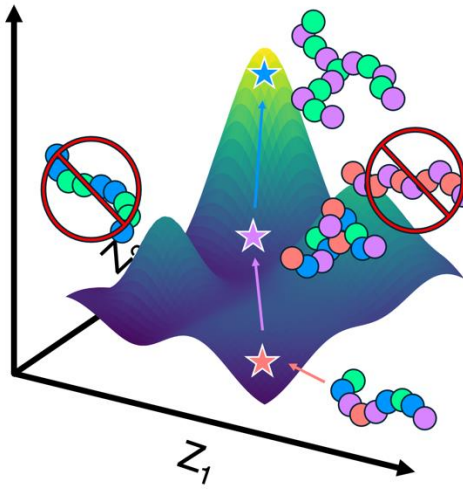
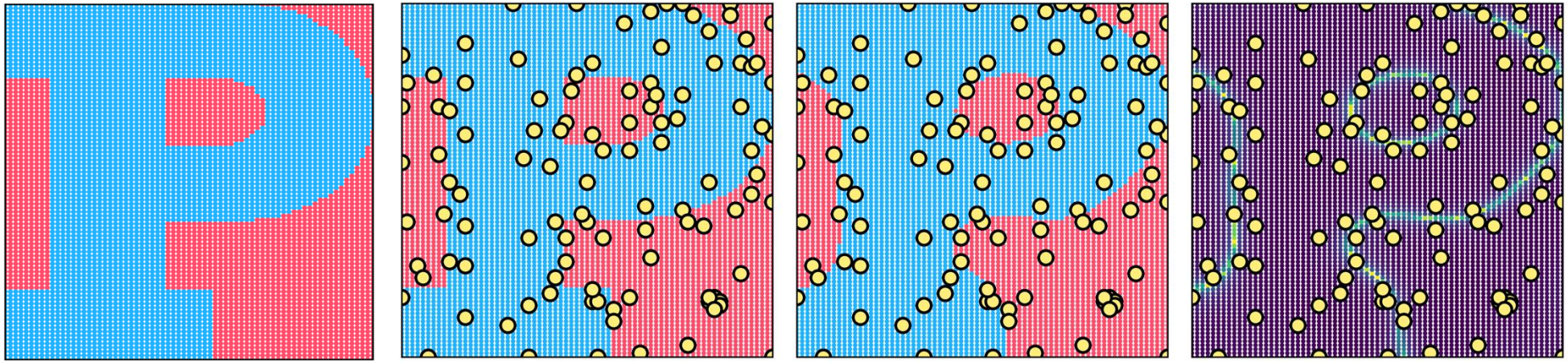
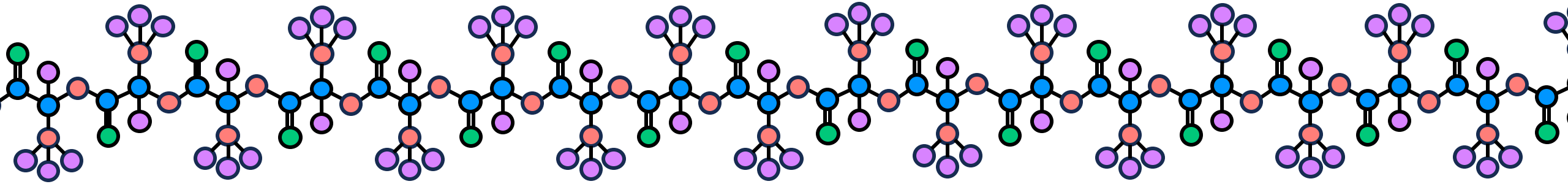
Prof. Mike Webb

webbgroup.princeton.edu, @xmwebb
cbe.princeton.edu/michael-webb

# Tutorial: Principles of Active Learning in Chemical and Materials Spaces



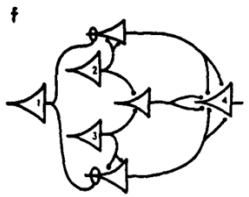





Artificial Intelligence for Biomaterials Design

Virtual Workshop for Materials Research Society
in partnership with Society for Biomaterials

# Machine learning is not new



**A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY**

WARREN S. MCCULLOCH and WALTER H. PITTS

Bulletin of Mathematical Biophysics, Vol. 5, 1943, p. 115-133

*McCulloch and Pitts publish paper establishing a mathematical network representation of nervous activity based on propositional logic (i.e., a **neural network**)*

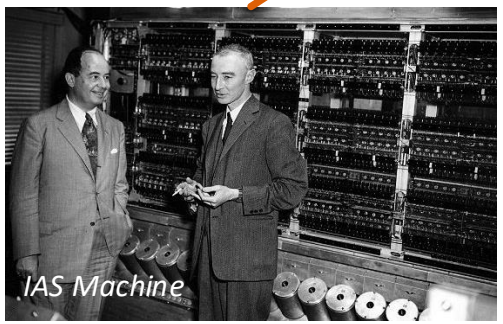
1943

Big Bang

1950

*Alan Turing proposes Turing Test*

1953




*Continuing advances in computing*

1957

*Frank Rosenblatt's Perceptron used for image recognition*


*Lighthill report*

Part I Artificial Intelligence

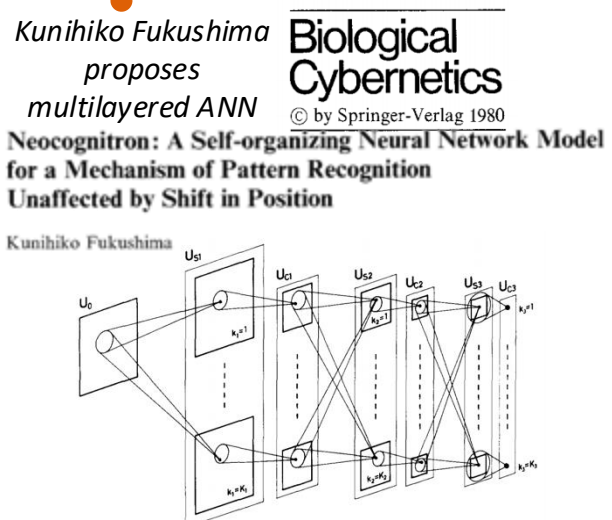
A general survey by Sir James Lighthill FRS Lucasian Professor of Applied Mathematics, Cambridge University. July 1972.

"Workers entered the field around 1950, and even around 1960, with high hopes that are very far from having been realised in 1972. In no part of the field have the discoveries made so far produced the major impact that was then promised."

1972

1979

*Kunihiko Fukushima proposes multilayered ANN*



Biological Cybernetics

© by Springer-Verlag 1980

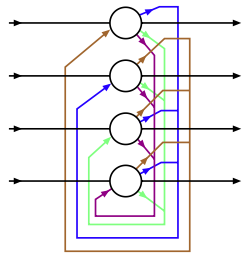
**Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position**

Kunihiko Fukushima

1982

*Hopfield network*

$$E = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j + \sum_i \theta_i s_i$$



