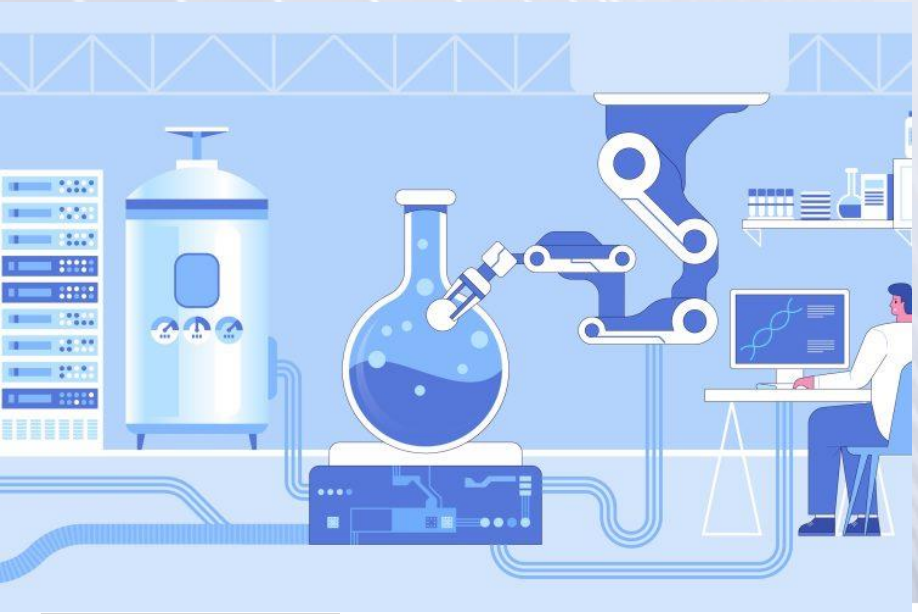
Sargent Centre for Process Systems Engineering

BO for Guided Chemical Design



Austin Mroz

Imperial College London

a.mroz@imperial.ac.uk

tutorial notebooks

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*	

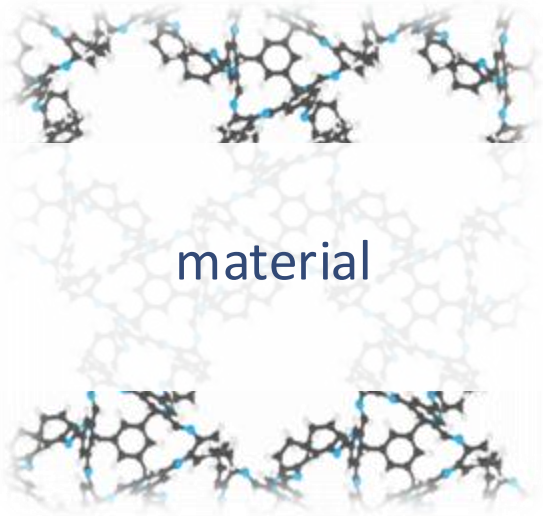
time	subject	learning objectives
16:00-16:15	Complex BO formulations in chemistry	MFBO, MOBO, MF-MO BO, etc.
16:15-17:00	***interactive***	code-your-own-adventure – ***MOBO, MFBO, GPs for molecules***
17:00-17:20	Complex BO formulations in chemistry	Let's dig into the literature
17:20-17:30	wrap-up discussion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

why BO for chemistry?

