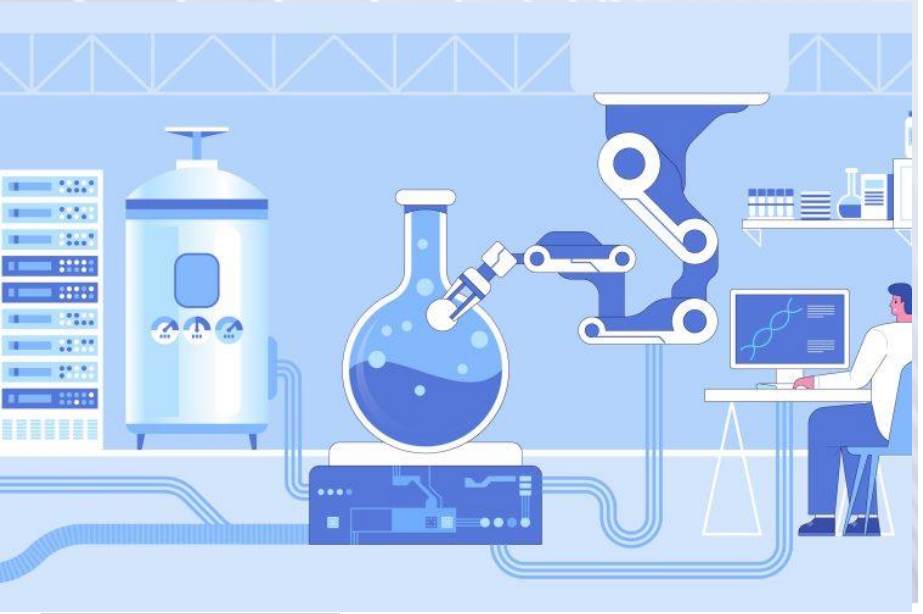
Sargent Centre for Process Systems Engineering



BO for Guided Chemical Design

Austin Mroz

Imperial College London

a.mroz@imperial.ac.uk

UCL IMPERIAL Schmidt Sciences Imperial-X

Sargent Centre for Process Systems Engineering

Agenda

time	subject	learning objectives
14:00-14:20	Why BO for chemistry	decision making in chemistry from OFAT & DOE to BO
14:20-14:30	Chemistry-specific considerations	surrogate model selection & chemical representations
14:30-14:40	Web-BO walkthrough	introduction to tool and notebooks we will be exploring
14:40-15:20	***interactive***	chemical representation comparison with SUMMIT csv
15:20-15:30	Web-BO: what's happening on the backend?	SOBO code walkthrough
15:30-16:00	*break*	

time	subject	learning objectives
16:00-16:15	Complex BO formulations in chemistry	MFBO, MOBO, MF-MO BO, etc.
16:15-17:00	***interactive***	code-your-own-adventure – ***MOBO, MFBO, GPs for molecules***
17:00-17:20	Complex BO formulations in chemistry	Let's dig into the literature
17:20-17:30	wrap-up discussion	summary and additional resources

complex BO formulations in chemistry

chemistry-specific considerations

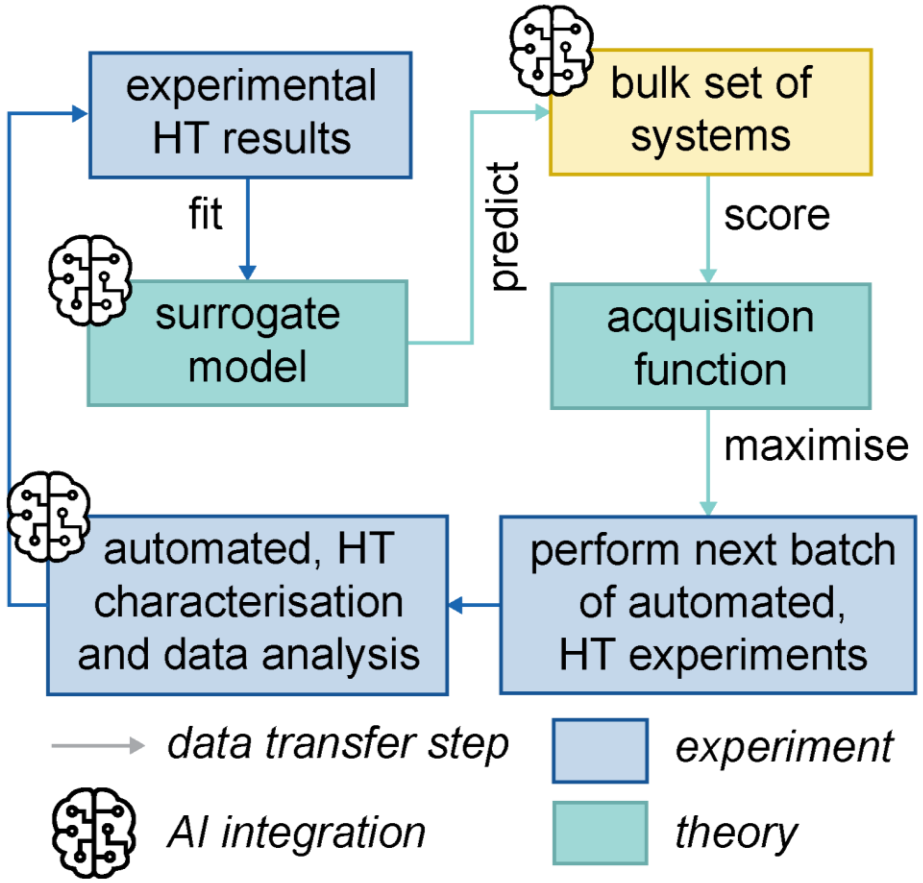
objective

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

multi-objective

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

CHEMISTRY OF MATERIALS

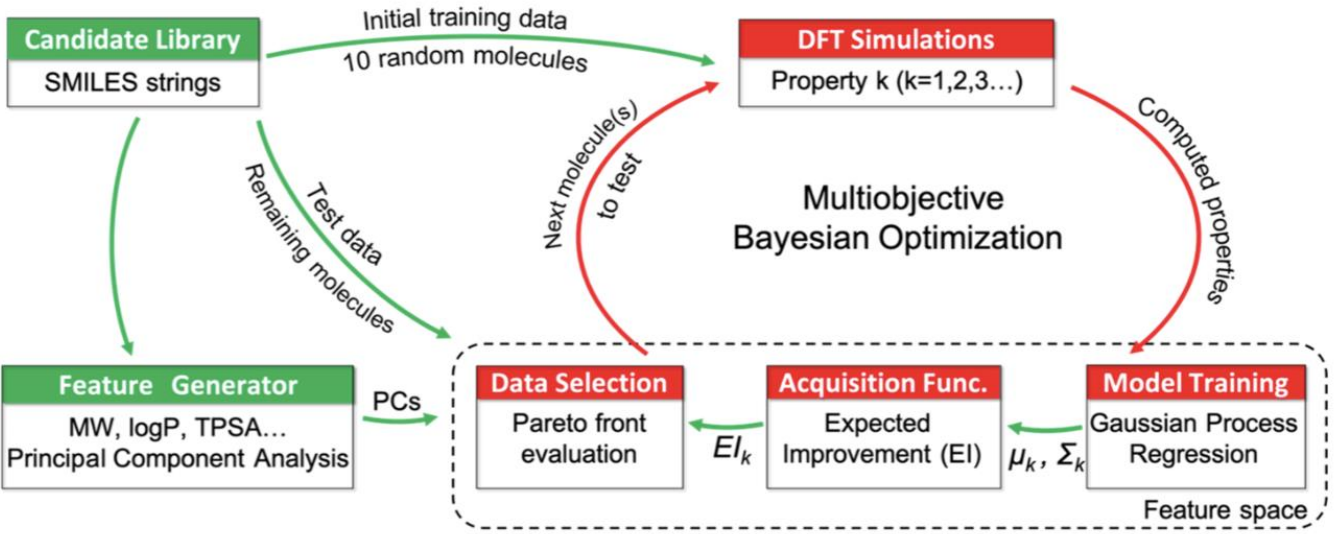
pubs.acs.org/cm Article

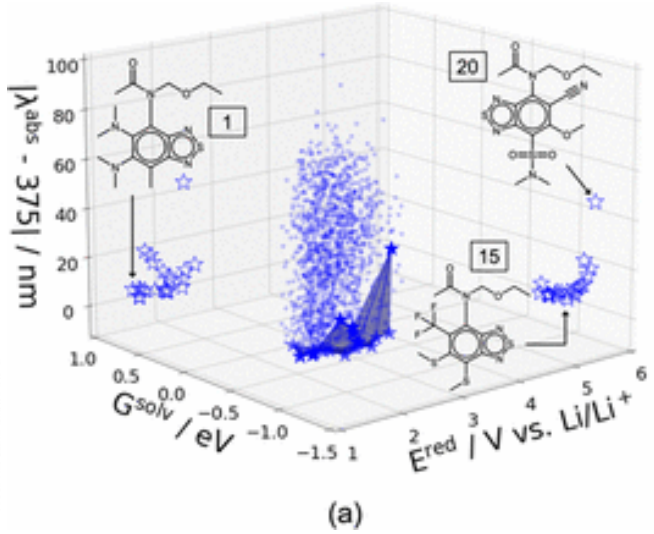
Discovery of Energy Storage Molecular Materials Using Quantum Chemistry-Guided Multiobjective Bayesian Optimization

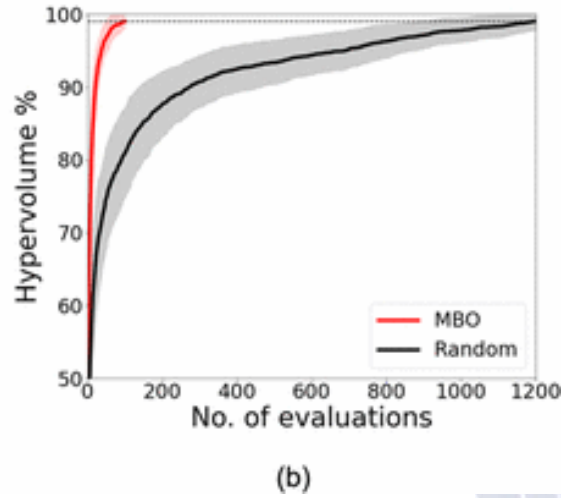
Garvit Agarwal, Hieu A. Doan, Lily A. Robertson, Lu Zhang, and Rajeev S. Assary*

Cite This: *Chem. Mater.* 2021, 33, 8133–8144 Read Online

challenge efficient discovery of redox-active molecules (redoxmers) for non-aqueous redox flow batteries (NRFBs) requires simultaneously optimize multiple properties: *reduction potential, solvation free energy, and absorption wavelength.* Traditional methods for identifying such molecules are computationally expensive and time-consuming.







Limitations

1.The study does not consider the stability of anolyte molecules, which is an important criterion for designing long-duration RFBs

2.The actual solubility measurements using computations are not possible due to the lack of sublimation energies of the molecular materials

3.Many of the suggested Pareto-optimal BzNSN molecules are complex and difficult to synthesize, which may limit their practical applicability

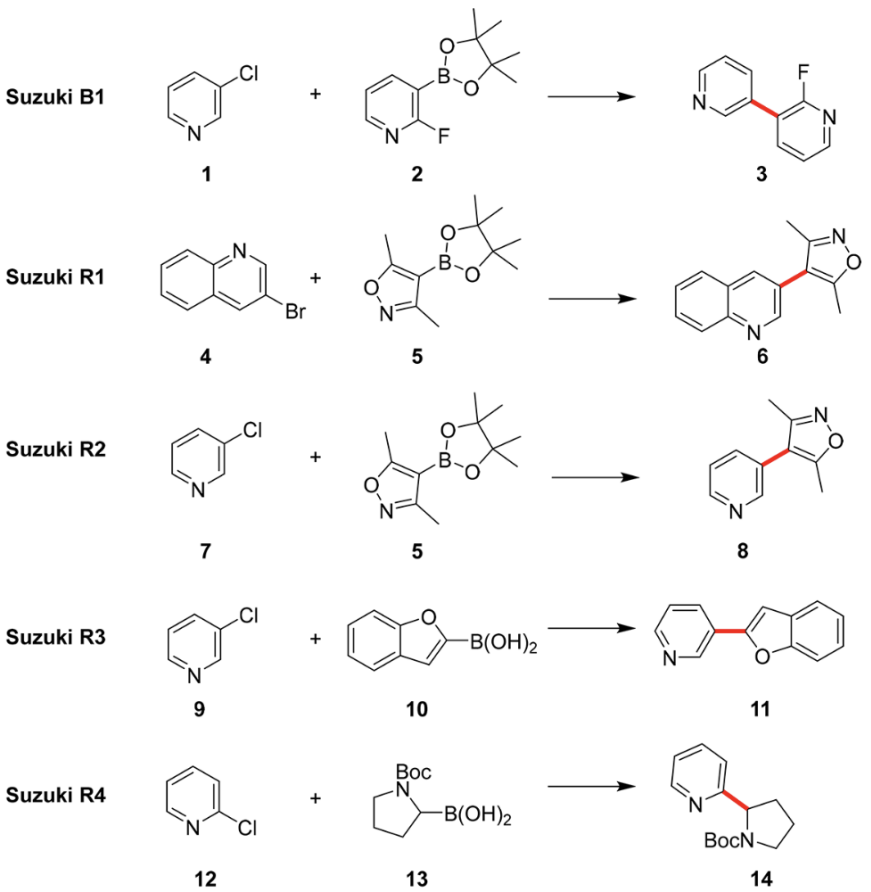
ACS central science
http://pubs.acs.org/journal/acscii
Research Article

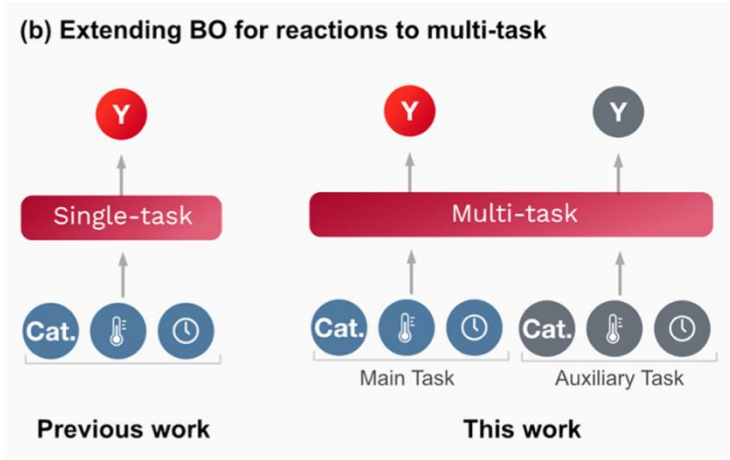
Accelerated Chemical Reaction Optimization Using Multi-Task Learning

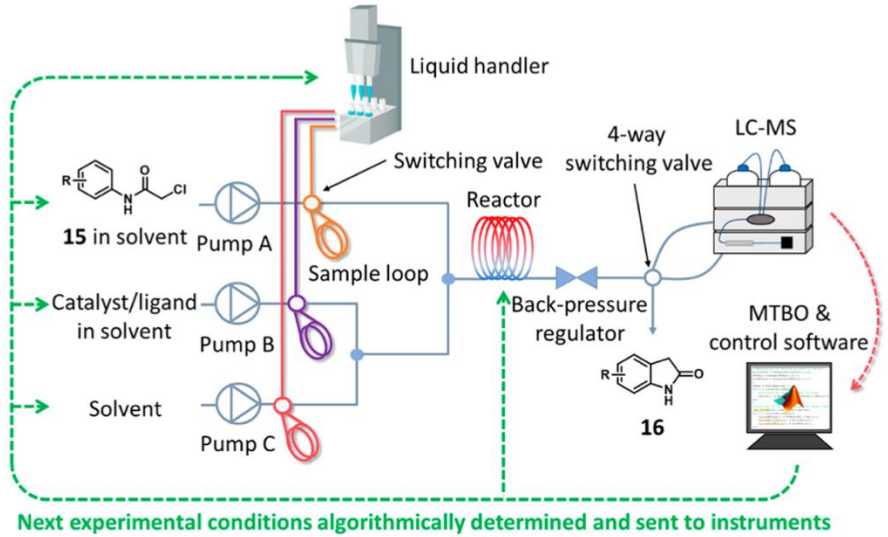
Connor J. Taylor,*,# Kobi C. Felton,# Daniel Wigh, Mohammed I. Jeraal, Rachel Grainger, Gianni Chessari, Christopher N. Johnson, and Alexei A. Lapkin*

