Imperial College London

Bayesian optimisation for chemical applications

explore **chemical representation, domain & sampling strategies** for BO

- **How do we formulate BO problems for chemistry?**
- **How can we integrate chemical knowledge into the BO formulation?**
- **How can we extend the conventional BO algorithm to more complex problems?**

Austin Mroz

a.mroz@imperial.ac.uk

February 2026

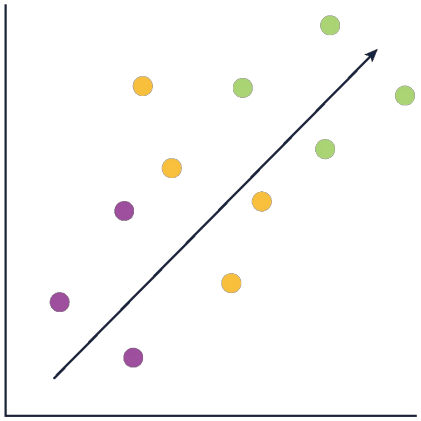
Imperial-X

Schmidt Sciences

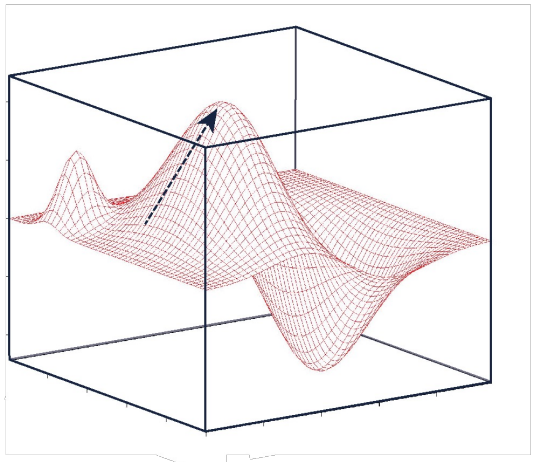
AI SuperConnector

Google Cloud

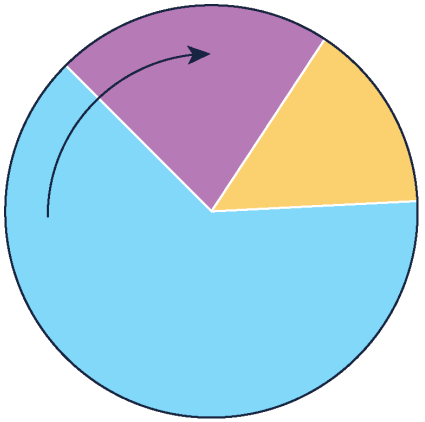
decisions lie at the core of experimentation



property

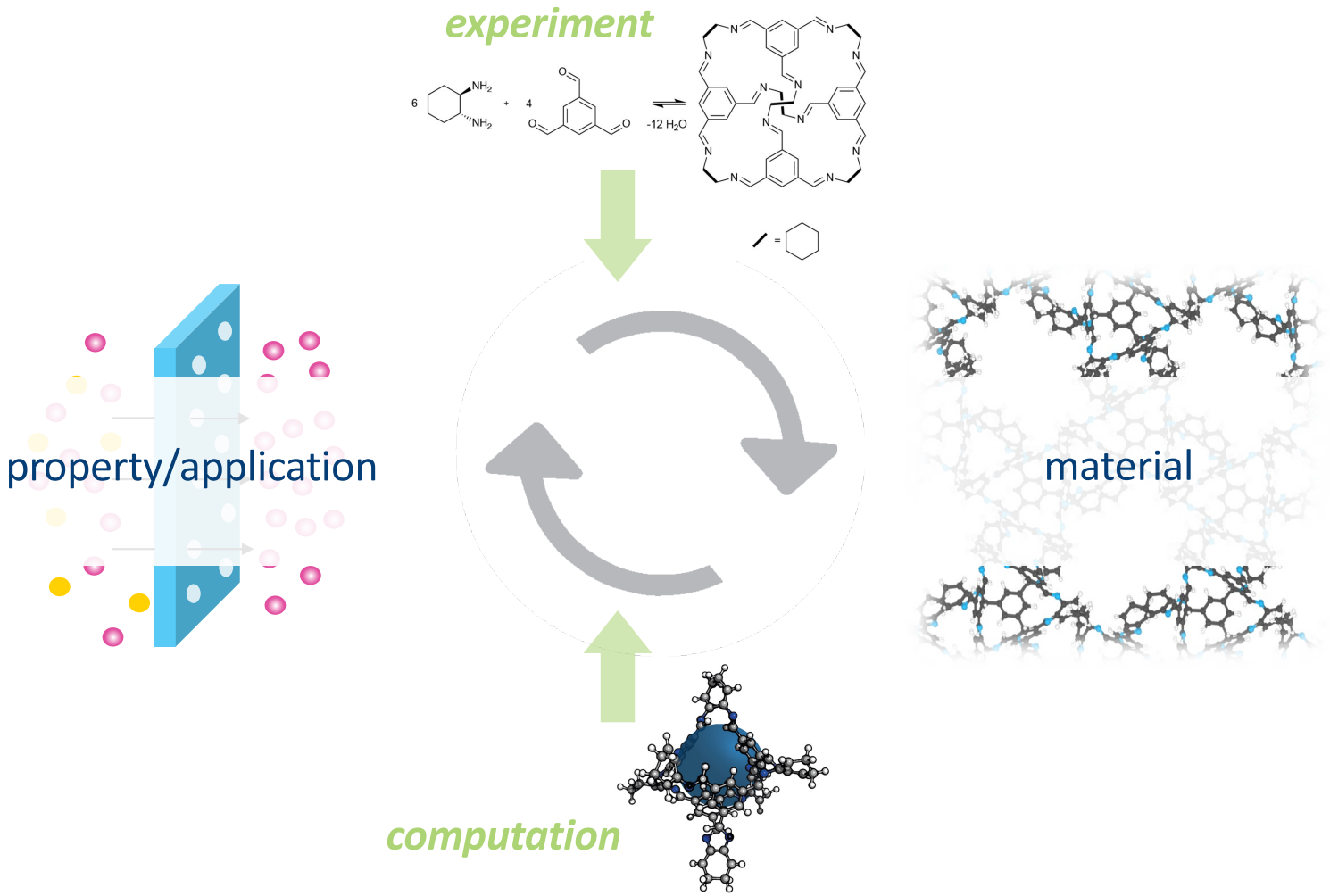


yield



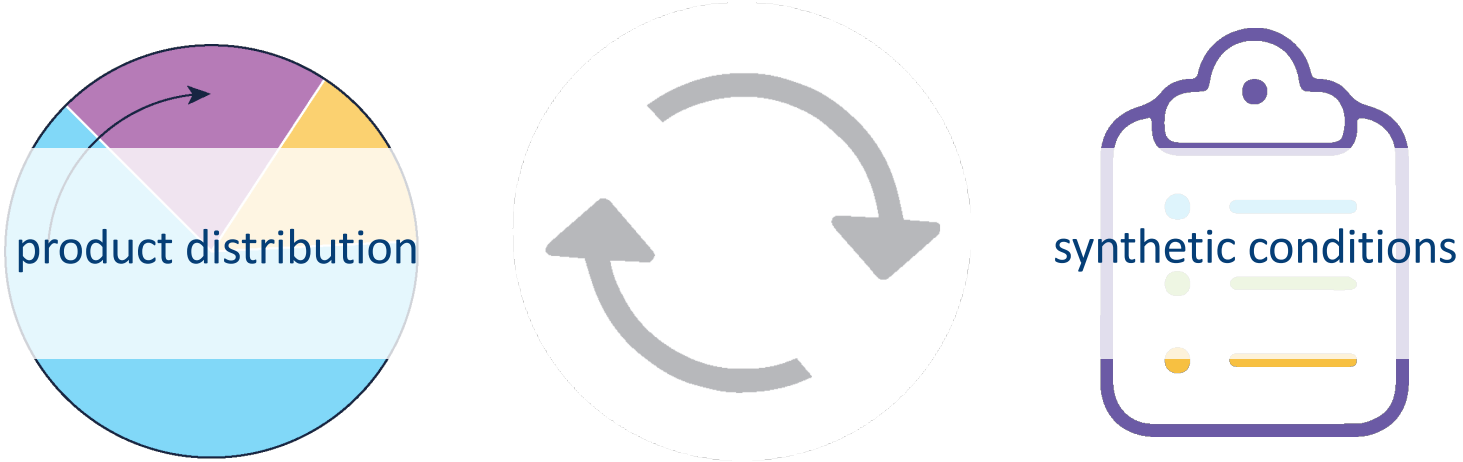
product distribution

decisions lie at the core of experimentation



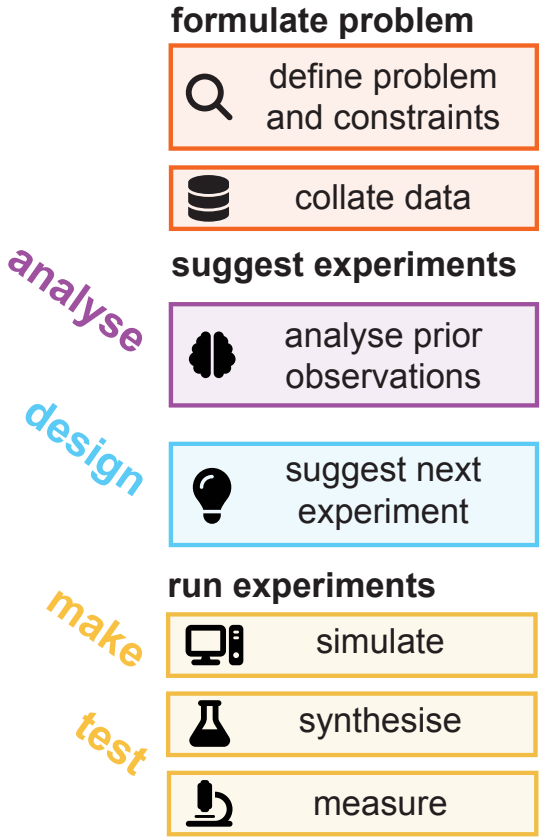
J. Am. Chem. Soc., **2022**, 144, 41, 18730

decisions lie at the core of experimentation

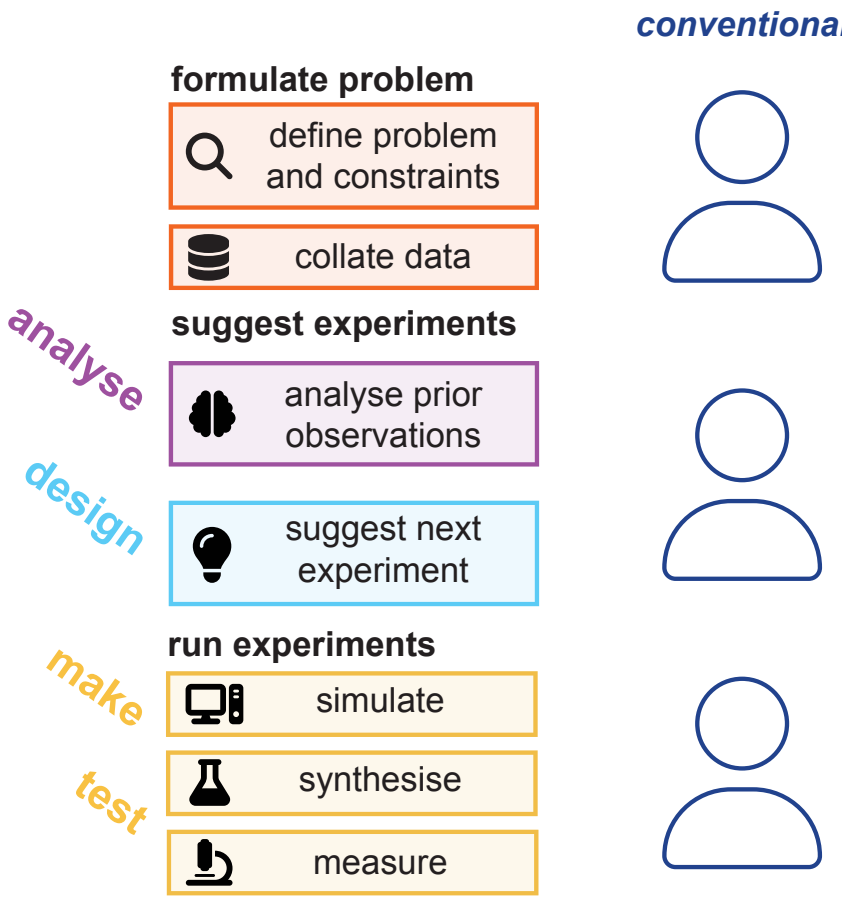


J. Am. Chem. Soc., **2022**, 144, 41, 18730; *Acc. Chem. Res.*, **2020**, 53, 599; *Energy Environ. Sci.*, **2022**, 15, 579

the design-make-test-analyse cycle in the digital age



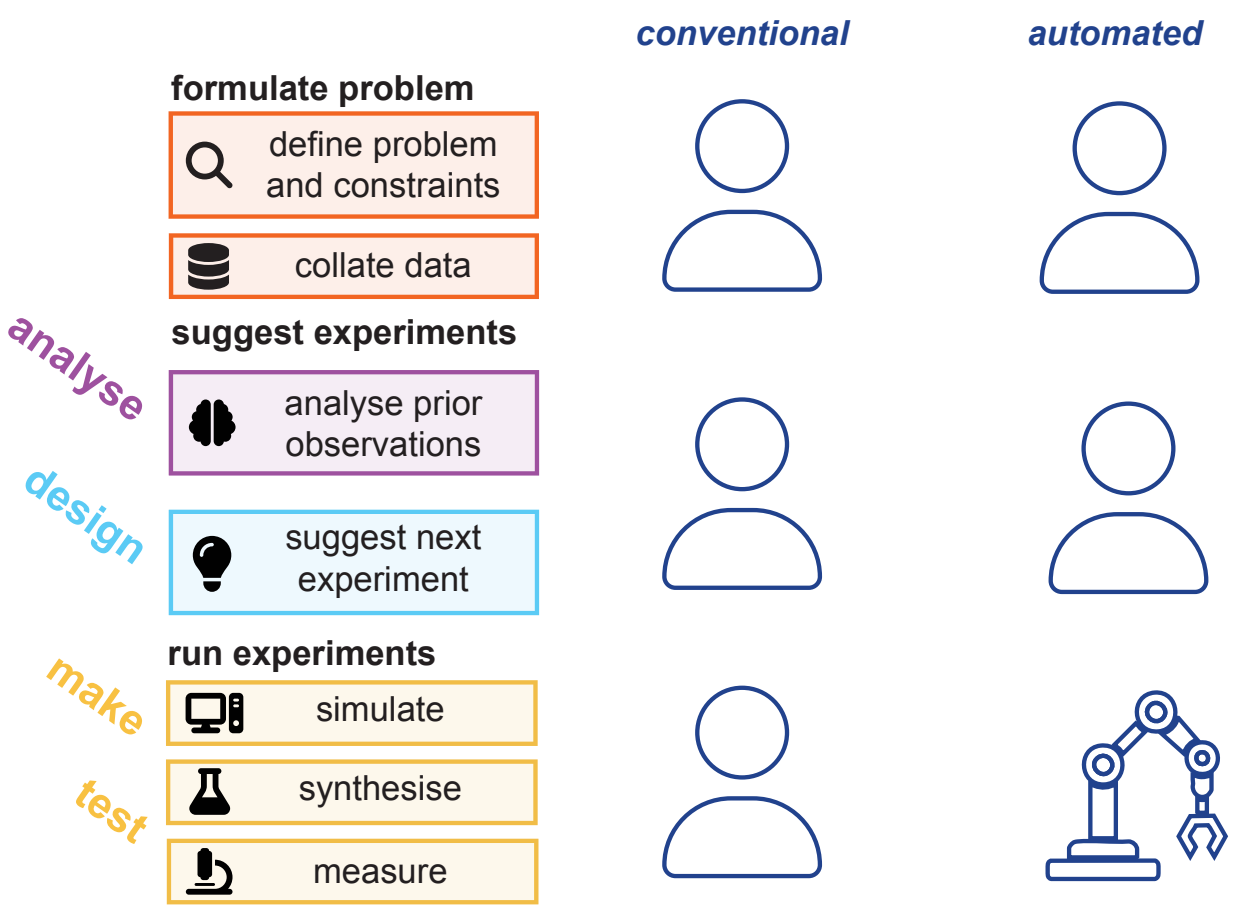
the design-make-test-analyse cycle in the digital age



where are the bottlenecks?