chemical design

INPUT

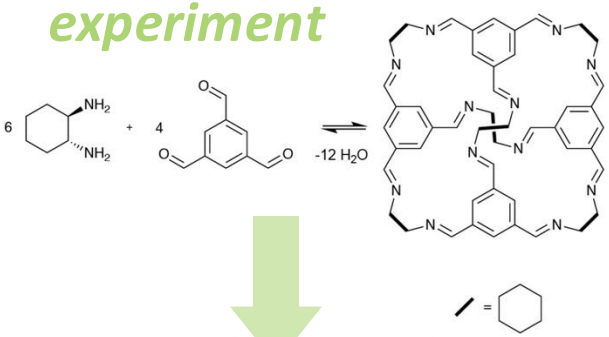
OUTPUT



Sargent Centre for Process Systems Engineering

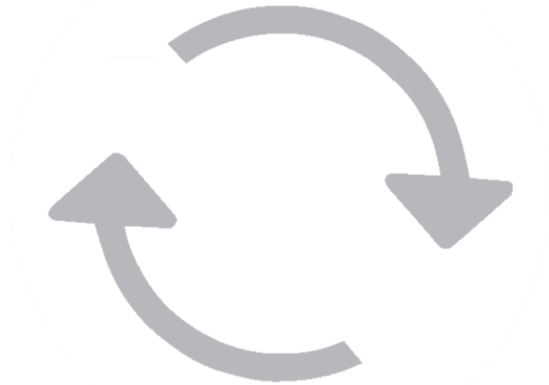
chemical design

experiment

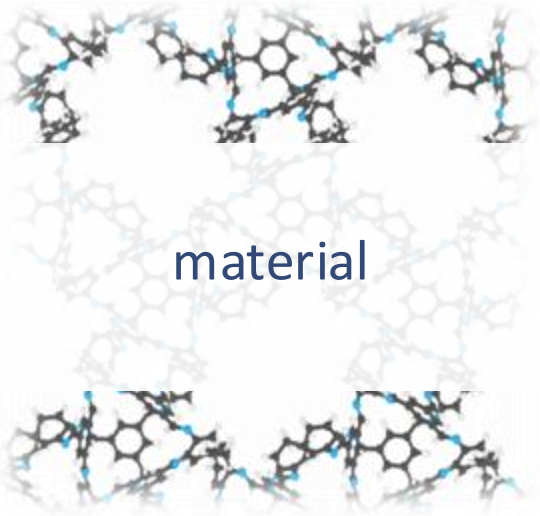


INPUT

property/application



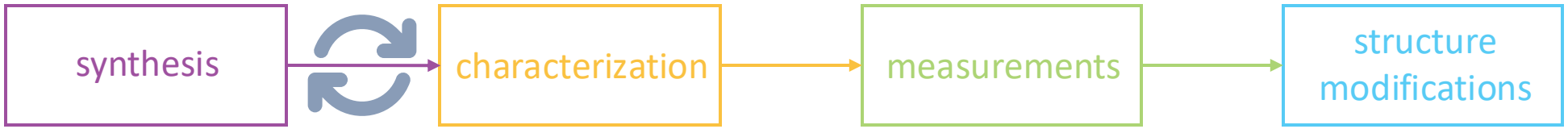
OUTPUT



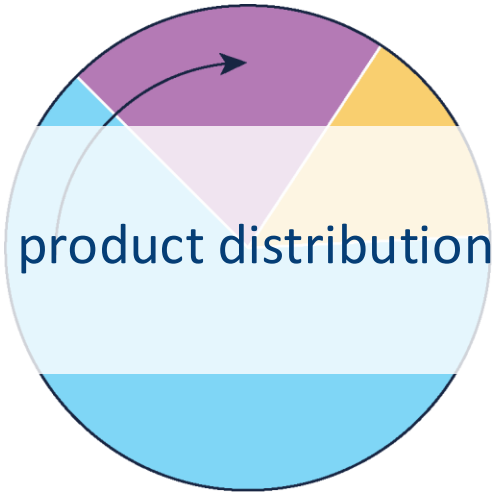
material

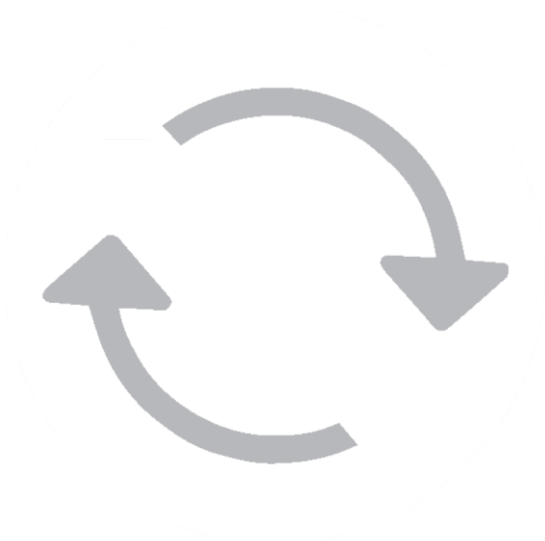
computation

conventional *experimental* chemical discovery



INPUT





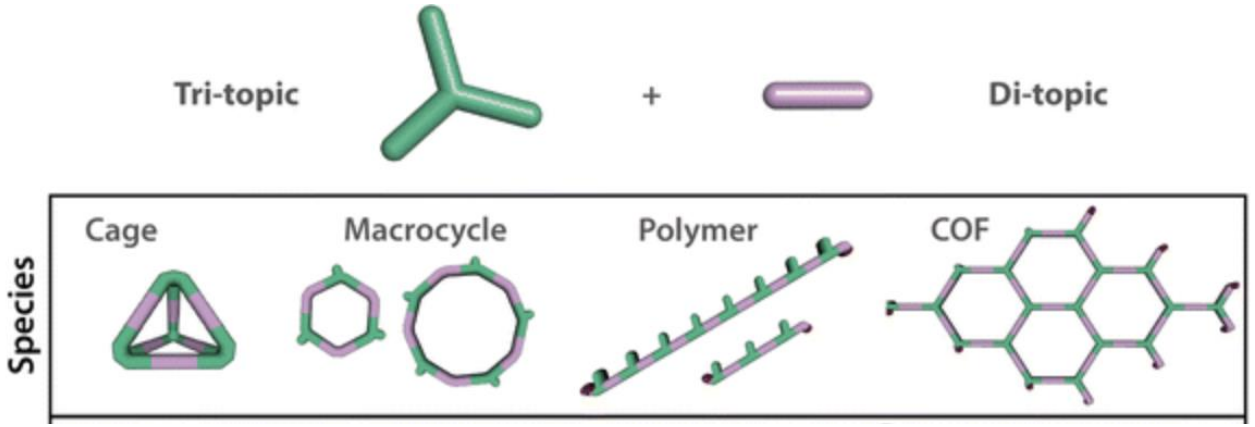
OUTPUT

conventional *experimental* chemical discovery

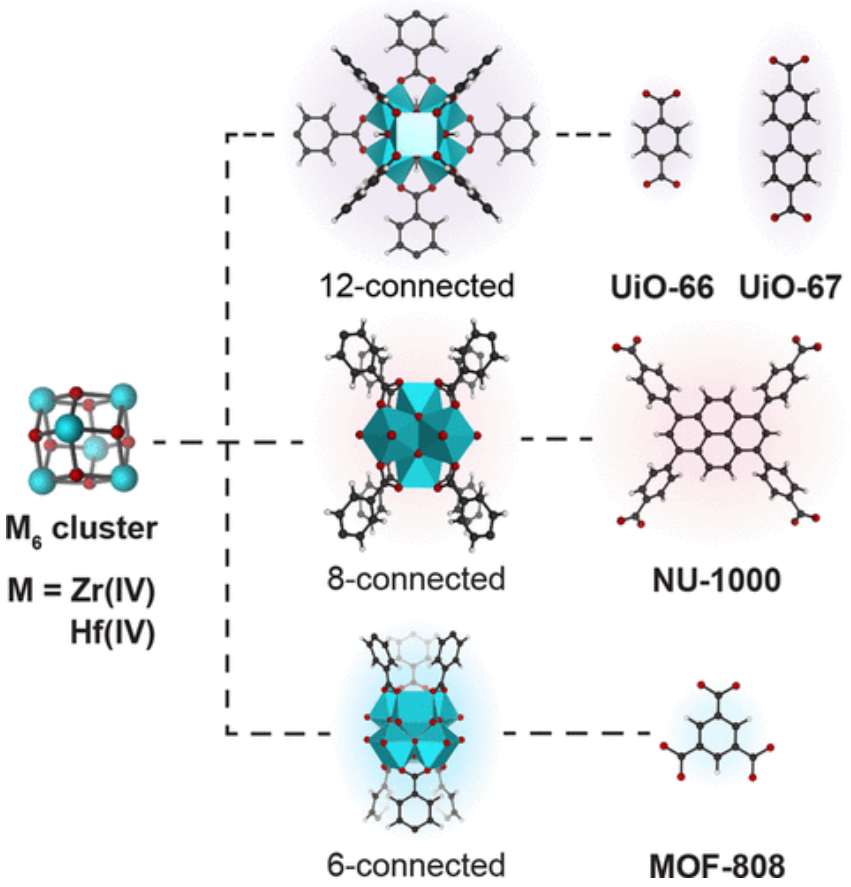
synthesis → characterization → measurements → structure modifications