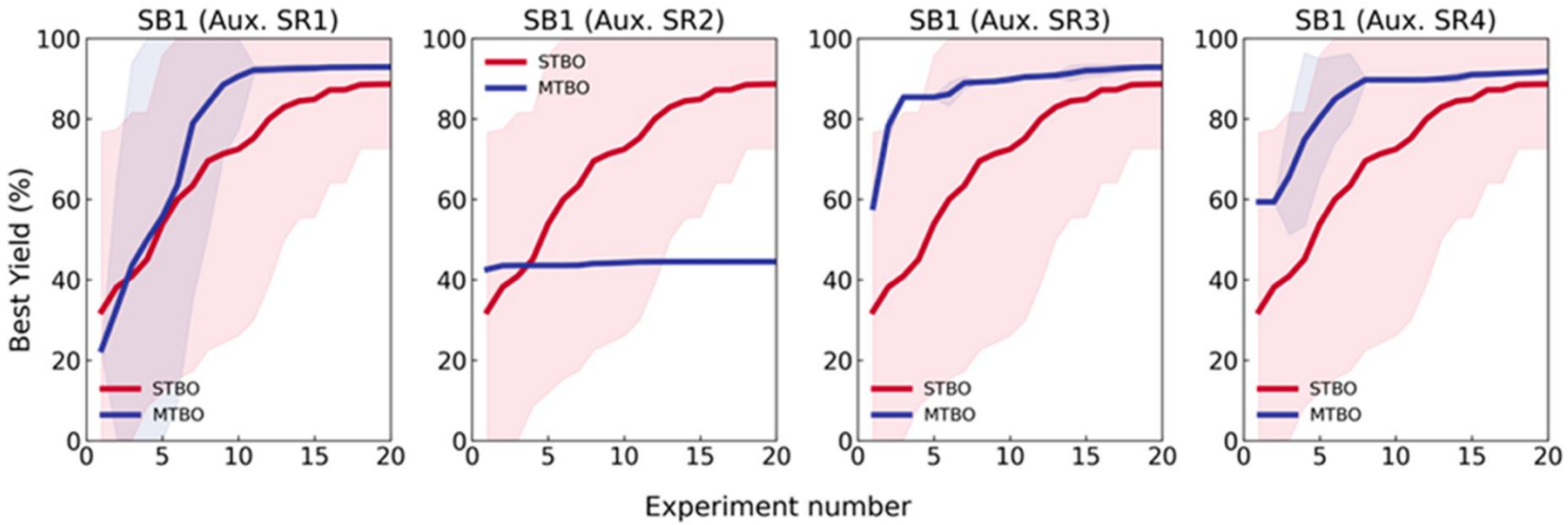
Cite This: ACS Cent. Sci. 2023, 9, 957–968
Read Online

aim – leverage pre-existing chemical knowledge to reduce the number of experiments and material required for optimizing pharmaceutically relevant transformations, specifically C-H activation reactions to form oxindoles.









chemistry-specific considerations

objective

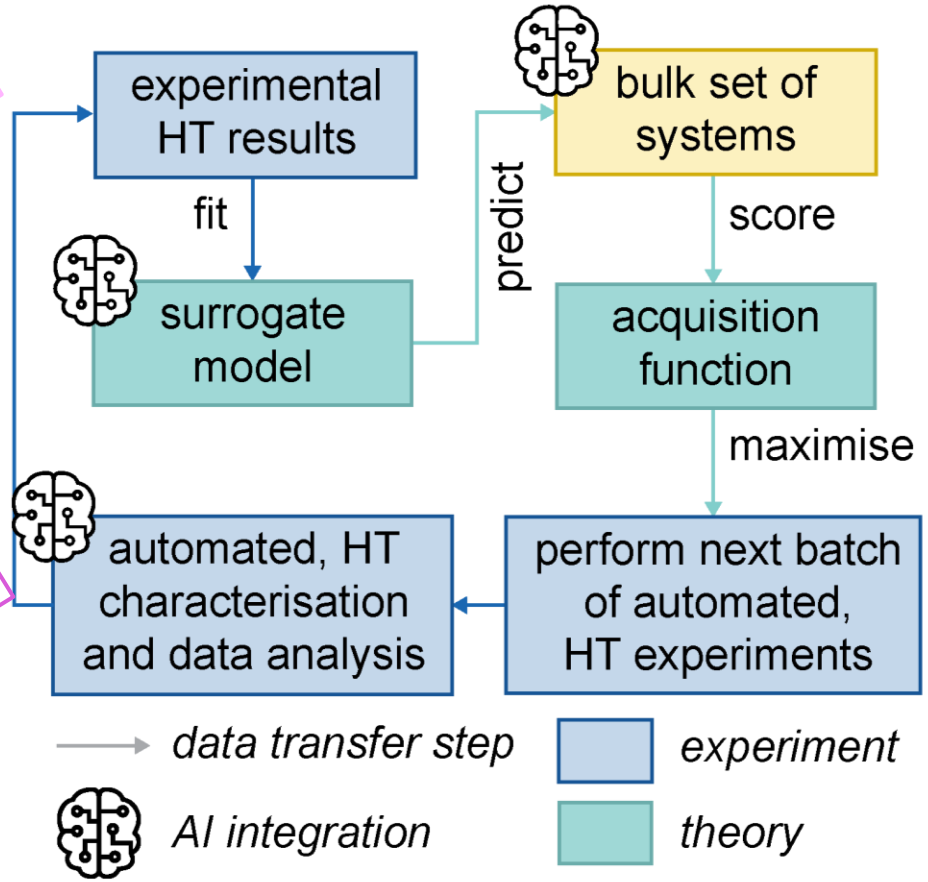
- maximise, minimise, combination
- new material
- improved performance
- experimental conditions

multi-objective

information streams

- experiment, theory, combination?
- time scales
- batch?

multi-fidelity



design space

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

stopping criteria

- precursor amounts
- number of experiments
- total time

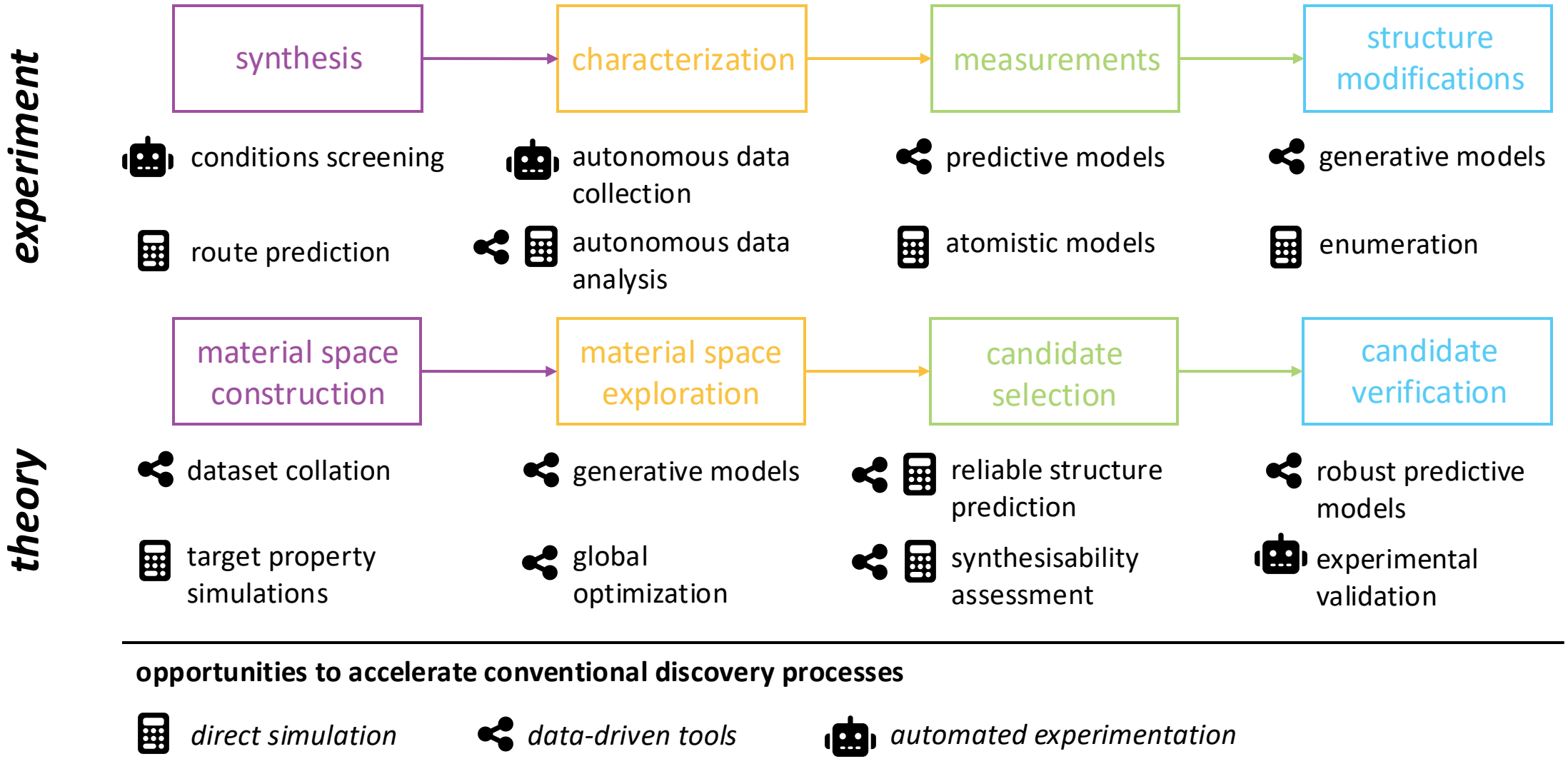
information streams and the advantages of theory

opportunities to accelerate conventional discovery processes

direct simulation *data-driven tools* *automated experimentation*

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

information streams and the advantages of theory

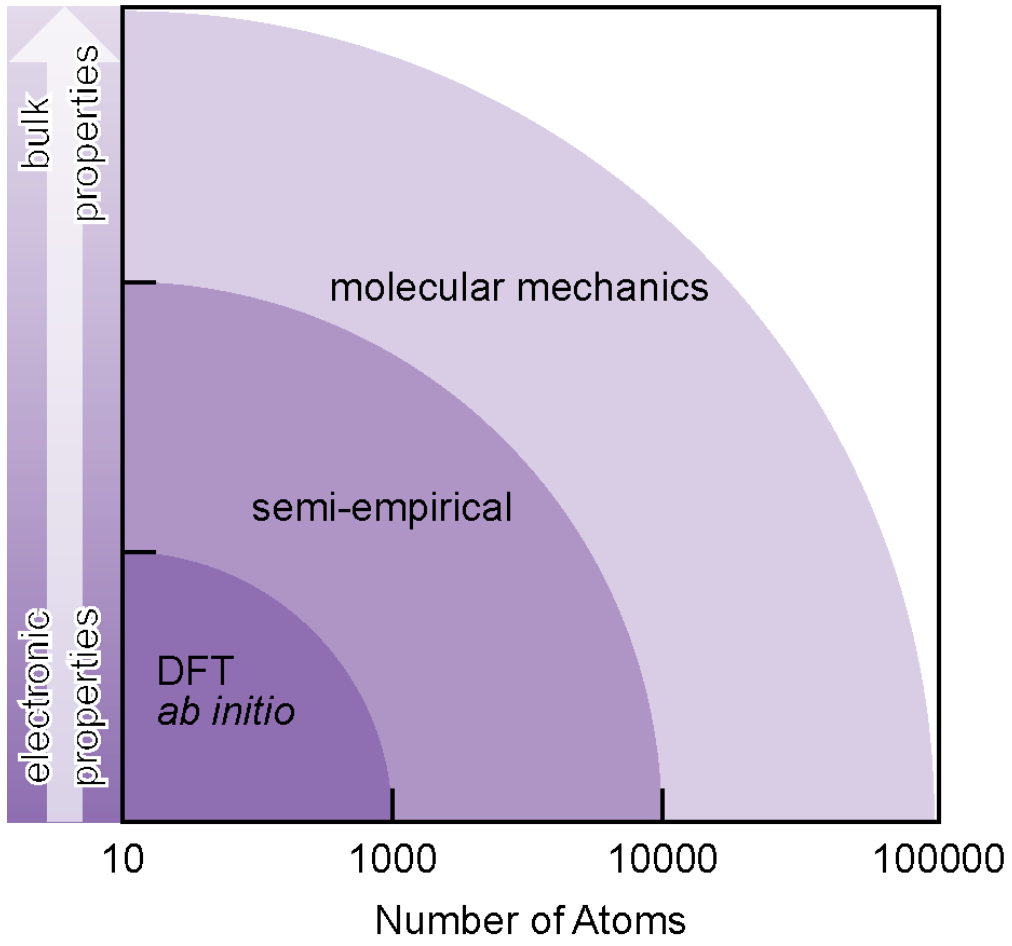


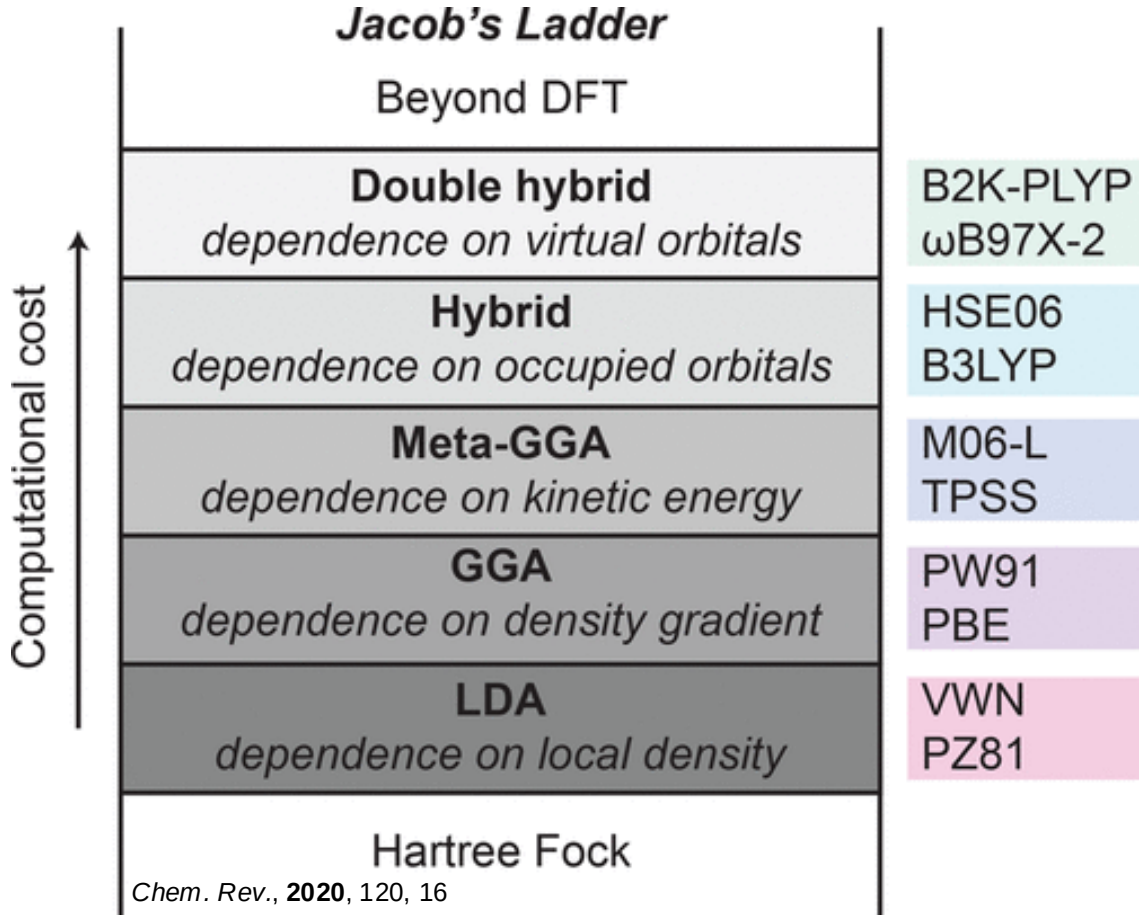
opportunities to accelerate conventional discovery processes