*LSTMs*

*Random Forests + SVM*

*Torch introduced*

2000's

1998

*MNIST database*

***2021: AlphaFold2***
***2022: Dall-E 2, Stable Diffusion***
***ChatGPT,...***

2014

*GANs*

***Explosion of interest and applications by BigTech***

*seminar today*

2024

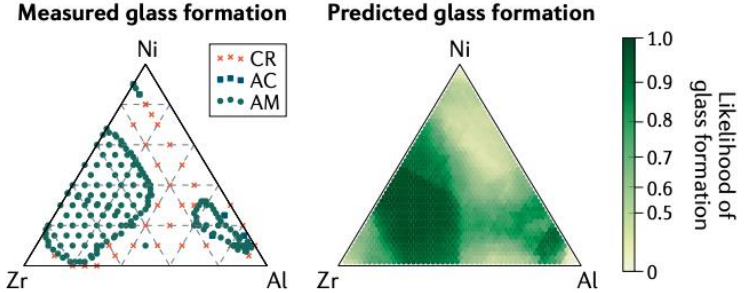
# We observe increasing utilization across materials science

nature reviews materials

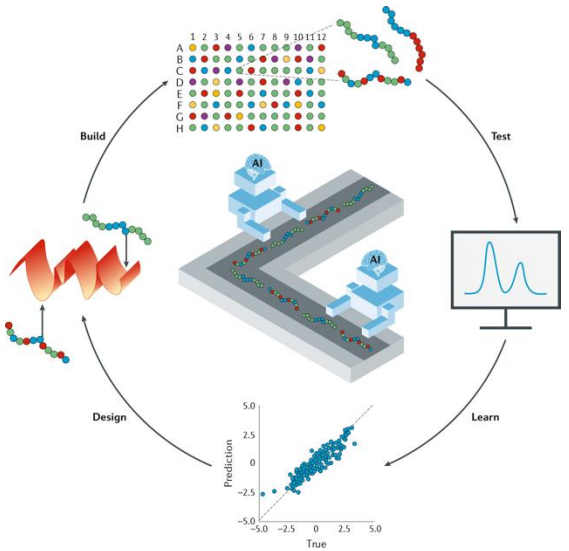
COLLECTION | 17 AUGUST 2021

## Machine learning in materials science

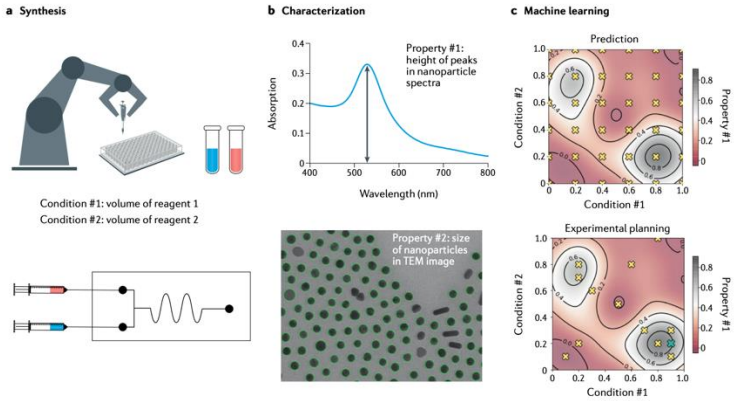
Machine learning is a powerful tool in materials research. Our collection of articles looks in depth at applications of machine learning in various areas of materials science.



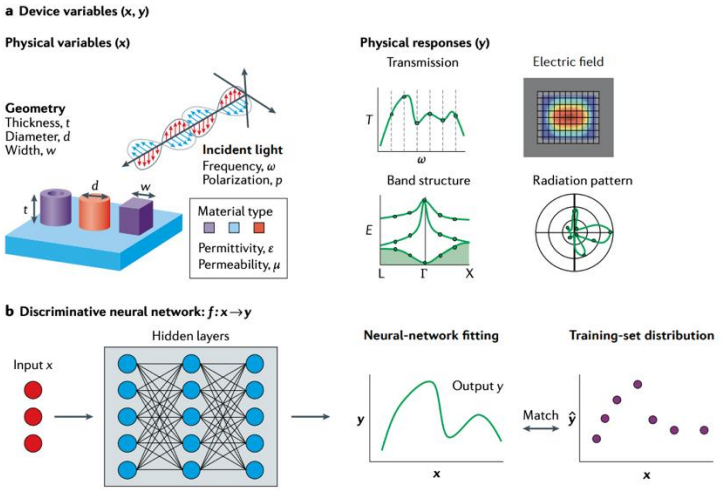

*"Machine learning for **alloys**"*
Hart, Mueller, Toher, Curtarolo



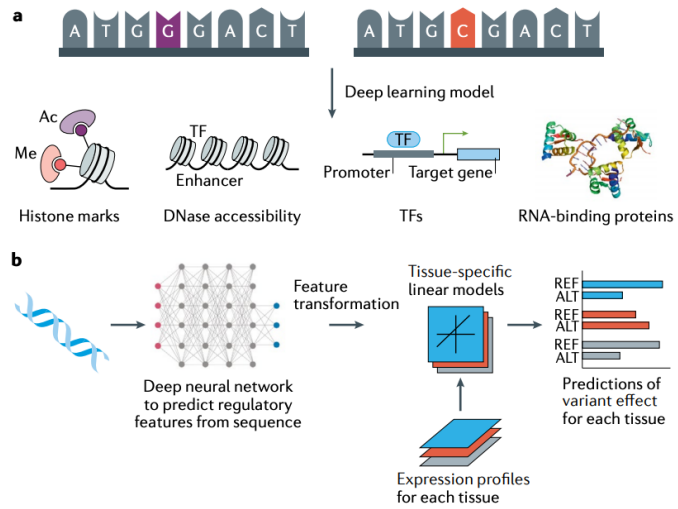

*"Machine Learning in combinatorial **polymer chemistry**"*
Gormley and Webb




*"**Nanoparticle synthesis** assisted by machine learning"*
Tao, Wu, Aldeghi, Aspuru-Guzik, abd Kumacheva

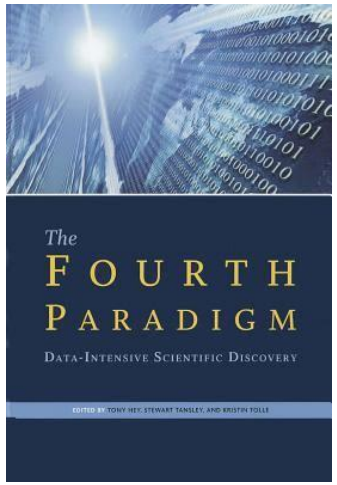



*"Deep neural networks for the evaluation and design of **photonic devices**"*
Jiang, Chen, and Fan

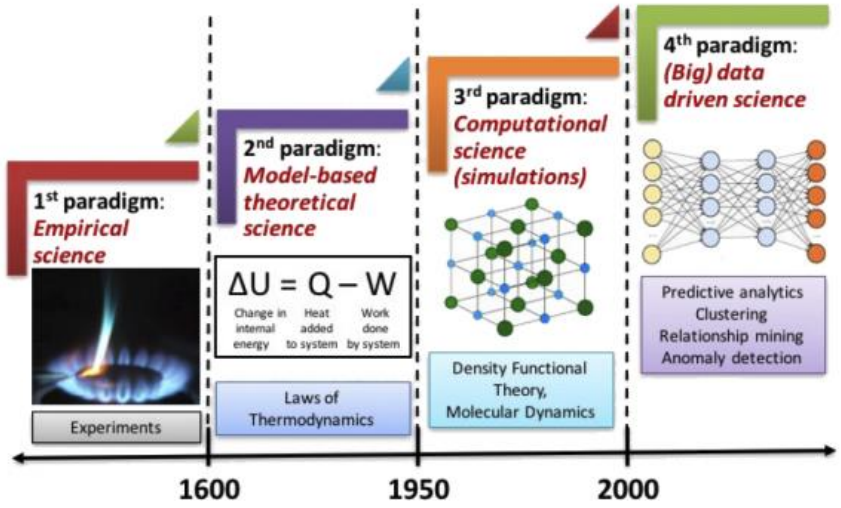



*"Machine learning methods to model **multicellular complexity and tissue specificity**"*
Sealfon, Wong, and Troyanskaya

# This reflects shifts in paradigms of scientific discovery




Microsoft Research (2004)




Agrawal & Choudary. *APL Mater.* **4** (2016)

AI4Science to empower the fifth paradigm of scientific discovery

Published July 7, 2022

By Christopher Bishop, Technical Fellow and Director, Microsoft Research AI4Science

Research Area

Artificial intelligence

> "Not only can AI learn from our past experiments, but, with each new iteration of designing and testing in the lab, the machine learning algorithms can identify new patterns and help guide the early drug discovery and development process. Hopefully in doing this we can augment our human scientists' expertise so they can design better molecules faster."