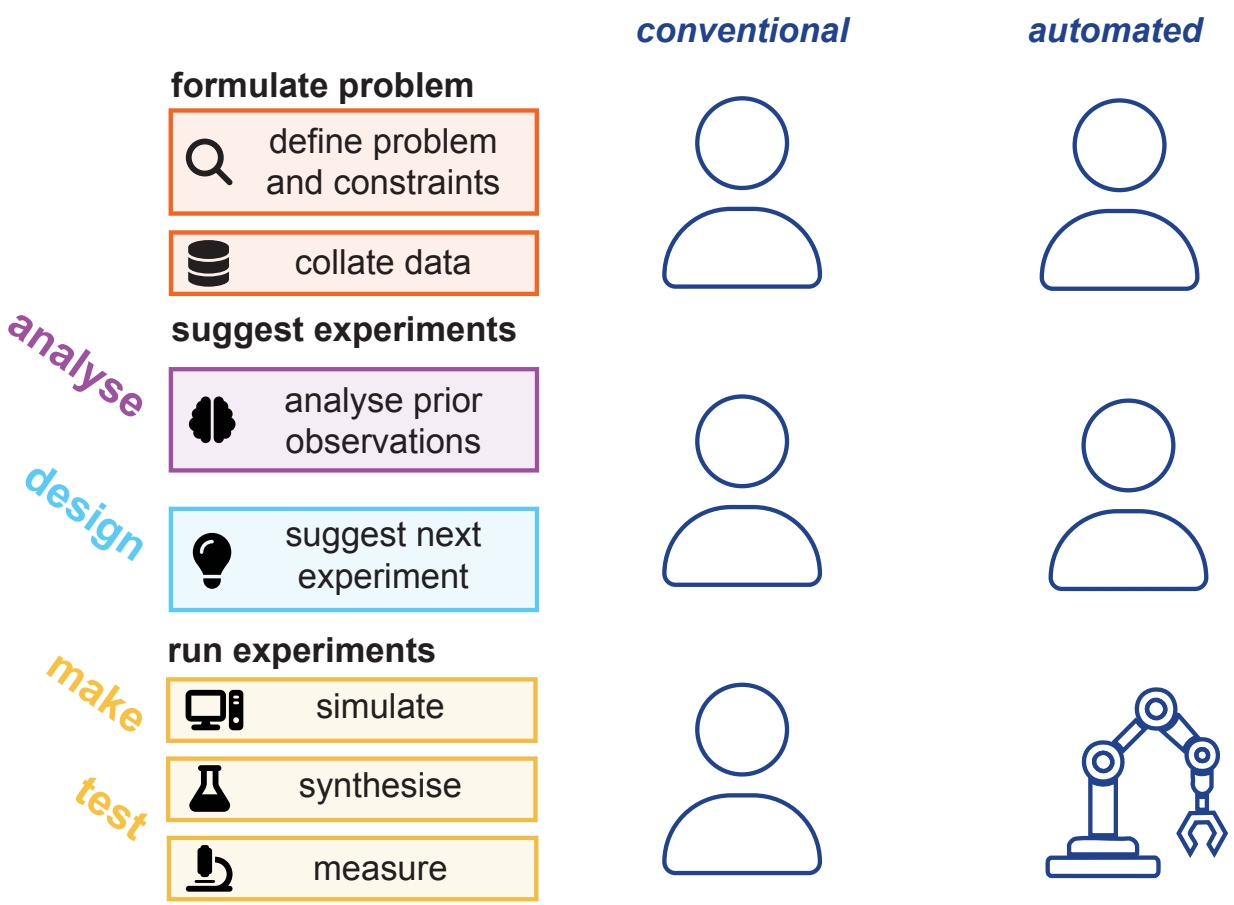
1. manual synthesis, characterization and target property measurements are time and resource intensive
2. chemical intuition is largely limited to *small* changes to *either* experimental conditions *or* structure

the design-make-test-analyse cycle in the digital age



Dr. Becky Greenaway

the design-make-test-analyse cycle in the digital age



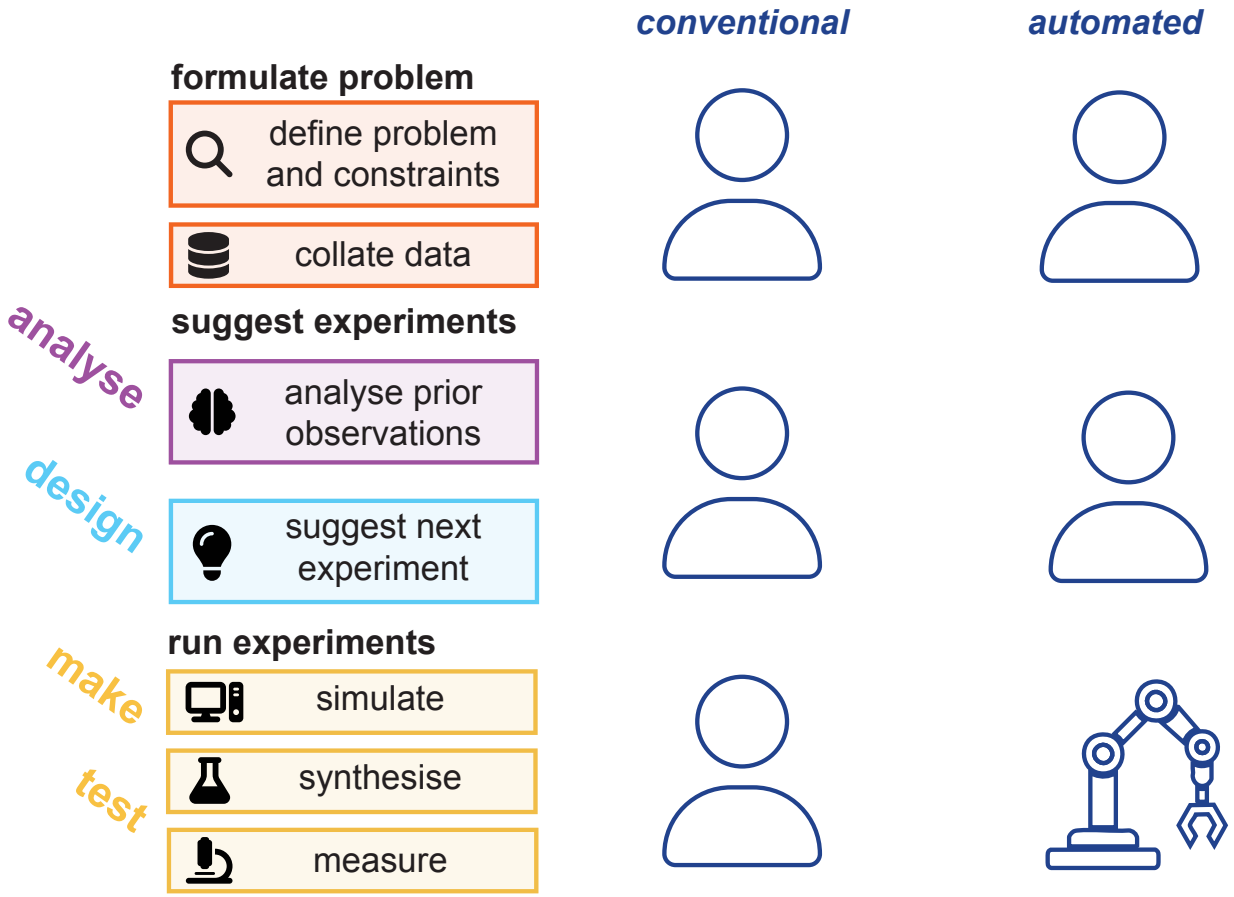
where are the bottlenecks?

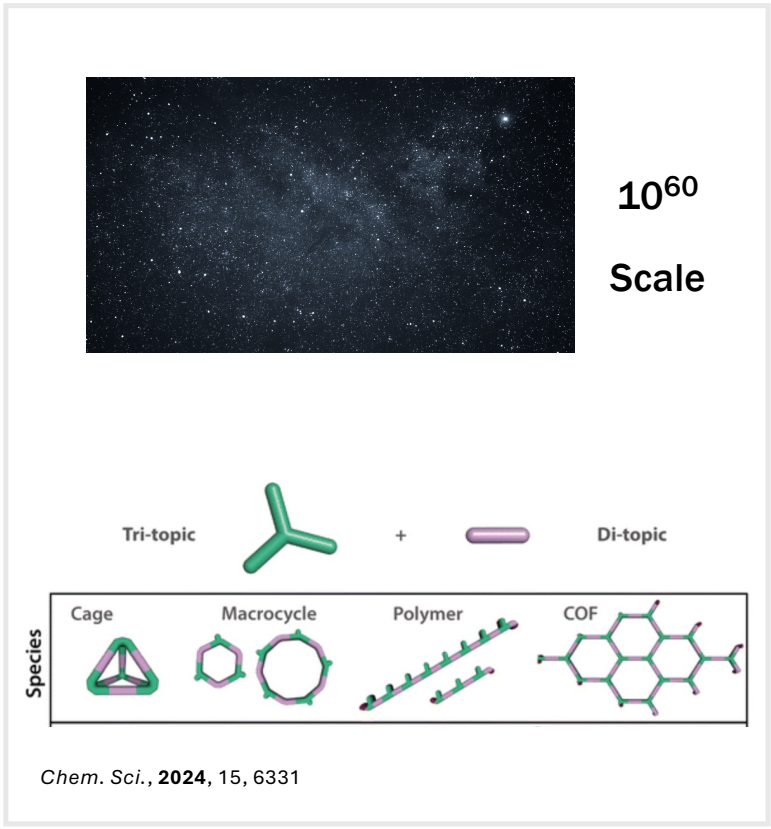
1. **manual synthesis, characterization and target property measurements are time and resource intensive**
2. chemical intuition is largely limited to *small* changes to *either* experimental conditions *or* structure

how does this help us?

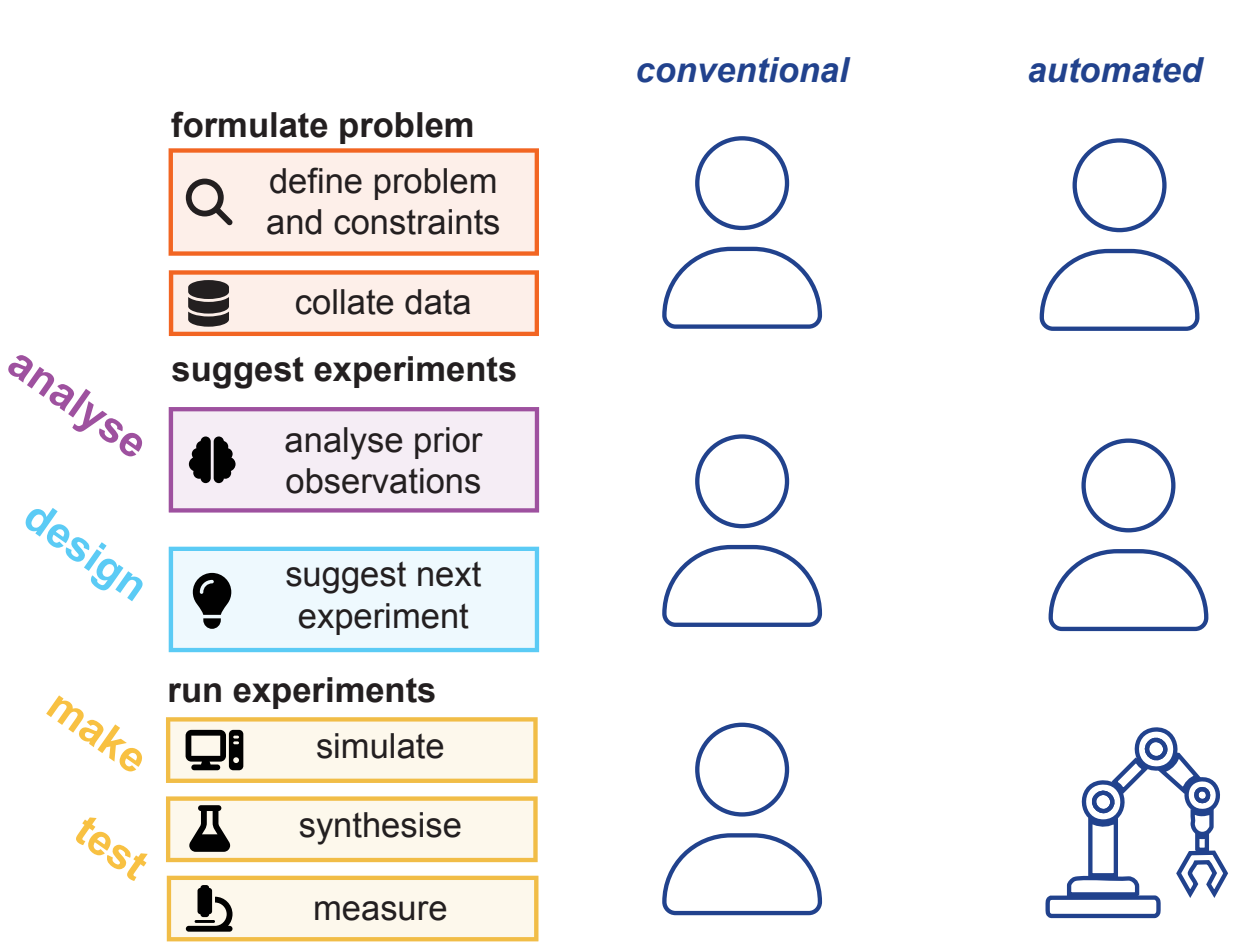
1. chemistry research 24/7!
2. improved reproducibility!
3. large batches!

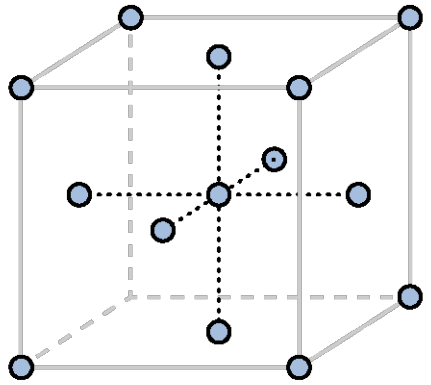
the design-make-test-analyse cycle in the digital age