direct simulation *data-driven tools* *automated experimentation*

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

levels of computation



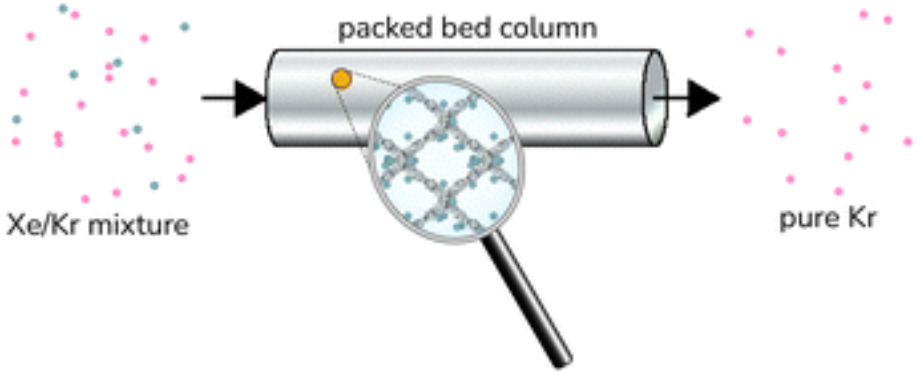


Chem. Rev., **2020**, 120, 16

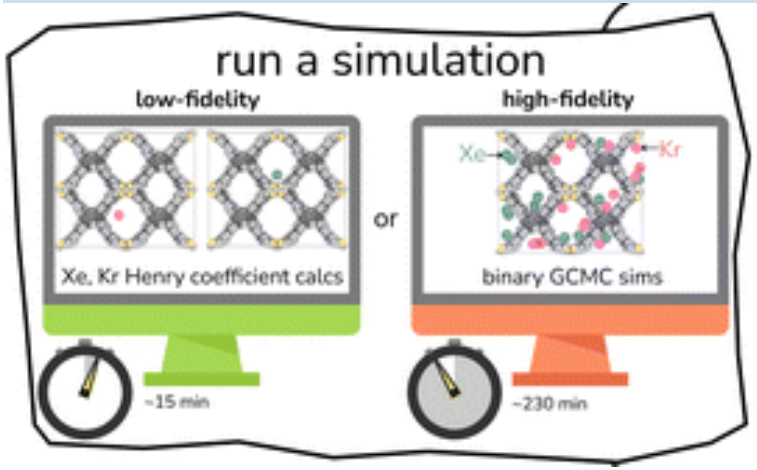
Sargent Centre for Process Systems Engineering