## Active learning (and Bayesian optimization) is a foundational strategy within the algorithmic scope of the 5th paradigm

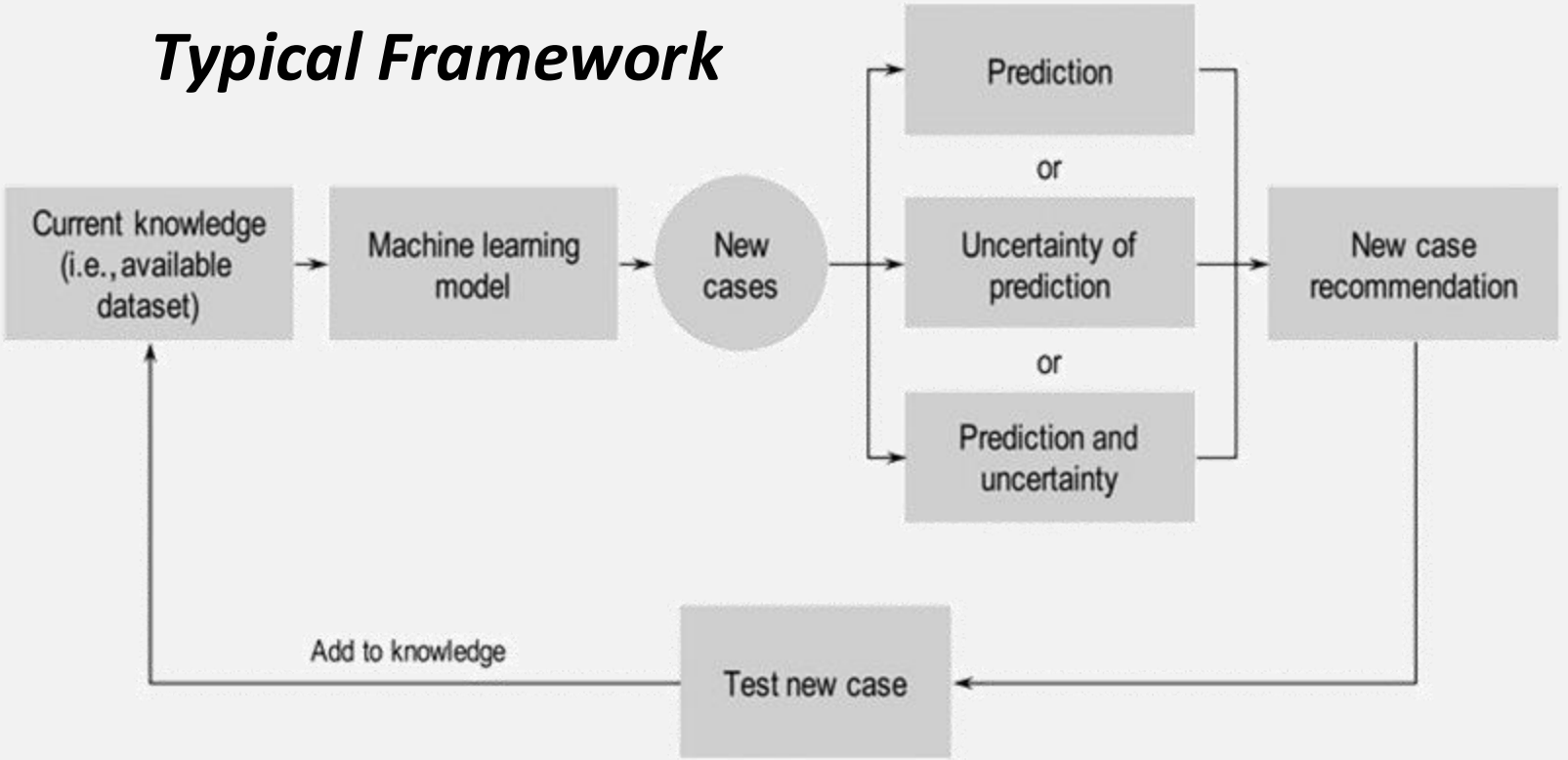
The **fifth paradigm of scientific discovery** is an emerging concept that extends beyond the existing four paradigms (empirical, theoretical, computational, and data-driven science). While it is still under discussion in scientific and technological circles, the fifth paradigm is often characterized by the **integration of machine learning, artificial intelligence (AI), automation, and autonomous discovery systems** to significantly accelerate the pace of scientific research. This paradigm envisions **self-driving laboratories** or **autonomous scientific workflows** where AI systems not only analyze data but also design, conduct, and optimize experiments in real-time.

The fifth paradigm is significant because it enables autonomous systems to design, conduct, and optimize experiments in real-time, accelerating the pace of scientific discovery and making research more efficient. Unlike the **fourth paradigm**, which focuses on using data-driven approaches and machine learning to analyze large datasets, the fifth paradigm goes further by integrating AI with robotics and automation, allowing AI to actively guide and execute the entire discovery process, from hypothesis generation to experiment execution. This shift from **data analysis** to **autonomous discovery** marks a critical leap in scientific capabilities.

# What is active learning?

***Active learning (AL)*** refers to a *mode/paradigm* of machine learning in which data is labeled (in ***iterative*** fashion) by ***interactive query***

***Typical Framework***




Kim et al. *MRS Communications* (2019)

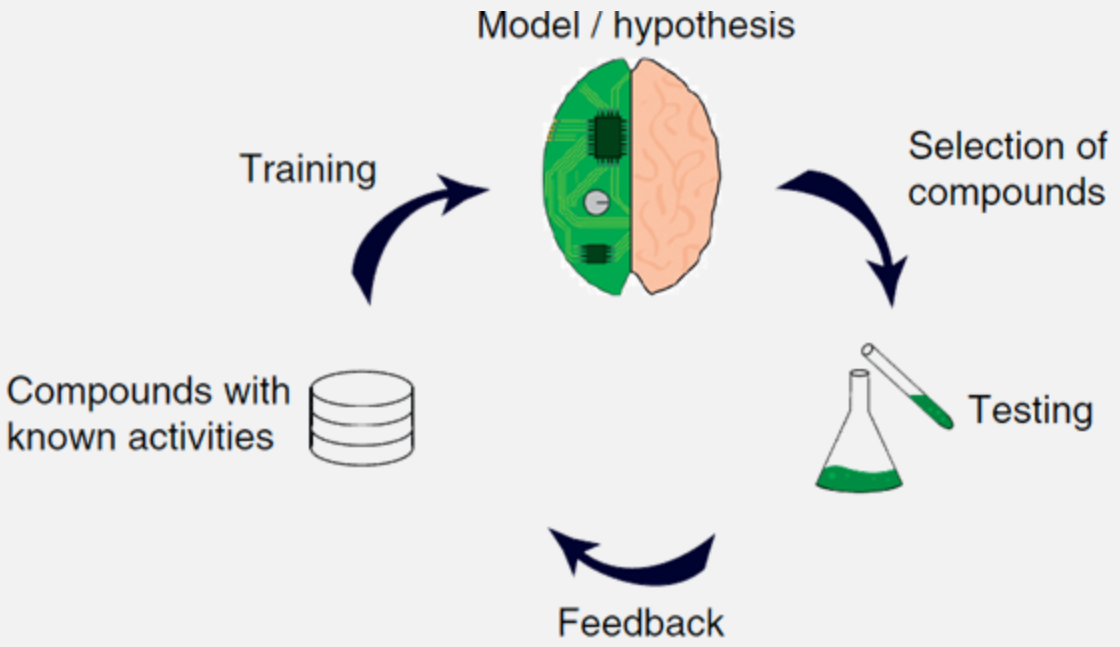
## Key Points:

- **AL** is essentially a decision-making formalism that dictates data acquisition; it relates to DOE
- **AL** is commonly considered in situations where labeling data is "expensive" (with respect to some important resource consideration)
- The objective is to be efficient! (create better models with less data, find optimal designs with limited resources)

