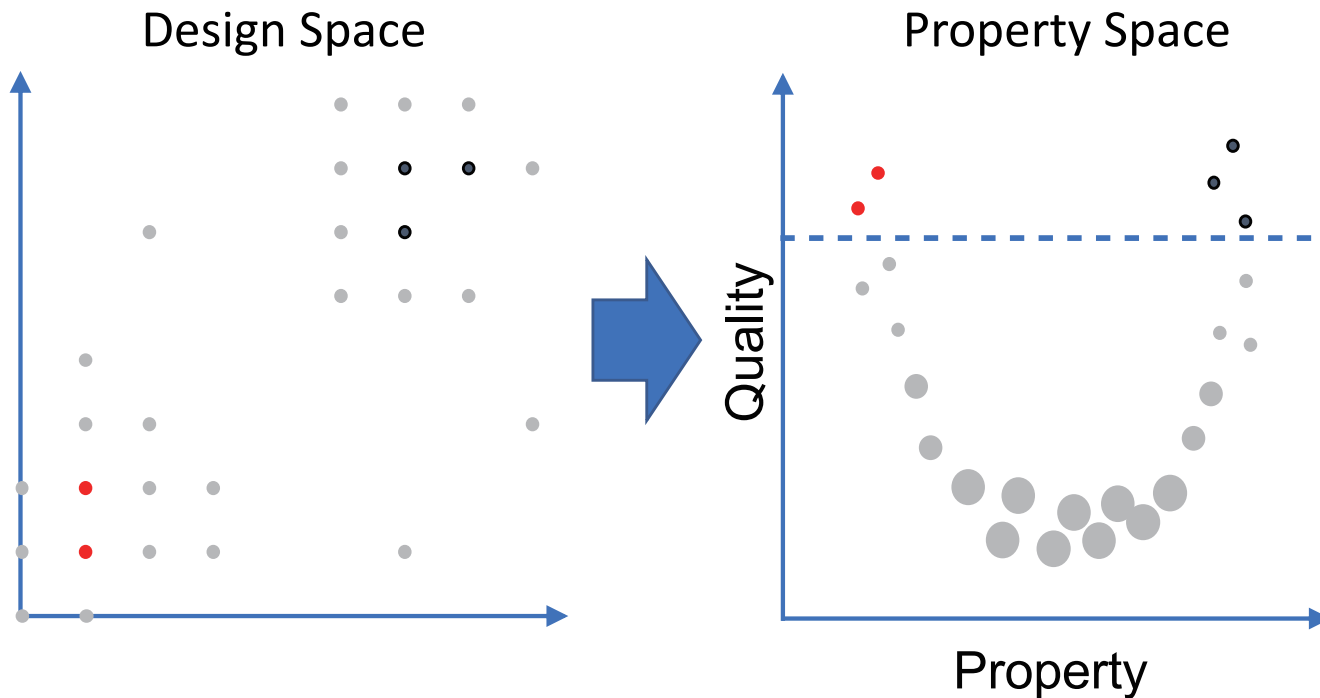
# Understanding a black box by complete profiling