multi-fidelity BO: COFs

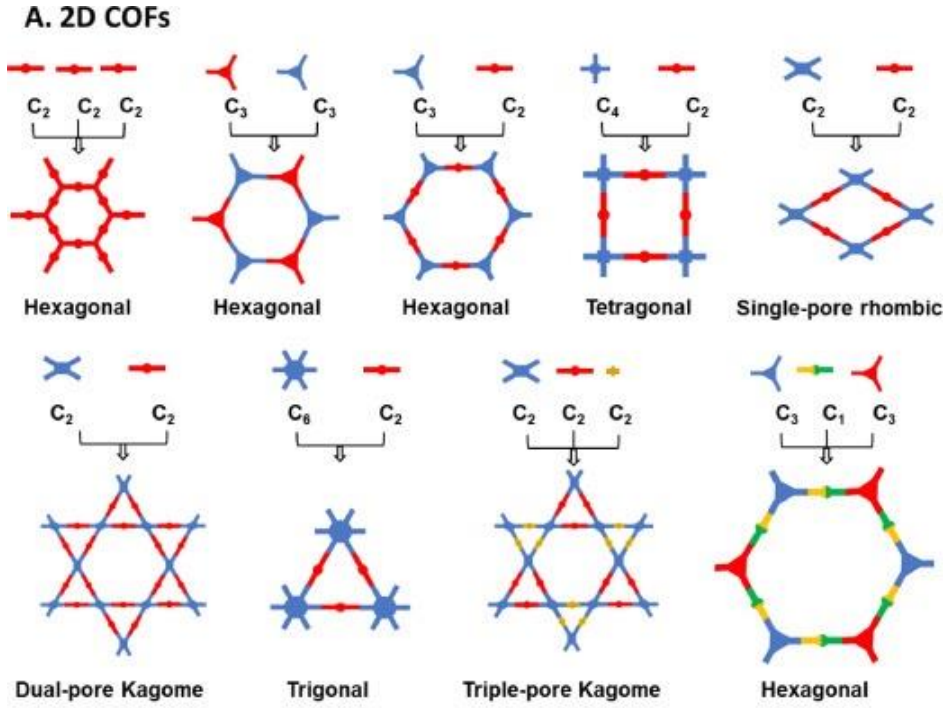
objective



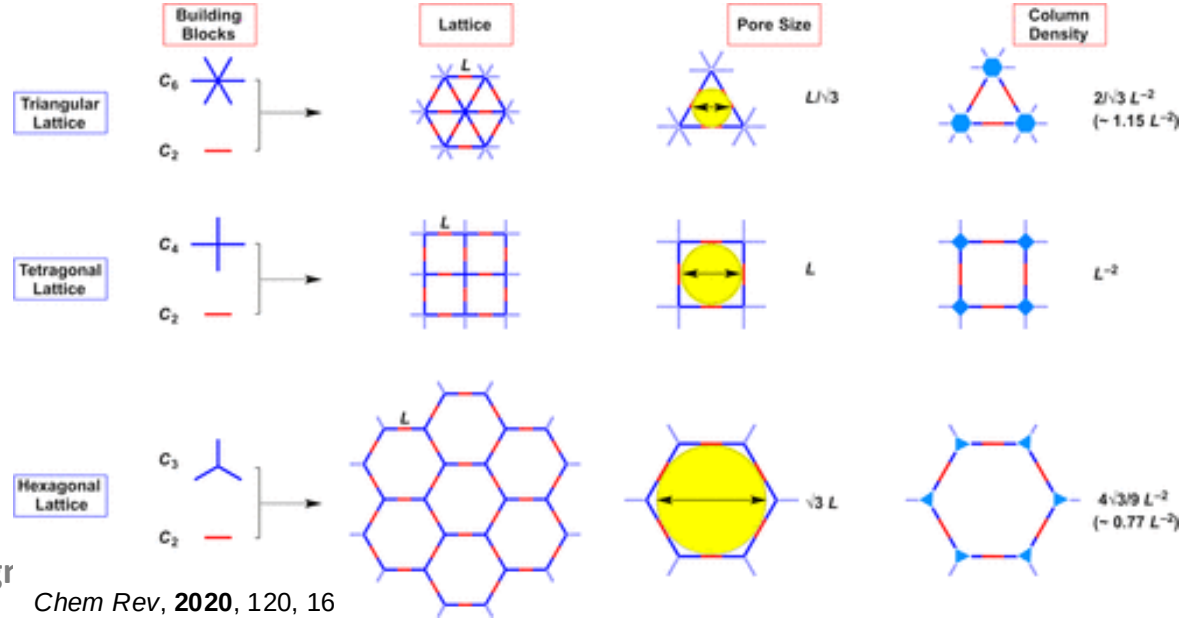
information streams



Digital Discovery, **2023**, 2, 1937



Giant, **2021,** 6, 100054



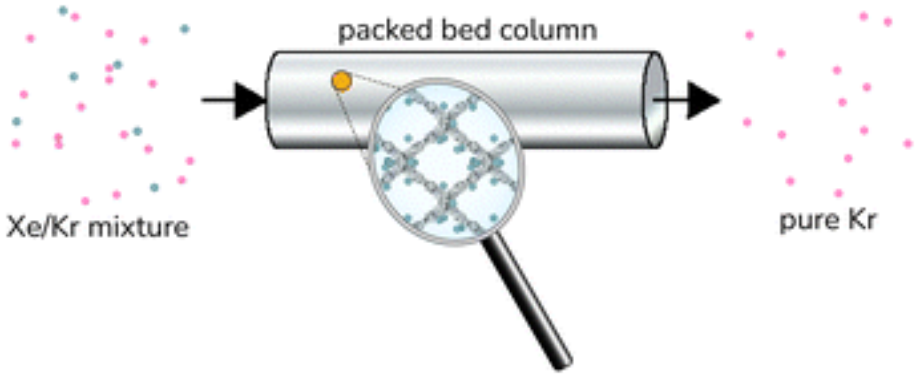
Chem Rev, **2020**, 120, 16

UCL IMPERIAL

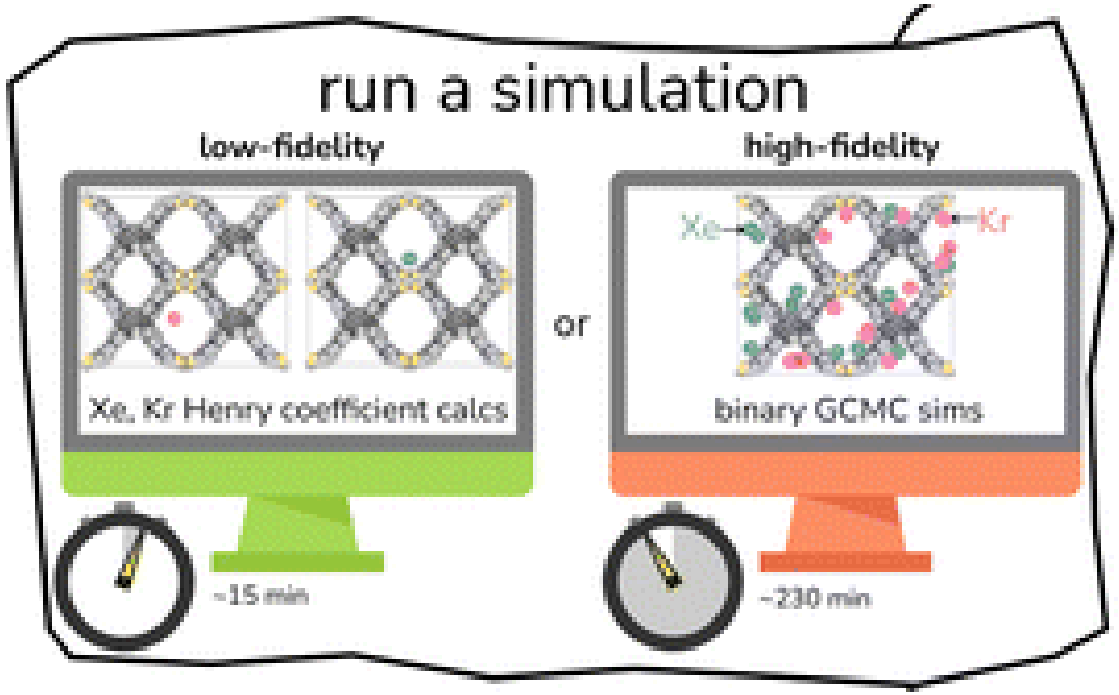
BO for Guided Chemical Design

multi-fidelity

objective

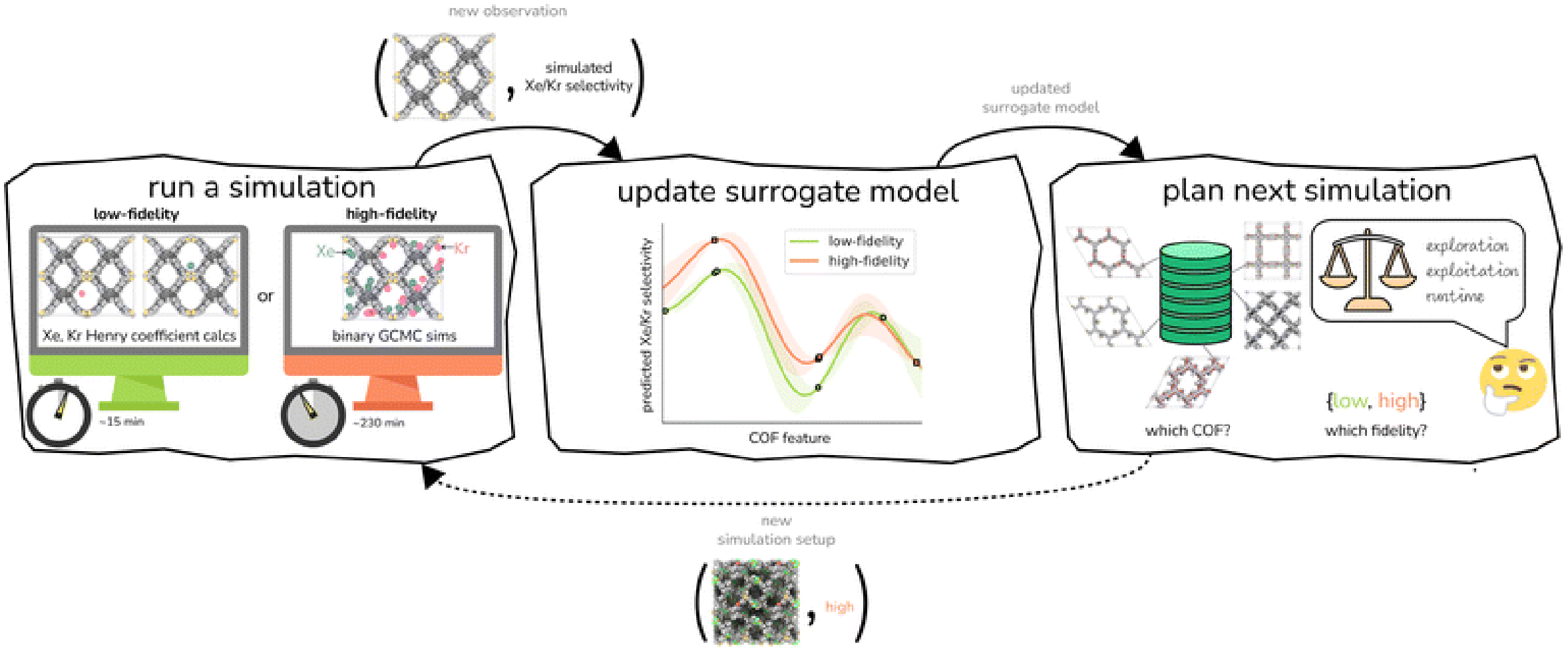


information streams



Digital Discovery, **2023**, 2, 1937

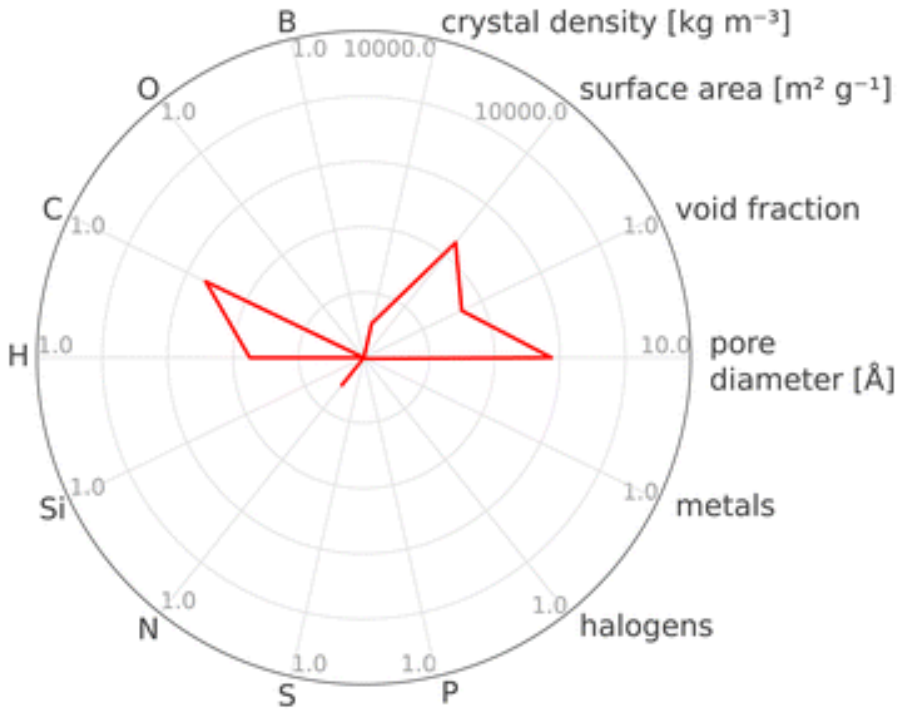
multi-fidelity



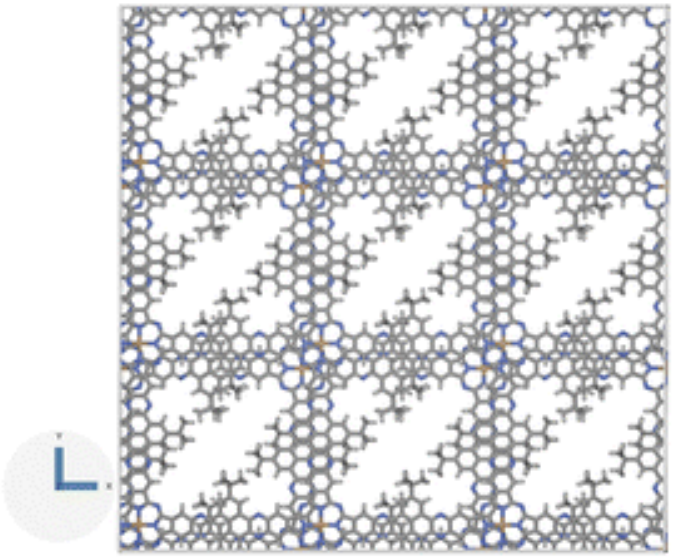
Digital Discovery, **2023**, 2, 1937

multi-fidelity

COF representation



(a) top COF - feature vector



(b) 19440N2

we expect similar performance from COFs with similar feature vectors

Digital Discovery, **2023**, 2, 1937

multi-fidelity

acquisition function – augmented, cost-aware EI

$$\left(\mathbf{x}_{[n+1]}, \ell_{[n+1]}\right) = \underset{(\mathbf{x},\ell)\in\mathcal{X}\times\{1/3,2/3\}}{\arg\max} \mathbb{E}\left[\max\left[0, Y^{(2/3)}(\mathbf{x})\middle|\mathcal{D}_{[n]} - \hat{y}^{(2/3)*}_{[n]}\right]\right] \times \operatorname{corr}\left[Y^{(\ell)}(\mathbf{x})\middle|\mathcal{D}_{[n]}, Y^{(2/3)}(\mathbf{x})\middle|\mathcal{D}_{[n]}\right] \times \left(\frac{\tau^{(2/3)}_{[n]}}{\tau^{(\ell)}_{[n]}}\right).$$

EI – the amount that the high-fidelity simulation of COF **x** is expected to improve upon the best high-fidelity simulation we have performed

correlation with high-fidelity data – low-fidelity data is less correlated with the high-fidelity data that we are after

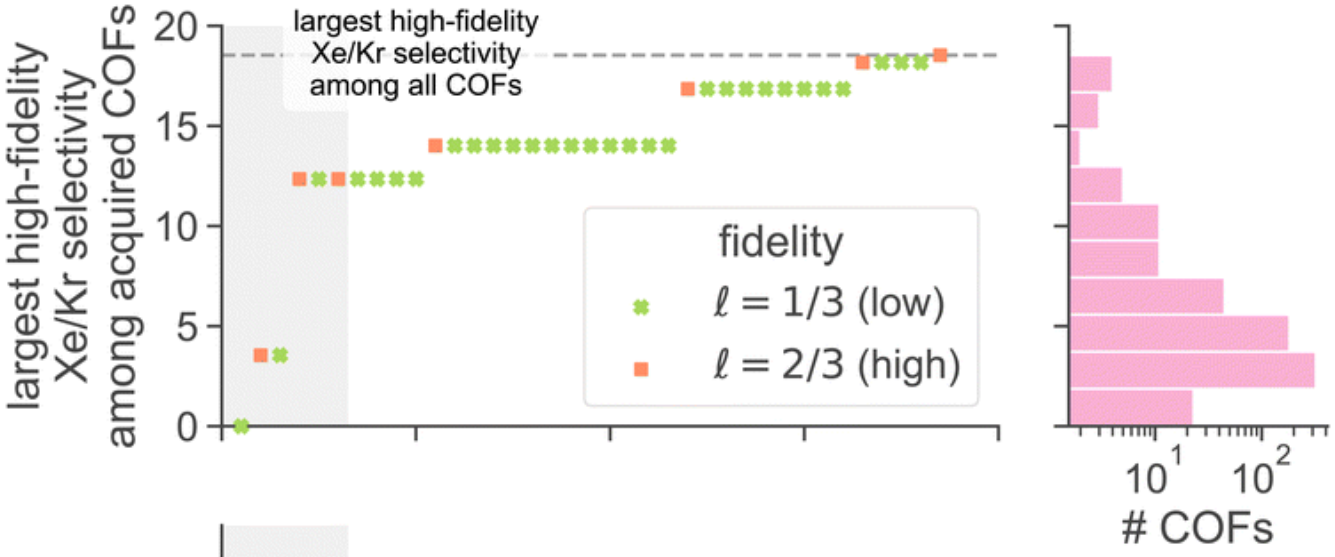
cost ratio – ratio of required simulation time; lower fidelity simulations are faster than high

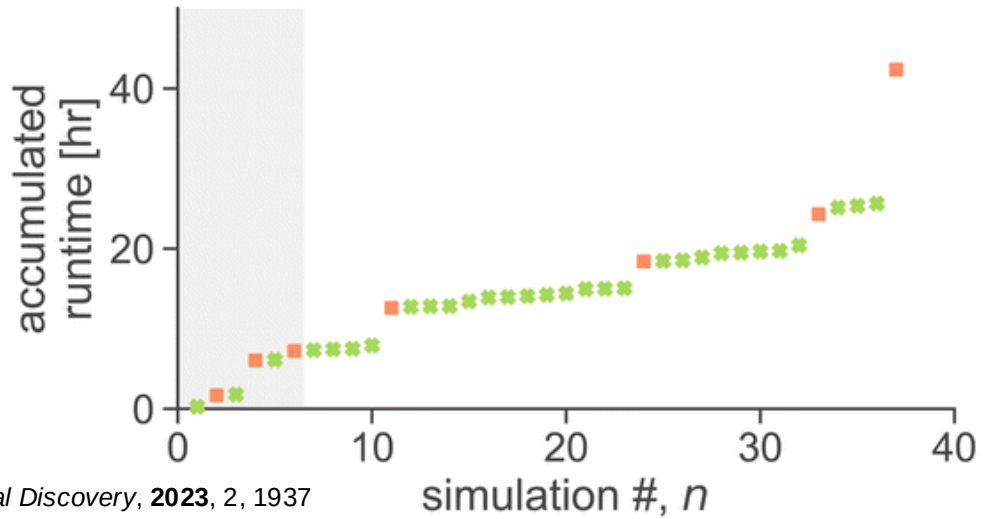
Digital Discovery, **2023**, 2, 1937

Sargent Centre for Process Systems Engineering

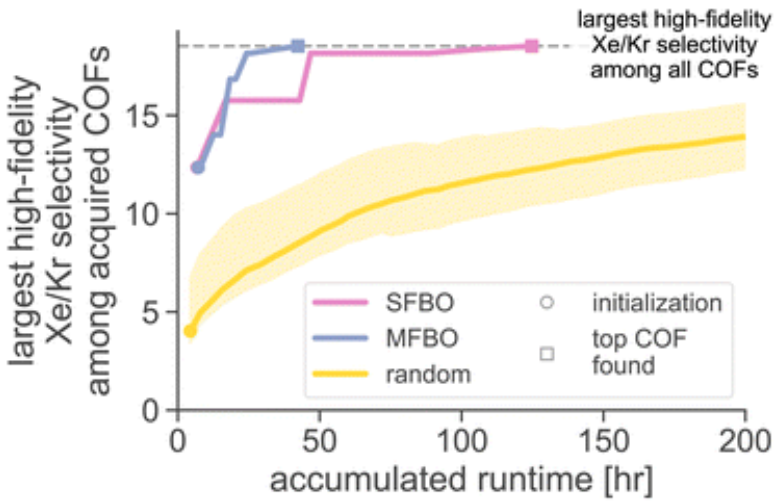
multi-fidelity

MFBO search efficiency

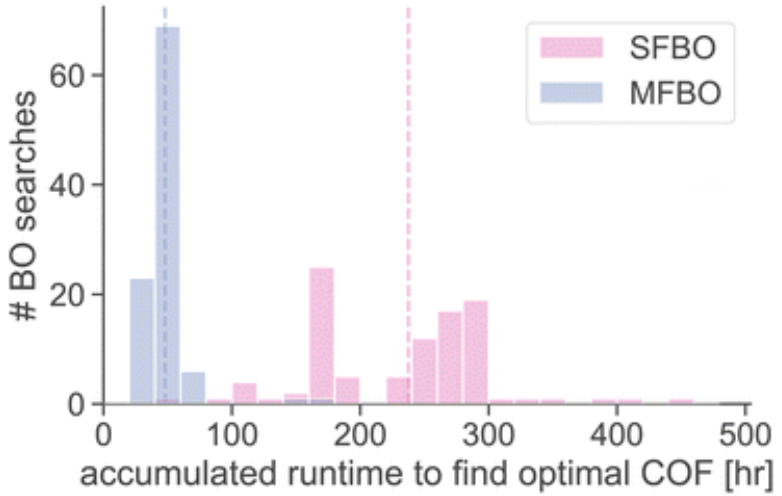




performance



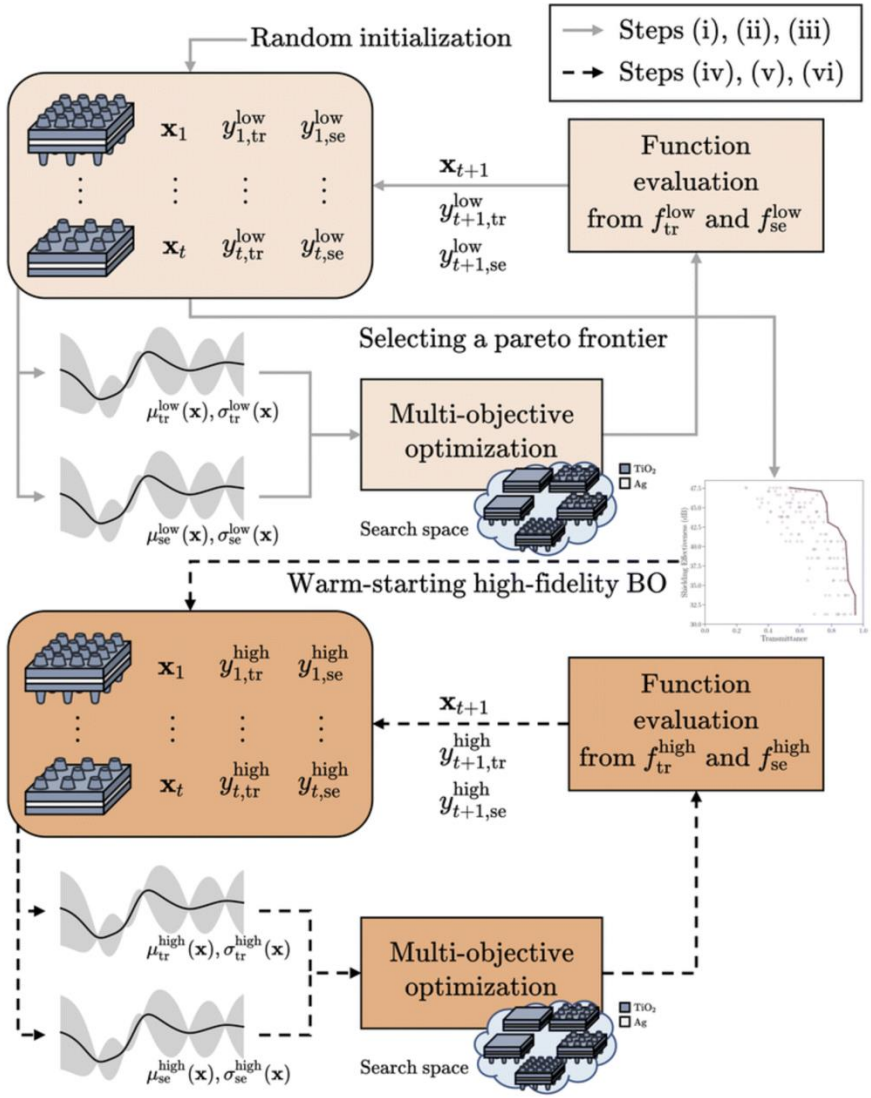
(a)



(b)

Sargent Centre for Process Systems Engineering

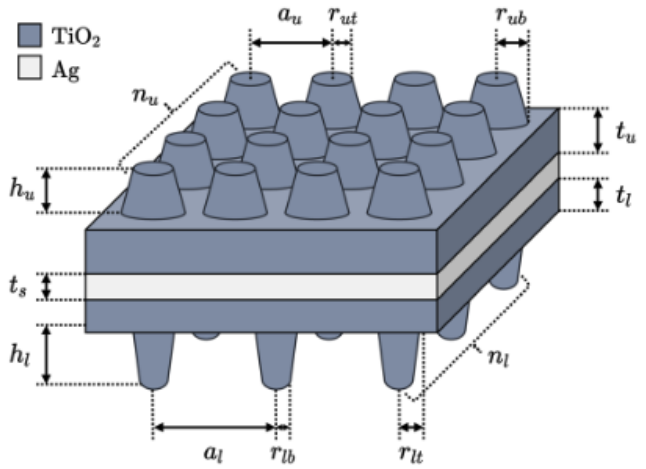
multi-fidelity multi-objective BO

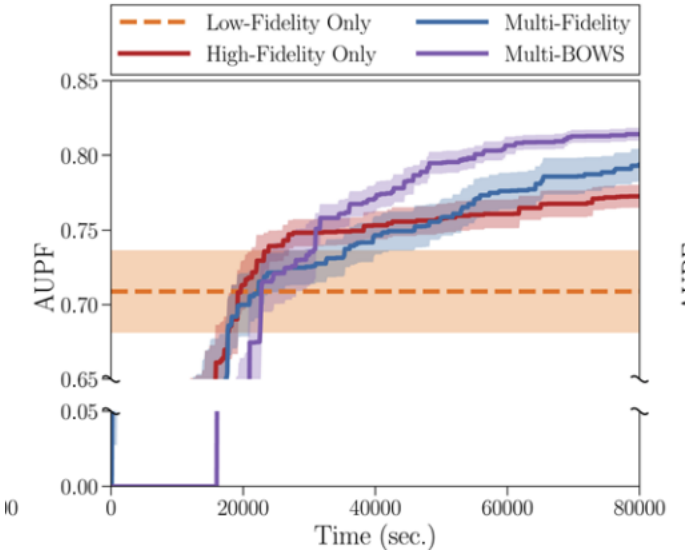


aim – design optical devices for transparent electromagnetic shielding
we need high visible transparency and good electromagnetic shielding

challenge – massive, constrained search space
high-fidelity (finer mesh) evaluations are resource intensive

how – use the Pareto frontier of low-fidelity data points to "kickstart" the high-fidelity BO
i.e. improve efficiency of search by ensuring they start with good initial high-fidelity points