# What is active learning?

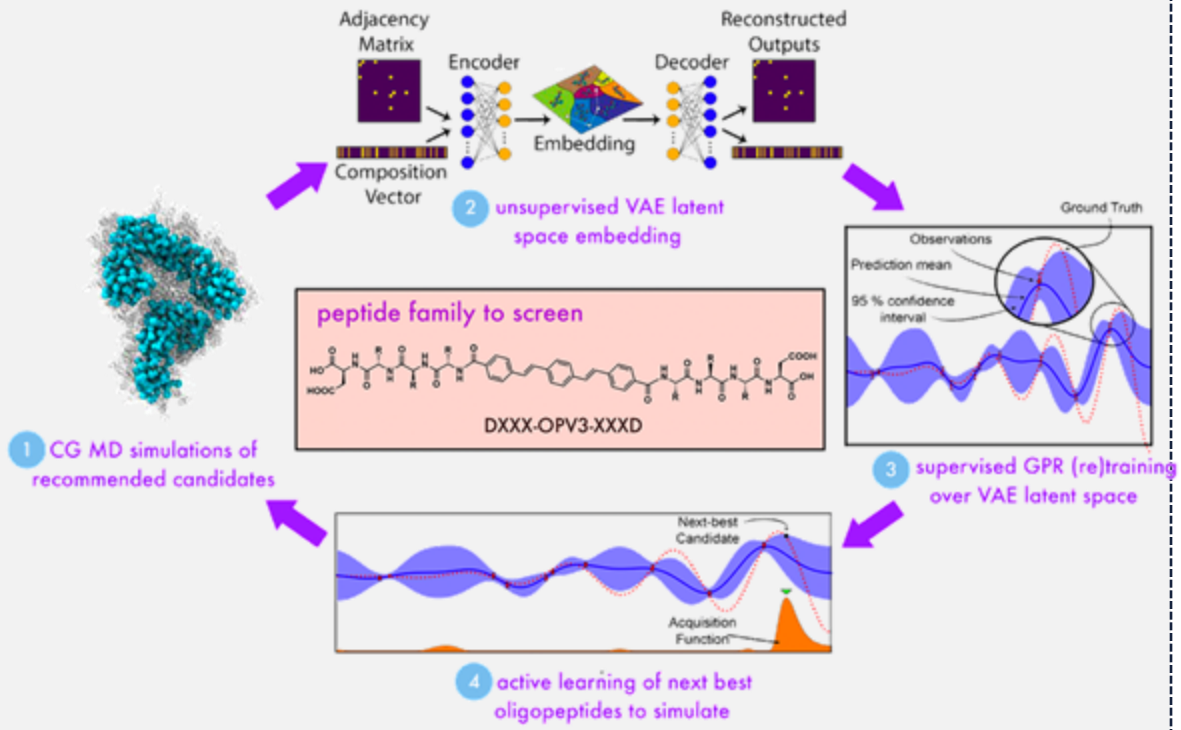
***Active learning (AL)*** refers to a *mode/paradigm* of machine learning in which data is labeled (in ***iterative*** fashion) by ***interactive query***

*"Active-learning strategies in computer-assisted drug discovery"*




Reker and Schneider. *Drug Discovery Today* **2015**

*"Discovery of Self-Assembling pi-Conjugated Peptides…"*

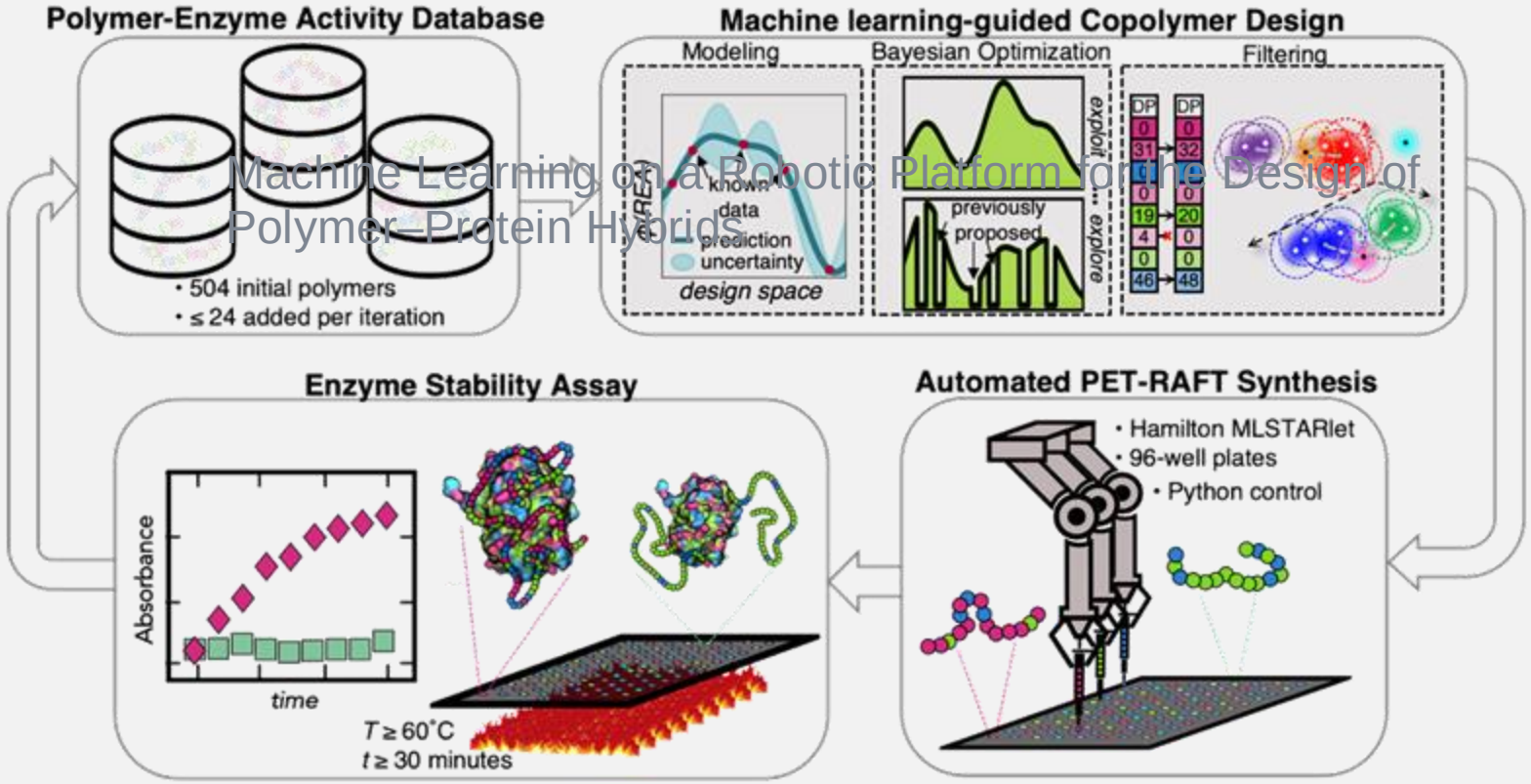



Shmilovich et al. *JPCB* **2020**

# What is active learning?

***Active learning (AL)*** refers to a *mode/paradigm* of machine learning in which data is labeled (in ***iterative*** fashion) by ***interactive query***

*"Machine Learning on a Robotic Platform for the Design of Polymer–Protein Hybrids"*






Tamasi, Patel, Borca, Kosuri, Mugnier, Upadhya, Murthy, Webb*, Gormley*. *Advanced Materials.* **2022**

# An active learning strategy is defined by few key elements

***Strategy = seed + model + acquisition function → performance***

Texts in Statistical Science



**Algorithm 6.1** Sequential Design/Active Learning

**Assume** a flexible surrogate, e.g., a GP model, but with potentially unknown hyper-parameterization.

**Require** a function $f(\cdot)$ providing outputs $y \sim f(x)$ for inputs $x$, either deterministic or observed with noise; a choice of initial design size $n_0$ and final size $N$; and **criterion** $J(x)$ to search for design augmentation.

Then

1. Run a small seed, or bootstrapping experiment.
   a. Create an initial seed design $X_{n_0}$ with $n_0$ runs. Typically $X_{n_0}$ is a model-free choice, e.g., derived from a static LHS or maximin design.
   b. Evaluate $y_i \sim f(x_i)$ under each $x_i^\top$ in the $i^{\text{th}}$ row of $X_{n_0}$, for $i = 1, \ldots, n_0$, obtaining $D_{n_0} = (X_{n_0}, Y_{n_0})$.
   c. Set $n \leftarrow n_0$, indexing iterations of sequential design.
2. Fit the surrogate (and hyperparameters) using $D_n$, e.g., via MLE.
3. Solve **criterion** $J(x)$ based on the fitted model from Step 2, resulting in a choice of $x_{n+1} \mid D_n$: $x_{n+1} = \text{argmax}_{x \in \mathcal{X}} J(x) \mid D_n$.
4. Observe the response at the chosen location by running a new simulation, $y_{n+1} \sim f(x_{n+1})$.
5. Update $D_{n+1} = D_n \cup (x_{n+1}, y_{n+1})$; set $n \leftarrow n+1$ and **repeat** from Step 2 unless $n = N$.