- Try all possible inputs
- Observe *density of states* in the property space



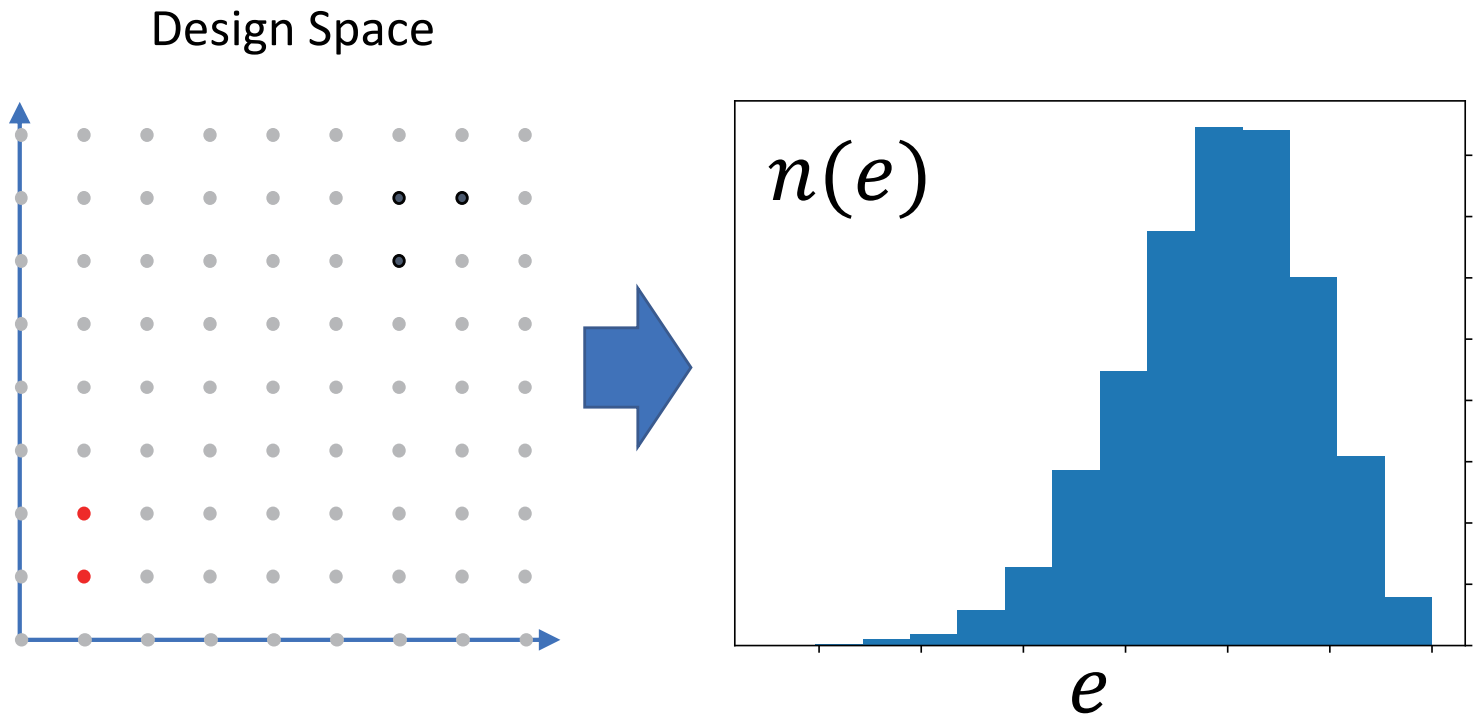
# Benefits of entropic sampling

- Minimize the number of samples (=experiments) using **weights**



# Density of states $n(e)$

- $X$ : Set of all possible inputs
- $e(x)$ : energy of black-box (i.e., a property of interest)
- $n(e)$ : Fraction of inputs whose energy is  $e$



# Markov Chain Monte Carlo (Metropolis)

- Distribution depending on energy alone

$\beta$ : inverse temperature

$$P_{\beta}(x) = \exp(-\beta e(x) + f)$$

$$f = -\log \sum_{x \in X} \exp(-\beta e(x))$$

Free energy

- Sample from  $P_{\beta}(x)$ 
  - 1. Particle  $x$  is perturbed to  $x'$
  - 2.  $x'$  is accepted with probability

$$\min\left(1, \frac{P_{\beta}(x')}{P_{\beta}(x)}\right)$$

# Single histogram method

- When sampled from  $P_\beta(x)$ , energy histogram is

$$h_\beta(e) \propto n(e)\exp(-\beta e)$$

- So, DoS is obtained as weighted histogram

$$n(e) \propto h_\beta(e)\exp(\beta e)$$

- Each sample with  $e$  is assigned a weight  $\exp(\beta e)$
- This method is **not efficient !**

# Multiple histogram method

(Ferrenberg and Swendsen, 1989)

- $N_i$  samples from inv. temp.  $\beta_i$
- Weight of sample at  $\beta_i$  with energy  $e$

$$r_i(e) = \frac{N_i \exp(-\beta_i e + f_i)}{\sum_i N_i \exp(-\beta_i e + f_i)}$$

- Proven optimal in terms of statistical error
- Free energy  $f_i$  are obtained from histograms by fixed-point iteration