Digital Discovery, **2024**, 3, 381

Sargent Centre for Process Systems Engineering

BO for latent space exploration

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

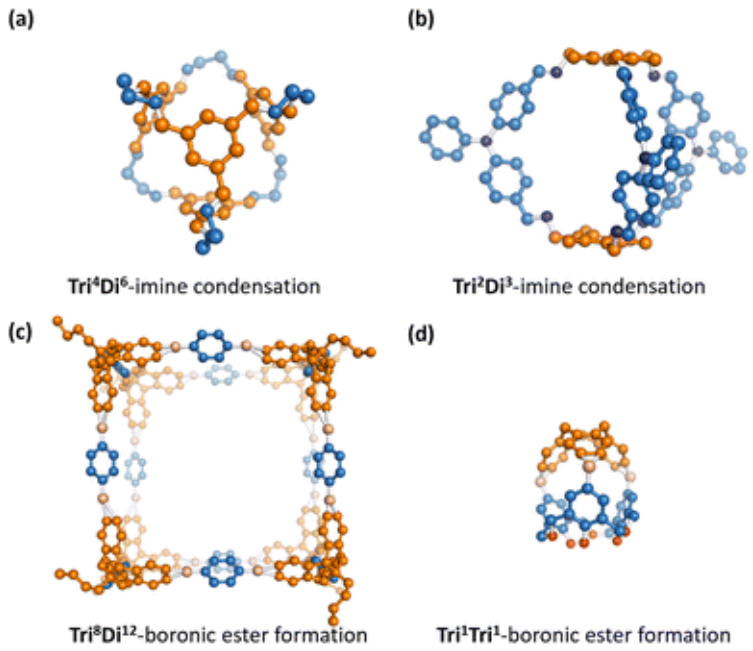
View Article Online
View Journal | View Issue

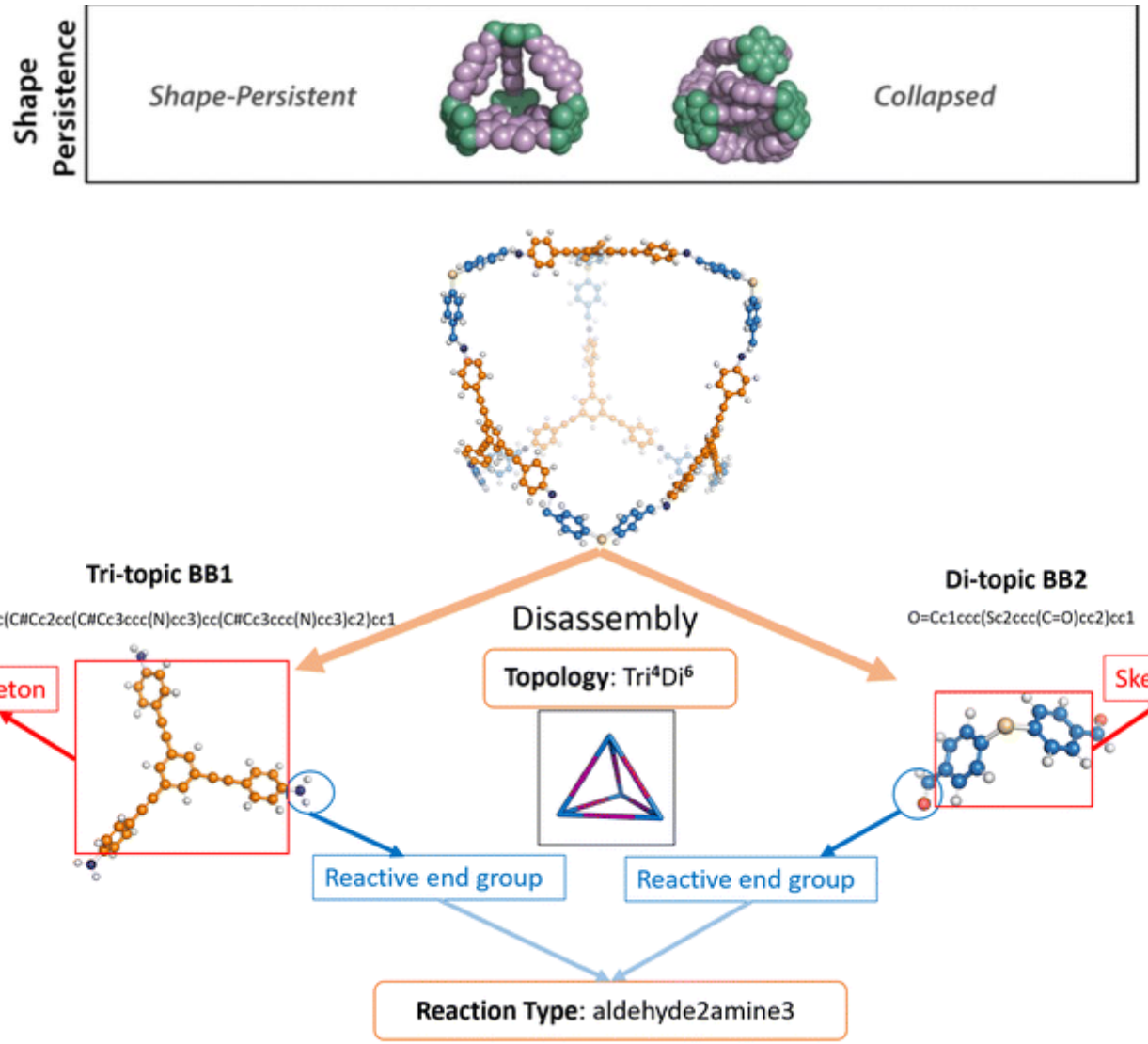
Check for updates

Cite this: *Digital Discovery*, 2023, 2, 1925

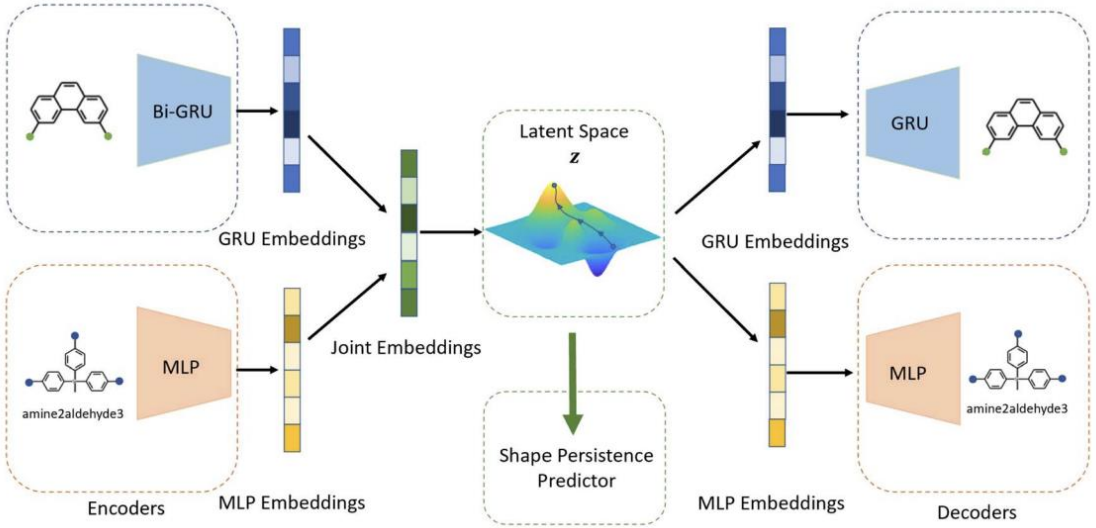
Deep generative design of porous organic cages *via* a variational autoencoder†

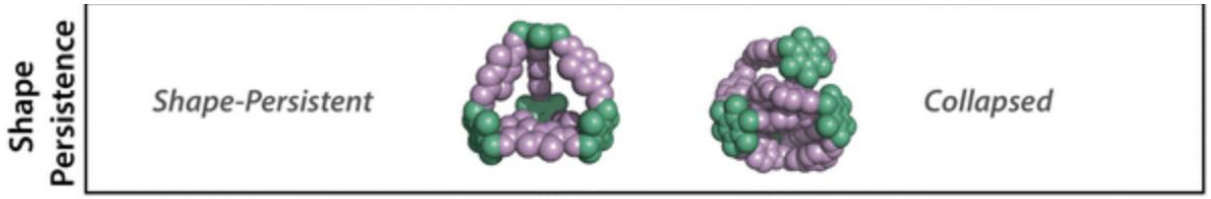
Jiajun Zhou, Austin Mroz and Kim E. Jelfs *

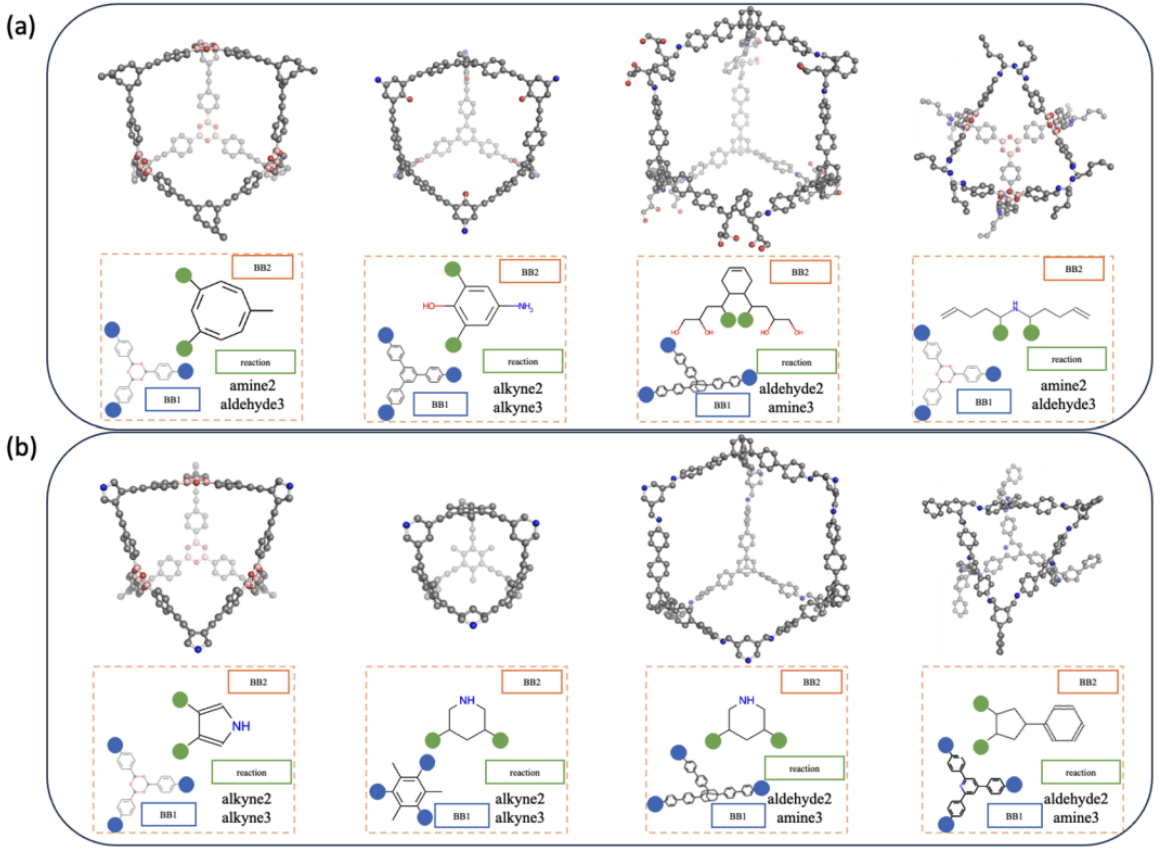




BO for latent space exploration





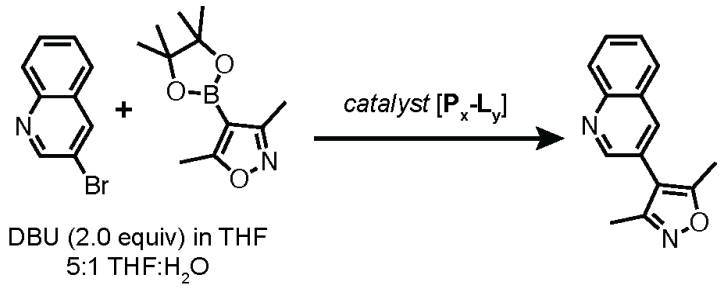


coding choose your own adventure – complex BO formulations in chemistry

Sargent Centre for Process Systems Engineering

from single-objective to multi-objective

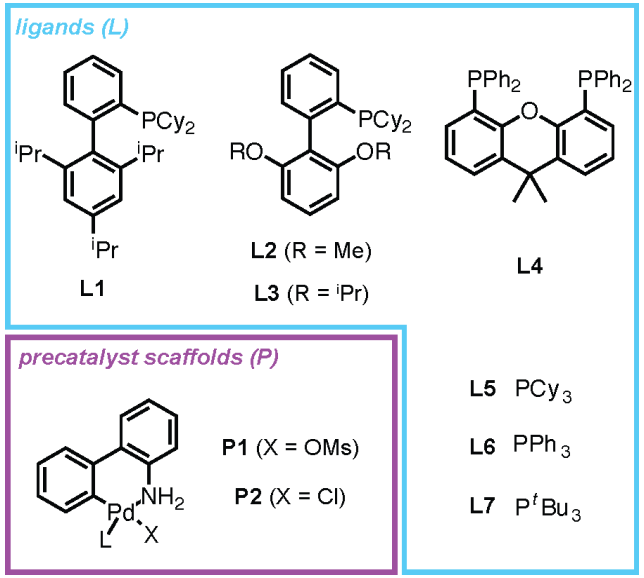
optimizing the coupling of 3-bromoquinoline with 3,5-dimethylisoxazole-4-boronic acid pinacol ester in the presence of 1,8-diazobicyclo[5.4.0]undec-7-ene (DBU) and THF/water



parameter	parameter type	parameter space
catalyst [P_x-L_y]	categorical	[P1-L1], [P2-L1], [P1-L2], [P1-L3] [P1-L4], [P1-L5], [P1-L6], [P1-L7]
catalyst loading	continuous	(0.5 - 2.0 %)
temperature	continuous	(30 - 110 °C)
residence time	continuous	(1 - 10 min.)