**Return** the chosen design and function evaluations $D_N$, along with surrogate fit (i.e., after a final application of Step 2).

*How will you choose the next measurement?*

This is formalized by defining an **acquisition function**

*How will you prepare the initial set of data for building your first model?*

Here we define a **seed design**.

- *random search*
- *systematic search*
- *existing dataset*
- *Latin hypercube*
- *Maximin*

*How will you model your data?*

This is just **model-building**, relying on other best-practices of ML

How will you find the next point?

This is just **optimization**

(should not have significant impacts)

# Visual progression of active learning strategy

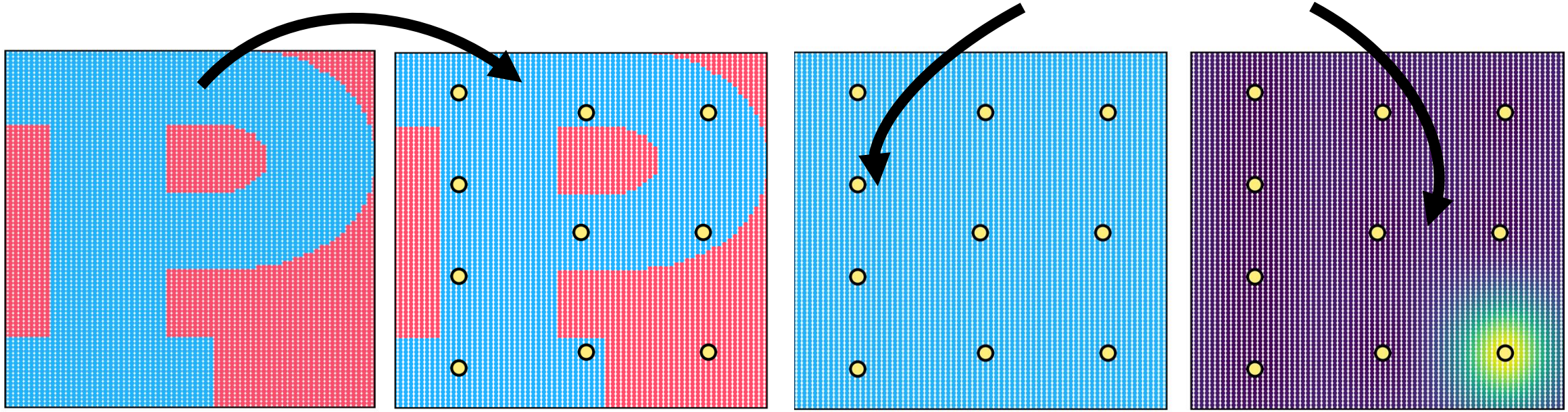
**Task:** *learn to classify points as part or not part of 'P'*

**Strategy:** (seed); (acquisition function); (model)

*medoids; maximum uncertainty; Gaussian Process*

The ***seed algorithm*** chooses a *small* set of points…

The ***model*** computes *predictions* and *uncertainties* on the domain...



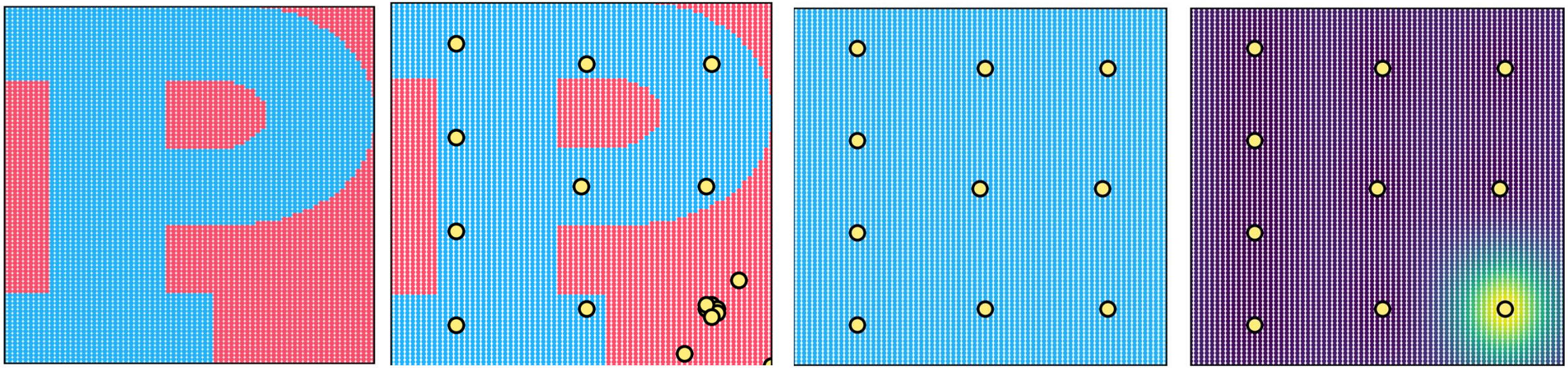
# Visual progression of active learning strategy

**Task:** *learn to classify points as part or not part of 'P'*

**Strategy:** (seed); (acquisition function); (model)

*medoids; maximum uncertainty; Gaussian Process*

uncertain points are measured…



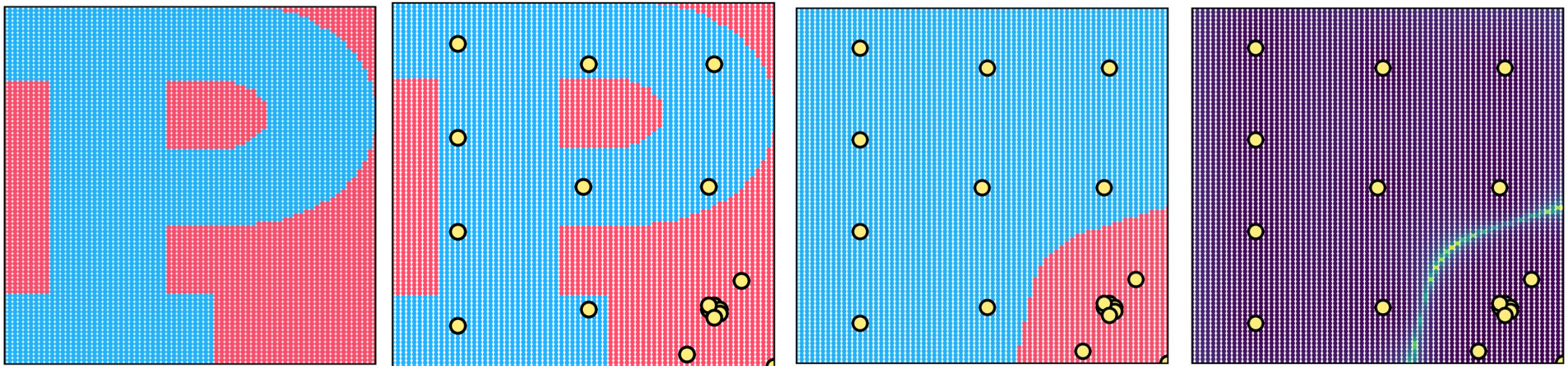
# Visual progression of active learning strategy