10^{60}

Scale



Chem. Sci., **2024**, 15, 6331

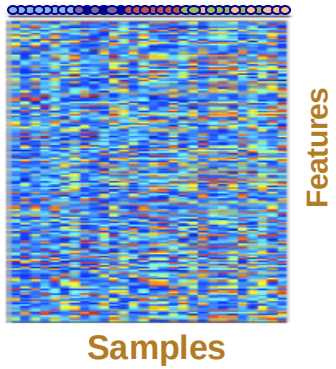


conventional *experimental* chemical discovery

synthesis → characterization → measurements → structure modifications

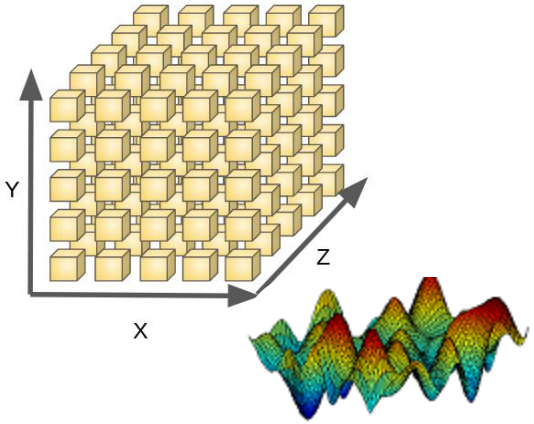
10^{60}

Scale



10^{6}

High-throughput technologies

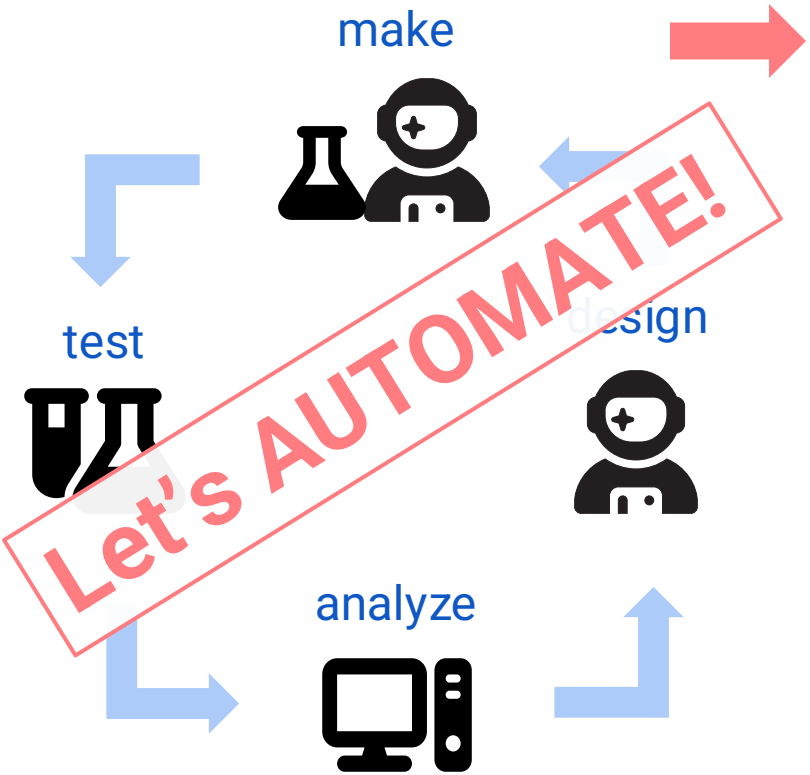


"Curse of dimensionality"

but given these near-infinite landscapes and limited resources, how do we efficiently move towards desired maxima?