Summit

Objective	Description	Goal
yield	reaction yield [%]	maximize
turnover	catalyst turnover number [prod/cat]	maximize



your task

1. select a BO package
2. code your own MOBO!

example from the literature

npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

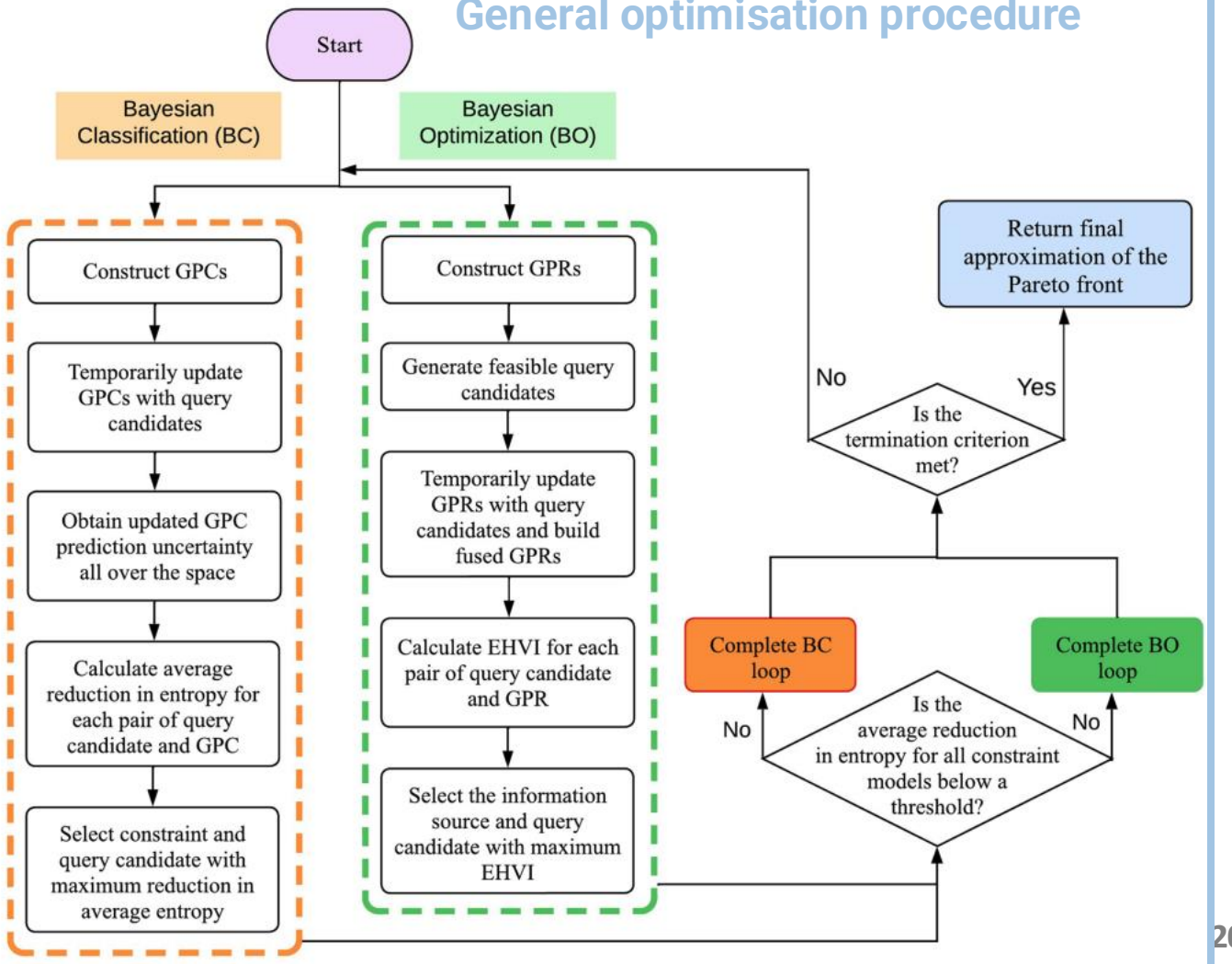
Check for updates

Bayesian optimization with active learning of design constraints using an entropy-based approach

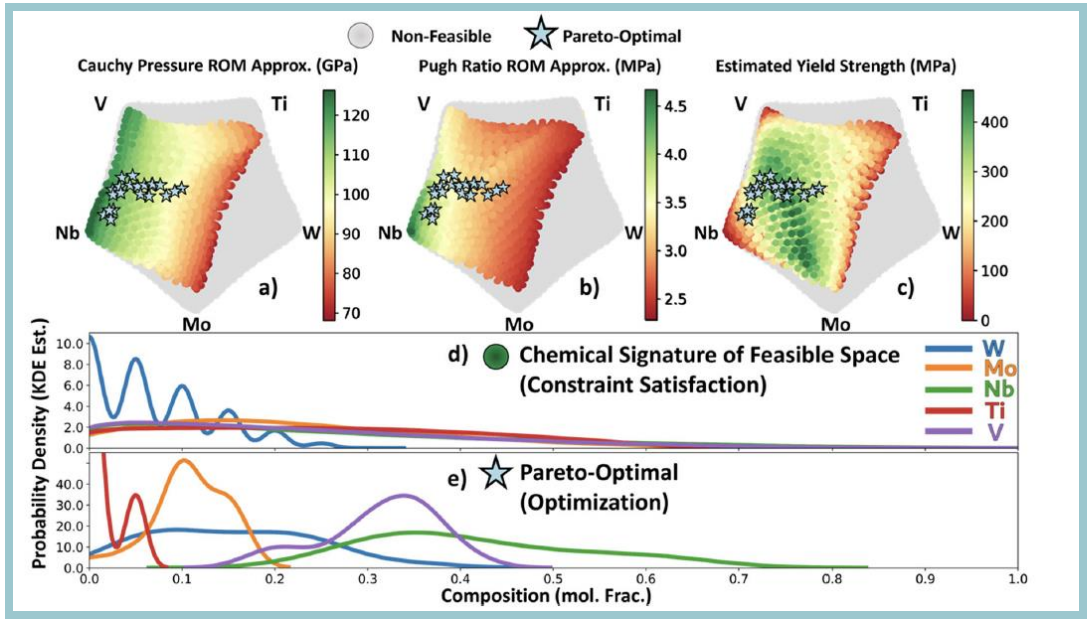
Danial Khatamsaz[1], Brent Vela[2✉], Prashant Singh[2,3], Duane D. Johnson[3,4], Douglas Allaire[1] and Raymundo Arróyave[1,2,5]

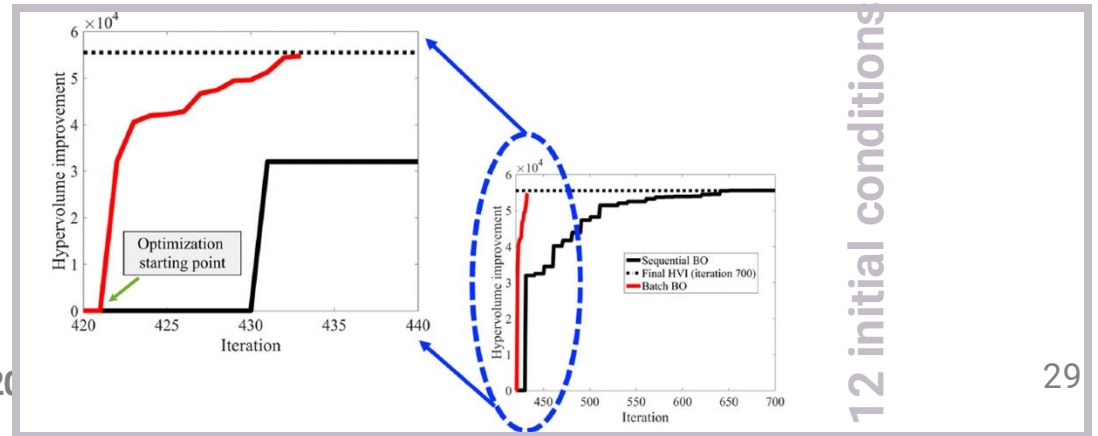
- Challenge – designing alloys for gas turbine engines requires balancing multiple objectives and constraints; the search space that these span is far too vast for conventional search techniques
- Propose – an entropy-based approach that actively learns the boundaries of contraints within a constrained multi-objective materials design problem.
- Demonstrate its utility on the Mo-Nb-Ti-V-W system and identify 21 Pareto-optimal alloys that satisfy all constraints.

General optimisation procedure



Explored initial sampling methods





npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Bayesian optimization with active learning of design constraints using an entropy-based approach

Danial Khatamsaz[1], Brent Vela[2✉], Prashant Singh[2,3], Duane D. Johnson[3,4], Douglas Allaire[1] and Raymundo Arróyave[1,2,5]

challenge	designing alloys for gas turbine engines requires balancing multiple objectives and constraints; the search space that these span is far too vast for conventional search technique
propose	an entropy-based approach that actively learns the boundaries of constraints with in a constrained multi-objective materials design problem
demonstrate	utility is demonstrated using the Mo-Nb-Ti-V-W system; identify 21 Pareto-optimal alloys that satisfy all constraints

The challenge

Imperial College London

designing alloys for gas turbine engines requires balancing multiple objectives and constraints (often at conflicting – *e.g.* must be ductile at room temperature, while maintaining yield strength at high temperature)

objectives

1. Low density
2. High-temperature yield strength
3. Creep resistance
4. Oxidation resistance
5. Etc.

constraints

1. Solidification range
2. Thermal conductivity
3. Solidus temperature
4. Linear CTE
5. Density

the search space that MPEAs span is far too vast for conventional search technique

npj Mater. Degrad., **2021**, 5, 14

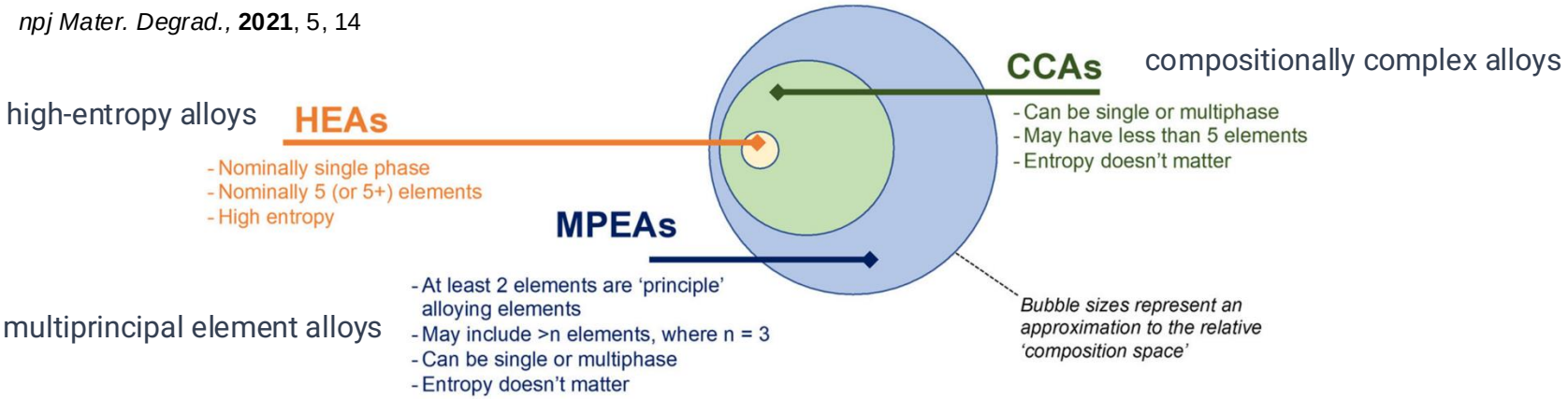


Fig. 1 A schematic representation of the classification notions for multi-principal element alloys (MPEAs). HEAs may be considered as a subset of CCAs; which in turn may also be considered a subset of the broader description of MPEAs—whereby the terminology of MPEAs encompasses a broad range of new and emerging metallic alloys.

Consider a 5-component alloy system sampled at 5% > 10,000 candidate alloys (excluding potential microstructures)

traditional methods (process-structure-property-performance chain) are resource-intensive
∴ need efficient search methods

Constraints pose a significant challenge

Imperial College London

Surrogate-based derivative-free optimization strategies are well-suited to a problem of this nature

Multi-objective BO – efficiently balances multiple goals

Multi-fidelity BO – cost-effective sampling approach

Existing methods do not consider cases where the constraints are actively learned in order to determine a feasible design space

- Often challenging to satisfy constraints while simultaneously maintaining feasibility
- Complicating this further, verifying a constraint is satisfied may be resource-intensive
- When learning a constraint space, we are focused on identifying the boundary of the design space, as opposed to quantifying the constraint.
- This can be done using a classifier (feasible vs. infeasible)

Introduce an entropy-based approach to decision-making process within MOBO

Imperial College London

Previous work

- Calculated entropy based on the difference between class membership probabilities predicted by Gaussian process classifiers – this results in higher entropy for designs close to the predicted boundary
- Here, the authors did not consider uncertainty in the probability predictions

This work

- Calculates entropy based on uncertainty in class membership probability predictions
- In this way, designs with higher uncertainty about their class membership possess higher entropy (and this is not correlated with their distance from the predicted boundary)

Utility is demonstrated using the Mo-Nb-Ti-V-W system; identify 21 Pareto-optimal alloys that satisfy all constraints

Imperial College London

Constraints are modeled using **high-fidelity** CALculation of Phase Diagrams (CALPHAD)-based **simulations**

Table 1. The five constraints and the three objectives associated with the design problem addressed in this work.

Property	Constraint/Objective
Solidus Temperature	$T_s \geq 2000\,^{\circ}C$
Solidification Range	$\Delta T \leq 400$
Thermal Conductivity	$\kappa \geq 20$
Linear CTE	$a_1 \leq 2\%$
Density	$\rho \leq 9 g/cc$
Ductility	Max $C_{12} - C_{44}$
Ductility	Max B/G
HT Yield Strength	Max σ_{HT}

Temperature below which a material is solid

Material must be compatible with thermal coatings and dissipate heat

Material must be lightweight

Inferred from ground state properties *(high-fidelity DFT-KKR-CPA)*

High temperature (HT) >> 1300℃
(HT Yield Strength evaluated using a previously reported physics-based model)

Objectives are modeled using **high-fidelity** CALculation of Phase Diagrams (CALPHAD)-based **simulations**