**Task:** *learn to classify points as part or not part of 'P'*

**Strategy:** (seed); (acquisition function); (model)

*medoids; maximum uncertainty; Gaussian Process*

Resulting in a new set of predictions and uncertainties from the ***model***…



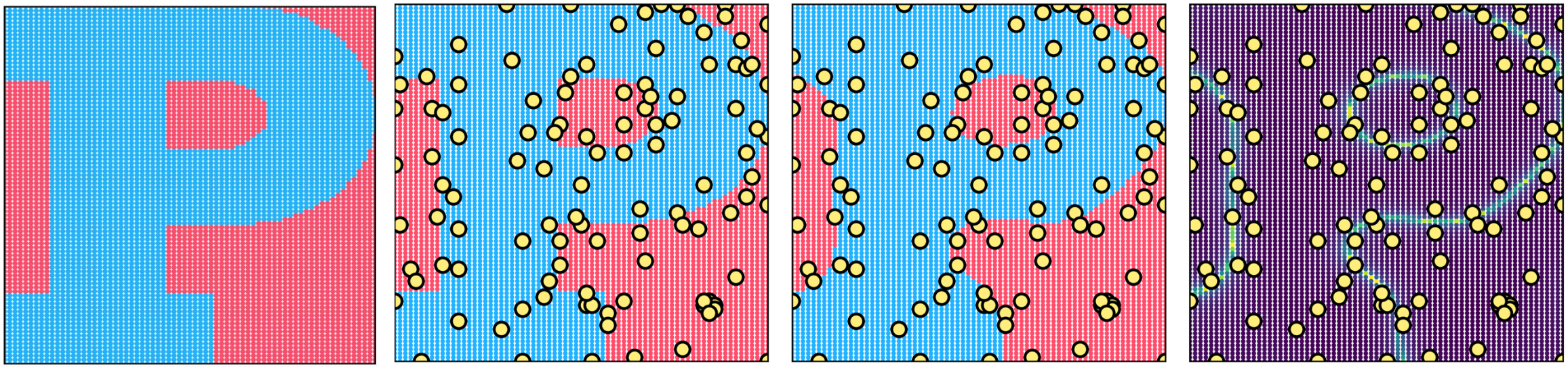
# Visual progression of active learning strategy

**Task:** *learn to classify points as part or not part of 'P'*

**Strategy:** (seed); (acquisition function); (model)

*medoids; maximum uncertainty; Gaussian Process*

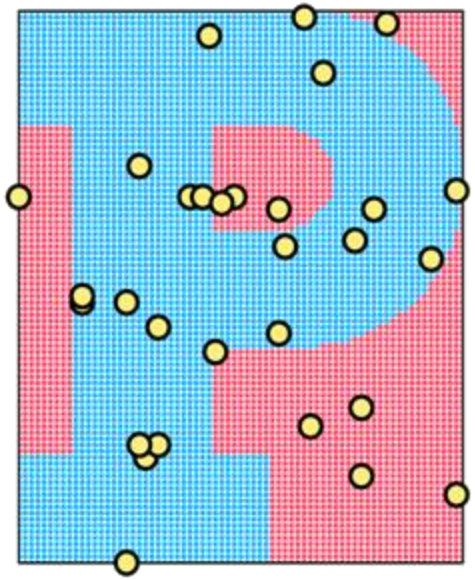
This process continues until resource budget is exhausted



# Algorithms for seed designs

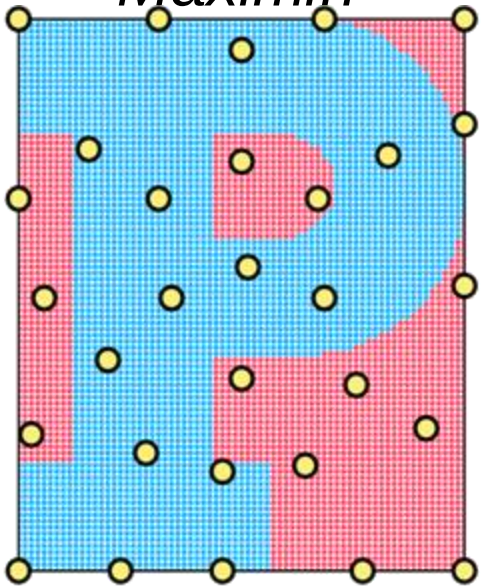
## The more iterations of data selection to follow, the less consequential the choice of initial seed is expected to be

Random

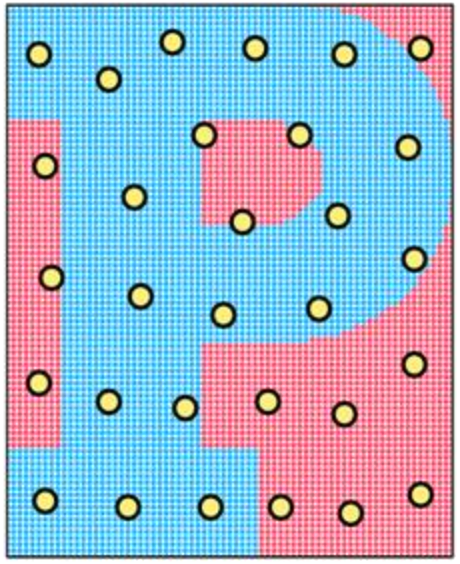


**Space-filling methods** *are geared towards generating dataset diversity, which is hypothesized to yield better initial models and therefore promote AL efficiency*

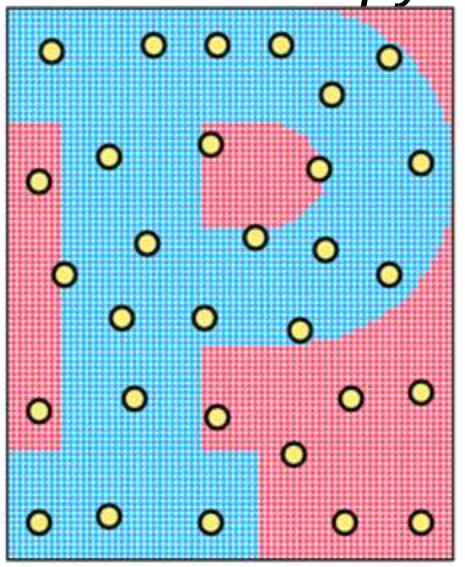
Maximin



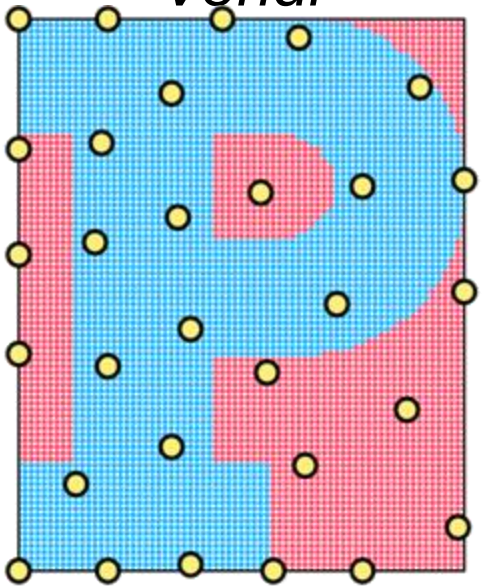
Medoids



Max Entropy



Vendi



Gallagher, Webb. "Data Efficiency of Classification Strategies for Chemical and Materials Design" *ChemRxiv.* **2024.**

# Common acquisition functions

The acquisition function is a more striking distinction in strategy

**Improvement-based acquisition functions**

$$f(\boldsymbol{x}^+); \boldsymbol{x}^+ = \arg\max_{\boldsymbol{x}_i \in \mathcal{D}} f(\boldsymbol{x}_i)$$ ***incumbent***