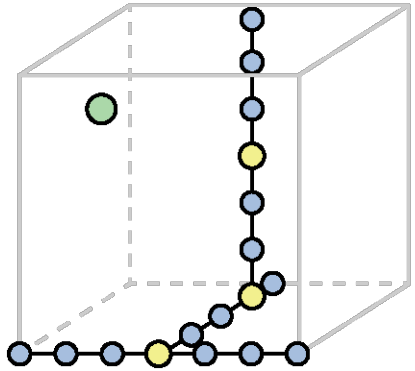


the design-make-test-analyse cycle in the digital age



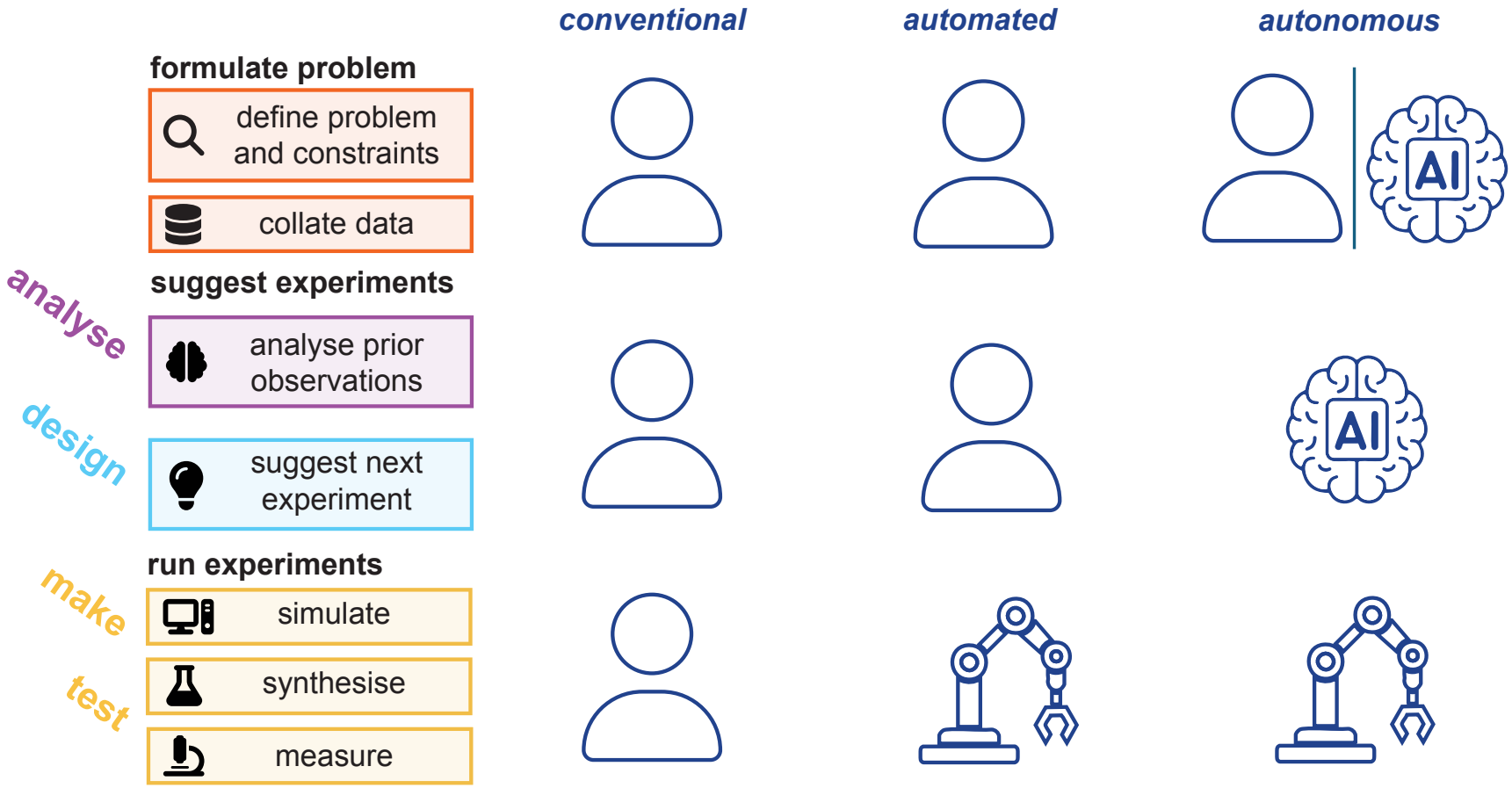


design of experiment

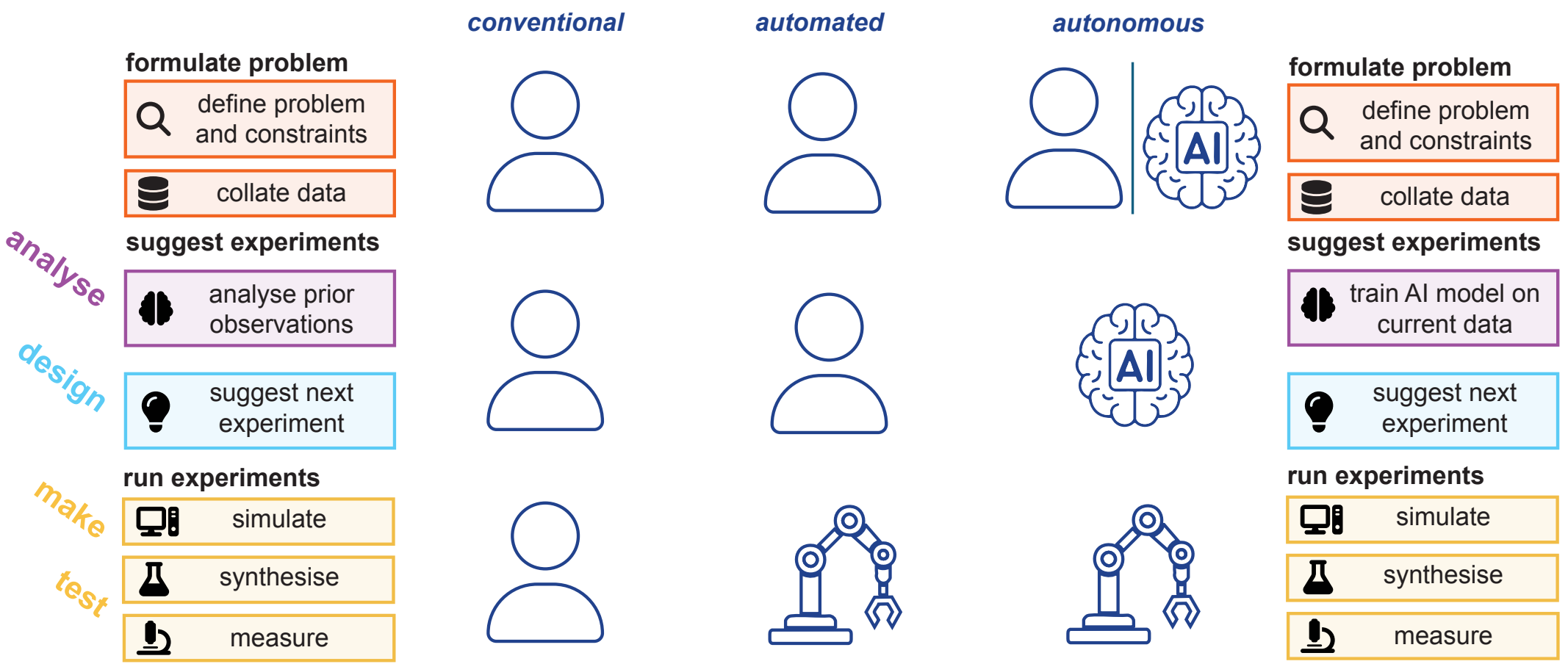


one-factor-at-a-time

the design-make-test-analyse cycle in the digital age

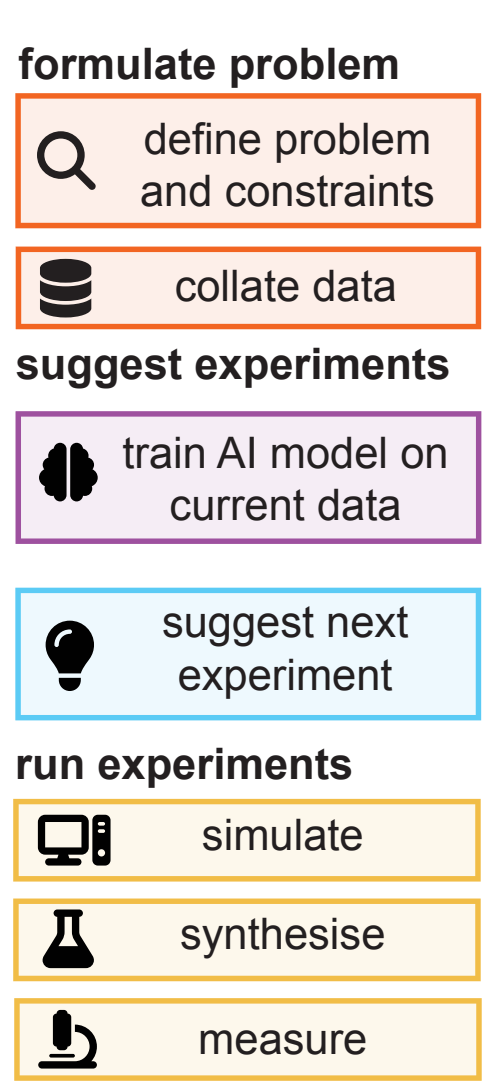


the design-make-test-analyse cycle in the digital age

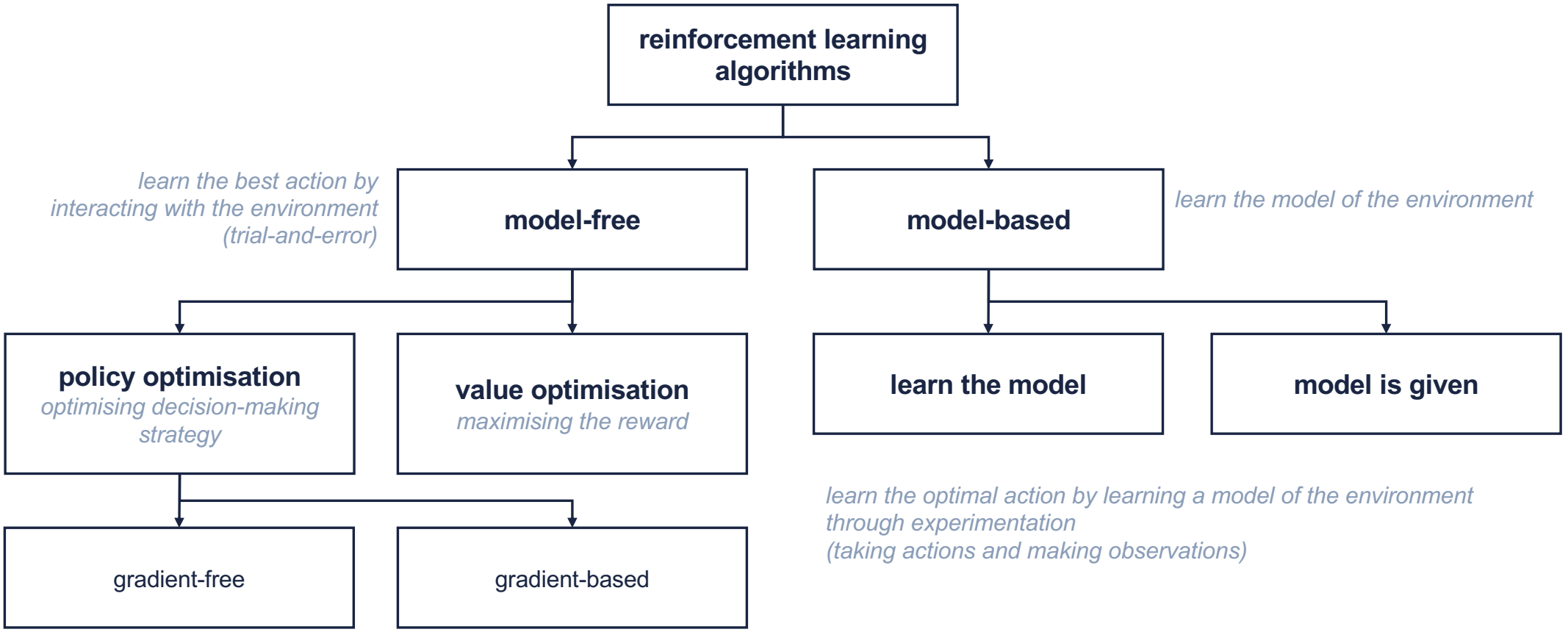


what does our solution need?

main considerations	*what does our solution need?*
1. we don't know what our objective function looks like	1. we need to be able to sample
2. we are often operating under budgetary and resource constraints	2. resource- and budget—aware
3. we often don't have a lot of starting data, if any at all	3. need to work well in the sparse/low data regime



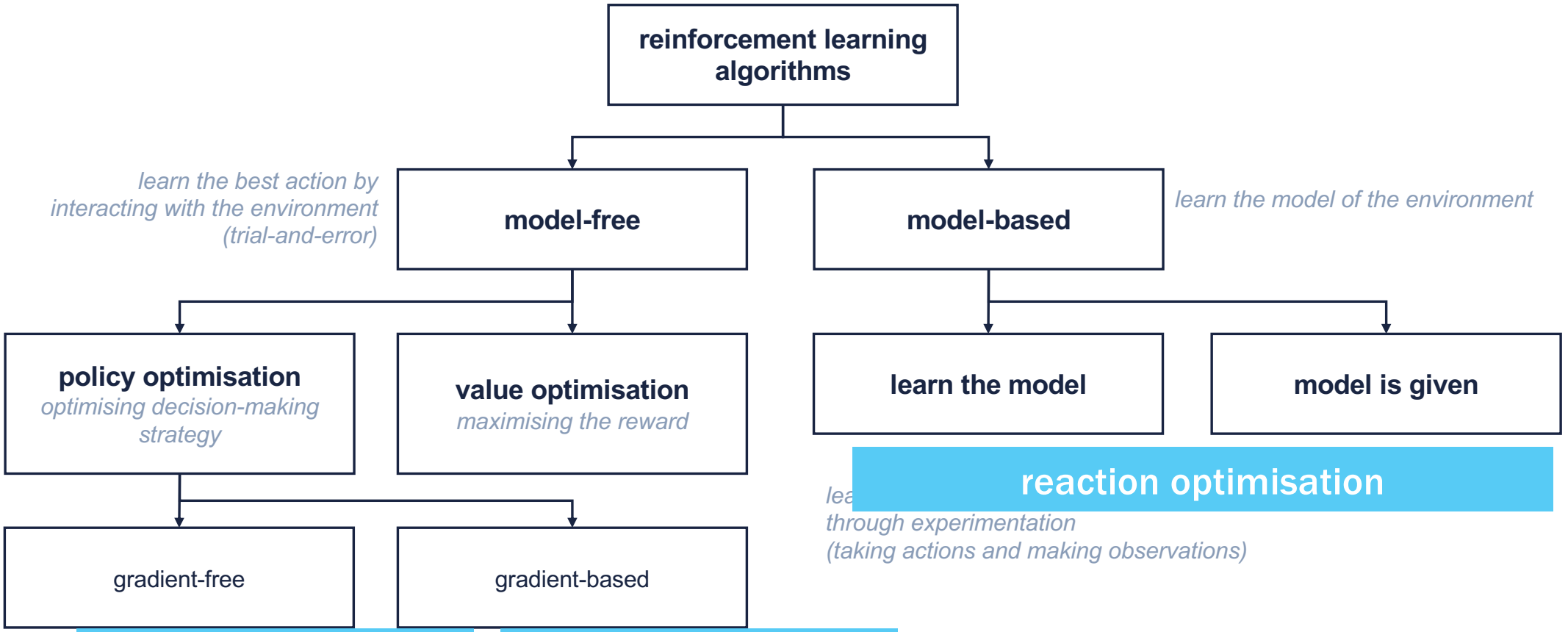
optimization and decision-making in chemistry



- ***Bayesian optimisation***
- *evolutionary algorithms*
- *…*

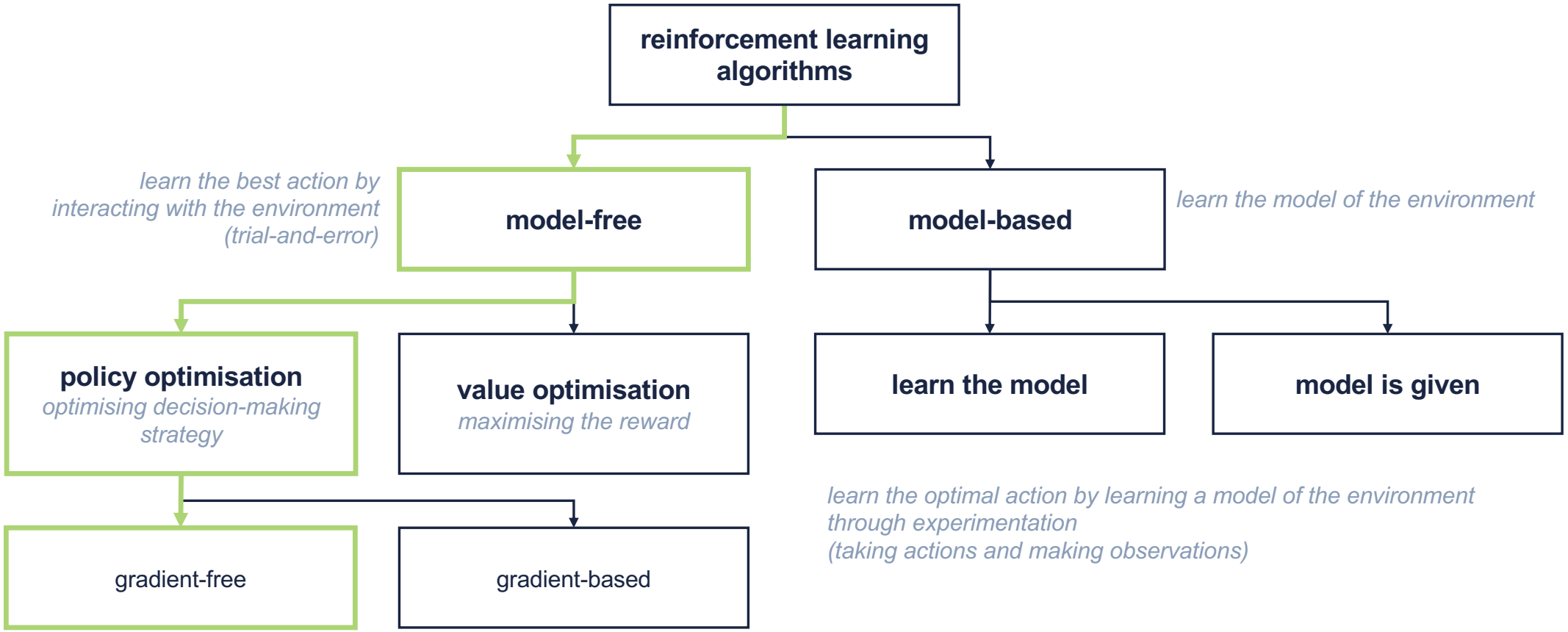
inspired by Antonio del Rio Chanona – ML for Chemical Engineering Lecture Series

optimization and decision-making in chemistry



inspired by Antonio del Rio Chanona – ML for Chemical Engineering Lecture Series

optimization and decision-making in chemistry

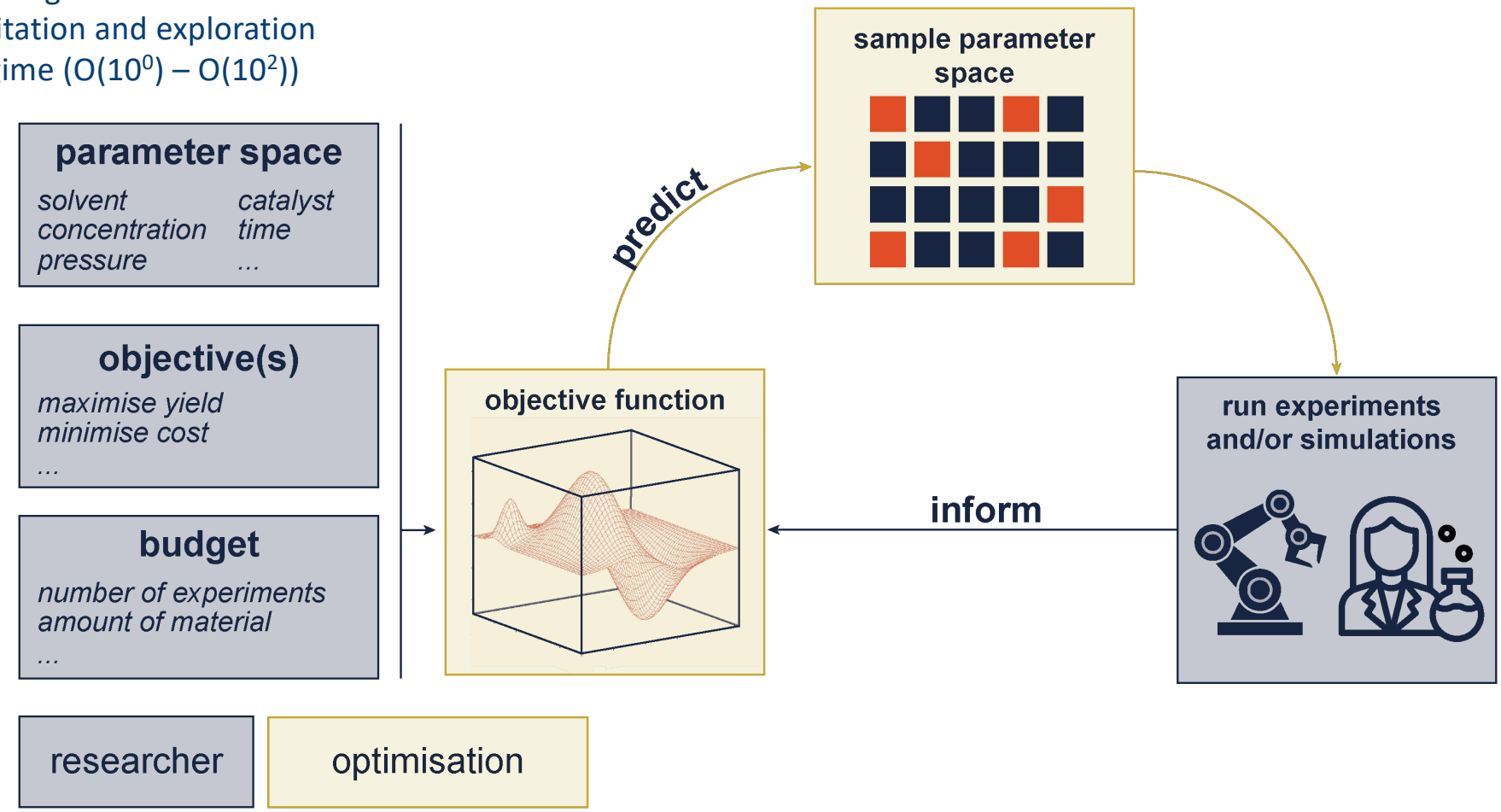


- ***Bayesian optimisation***
- *evolutionary algorithms*
- *…*

inspired by Antonio del Rio Chanona – ML for Chemical Engineering Lecture Series

Bayesian optimization (BO)