Conventional design in chemistry

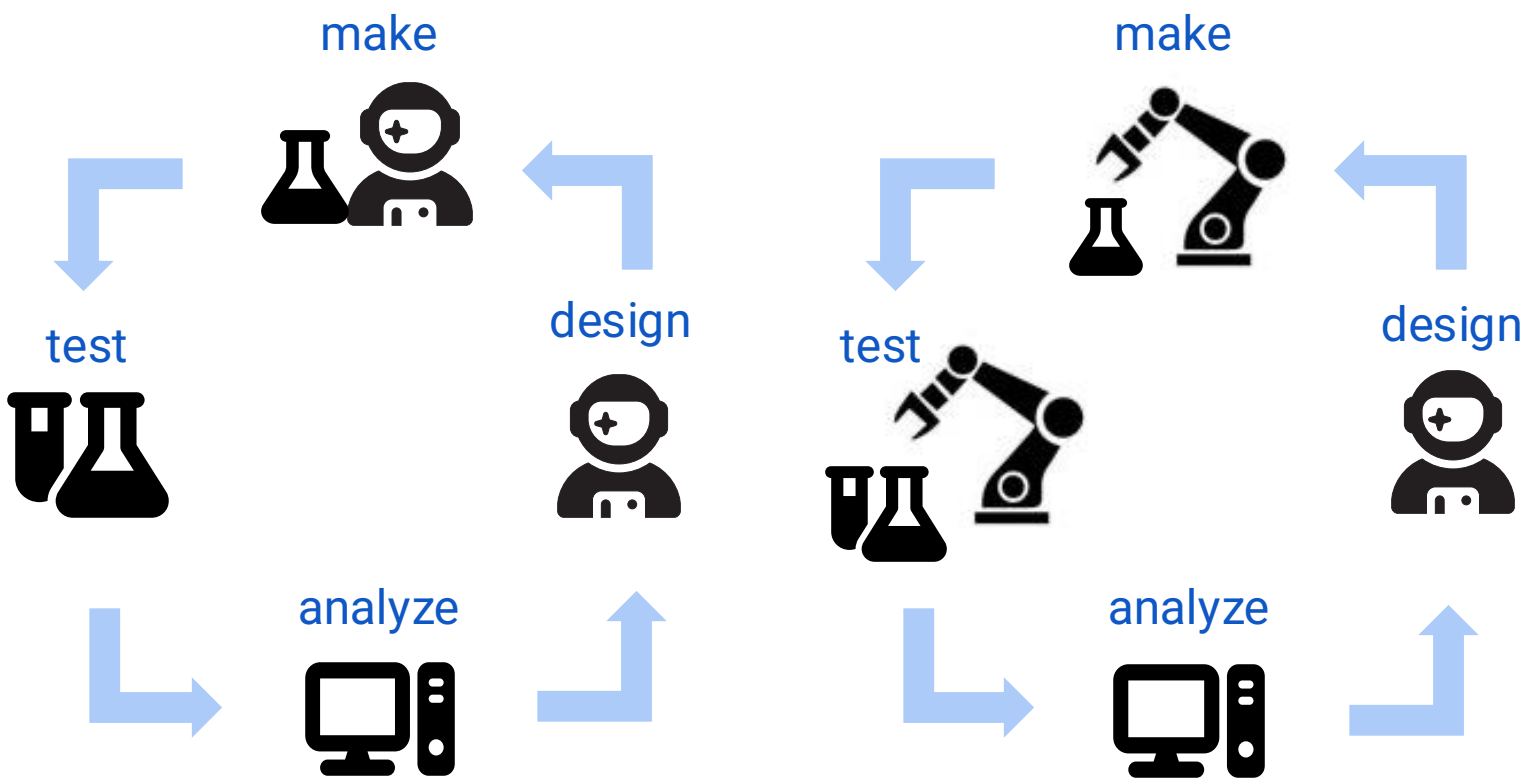
conventional



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

Design-make-test-analyze



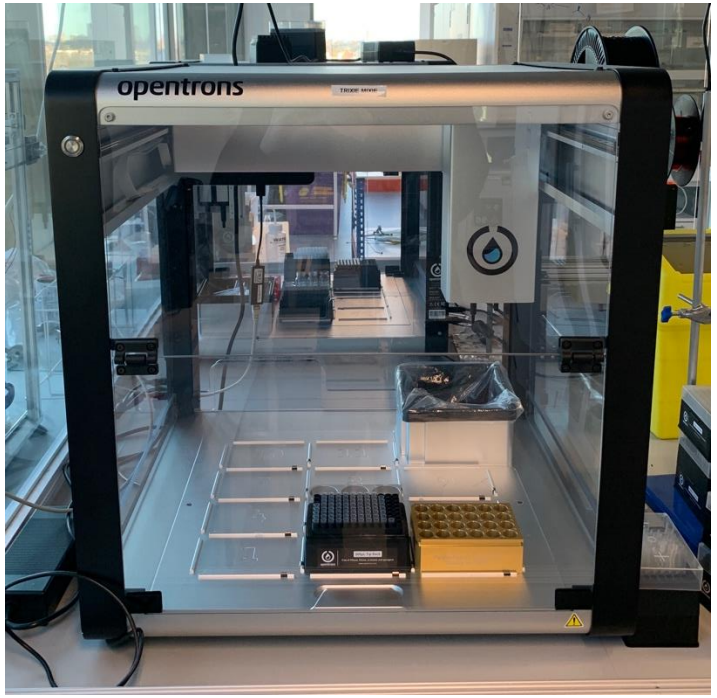
where are the bottlenecks?

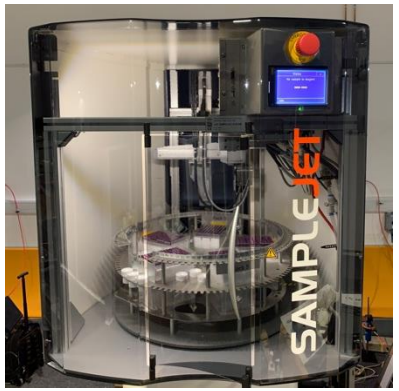
1. **manual synthesis, characterization and target property measurements are time and resource intensive**
2. chemical intuition is largely limited to *small* changes to *either* experimental conditions *or* structure

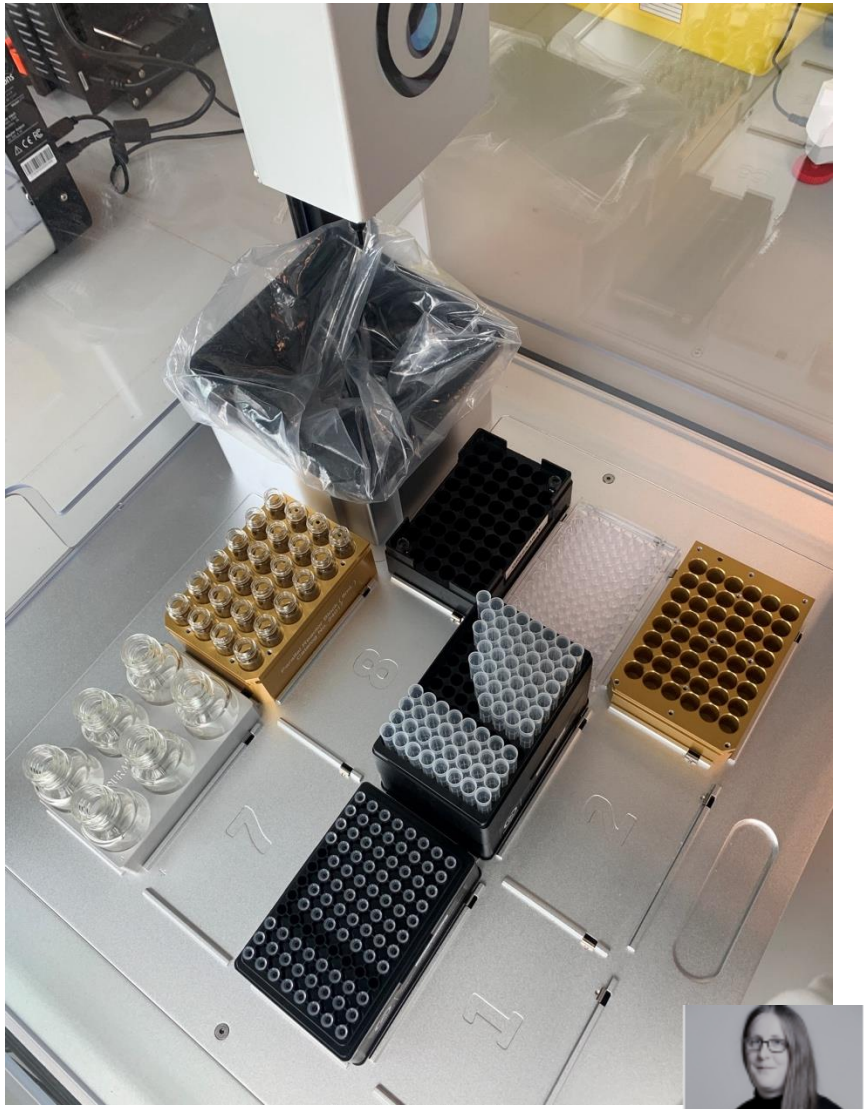
how does this help us?

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

Automation in chemical sciences

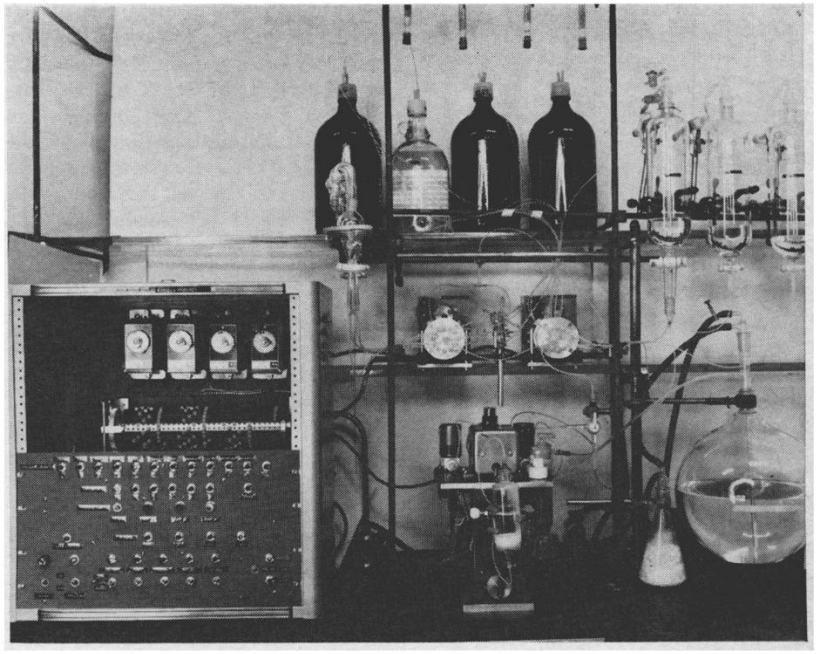






Dr. Becky Greenaway

Automation in chemistry is not new



Science, **1965**, 150, 178

Science, **2020**, 368, 980

Automated Synthesis of Peptides

Solid-phase peptide synthesis, a simple and rapid synthetic method, has now been automated.