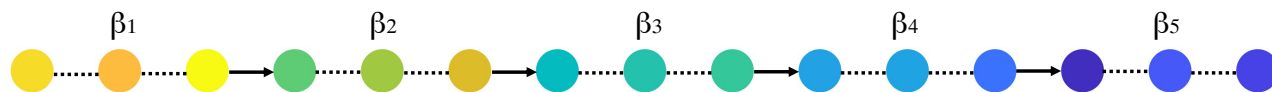


# Population annealing

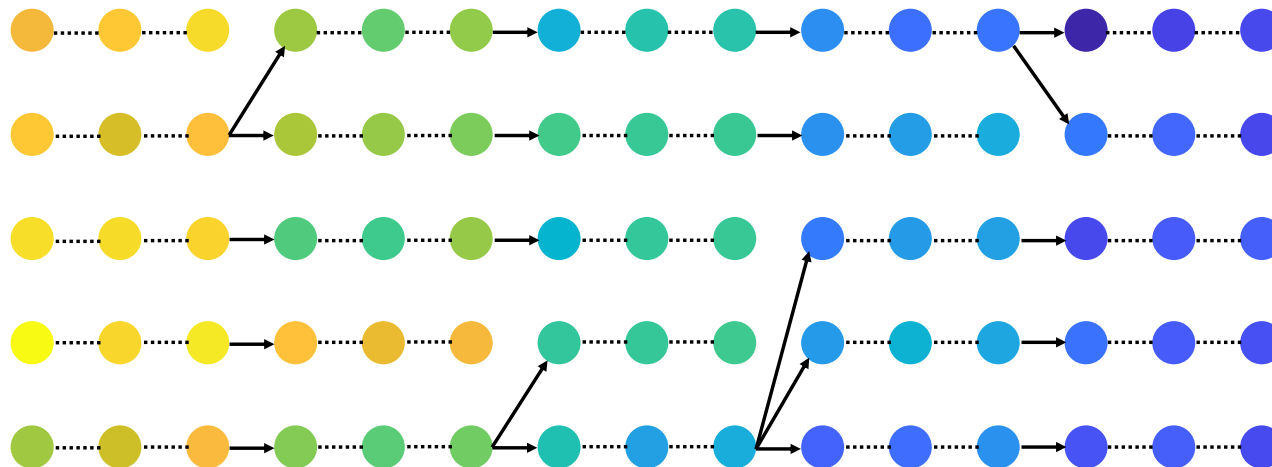
(Hukushima and Iba, 2003)

- Create samples at multiple temperatures by gradually decreasing temperature
- Update the sample set with **resampling**

(a) Simulated annealing with single particle



(b) Population annealing



# Resampling

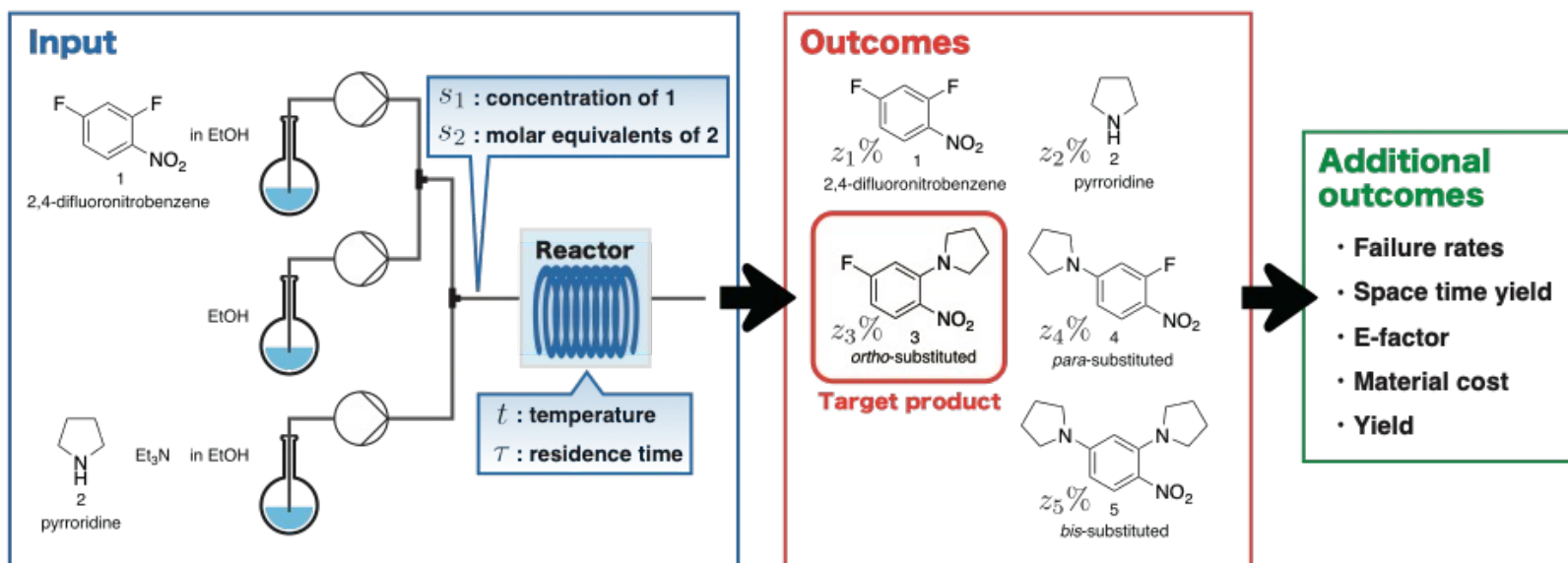
- M particles  $x_1, \dots, x_M$  at temp  $\beta_i$
- Adapt the particle set for next temp  $\beta_{i+1}$
- Probability for  $x_m$