- ✓ optimization via sampling (black-box optimization)
- ✓ resource- and budget-aware
- ✓ balances exploitation and exploration
- ✓ sparse data regime ($O(10^0)$ – $O(10^2)$)



Bayesian optimization for chemistry

the general algorithm *part 1*

while {resources available}

fit a Bayesian ML model on current data points/observations $\{x, f(x)\}$ *first*

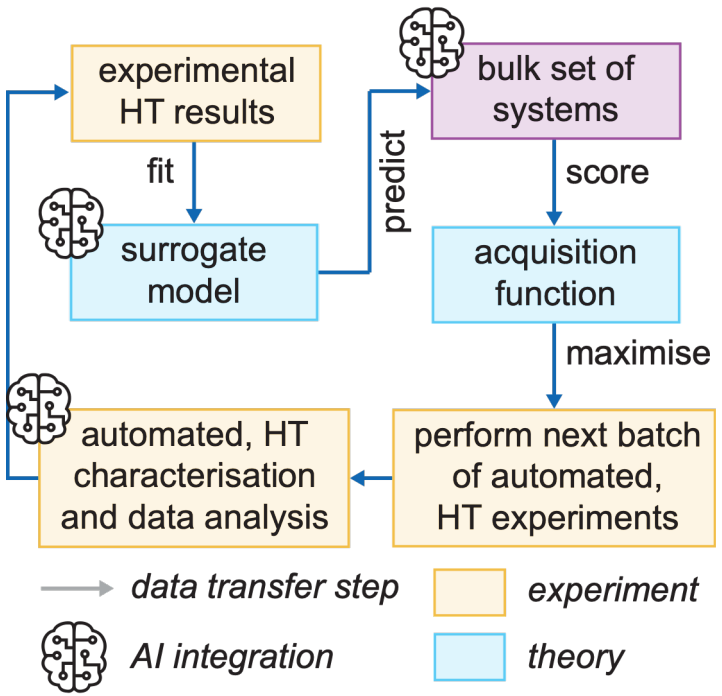
find x that maximizes the acquisition function *second*

take measurement (sample x & observe $f(x)$)

end *third*

inspired by Joel Paulson – 2024 Sargent Process Centre for Process Engineering BO Summer School

through the chemistry lens *part 2*



the role of the surrogate model

surrogate models reflect our best approximation of the underlying objective function that we are aiming to optimize

what do we need for our model?

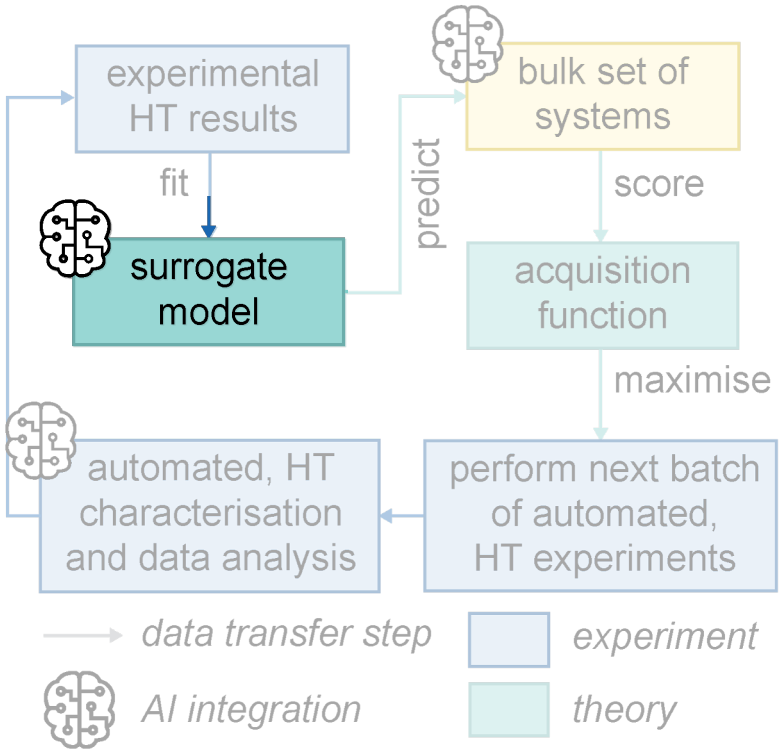
probabilistic (quantitative uncertainty)

what are our options?

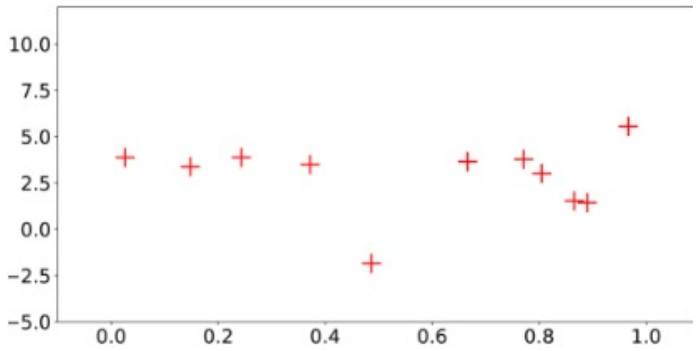
1. Bayesian linear regression (BNNs)
2. Random forests
3. Tree-structured Parzen estimators
4. Gaussian processes (GPs)

why do people focus largely on GPs?

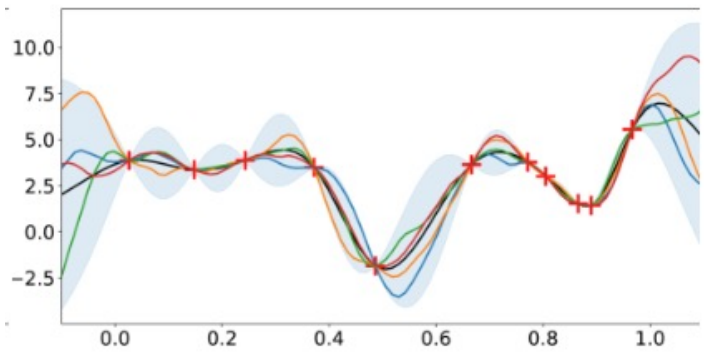
GPs are **flexible** and **simple**
can model functions on an infinite domain
can make predictions about our data using prior knowledge



a crash course on GPs



(a) Data point observations



(b) Five possible functions by GPR

CiSE, **2023,** 25, 4-11

the main idea

- for a given set of data points, there are an infinite number of functions that might fit the data
- GPs assign a probability to each of these functions!
- The mean of all the possible functions represents the most likely underlying model **< *this is the function that is used for predictions!***
- The variance represents our confidence in the mean (predictions)

our problem statement

parameter vectors *measurements* *gaussian process*

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad f = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \qquad p(f) = \mathcal{GP}(f; \mu, k)$$

let's unpack these functions!

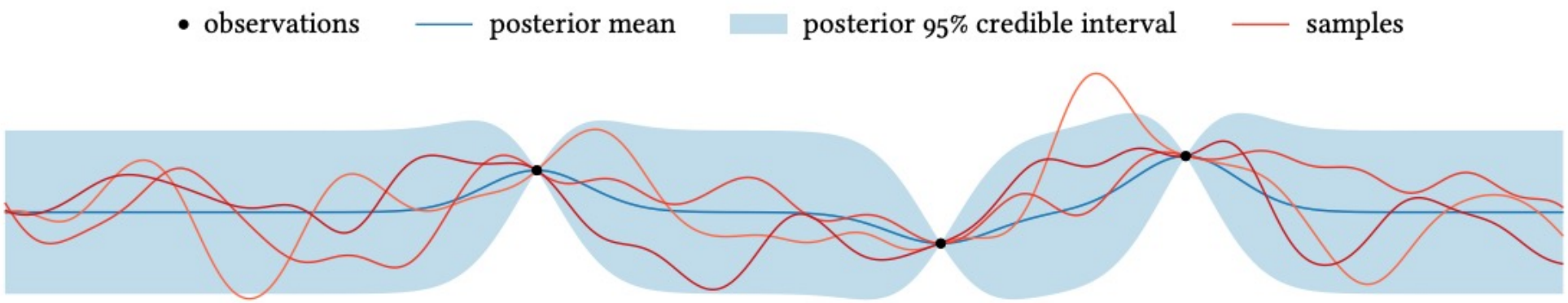
mean $\mu = \begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}$

covariance $K_{ij}(x, x') = k(x_i, x_j)$

what do you mean by “mean”?

what does $\mu(x)$ tell us?

the expected value (“measurement”) at a location in our parameter space (“experimental conditions”)



Garnett, Roman; *Bayesian Optimization*, **2023**, Cambridge University Press

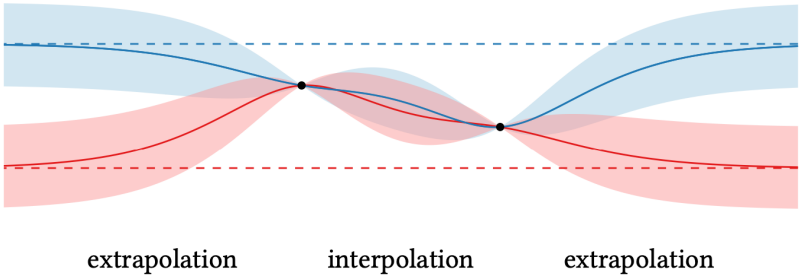
what do you mean by “mean”?

what does $\mu(x)$ tell us?

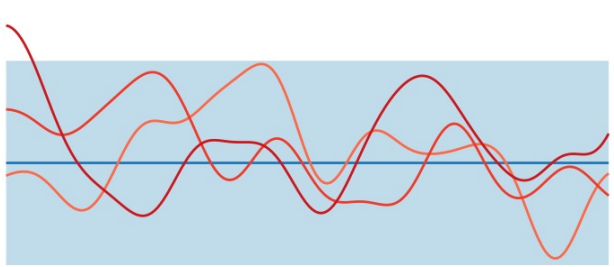
the expected value (“measurement”) at a location in our parameter space (“experimental conditions”)

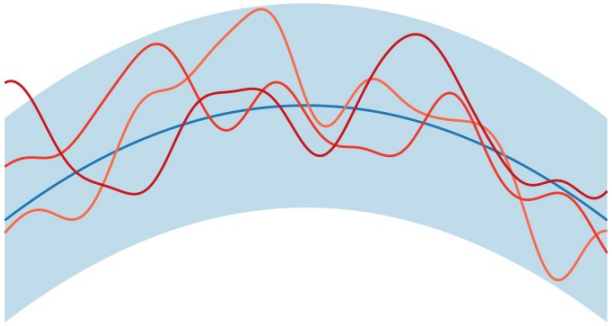
is this a way to introduce chemical knowledge?

Yes!
but be careful!



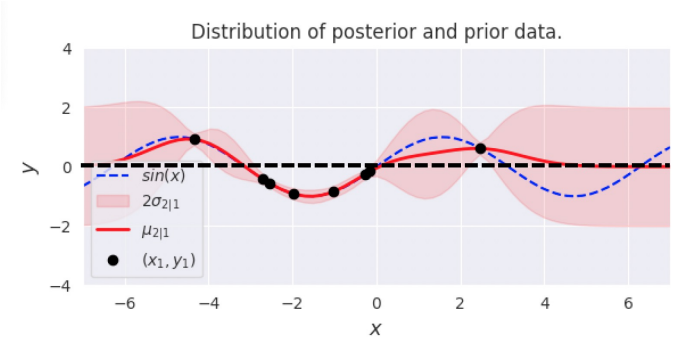
we can modify the functional form of the mean



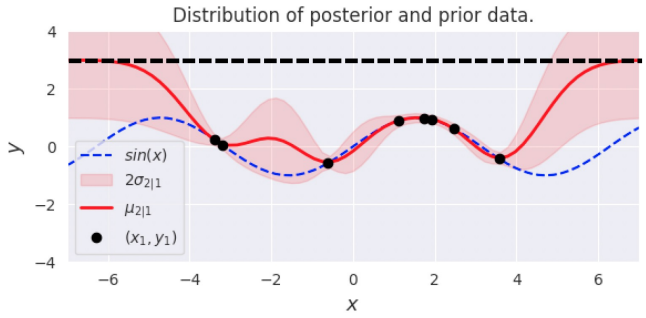


Garnett, Roman; *Bayesian Optimization*, **2023**, Cambridge University Press

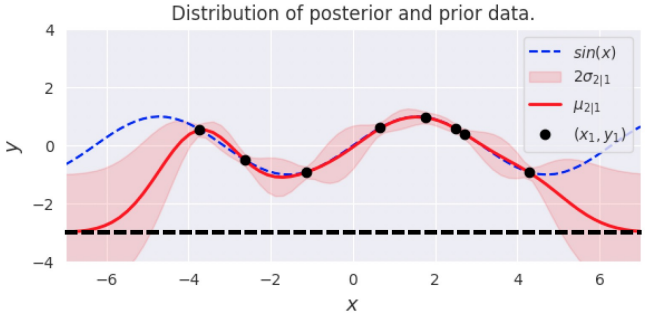
what do you mean by “mean”?



$m(x) = 0$

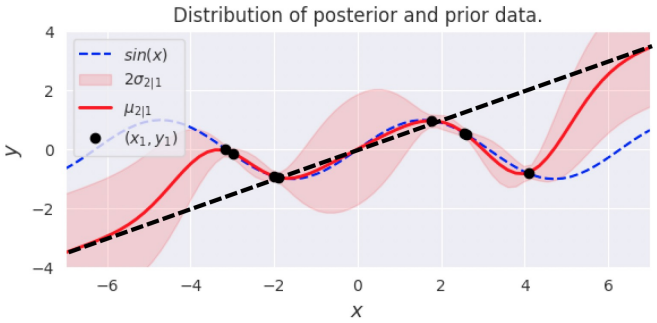


$m(x) = 3$

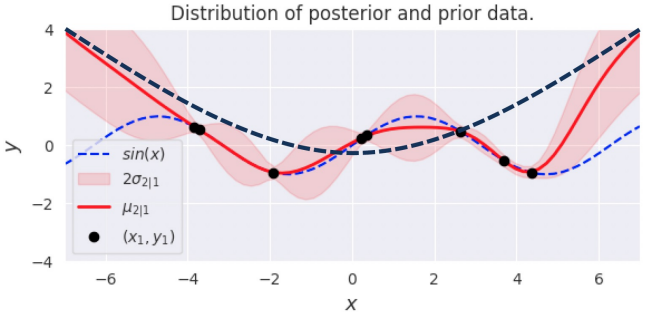


$m(x) = -3$

best practice is to take $m(x) = 0$ unless you have you have a *very* good reason



$m(x) = \frac{1}{2}x$



$m(x) = \frac{2}{25}x^2$

from Antonio del Rio Chanona – ML for Chemical Engineering Lecture Series

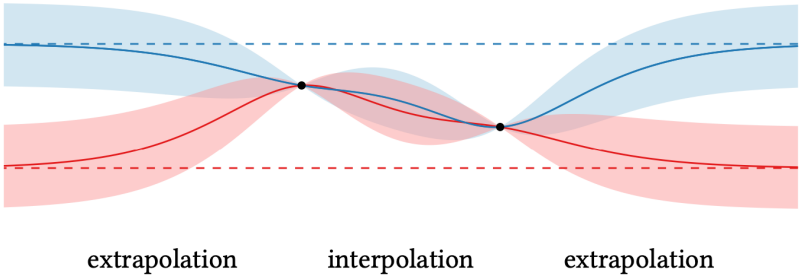
what do you mean by “mean”?

what does $\mu(x)$ tell us?

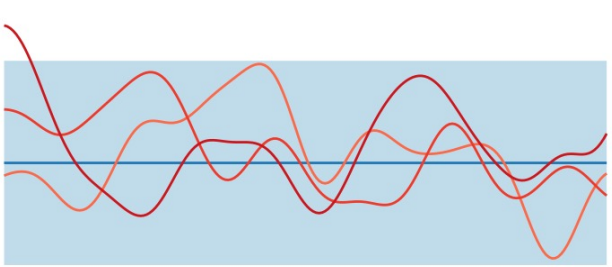
the expected value (“measurement”) at a location in our parameter space (“experimental conditions”)

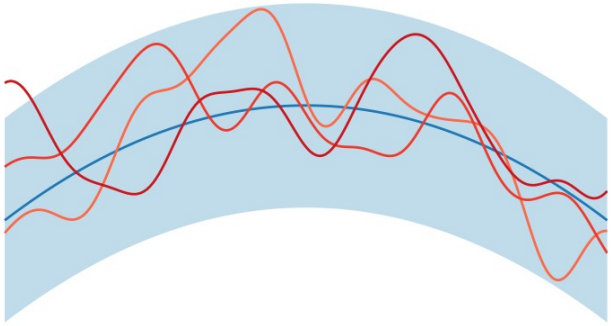
is this a way to introduce chemical knowledge?

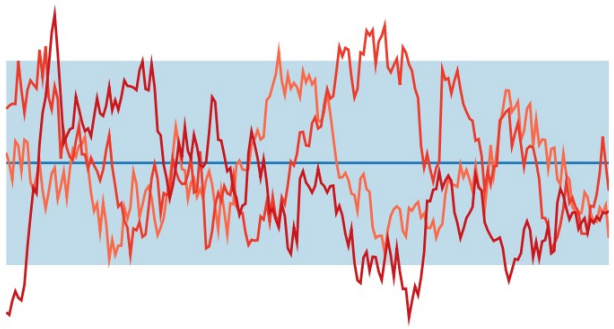
Yes!
but be careful!



does $\mu(x)$ or $k(x)$ have a larger impact?