R. B. Merrifield

PEPTIDE SYNTHESIS

Synthesis of proteins by automated flow chemistry

N. Hartrampf[1]*, A. Saebi[1]†, M. Poskus[1]†, Z. P. Gates[1]‡, A. J. Callahan[1], A. E. Cowfer[1], S. Hanna[1], S. Antilla[1,2], C. K. Schissel[1], A. J. Quartararo[1], X. Ye[1], A. J. Mijalis[1]§, M. D. Simon[1]¶, A. Loas[1], S. Liu[1]#, C. Jessen[3], T. E. Nielsen[3], B. L. Pentelute[1]**

Levels of automation in chemistry

manual	***standalone***	***integrated***	***autonomous***
single ***step*** at a time carried out by hand	single ***process*** at a time automated on one machine	multiple ***processes*** automated workflow no human intervention	***fully integrated*** workflow closed-loop ***algorithmic feedback***

Levels of automation in chemistry

TIMESCALE

manual ***standalone*** *integrated* *autonomous*

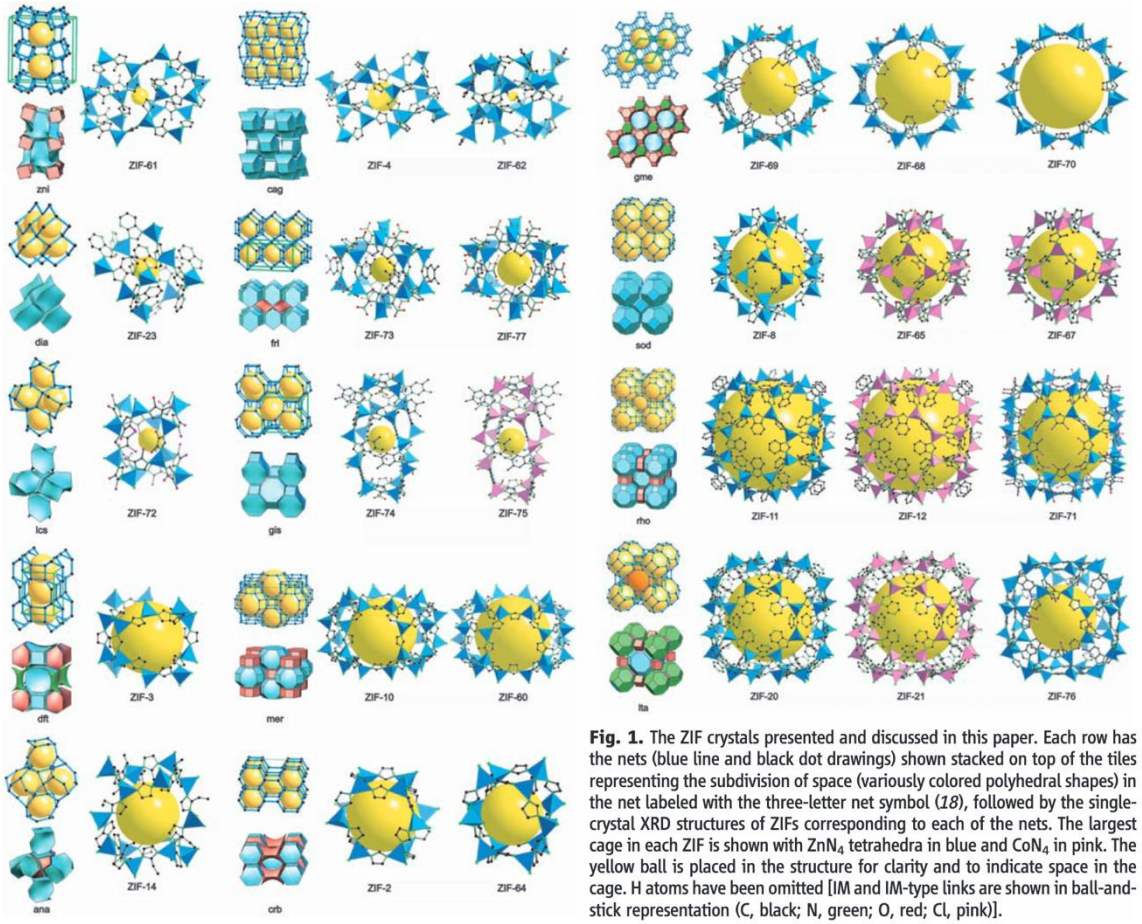


Fig. 1. The ZIF crystals presented and discussed in this paper. Each row has the nets (blue line and black dot drawings) shown stacked on top of the tiles representing the subdivision of space (variously colored polyhedral shapes) in the net labeled with the three-letter net symbol (*18*), followed by the single-crystal XRD structures of ZIFs corresponding to each of the nets. The largest cage in each ZIF is shown with ZnN_4 tetrahedra in blue and CoN_4 in pink. The yellow ball is placed in the structure for clarity and to indicate space in the cage. H atoms have been omitted [IM and IM-type links are shown in ball-and-stick representation (C, black; N, green; O, red; Cl, pink)].