$$q_m \propto \exp(-(\beta_{i+1} - \beta_i)e(x_m))$$

- Draw M particles with replacement
- Some particles multiply, some vanish

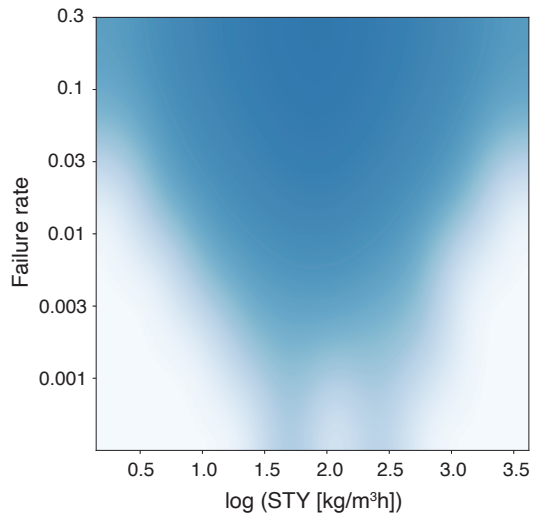
# Optimizing and Understanding Chemical Process

- Kinetic model of multi-step aromatic nucleophilic substitution reaction
- Design Space
  - Concentration of 2,4-difluoronitrobenzene
  - Molar equivalent of pyrrolidine
  - Resident time, temperature
- Property Space
  - Failure rate (**energy**), Space time yield, etc.