Imperial College London

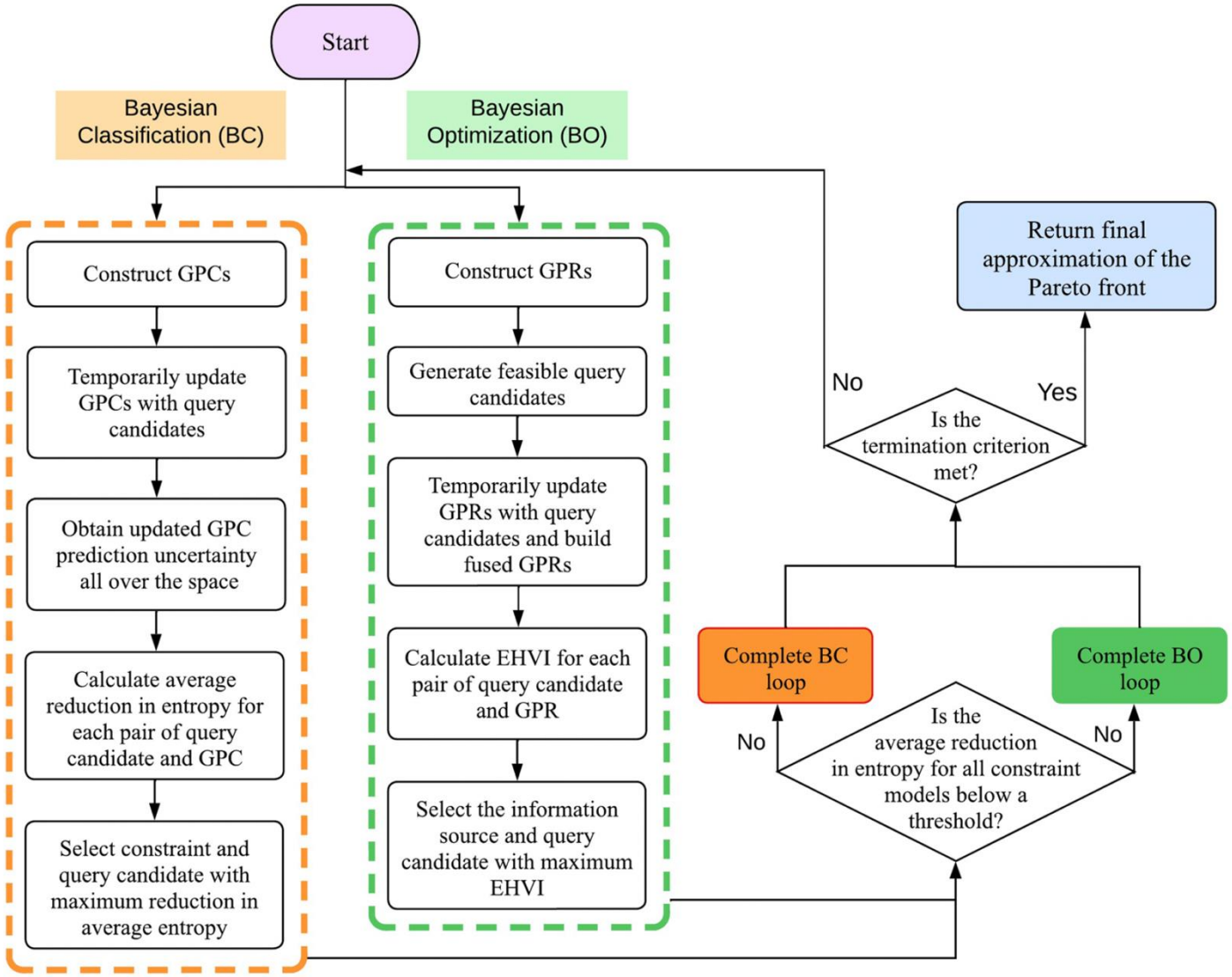
BC >> learn boundaries separating feasible and infeasible design regions

- GPCs provide uncertainty in class membership predictions – crucial to a Bayesian model
- Entropy enables uncertainty to be a comparable quantity – to select optimal sequence of informative experiments
- Uncertainty of a classifier is calculated from a set of randomly generated samples
- Optimization is about reducing entropy associated with the classifiers
- Once entropy reduction is below a threshold, the predicted feasible regions are delivered to the BO loop – but the framework calculates entropy for all constraints at each iteration
- "dynamic decision-making approach"

Then, we employ the discrete entropy formula to determine the entropy:

$$H = -\sum_{j=1}^{k} \sigma_j \log(\sigma_j) \quad (5)$$

where we have predicted the labels of k samples randomly generated, and σ_j is the standard deviation of the predicted class membership provided by the GPC. The more accurate a classifier is about the boundary, the less uncertain it will be about the assigned labels.



Imperial College London

BO > learn boundaries separating feasible and infeasible design regions

- Surrogate model = GPRs
- Each model is associated with its own GPR – surrogates are formulated by assuming there is multiple available information sources to estimate a quantity of interest

Surrogate model (GP) for information source (f(x)) of quantity of interest (i)

$$f_{\mathrm{GP},i}(\mathbf{x})|\mathbf{X}_{N_i}, \mathbf{y}_{N_i} \sim \mathcal{N}\left(\mu_i(\mathbf{x}), \sigma^2_{\mathrm{GP},i}(\mathbf{x})\right) \quad (1)$$

Posterior distribution of information source *I* at design point **x**

where

$$\mu_i(\mathbf{x}) = K_i(\mathbf{X}_{N_i}, \mathbf{x})^T \left[K_i(\mathbf{X}_{N_i}, \mathbf{X}_{N_i}) + \sigma^2_{n,i} I\right]^{-1} \mathbf{y}_{N_i}$$

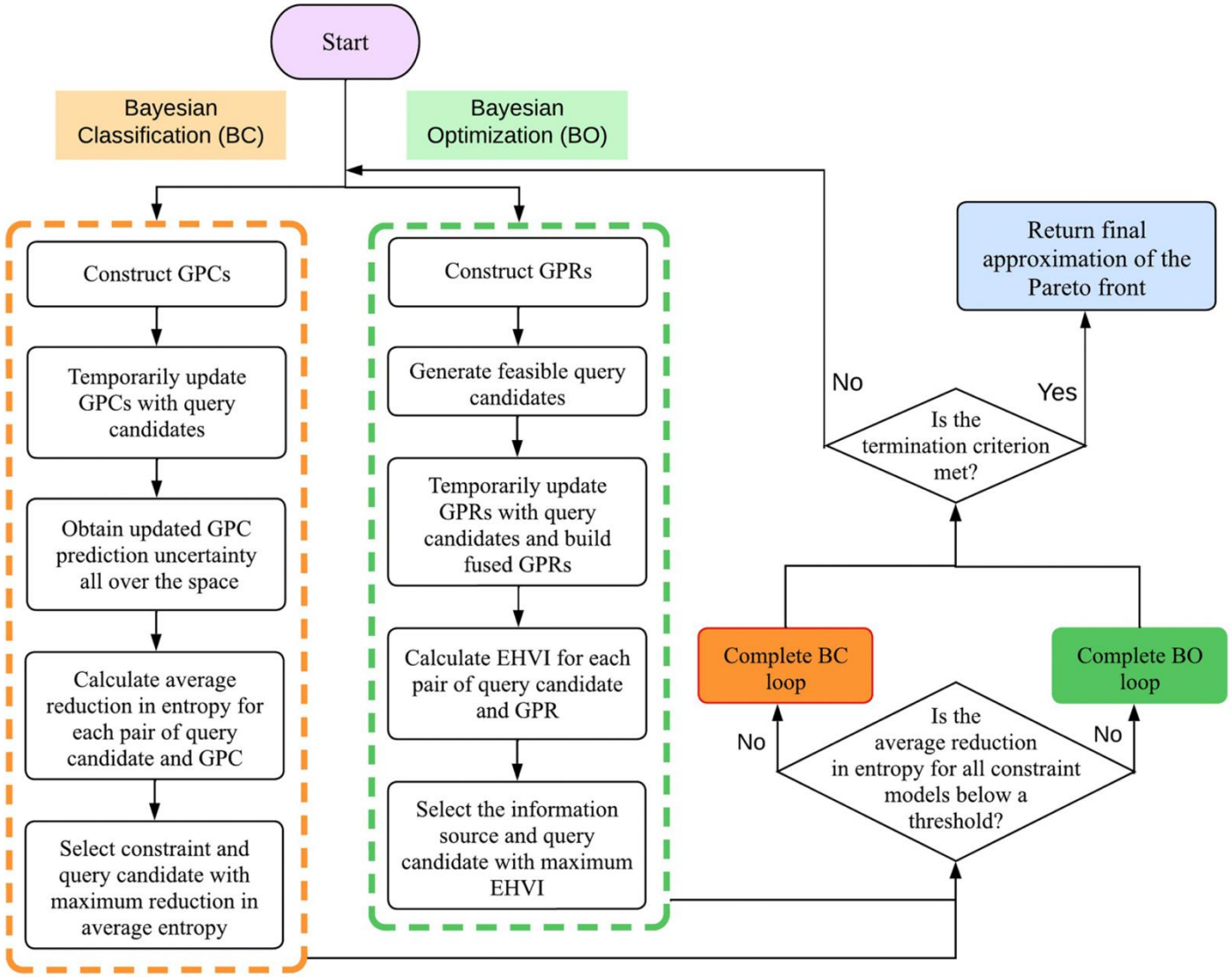
$$\sigma^2_{\mathrm{GP},i}(\mathbf{x}) = k_i(\mathbf{x}, \mathbf{x}) - K_i(\mathbf{X}_{N_i}, \mathbf{x})^T \left[K_i(\mathbf{X}_{N_i}, \mathbf{X}_{N_i}) + \sigma^2_{n,i} I\right]^{-1} K_i(\mathbf{X}_{N_i}, \mathbf{x}) \quad (2)$$

Model of observation error (based on experiments or expert opinion – human-in-the-loop)

"signal variance" covers 2 sources of uncertainty –

1. GPR estimation of objective function
2. Variance associated with information source and highest fidelity model

- Information from multiple sources is fused using a reification approach > estimates correlation coefficients between different information sources
- Expected HyperVolume Improvement (EHVI) as the acquisition function



BC runs in parallel to BO

Results of the 3-objective, 5-constraint MPEA design problem (sequential BO)

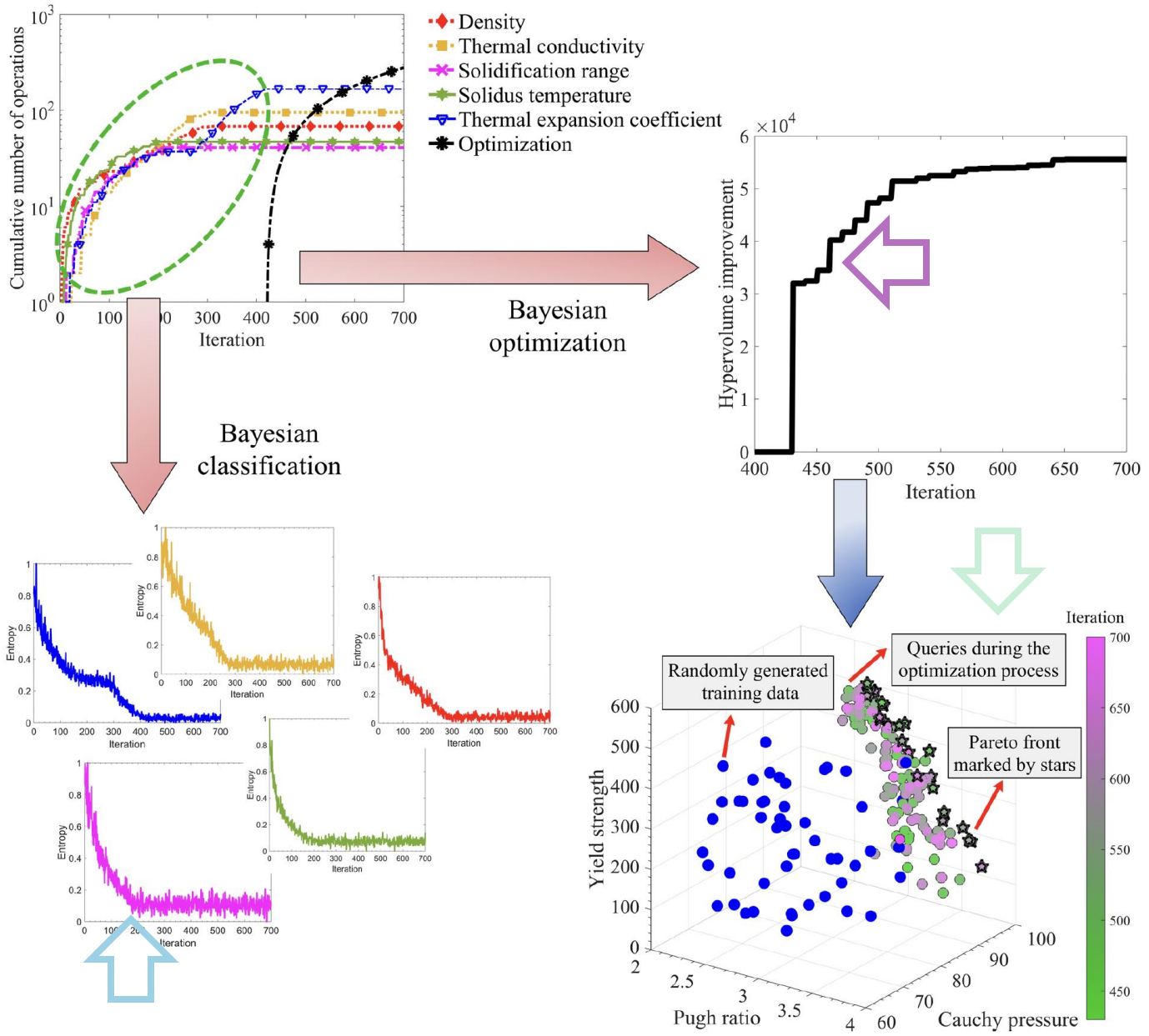
Imperial College London

- 700 iterations – total
- BO began after average entropy reduction plateaued for constraints

Increases in EHVI indicate better estimations of the ***Pareto front***

set of "degenerate" solutions that are superior to the rest of the search space

Queries are localized to one corner of the design space – indicates that the framework identified the optimal design region and is searching this space further



Results of the 3-objective, 5-constraint MPEA design problem (batch BO)

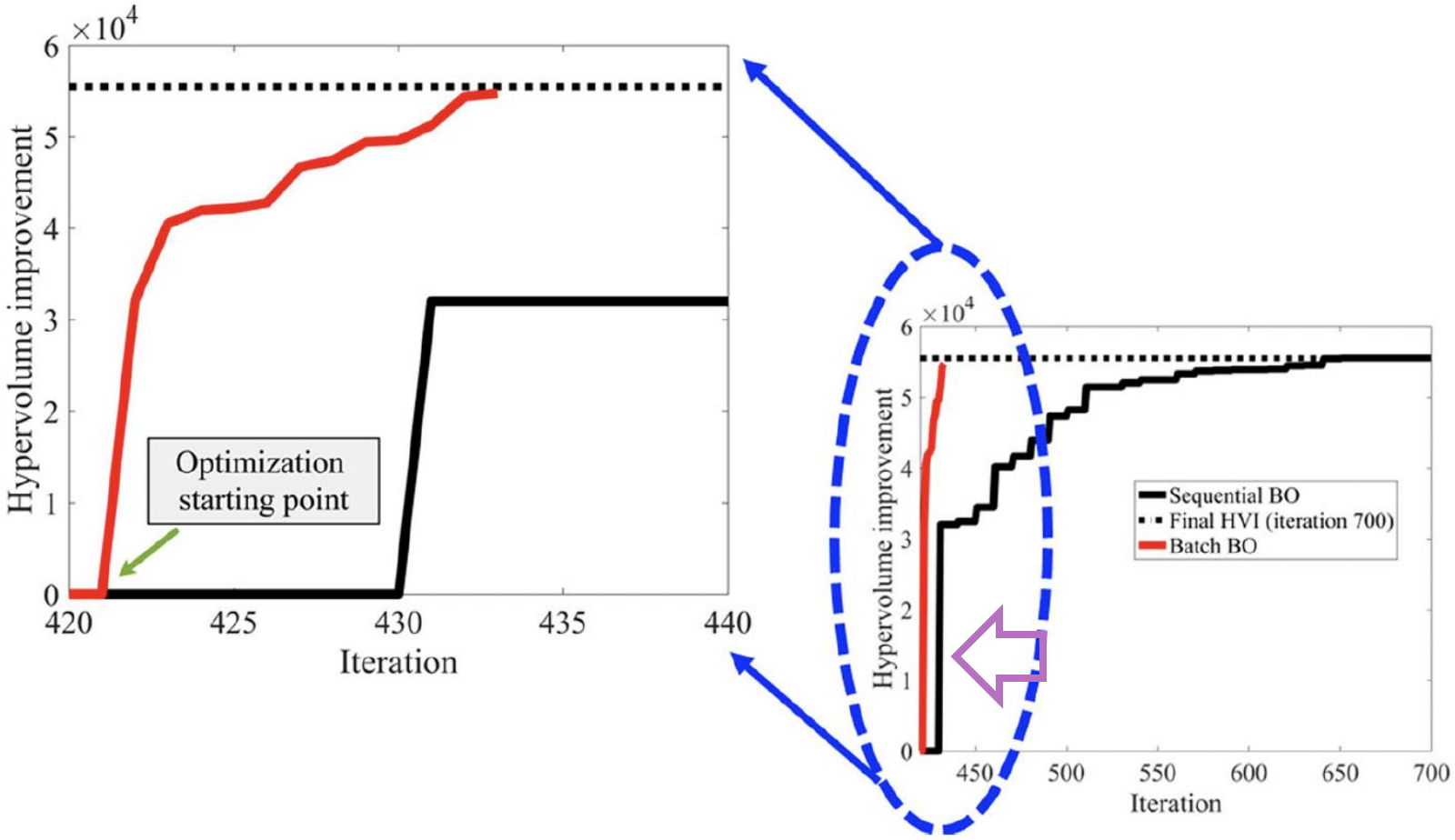
Imperial College London

- Batch BO allows 48 experiments to be performed in parallel (batch)

13 iterations necessary to converge (as opposed to 280 iterations in sequential MOBO)

*Total computational cost is *roughly* equivalent between sequential and batch MOBO*

Similar hypervolume improvement between batch and sequential MOBO indicates they learn Pareto fronts of similar quality
(yet material suggestions/predictions may not be the same!)