Garnett, Roman; *Bayesian Optimization*, **2023**, Cambridge University Press

kernels, kernels, kernels…

what is powerful about covariance?

the covariance function impacts the **shape of the distribution** and represents the **characteristics of the underlying function**

covariance matrix describes how the data points are correlated

covariance function vs. kernel

covariance matrix is generated by evaluating the kernel (*k*)

the **kernel** is a **function;** this is what we get to choose

what does the kernel tell us?

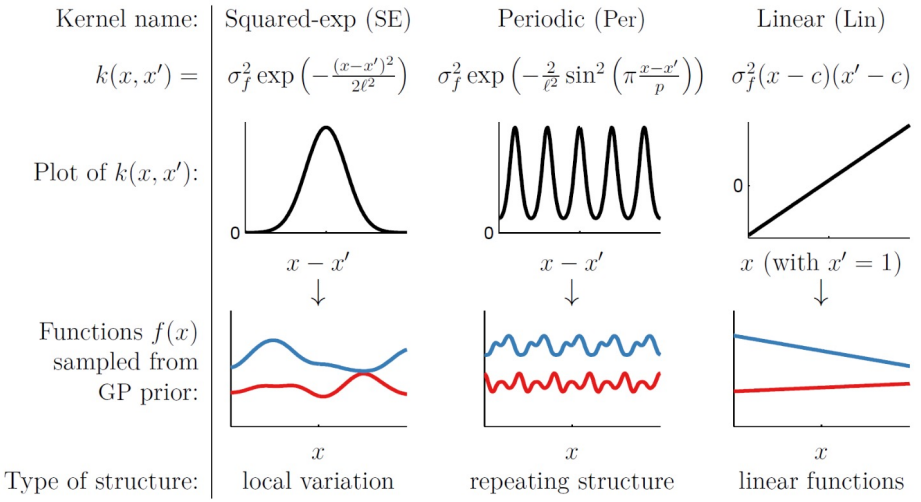
input – two points
output – measure of similarity between the two points

$$K_{ij}(x, x') = k(x_i, x_j)$$

describes how much influence *i* and *j* have on each other

within the context of regression, our measurements should be similar when our conditions are close together!

are kernels a way to introduce chemical knowledge?

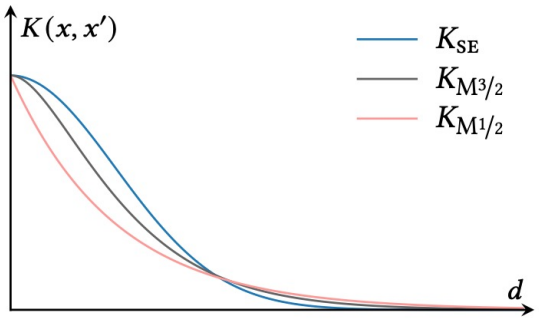


ekamperi.github.io/mathematics/2021/03/30/gaussian-process-regression.html

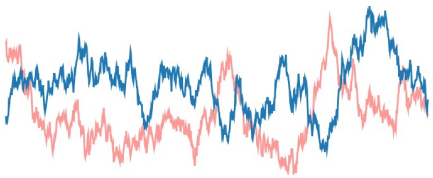
kernels, kernels, kernels…

the Matérn family

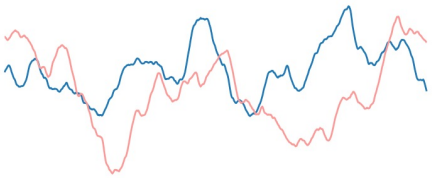
$K_{M(\nu)}$ $\nu \rightarrow$ smoothness



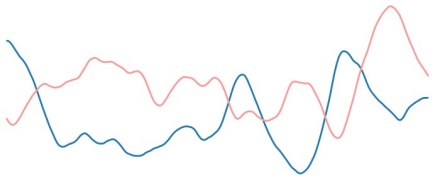
Garnett, Roman; *Bayesian Optimization*, **2023**



$\nu = 1/2$, (3.11)



$\nu = 3/2$, (3.13)

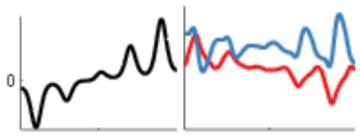


$\nu = 5/2$, (3.14)

the squared exponential (RBF) covariance is the most prevalent covariance function in the ML literature – however, this may not always be the best choice!

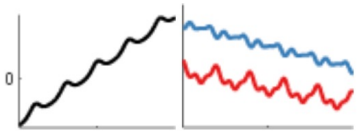
combining kernels

multiplying



linear * periodic

adding kernels



linear + periodic

deep kernel learning

combines NNs for feature learning with GPs for uncertainty quantification

beware of the tendency to ***overfit***

arXiv:2102.12108; *Comm. Chem.*, **2024**, 7, 136

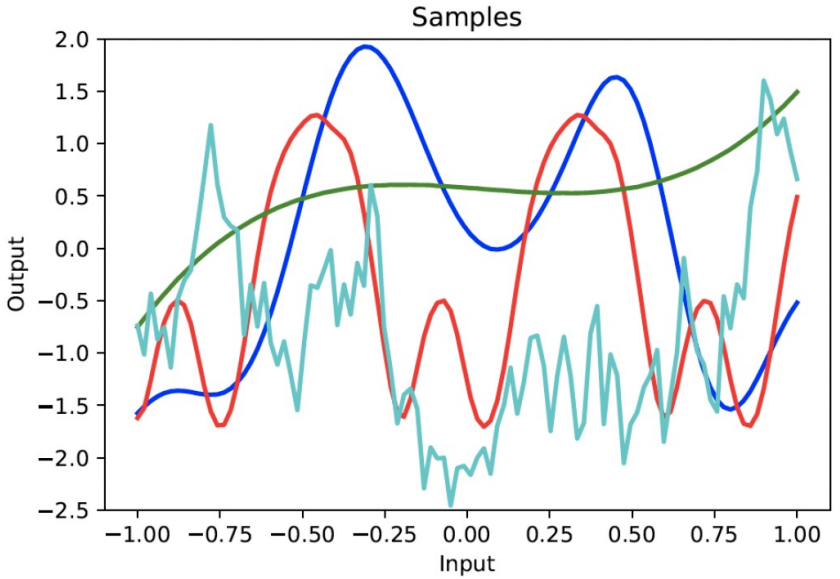
different kernels yield different predictions

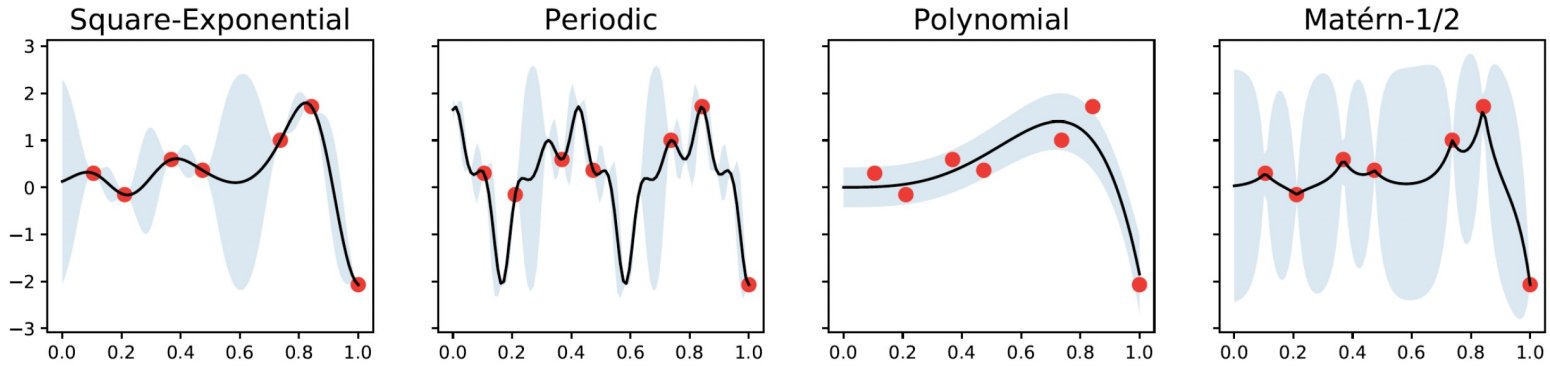
$$k_{SE}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

$$k_{per}(x, x') = \sigma^2 \exp\left(-\frac{2\sin\left(\frac{\pi|x - x'|}{p}\right)}{l^2}\right)$$

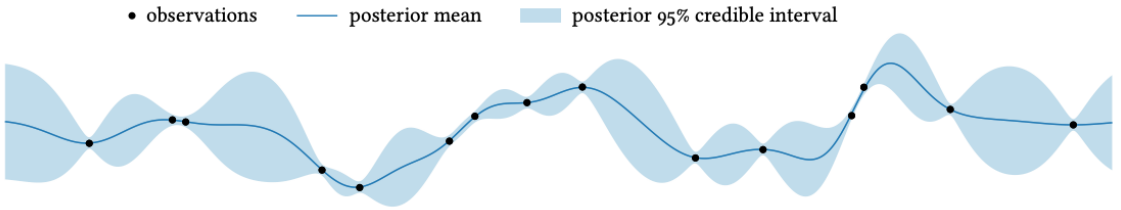
$$k_{M(1/2)}(x, x') = \sigma^2 \exp\left(-\frac{|x - x'|}{2l^2}\right)$$

$$k_{pol}(x, x') = (x^\intercal x' + c)^d$$

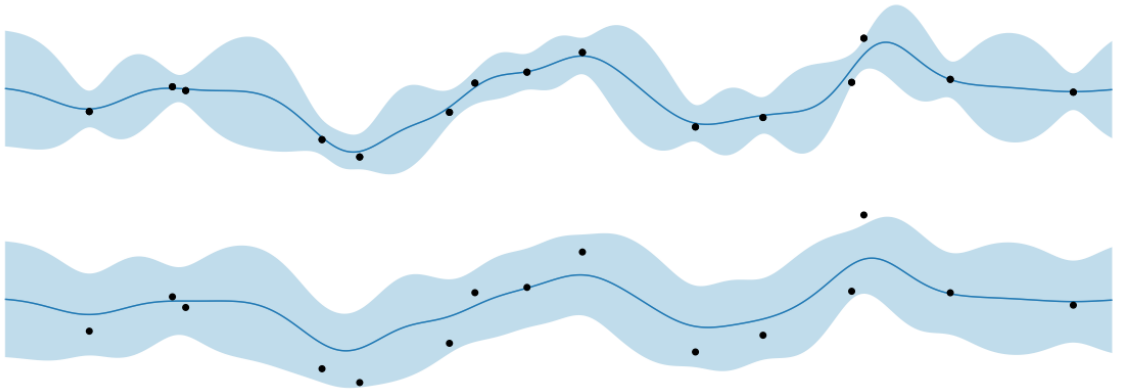




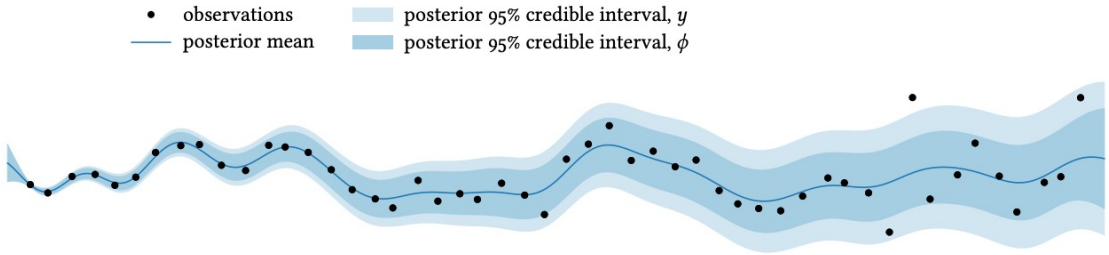
we can also account for noisy observations



independent homoskedastic noise



heteroskedastic noise



Garnett, Roman; *Bayesian Optimization*, **2023**, Cambridge University Press

Bayesian optimization for chemistry

the general algorithm *part 1*

while {resources available}

fit a Bayesian ML model on current data points/observations $\{x, f(x)\}$

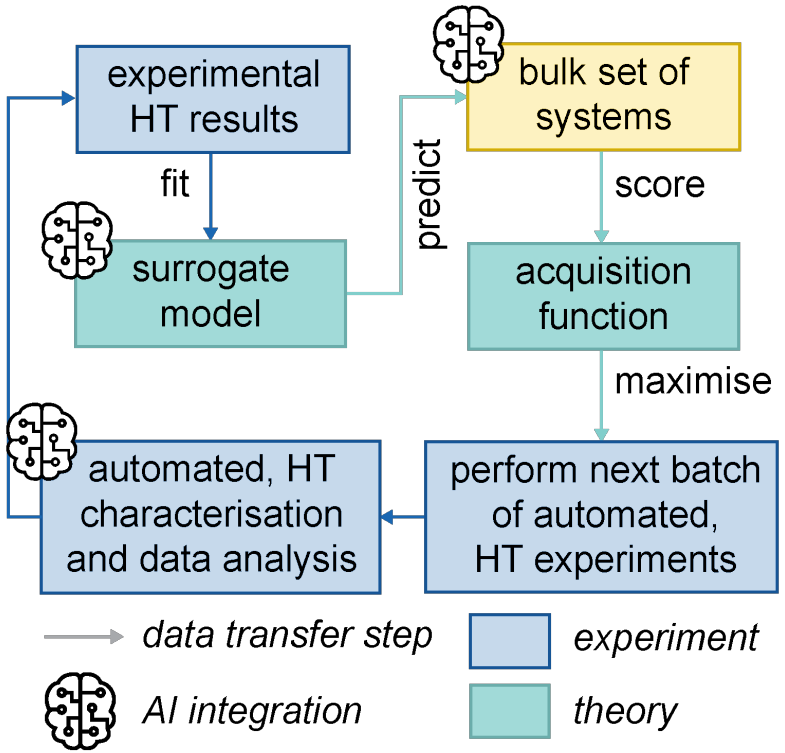
find x that maximizes the acquisition function *second*

take measurement (sample x & observe $f(x)$)

end

inspired by Joel Paulson – 2024 Sargent Process Centre for Process Engineering BO Summer School

through the chemistry lens *part 2*



the role of the acquisition function

acquisition functions are responsible for scoring our parameter/candidate space by balancing exploitation and exploration

from a decision theory perspective...

acquisition functions are the **first step** to defining our ***policy***!

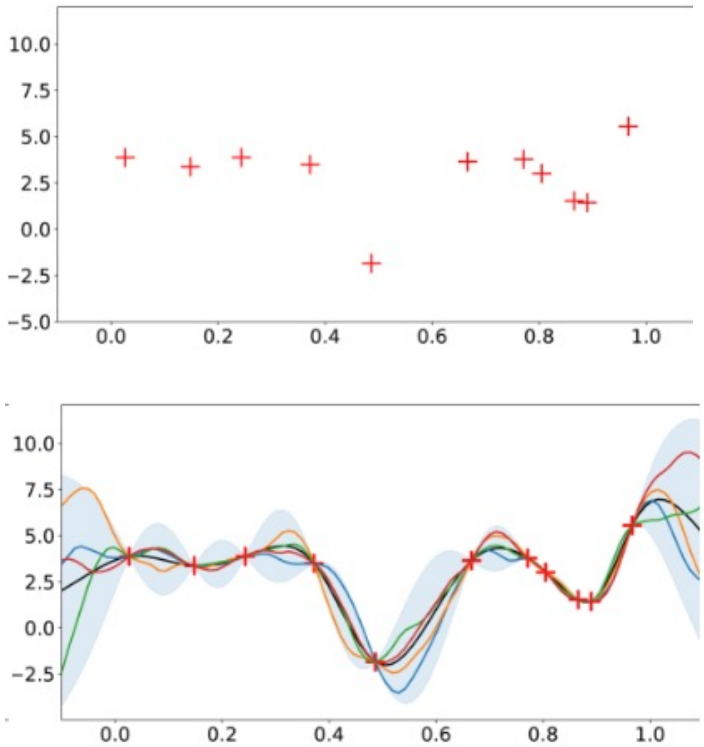
essentially the rules by which we make our decision

we are defining a ***score*** to each potential observation we ***could*** make (*e.g.* potential experimental conditions) with respect to the overall optimization objective

we then define a ***policy*** by observing at the point(s) defined most promising by the acquisition function

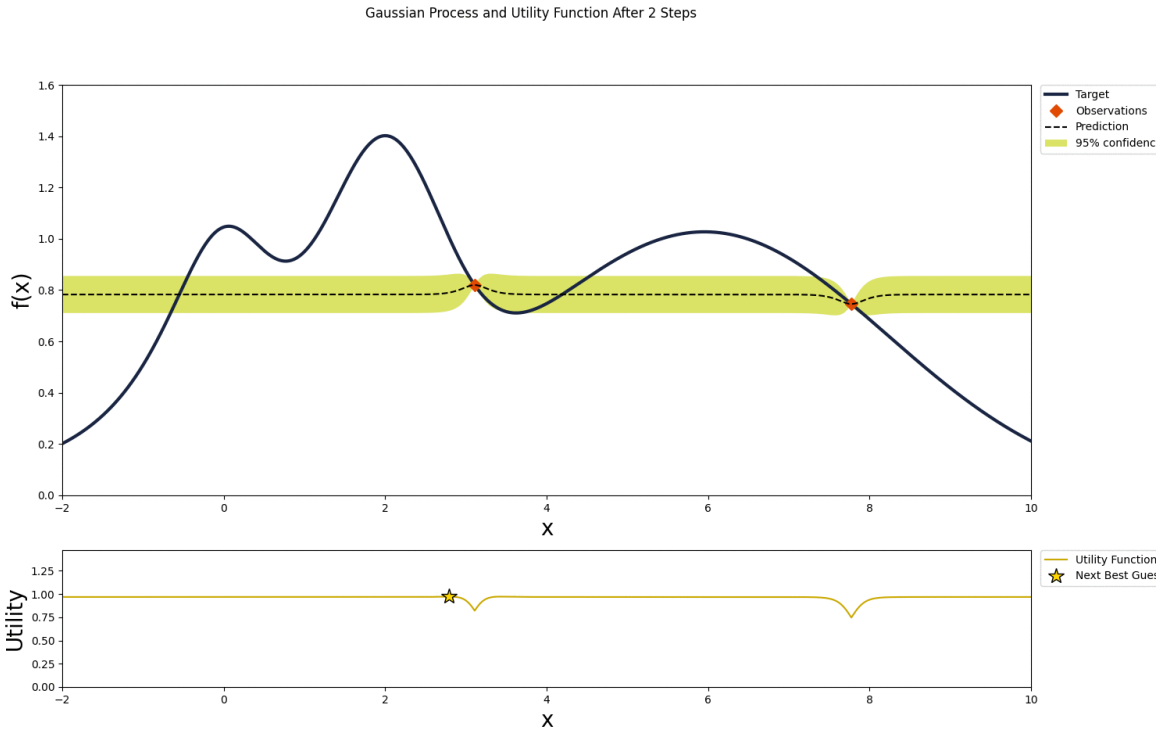
what does this look like in practice?