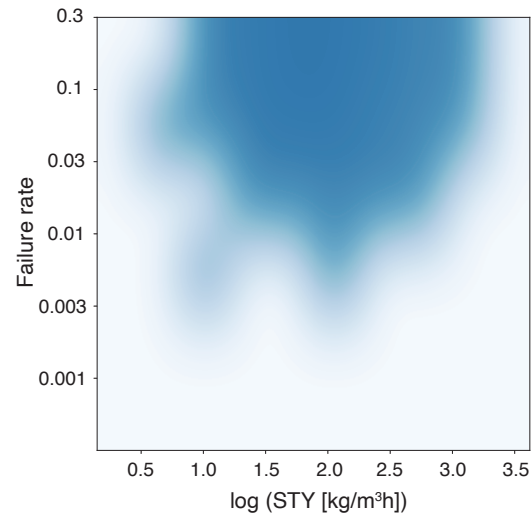


# Failure rate – space time yield (20,000 samples)

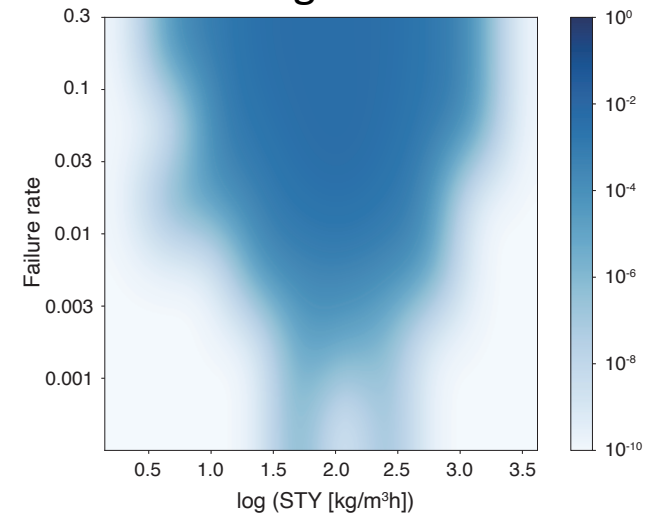
Complete profiling



Random sampling

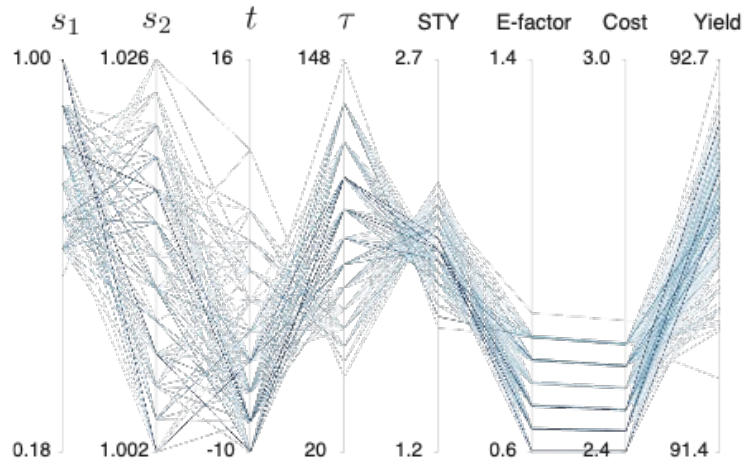


Entropic population  
annealing

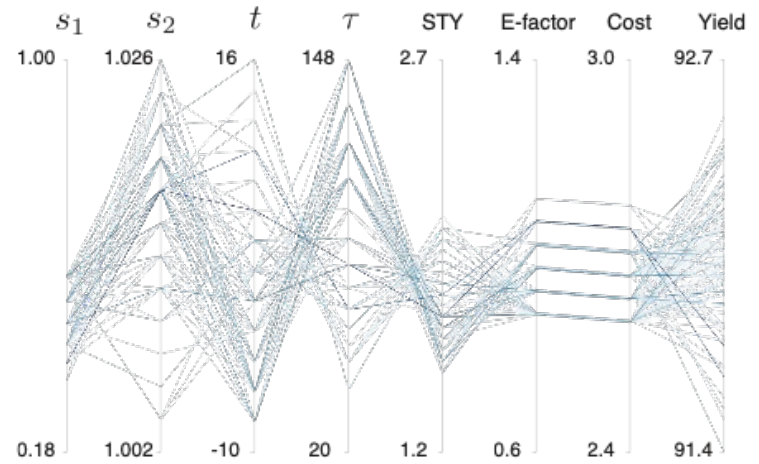


# Clusters of high-quality samples (failure rate < 0.01)

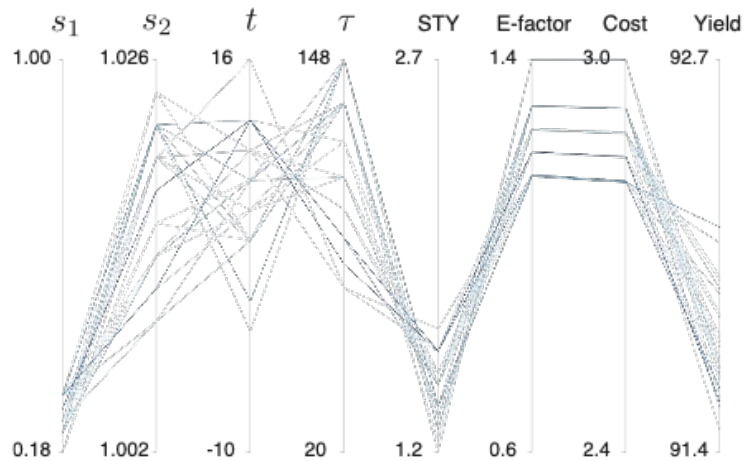
## Cluster 1



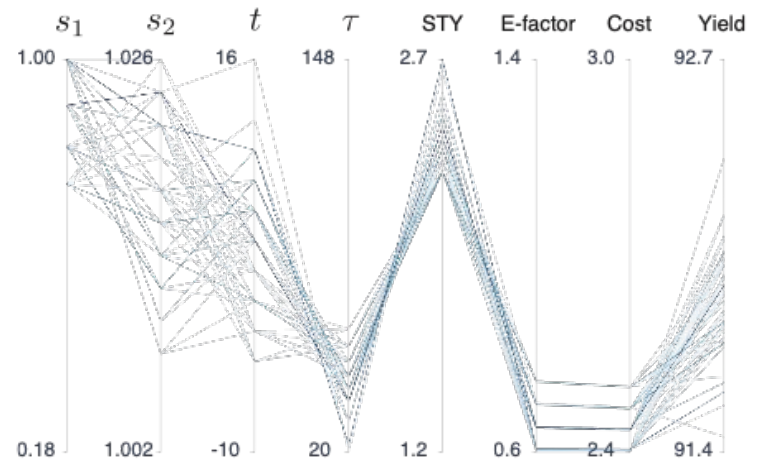
## Cluster 2



## Cluster 3



## Cluster 4



# Entropic population annealing as an optimization method

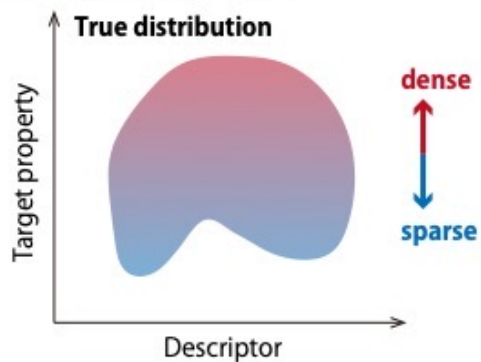
- EPA is very much like a genetic algorithm
  - MCMC = mutations
  - Resampling = selection
- Why don't we apply it to materials design?
  - Density of states comes as a bonus (!)
  - Increased **interpretability**
  - Need to reduce the number of black-box accesses

# Self-learning entropic population annealing (SLEPA)

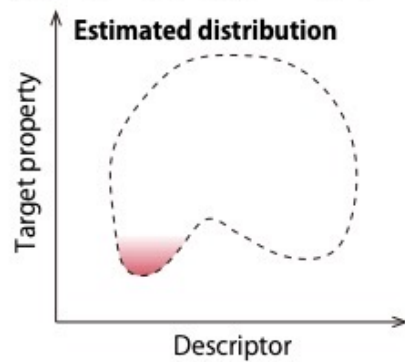
- Surrogate machine learning model of energy  $\tilde{e}(x)$
- MCMC is done with surrogate energy
- At temperature update, true energy is obtained, used for training.
  