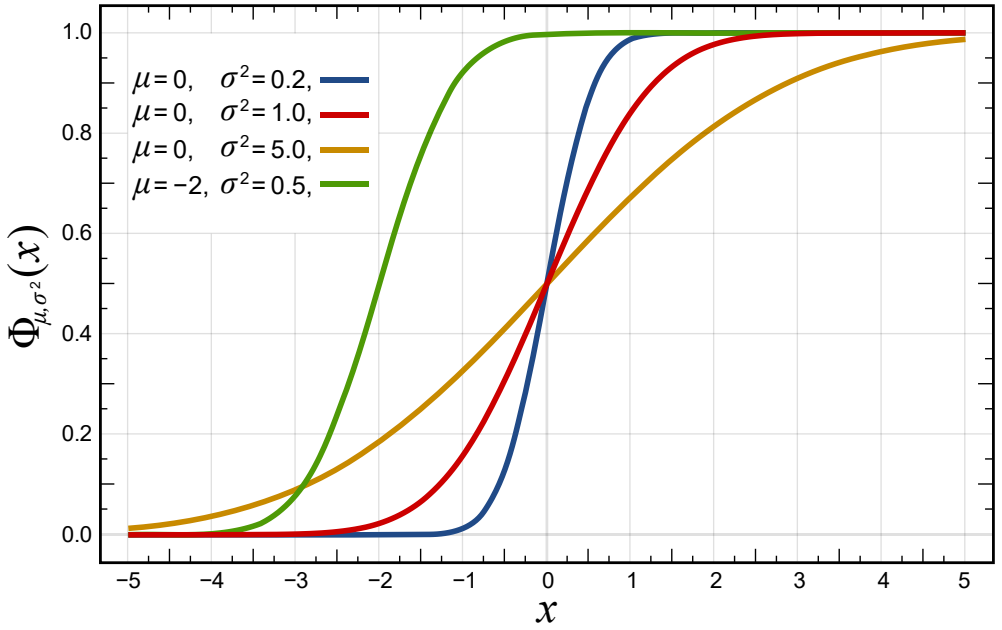
***probability of improvement***

*"maximize the probability of improvement over the* ***incumbent****"*

$$\mathrm{PI}(\boldsymbol{x}) = \Phi\left(\frac{\mu(\boldsymbol{x}) - f(\boldsymbol{x}^+) - \xi}{\sigma(\boldsymbol{x})}\right)$$

***standard normal cumulative distribution***

$$\Phi(z) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$




$\xi \geq 0$ *generally and often not used*

# Common acquisition functions

The acquisition function is a more striking distinction in strategy

**Improvement-based acquisition functions**

$$f(\boldsymbol{x}^+); \boldsymbol{x}^+ = \arg\max_{\boldsymbol{x}_i \in \mathcal{D}} f(\boldsymbol{x}_i)$$

***incumbent***

***expected improvement***

*"maximize the probability of improvement but also consider the magnitude of that improvement"*

***standard normal probability density distribution***

$$\mathrm{EI}(\boldsymbol{x}) = \sigma(\boldsymbol{x}) \left[ Z\Phi(Z) + \phi(Z) \right]$$

$$Z = \frac{\mu(\boldsymbol{x}) - f(\boldsymbol{x}^+) - \xi}{\sigma(\boldsymbol{x})}$$

$\xi$ *roughly controls the exploration-exploitation tradeoff*

- →0, more towards exploitation
- →large, more towards exploration

used by Jones *et al.* [2001]. Lizotte's experiments suggest that setting $\xi = 0.01$ (scaled by the signal variance if necessary) works well in almost all cases, and interestingly, setting a cooling schedule for $\xi$ to encourage exploration early and exploitation later does *not* work well empirically, contrary to intuition (though Lizotte did find that a cooling schedule for $\xi$ might slightly improve performance on short runs ($t < 30$) of PI optimization).

# Common acquisition functions

The acquisition function is a more striking distinction in strategy

**Confidence-bound acquisition functions**

$$f(\boldsymbol{x}^+); \boldsymbol{x}^+ = \underset{\boldsymbol{x}_i \in \mathcal{D}}{\arg\max}\, f(\boldsymbol{x}_i)$$

***incumbent***

***lower confidence bound***
(typically chosen when minimizing)

$$\mathrm{LCB}(\boldsymbol{x}) = \mu(\boldsymbol{x}) - \xi\sigma(\boldsymbol{x})$$

***upper confidence bound***
(typically chosen when maximizing)

$$\mathrm{UCB}(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \xi\sigma(\boldsymbol{x})$$

*variation for fixed number of iterations*

$$r(\boldsymbol{x}) = f(\boldsymbol{x}^*) - f(\boldsymbol{x})$$

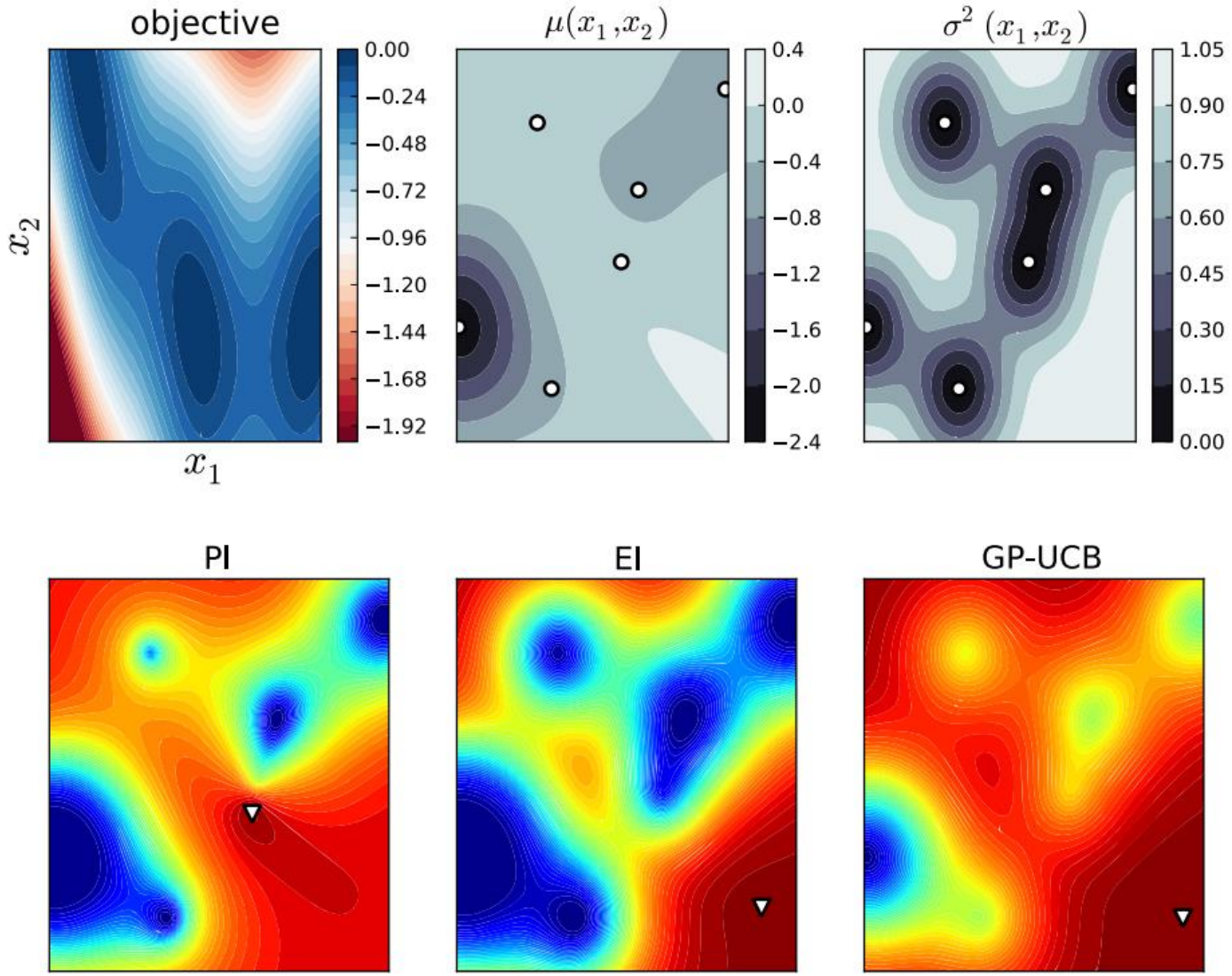
$$\min \sum_t^T r(\boldsymbol{x}_t)$$

related to ***multi-armed bandit*** problem

$$\text{GP-UCB}(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \sqrt{\nu\tau_t}\sigma(\boldsymbol{x})$$