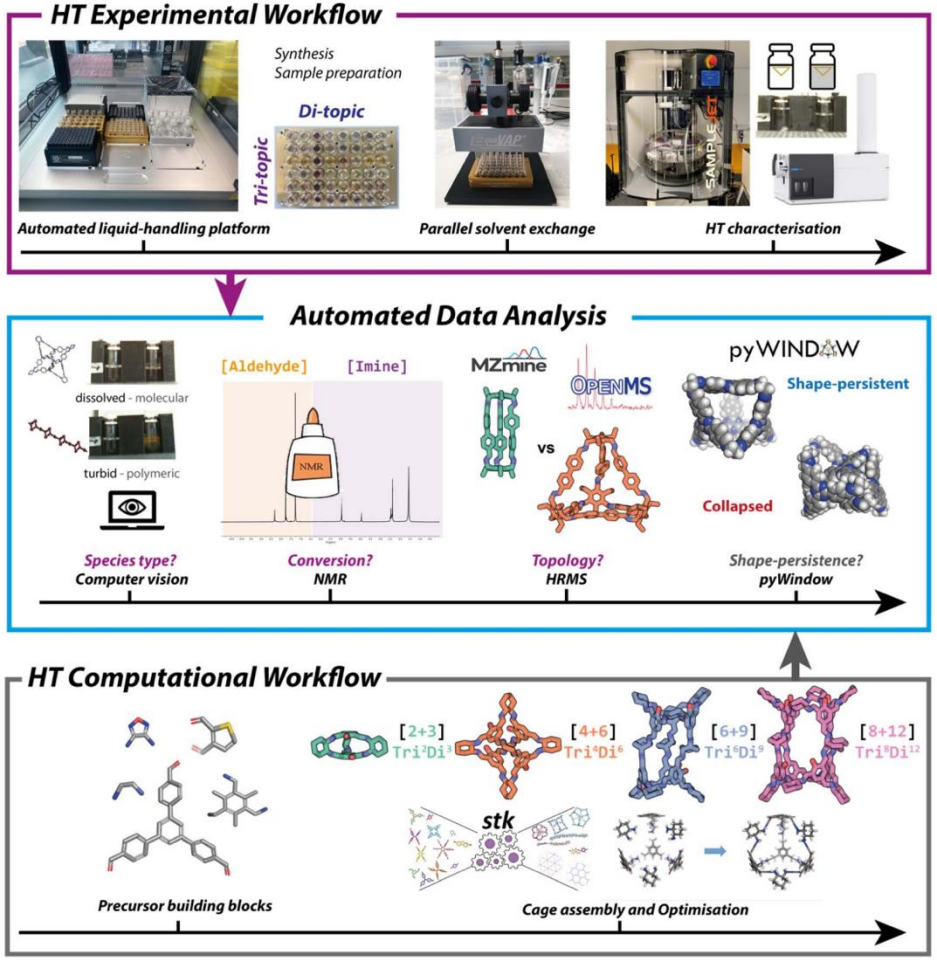
Zeolitic imidazole framework (ZIF) discovery

aim – high throughput synthesis of ZIFs
how – 96-well glass plate to screen 9600 microreactions
so? – 10 new ZIFs, featuring several high performing candidates

Science, **2008**, 319, 939

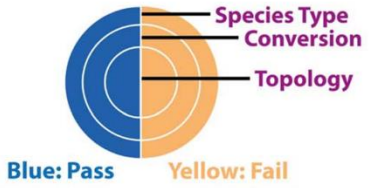
Levels of automation in chemistry

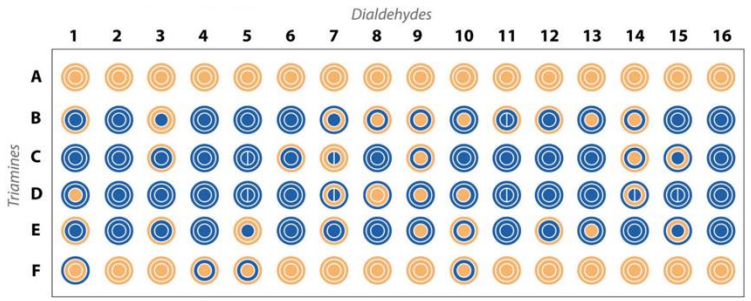
manual → *standalone* → ***integrated*** → *autonomous*



porous organic cage (POC) discovery

aim – high throughput synthesis of POCs
how – combined experimental and computational workflows
  – automated data analysis
  – computer vision for on-the-fly reaction feedback
so? – 225 POCs identified
  – 350-fold decrease in time required for data analysis

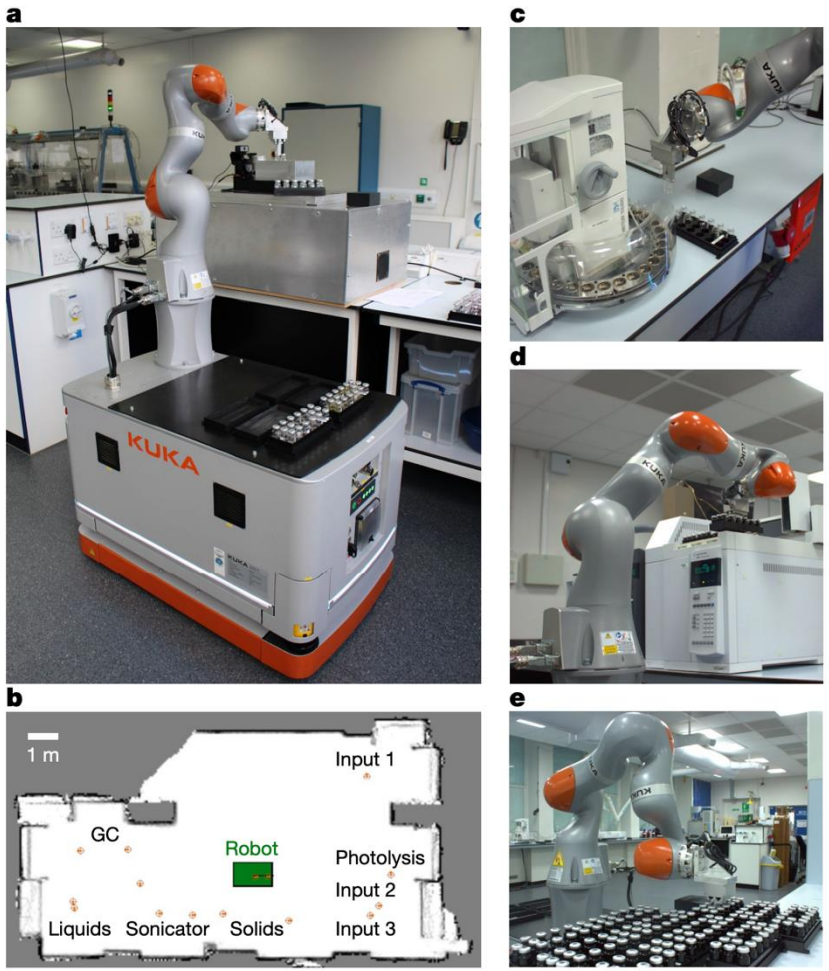




Chem. Sci., **2024**, 15, 6331

Levels of automation in chemistry

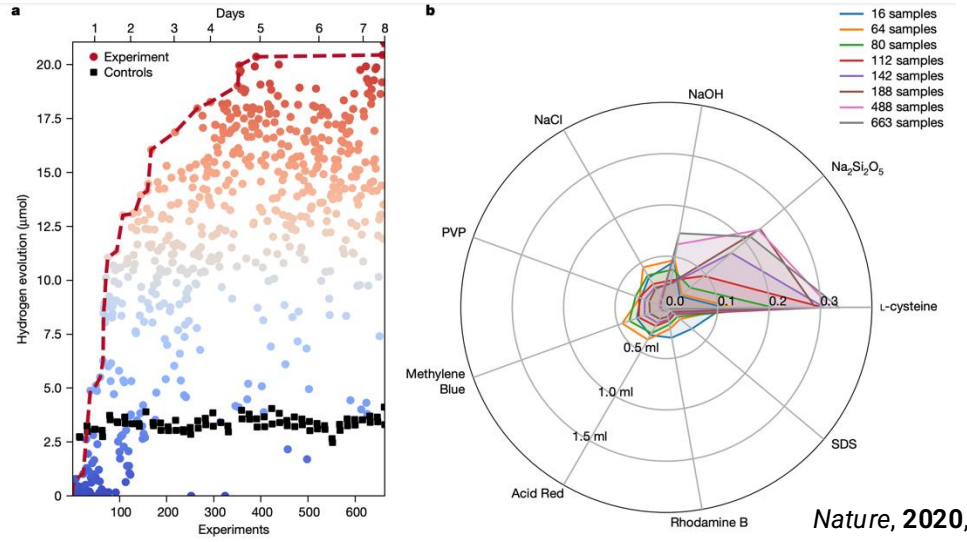
manual → *standalone* → *integrated* → ***autonomous***