- Before applying multiple histogram method, distribution is corrected via resampling



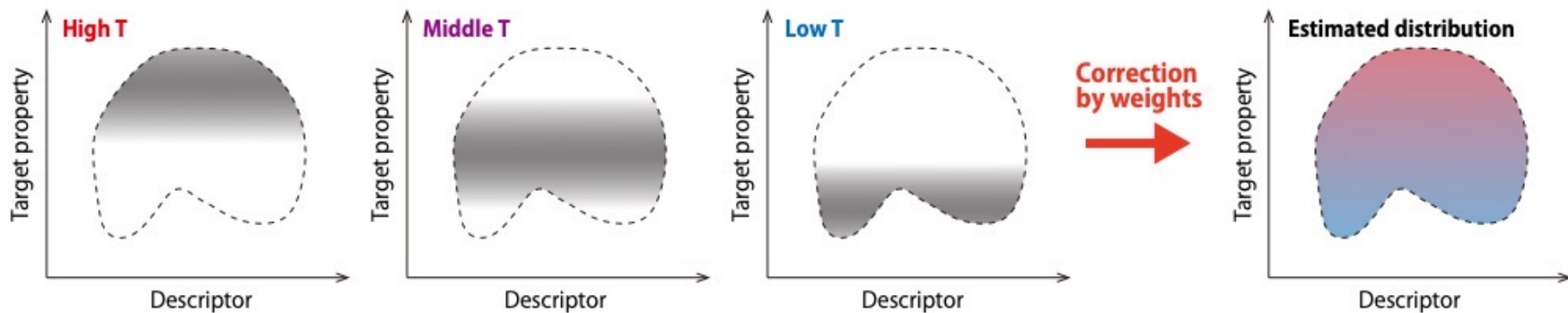
(a) Target distribution



(b) Black-box optimization



(c) Self-learning entropic population annealing

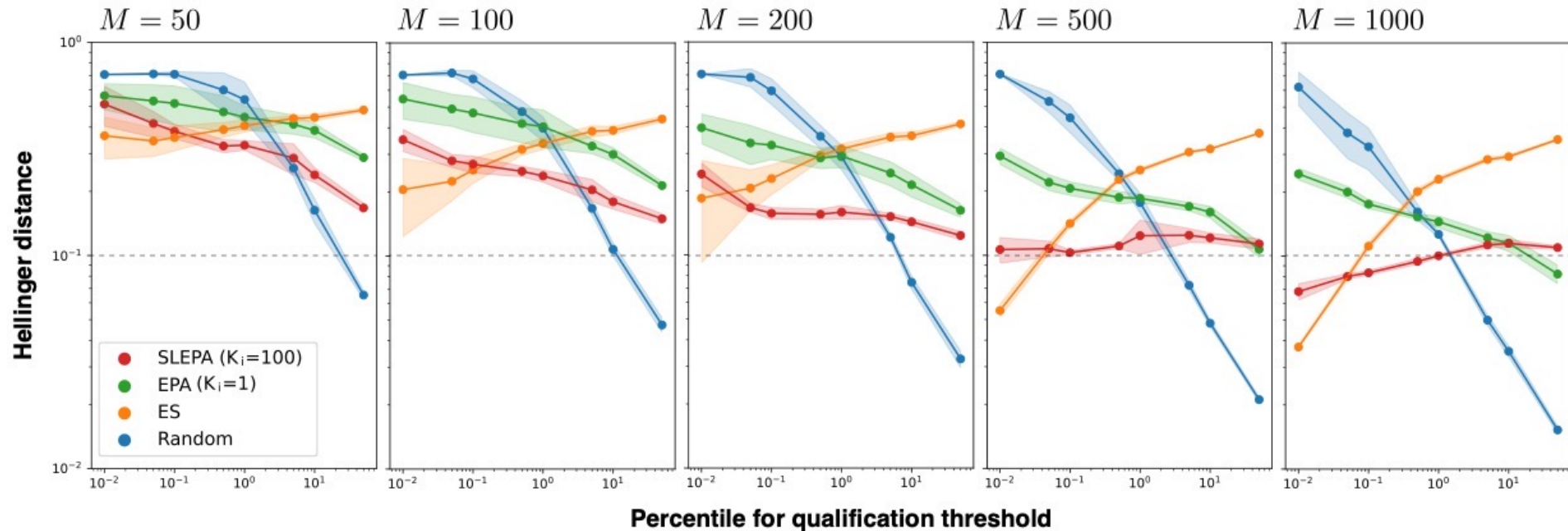


# Applying SLEPA to peptide design

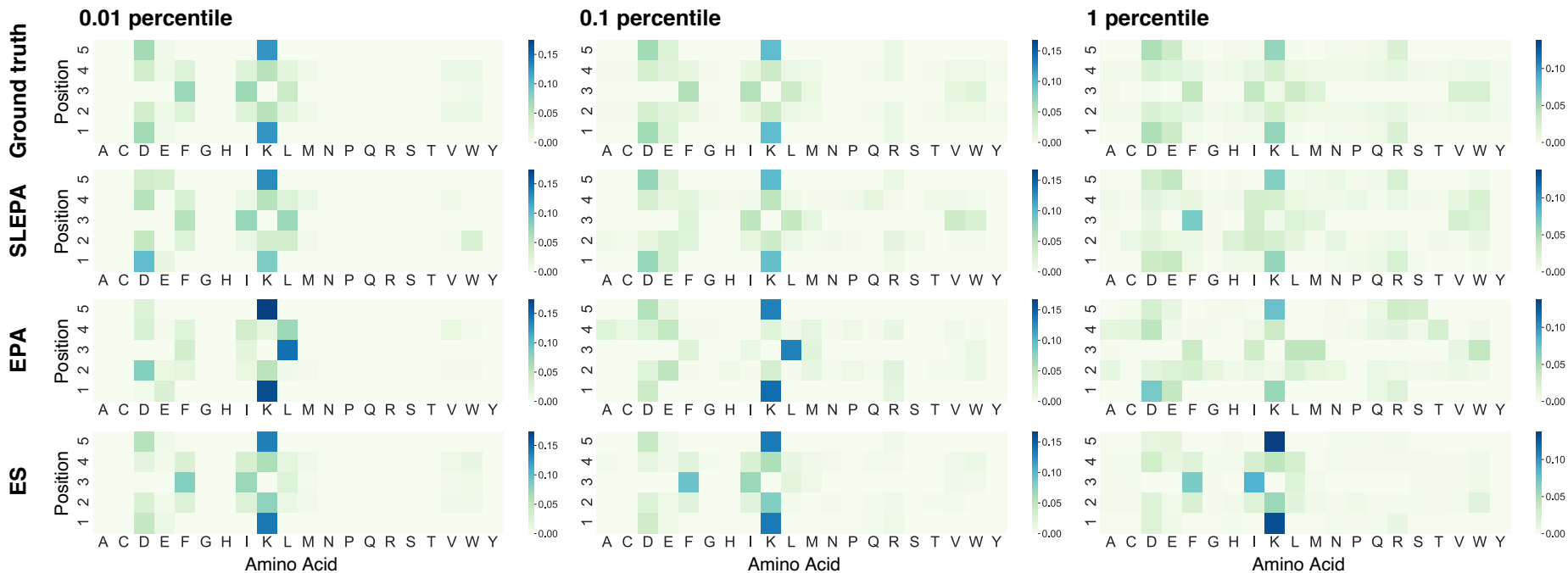
- Designing peptide of length 5
- **Target:** Hydrophobic moment (modIAMP)
- **MCMC:** one-character flip
- **Surrogate:** Gaussian process
- **DoS:** Target property and amino acid composition at a position
- **Comparison:** SLEPA, EPA, Evolution Strategy (ES) at the same number of observations