1. compute the ***score*** for all candidate data points
2. identify the candidate data points deemed most promising by ***maximizing*** the acquisition function



CiSE, **2023,** 25, 4-11

the role of the acquisition function



the role of the acquisition function

what does this look like in practice?

1. compute the ***score*** for all candidate data points
2. identify the candidate data points deemed most promising by ***maximizing*** the acquisition function

here, we are first estimating the objective function based on the data, and then setting preferences according to this estimate

putting the "function" in acquisition function

let $\alpha(x; \mathcal{D})$ be our acquisition function

if $\alpha(x; \mathcal{D}) > \alpha(x'; \mathcal{D})$
then x is preferred over x'

available data ($\mathcal{D}$) shape our preferences (*posterior*)

the BO paradox...

global optimization via global optimization

but this is okay!

acquisition functions possess two main properties:

1. cheap to evaluate
2. analytically differentiable
   (*i.e.* possess tractable gradients)

improvement-based acquisition functions

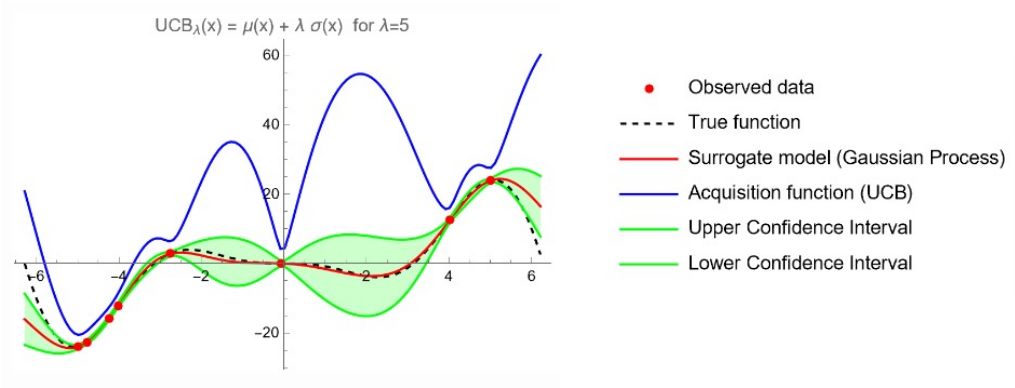
upper confidence bound (UCB)

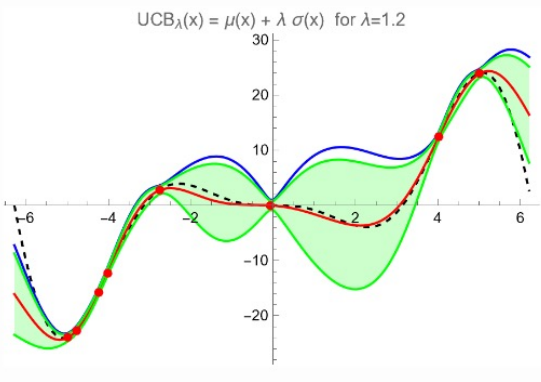
weighted sum of expected performance predicted by the GP, and the uncertainty

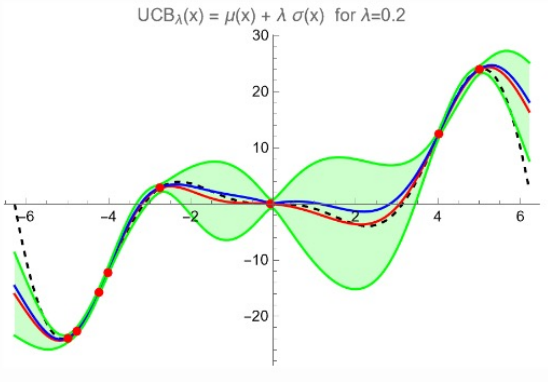
$$\alpha(x;\lambda) = \mu(x) + \lambda\sigma(x)$$

exploitation | exploration

exploration vs. exploitation tradeoff is tuned via λ







ekamperi.github.io/machine%20learning/2021/06/11/acquisition-functions.html

beware

if λ is too conservative, you may not see great short-term performance

in general, performance is sensitive to choice of λ

improvement-based acquisition functions

probability of improvement (PI)

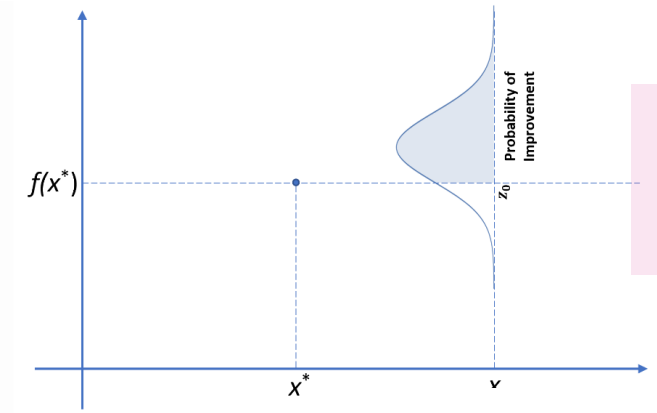
estimates the likelihood that a new design point (x) will improve upon the current best-known value (x^*)

want to maximize $f(x)$ and our best solution is $f(x^*)$

$$\mathrm{I}(x) = \max(f(x) - f(x^*), 0)$$

each sampled from normal distribution $f(x) \sim \mathcal{N}(\mu(x), \sigma^2(x))$

$$\mathrm{I}(x) = \max(\mu(x) + \sigma^2(x)z - f(x^*), 0)$$



doesn't factor in the magnitude of the improvement

ekamperi.github.io/machine%20learning/2021/06/11/acquisition-functions.html

expected improvement (EI)

selects the next point to evaluate by maximizing the expected positive increase in the objective function, based on current observations

$$\mathrm{EI}_n(x) = \mathbb{E}[\mathbf{I}(x)] = \int_{-\infty}^{\infty} \mathbf{I}(x)\phi(z)\mathrm{dz}$$

arXiv:1807.02811v1

sometimes you may see "qEI"

calculating expectations often involves solving an integral over the posterior distribution…

this is typically too complex to solve exactly (especially true when using batches!)

q indicates that Monte Carlo sampling is used to approximate the integrals!

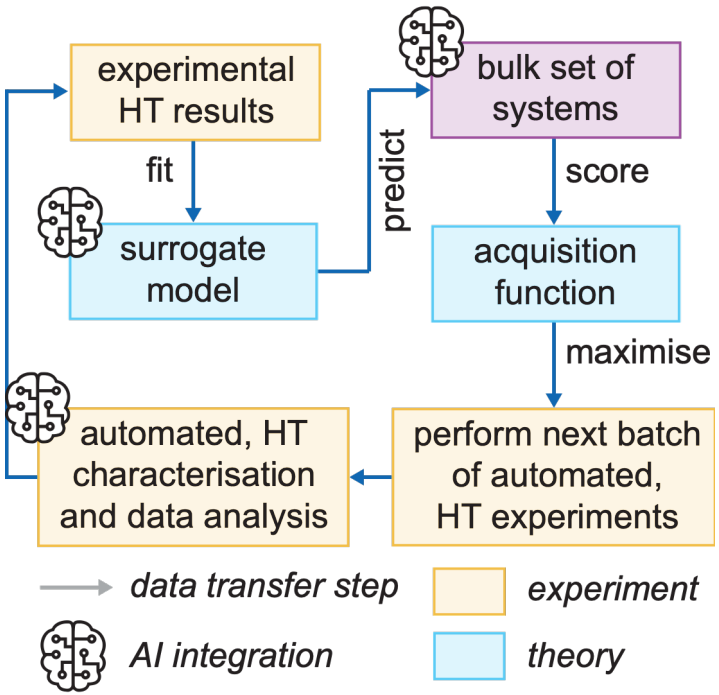
chemistry-specific considerations

objective

- maximise, minimise, combination?
- new material
- improved performance
- experimental conditions

information streams

- experiment, theory, combination?
- timescales
- batch?



design space

- integer, continuous, categorical, chemical?
- chemical representation / encoding

stopping criteria

- precursor amounts
- number of experiments
- total time

domain-specific parameter spaces

Example experimental domain:

Continuous parameter

Temperature:

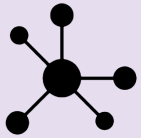
$T \in [0, 80]\,^{\circ}C$

Discrete numerical parameter

Stirring rate:

$r \in \{100, 500, 1000\}$ rpm

Categorical parameter



Ligand type:

$x \in \{L_1, L_2, L_3\}$

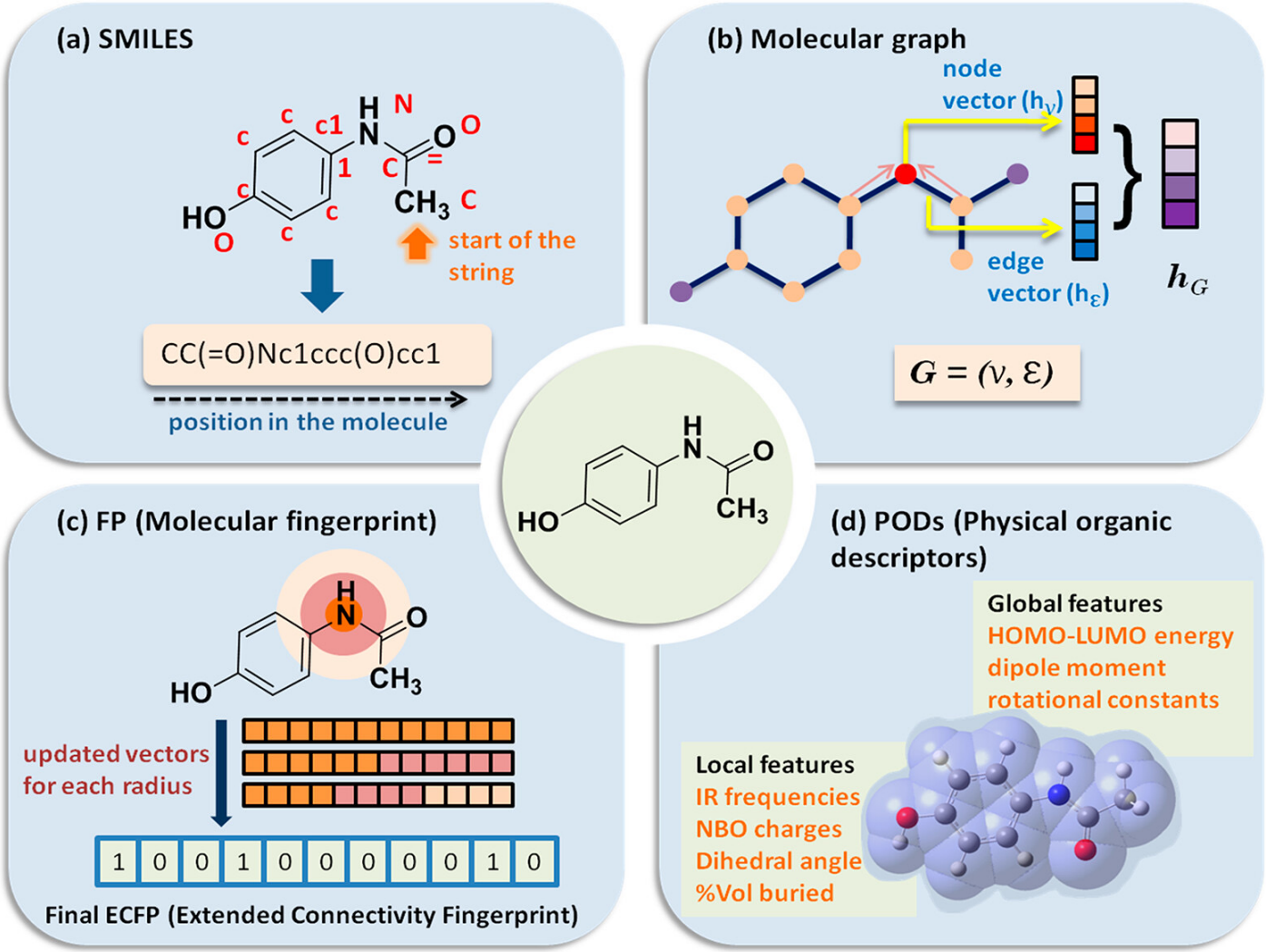
Chemical/molecular parameter

Solvent choice*:

$x \in \mathcal{M}$

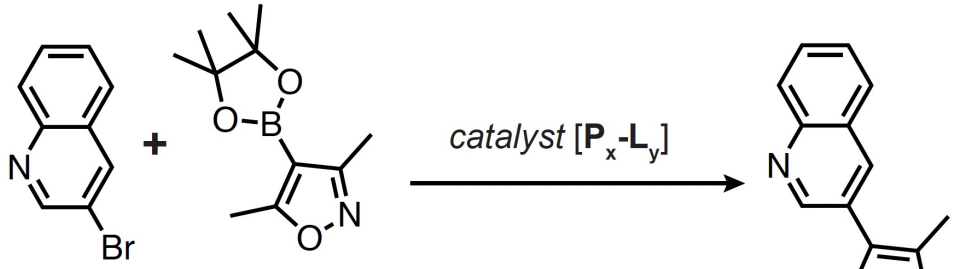
*where $\mathcal{M}$ denotes the valid set of molecular structures

some examples of chemical representations



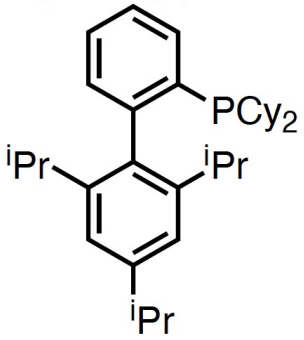
J. Phys. Chem. A, **2023**, 127, 8253

example chemical optimization problem

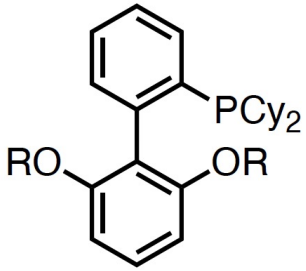


parameter
catalyst [P_x-L_y]
catalyst loading
temperature
residence time

ligands (L)

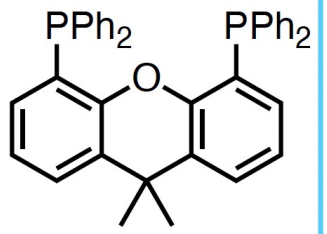


L1



L2 (R = Me)

L3 (R = ^{i}Pr)



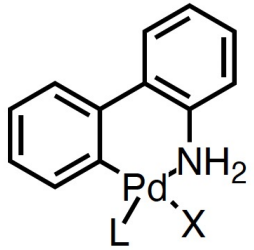
L4

L5 PCy_3

L6 PPh_3

L7 P^tBu_3

precatalyst scaffolds (P)



P1 (X = OMs)

P2 (X = Cl)

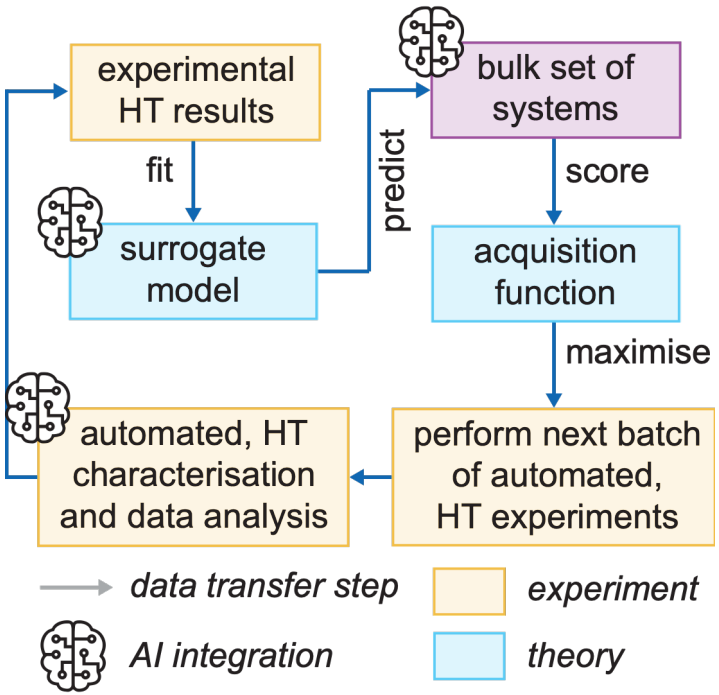
chemistry-specific considerations

objective

- maximise, minimise, combination?
- new material
- improved performance
- experimental conditions

information streams

- experiment, theory, combination?
- timescales
- batch?