mobile robotic chemist

aim – automating the researcher instead of the instruments

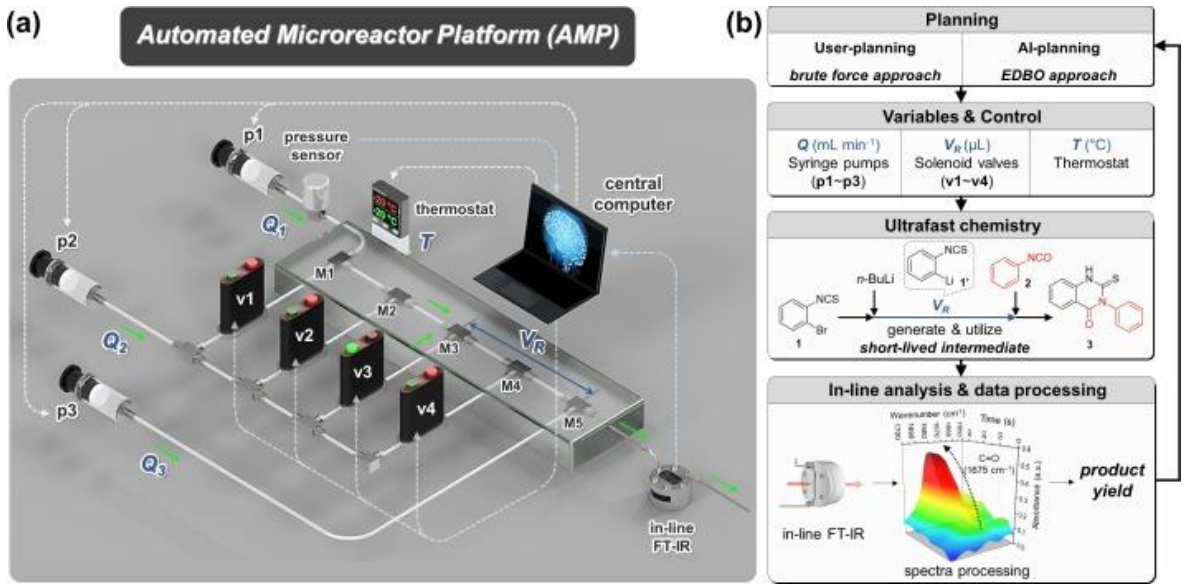
how – 8 days
688 experiments
10-variable parameter space