Constraint-satisfying candidate alloys

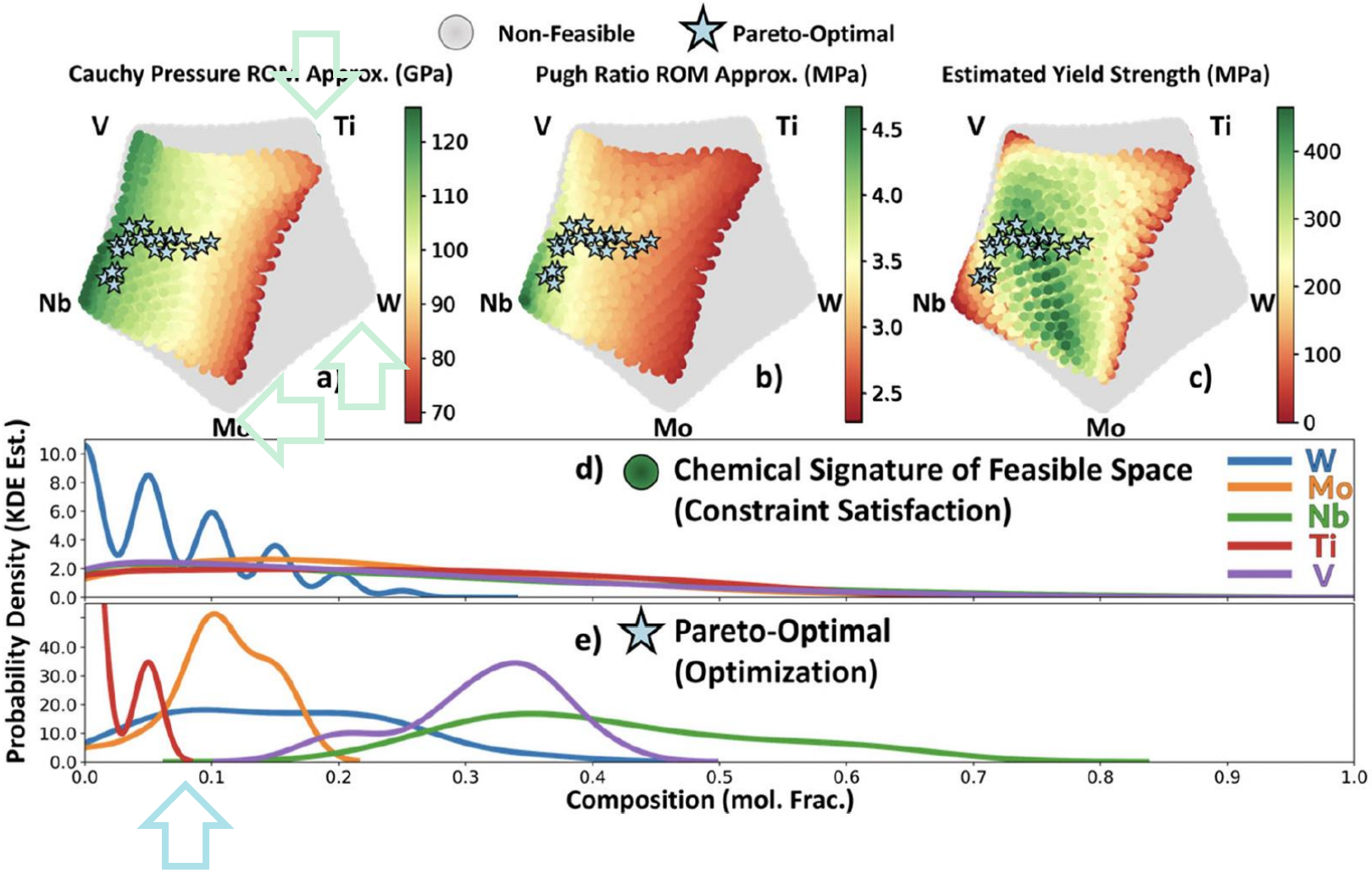
Imperial College London

UMAP plots yield design rules

- Ti-, Mo-, and W-rich alloys fail ≥ 1 constraint

KDEs reveal chemical signature of the feasible design space

- *Ti and Mo signatures shifted – indicates depletion in these elements*



Benchmarking constraint-satisfaction performance

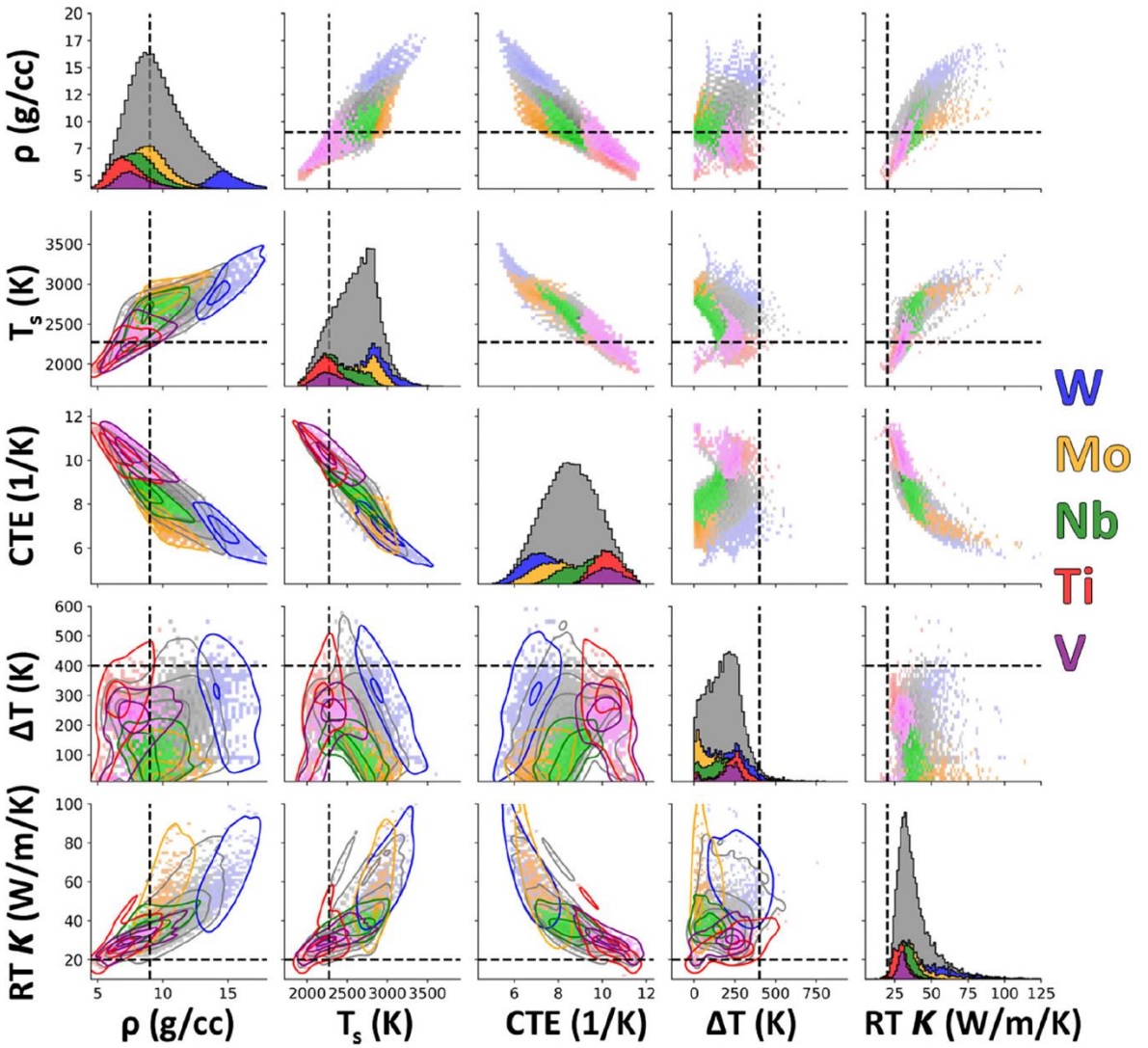
Imperial College London

Perform a factorial exploration of the design space

- This involved querying the 5 constraint information sources at increments of 5 > 53,130 total queries
- Within batch active learning, only 420 queries were necessary to learn the feasible design space

Instances where constraints are not satisfied can be explained by the chemistry

- Alloys that fail the thermal conductivity constraint are Ti- and/or V-rich; these elements possess the lowest thermal conductivity of those represented in the design space
- Further, these alloys are often compositionally complex and may suffer from enhanced phonon and electron scattering (which would decrease observed thermal conductivity)



Further benchmarking by DFT analysis of the Pareto-front

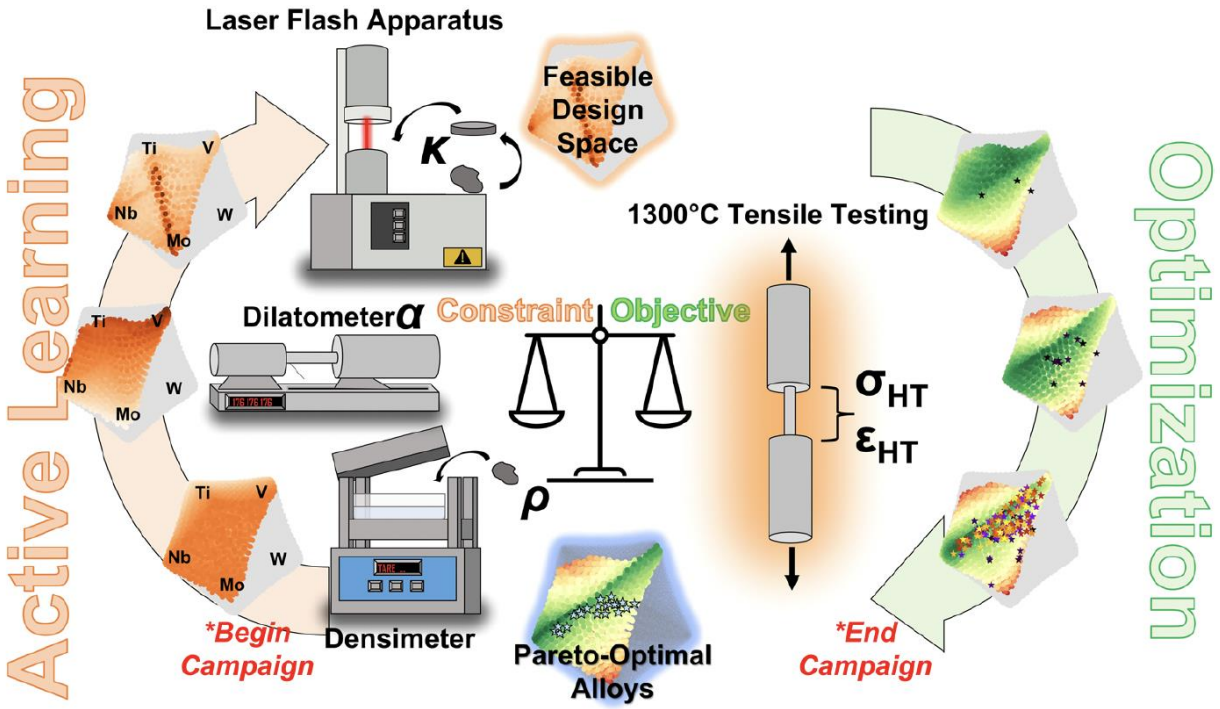
Moving towards autonomous closed-loop materials design

Imperial College London

1. Simulations may be used to initially constrain *(i.e., filter)* the design space and more cost-effectively eliminate materials

2. Experiments to actively learn stricter constraints

3. Optimize for target properties



final thoughts, considerations, and resources

We've spent this whole week talking about BO, but you need to ask yourself, is BO right for my problem?
In most chemical cases, the answer is yes, but not all

Zhu, M. *et al.*, **2024,** DOI:10.26434/chemrxiv-2024-h37x4

Sargent Centre for Process Systems Engineering

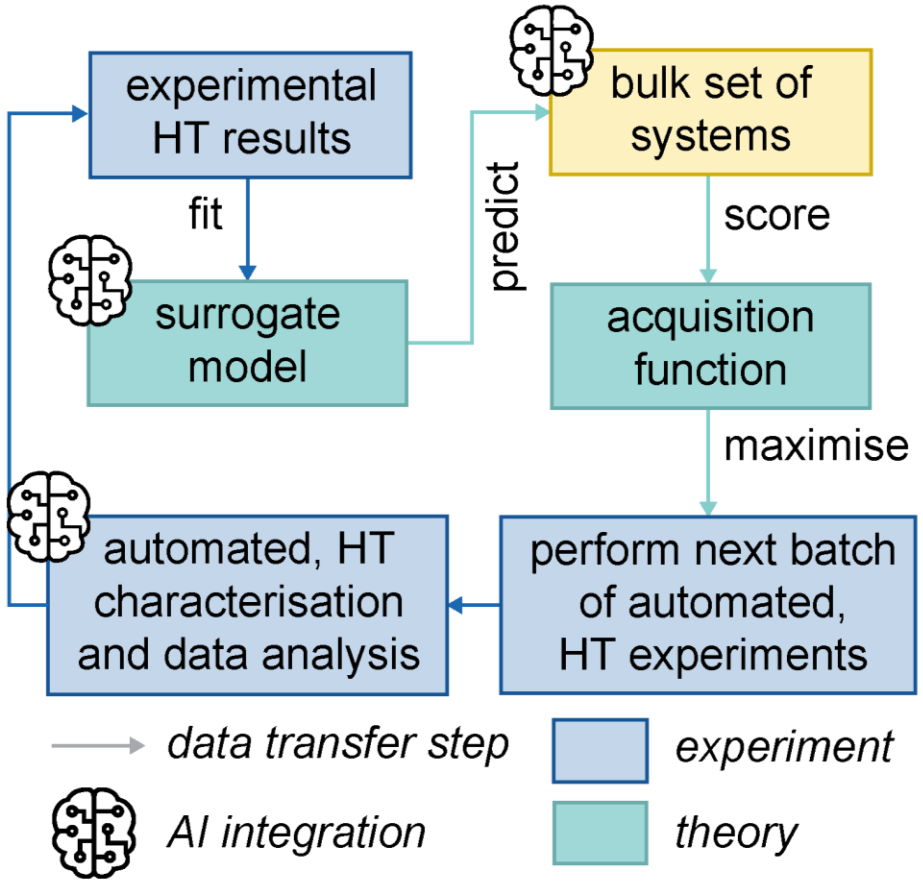
things to consider…

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

Thank you
a.mroz@imperial.ac.uk

UCL IMPERIAL

Schmidt Sciences Imperial-X