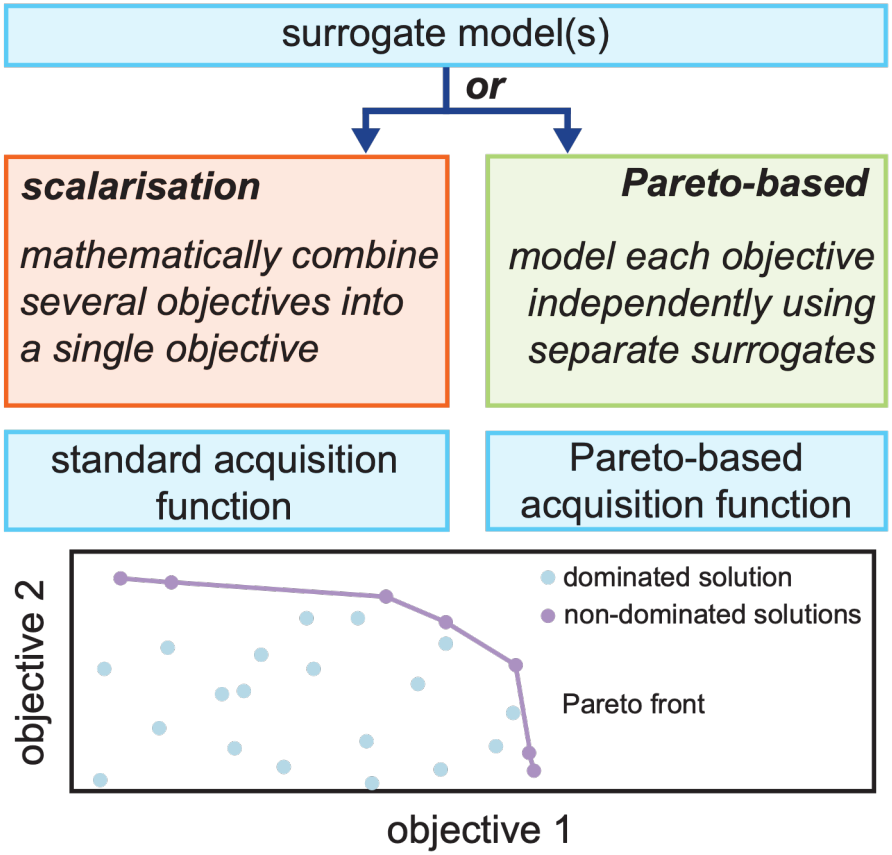
design space

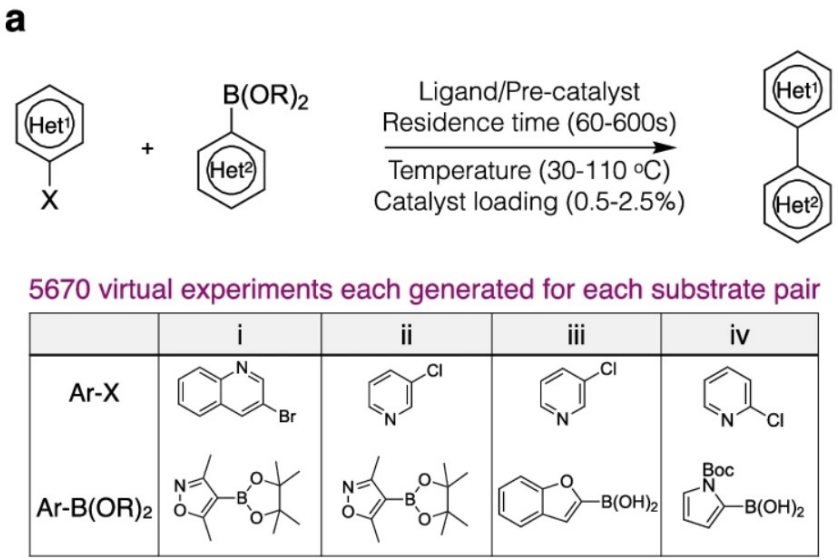
- integer, continuous, categorical, chemical?
- chemical representation / encoding

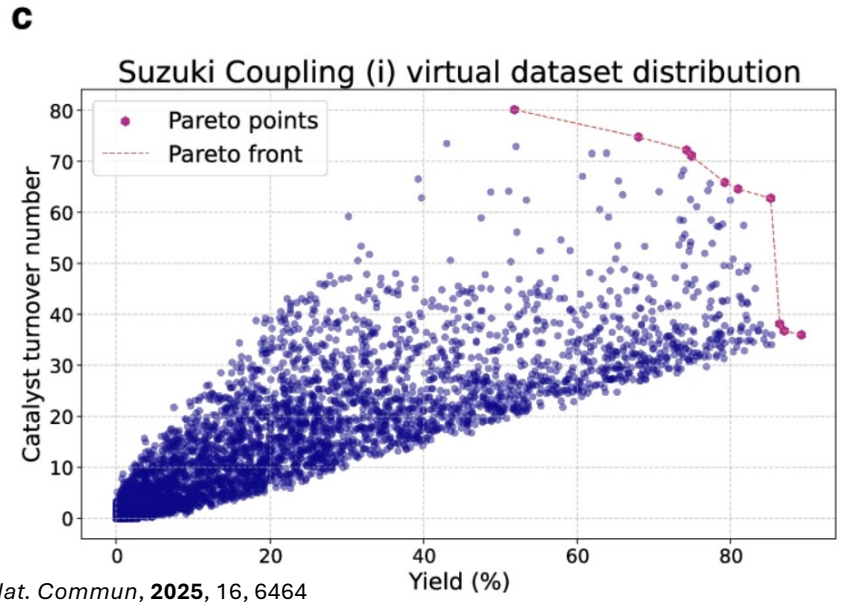
stopping criteria

- precursor amounts
- number of experiments
- total time

multi-objective BO







Nat. Commun, **2025**, 16, 6464

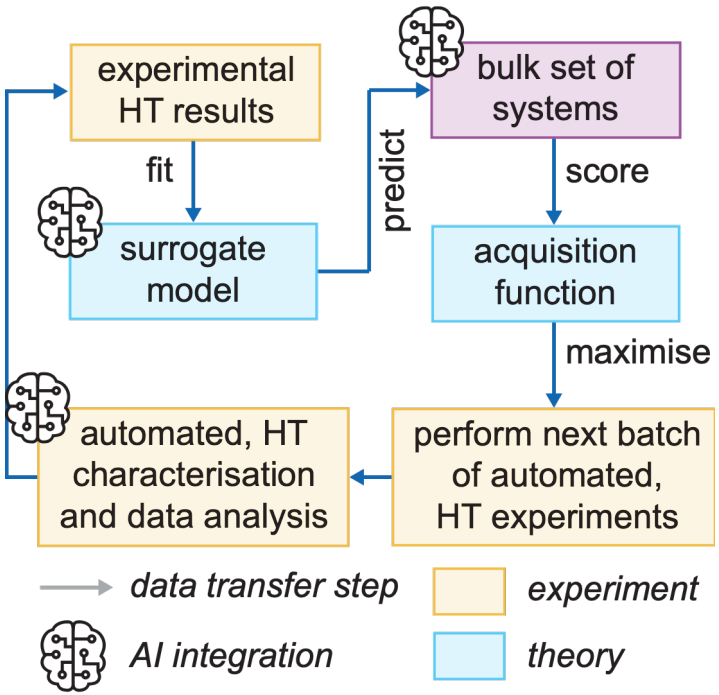
chemistry-specific considerations

objective

- maximise, minimise, combination?
- new material
- improved performance
- experimental conditions

information streams

- experiment, theory, combination?
- timescales
- batch?



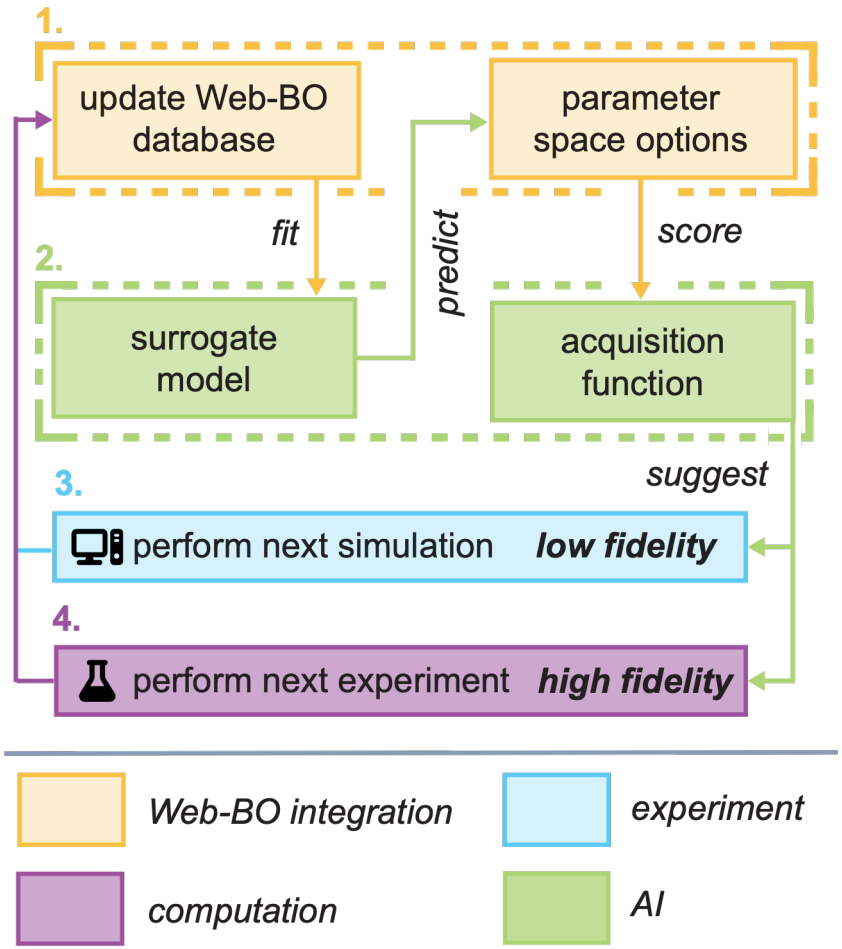
design space

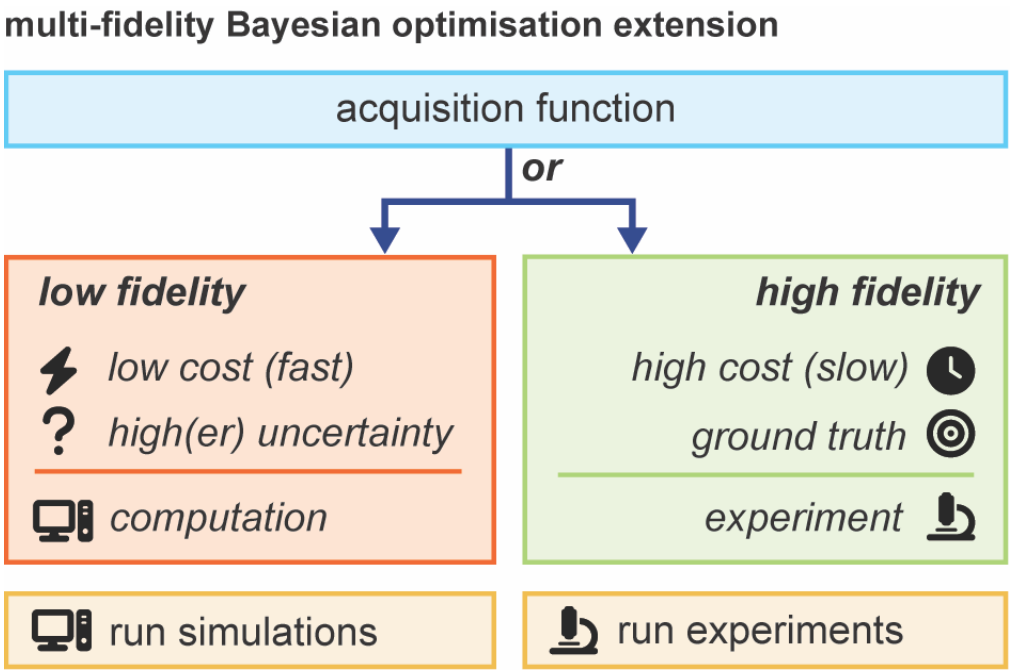
- integer, continuous, categorical, chemical?
- chemical representation / encoding

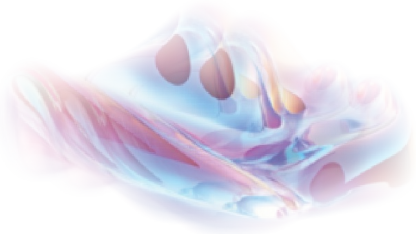
stopping criteria

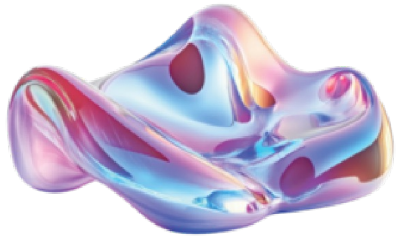
- precursor amounts
- number of experiments
- total time