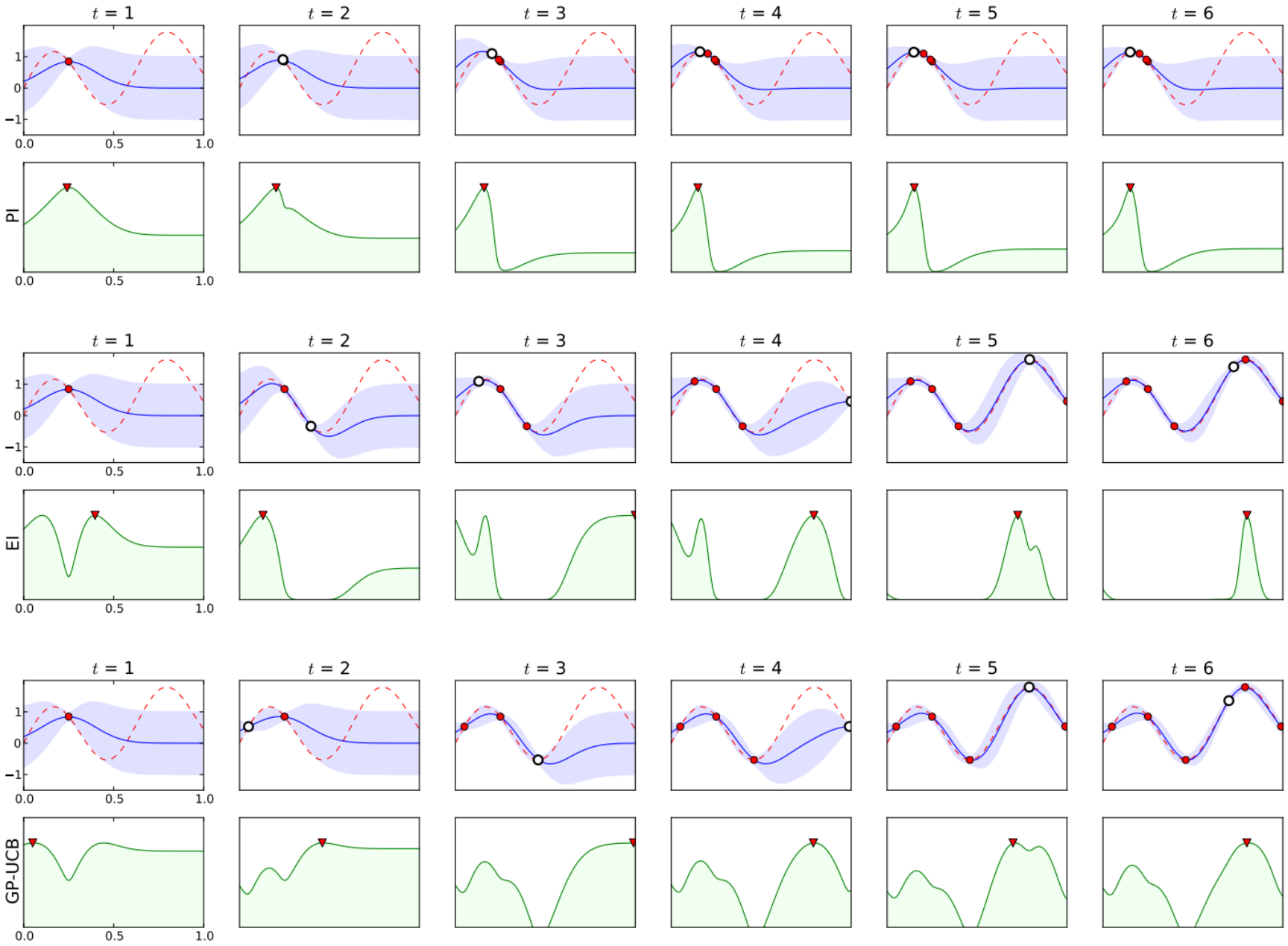
select schedule on constants that approximate limiting case of ***no regret*** (guaranteed solution/lower bound on convergence rate)

# Visual Comparison of Acquisition Functions




Brochu et al.,
ArXiv:1012.2599

# Visual Comparison of Acquisition Functions




Brochu et al.,
ArXiv:1012.2599

# Colab Notebook Demonstration

https://shorturl.at/B9SST

# Acknowledgments and Resources

***Quinn Gallagher***

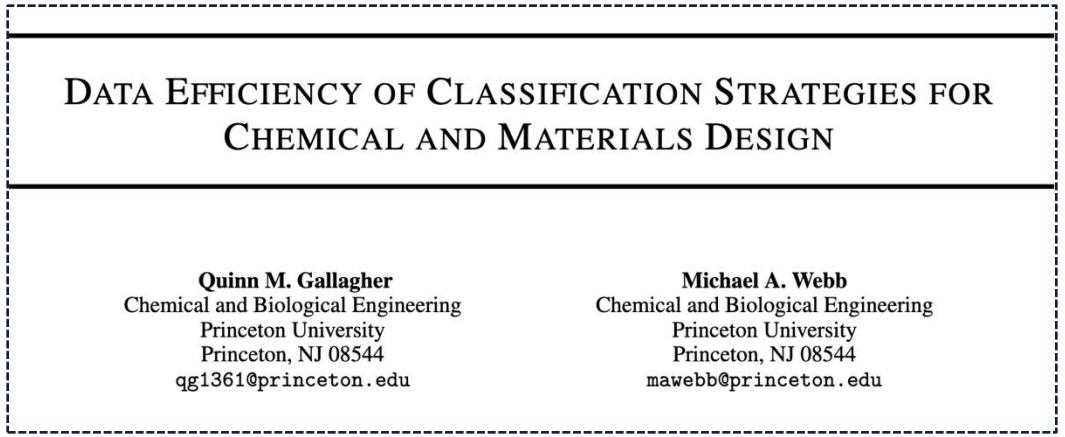
***Prof. Adam Gormley***


- GRFP (DGE-2039656)
- DMREF (2118861)


**https://github.com/webbtheosim/CBE512-MLinChmSciEng**

**Basic takeaways…**

- Active learning is a representative strategy of the 5th paradigm of scientific discovery (ML guides data acquisition)
- The objective active learning is usually sample efficiency
- Active learning strategies are defined by choice of



DATA EFFICIENCY OF CLASSIFICATION STRATEGIES FOR CHEMICAL AND MATERIALS DESIGN


**Quinn M. Gallagher**
Chemical and Biological Engineering
Princeton University
Princeton, NJ 08544
qg1361@princeton.edu

**Michael A. Webb**
Chemical and Biological Engineering
Princeton University
Princeton, NJ 08544
mawebb@princeton.edu

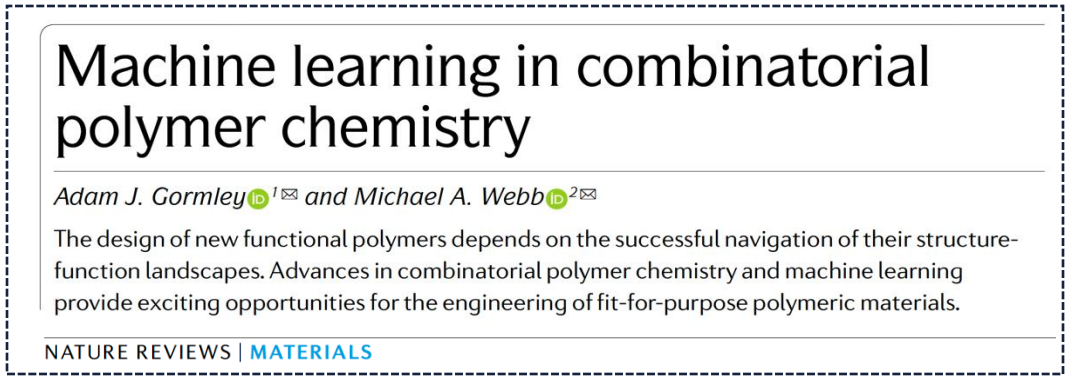

ACS APPLIED BIO MATERIALS

www.acsabm.org Spotlight on Applications

**Data-Driven Design of Polymer-Based Biomaterials: High-throughput Simulation, Experimentation, and Machine Learning**

Roshan A. Patel and Michael A. Webb*

Cite This: https://doi.org/10.1021/acsabm.2c00962 Read Online



Machine learning in combinatorial polymer chemistry

Adam J. Gormley[1]✉ and Michael A. Webb[2]✉

The design of new functional polymers depends on the successful navigation of their structure–function landscapes. Advances in combinatorial polymer chemistry and machine learning provide exciting opportunities for the engineering of fit-for-purpose polymeric materials.

NATURE REVIEWS | MATERIALS