# Accuracy of DoS estimation

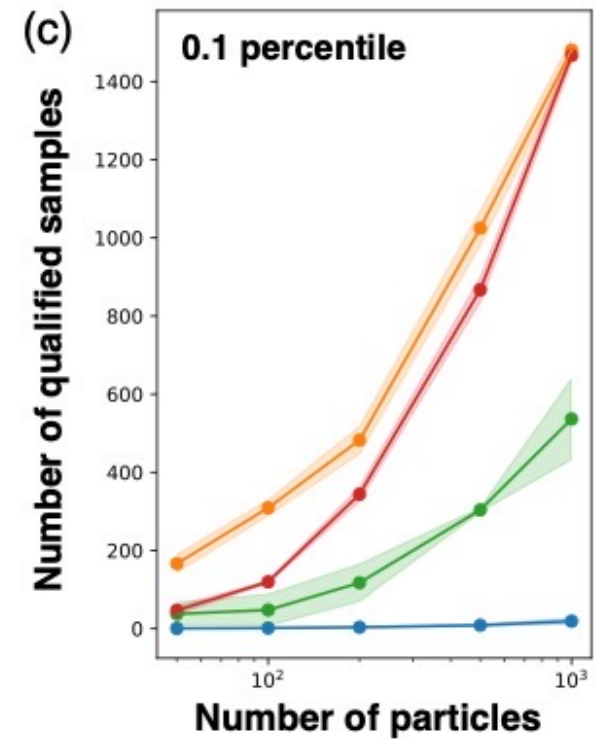
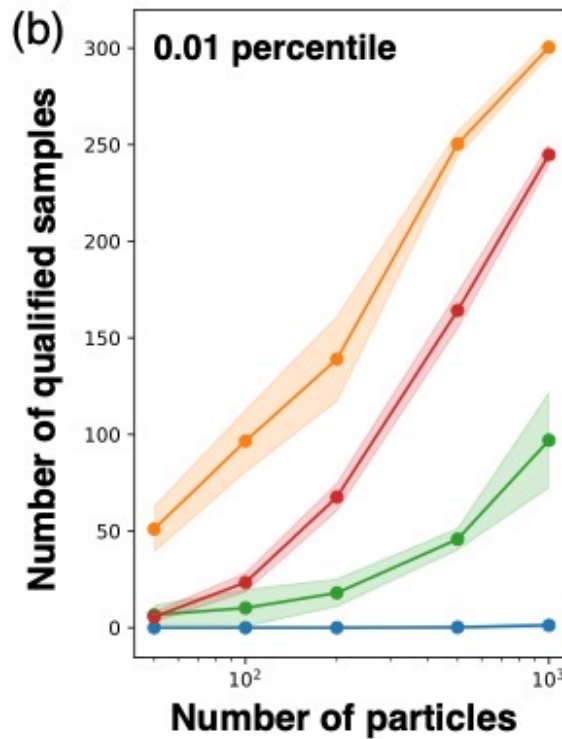
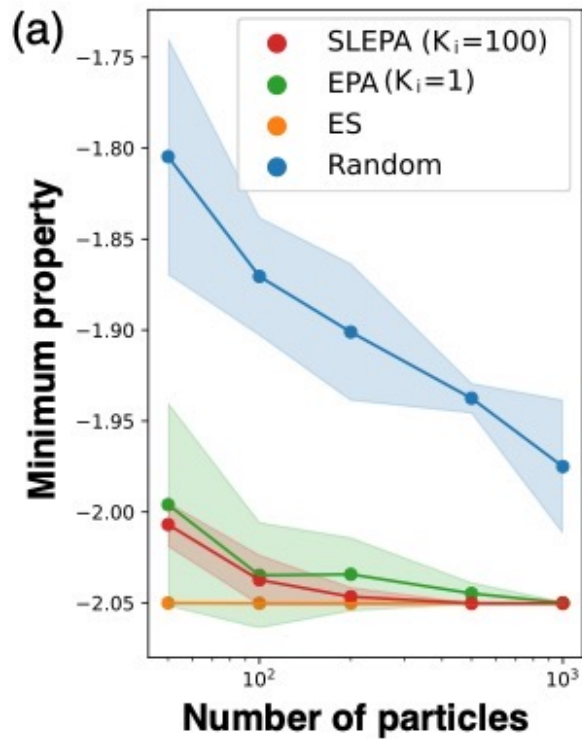
- SLEPA better at strict thresholds of target property



# DoS at different thresholds of target property



# Finding high-quality samples



# Summary

- In scientific studies, what to optimize is not obvious initially: BLOX or Entropic Sampling
- In quantum CAE, sampling is also possible
  - Quantum enhanced MCMC
  - Use quantum circuit to realize a proposal distribution
  - Reverse annealing can also be used

## Article


# Quantum-enhanced Markov chain Monte Carlo

<https://doi.org/10.1038/s41586-023-06095-4>

Received: 6 May 2022

Accepted: 18 April 2023

Published online: 12 July 2023

 Check for updates

David Layden<sup>1</sup>, Guglielmo Mazzola<sup>2,4</sup>, Ryan V. Mishmash<sup>1,5</sup>, Mario Motta<sup>1</sup>, Pawel Wocjan<sup>3</sup>, Jin-Sung Kim<sup>1,6</sup> & Sarah Sheldon<sup>1</sup>

Quantum computers promise to solve certain computational problems much faster than classical computers. However, current quantum processors are limited by their modest size and appreciable error rates. Recent efforts to demonstrate quantum speedups have therefore focused on problems that are both classically hard and naturally suited to current quantum hardware, such as sampling from complicated—although not explicitly useful—probability distributions<sup>1–3</sup>. Here we introduce and experimentally demonstrate a quantum algorithm that is similarly well suited to current hardware, but which samples from complicated distributions arising in several applications. The algorithm performs Markov chain Monte Carlo (MCMC),