multi-fidelity BO



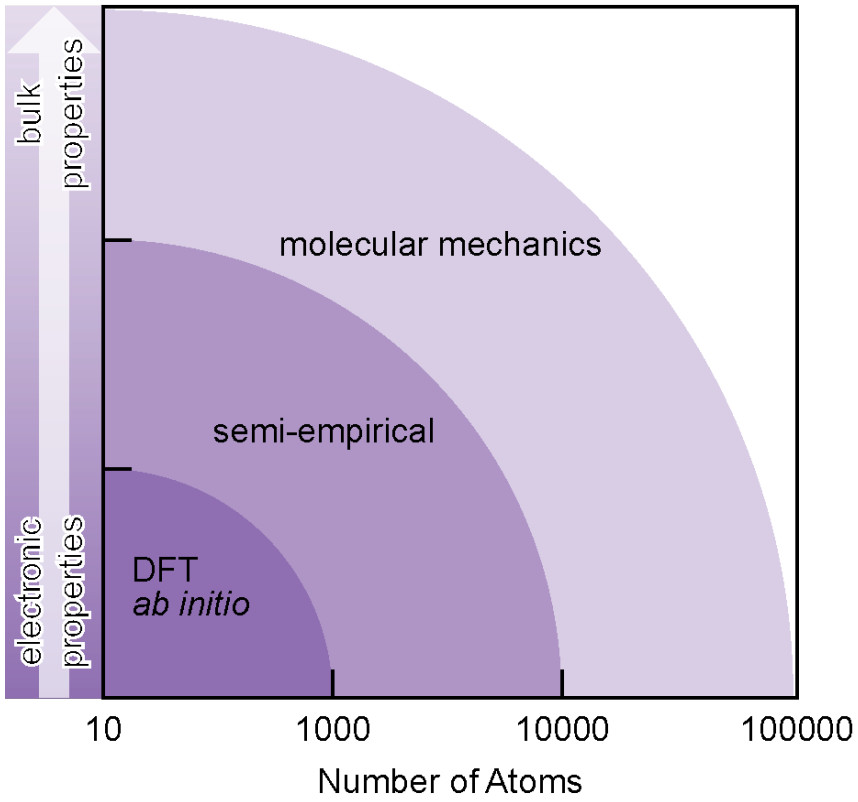


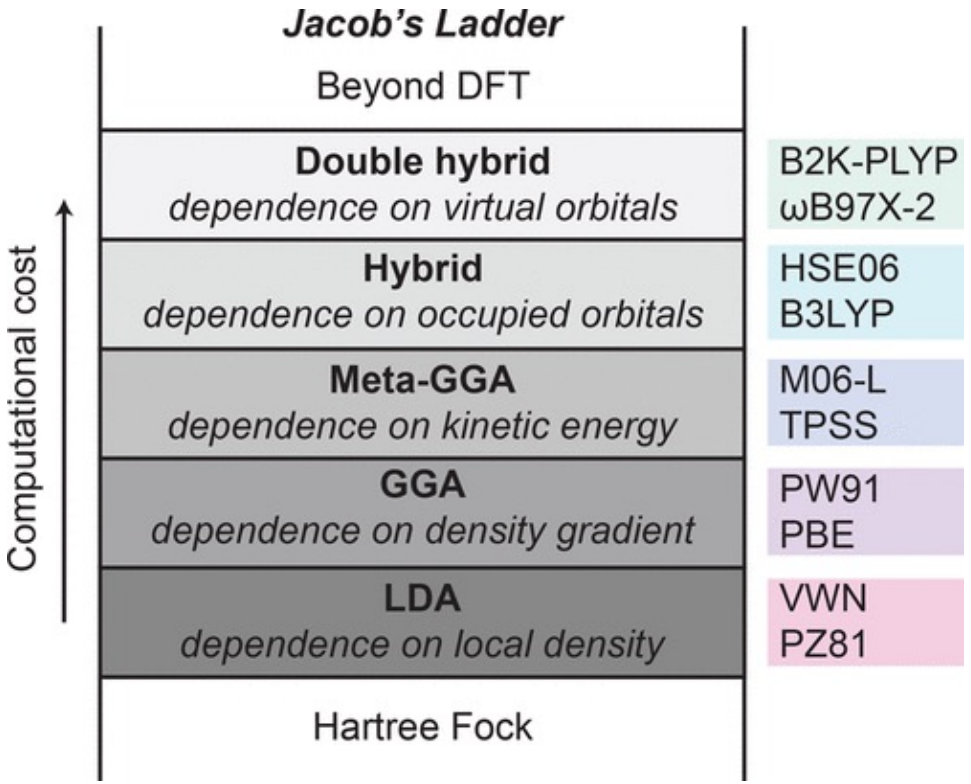




J. Phys. Chem. A, **2023**, 127, 8253

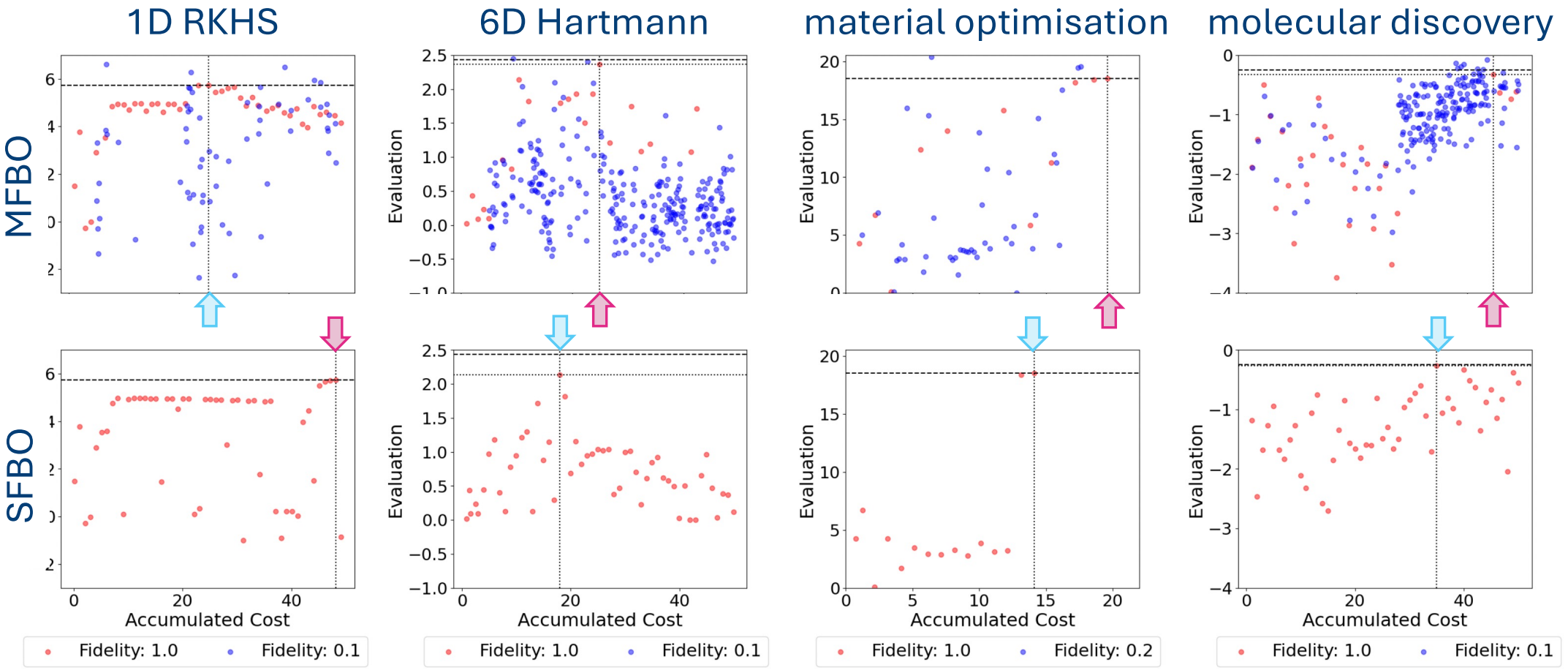
we can also exploit different levels of theory





Chem. Rev., **2020**, 120, 16

MFBO vs SFBO



E. Judge, et al, NeurIPS 2024 Workshop on AI for Accelerated Materials Design

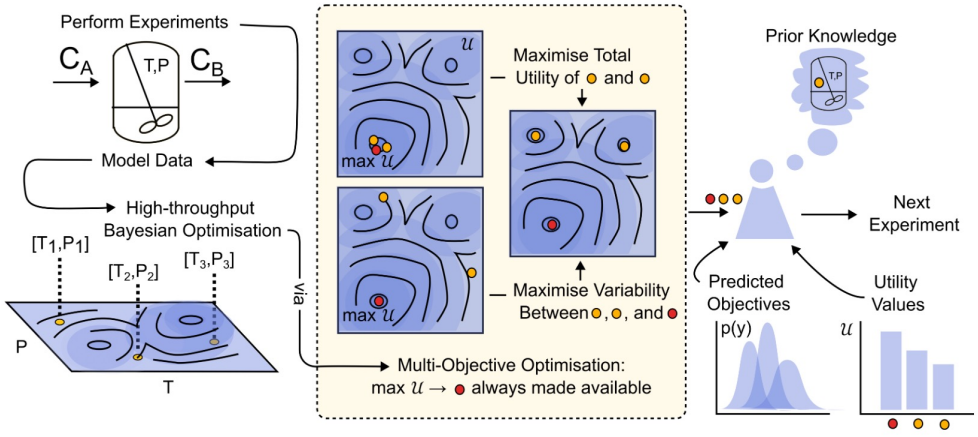
including expert knowledge

human-in-the-loop

aim - integrate expert knowledge into data-driven decision making

challenge - effectively incorporating domain-specific knowledge

how - multi-objective approach 3 step approach – expert-informed initial design, batch BO, human decision-making multi-objective formulation for distinct solutions



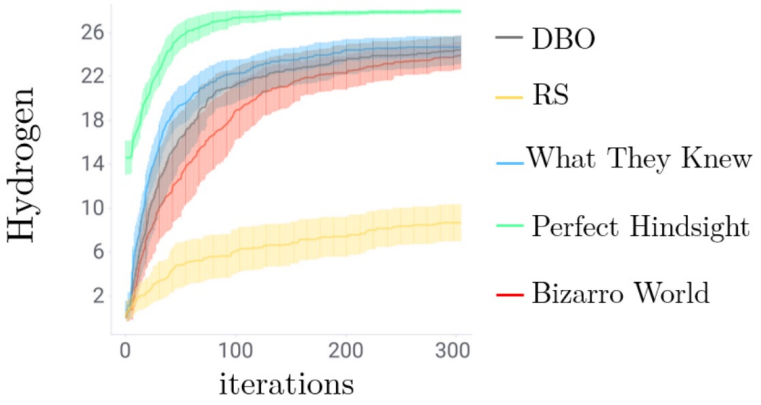
Comput. and Chem. Eng., **2024**, 189, 108810

ML-in-the-loop

aim - optimize black-box scientific experiments using expert hypotheses

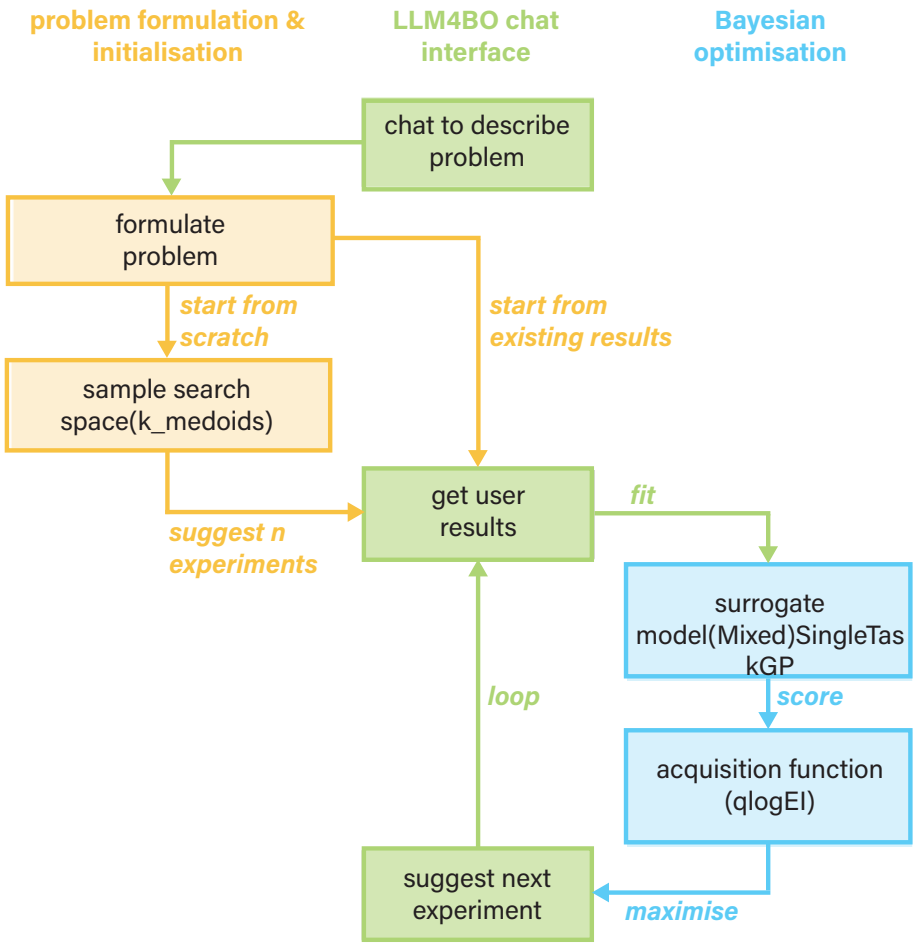
challenge - effectively incorporating domain-specific knowledge

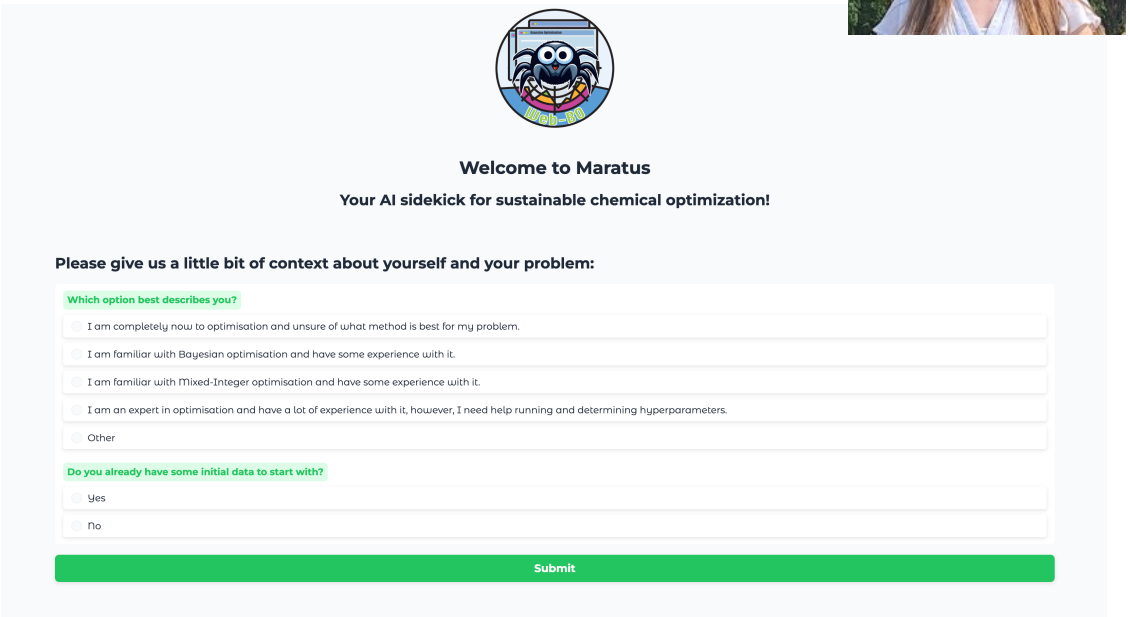
how - use expert hypotheses as intervals of confidence local GPs are used for each hypothesis – these are then evaluated by a global GP (bilevel optimization framework)



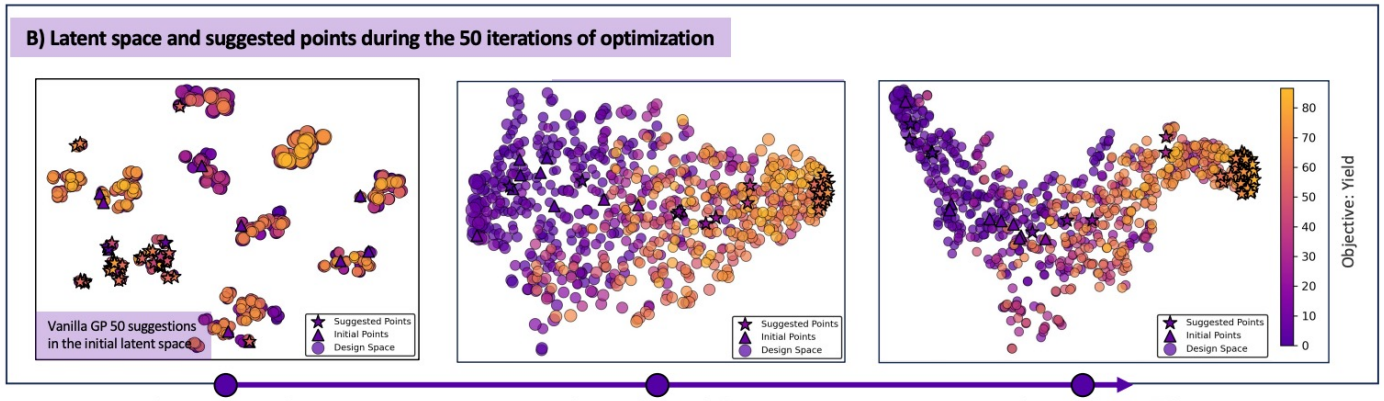
arXiv:2308.11787v3

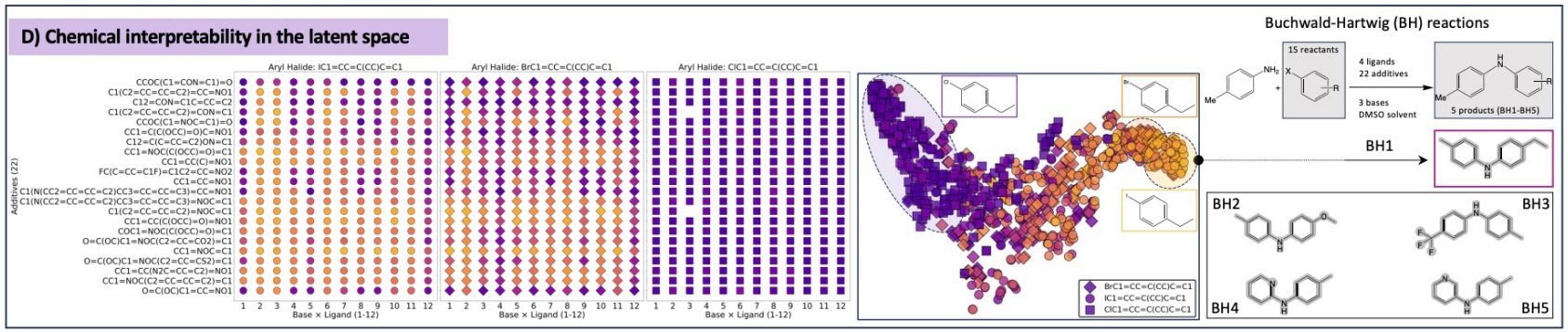
discussing your problem formulation with Web-BO

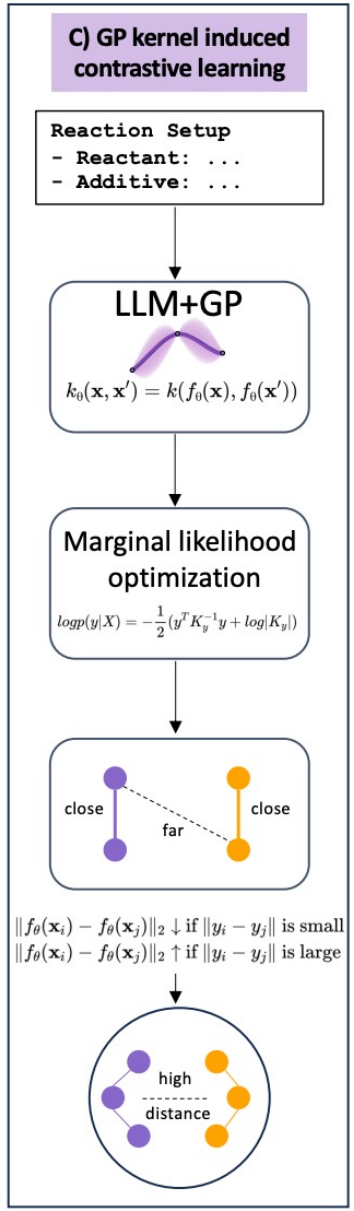




LLMs for parameter embeddings







GOLLuM, **2025**, https://openreview.net/forum?id=2ORViHAUbf

project outcomes

explore **chemical representation strategies** with

BayBE
Bayesian Back End

- **How do different catalyst encodings affect optimization?**
  - What are the conceptual differences between OHE, fingerprints, and Mordred descriptors, and why might they lead to different BO performance?
  - How does this change the BO behaviour and the quality/speed of the optimization?
- **How does the choice and number of initial experiments affect:**
  - the speed at which yield improves, and
  - the best yield reached within ~200 experiments?
- **How would you extend to a multi-objective problem?**
  - If you also want to maximize TON, how would you change the objective / targets in BayBE?
  - What kind of multi-objective strategy would you use, and how would you interpret a Pareto front of yield vs TON?

explore **domain & sampling strategies** with

BoFire

- **Are the chosen input bounds (temperature, catalyst loading, residence time) and categorical options (catalysts) chemically sensible?**
- **How do different choices of this initial random set affect:**
  - the quality of the first surrogate model, and
  - the subsequent BO trajectory?
  - Can you implement and diagnose the full BO loop?
- **How would you extend to a multi-objective problem?**
  - If you also want to maximize TON, how would you change the objective / targets in BayBE?
  - What kind of multi-objective strategy would you use, and how would you interpret a Pareto front of yield vs TON?