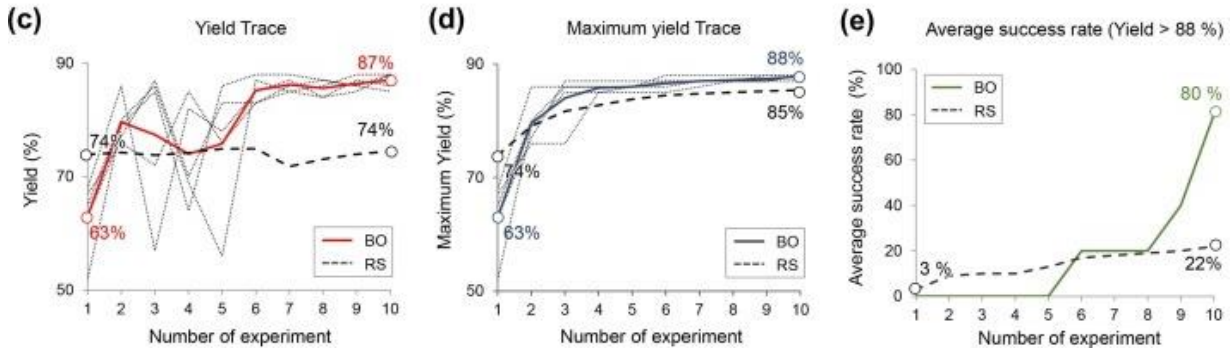


Nature, **2020**, 583, 237

Levels of automation in chemistry

manual → *standalone* → *integrated* → ***autonomous***





FLOW CHEMISTRY

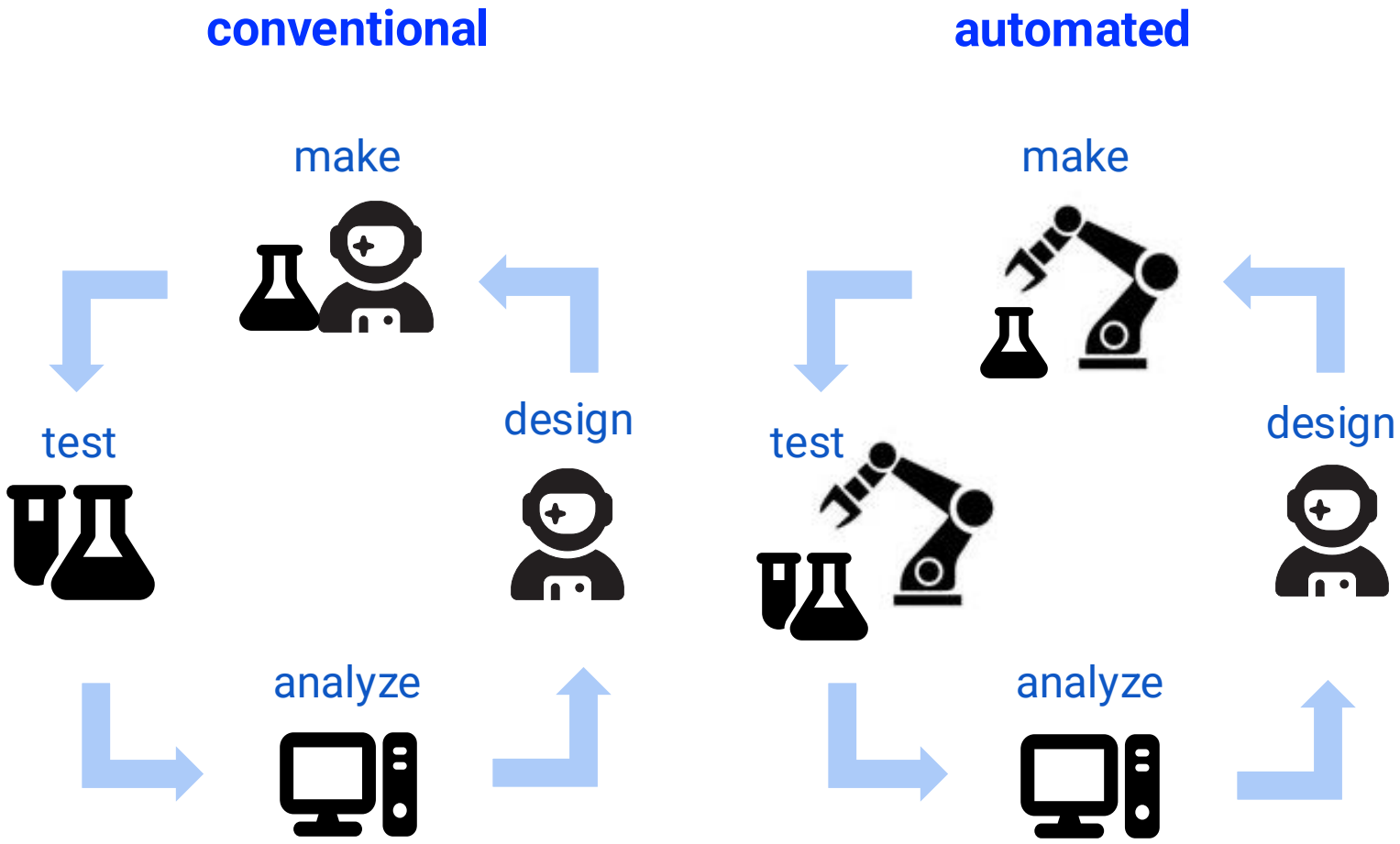
aim – ultrafast synthesis of biologically active compounds

how – exploits user-planned and AI-planned modes to accelerate synthesis

so – efficiency in chemical synthesis via improved precision and control offered by AMP platflorm scalability and flexibility

Chem. Eng. J., **2023**, 453, 139707

Design-make-test-analyze



it's not all rainbows…

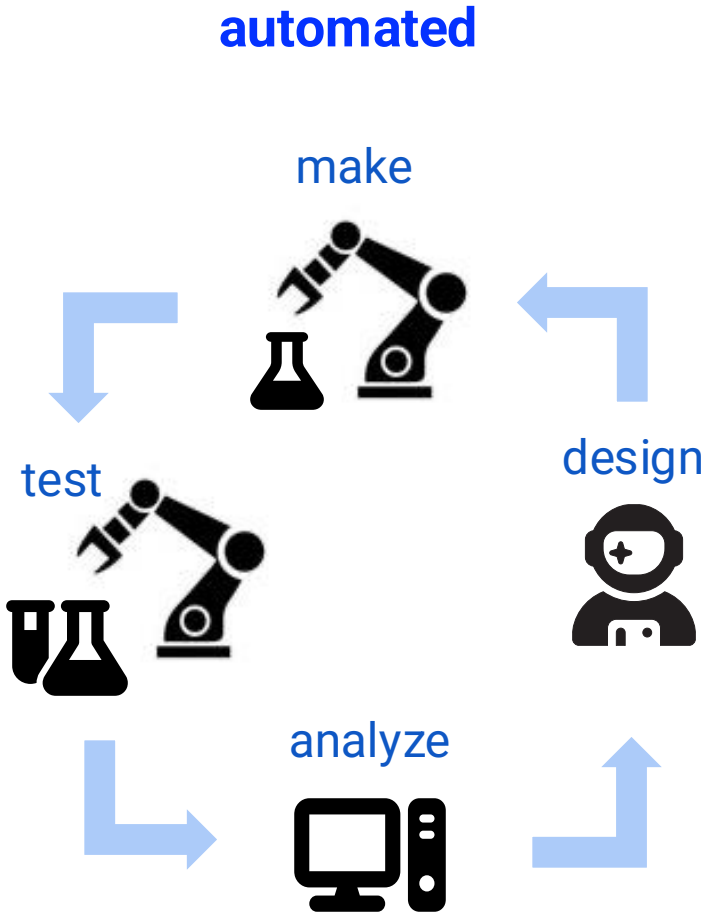
automation isn't automatic

Chem. Sci., **2021**, 12, 15473

where are the bottlenecks?

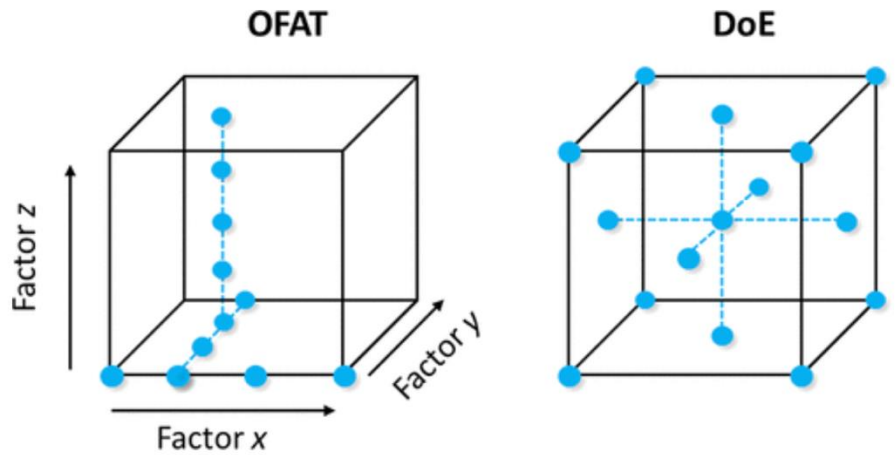
1. ~~manual synthesis, characterization and target property measurements are time and resource intensive~~
2. **chemical intuition is largely limited to *small* changes to *either* experimental conditions *or* structure**

Conventional design in chemistry



where are the bottlenecks?

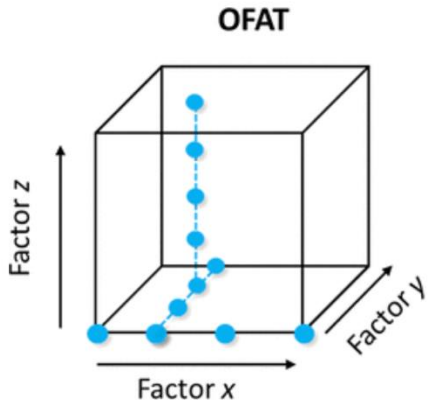
1. manual synthesis, characterization and target property measurements are time and resource intensive
2. **chemical intuition is largely limited to *small* changes to *either* experimental conditions *or* structure**



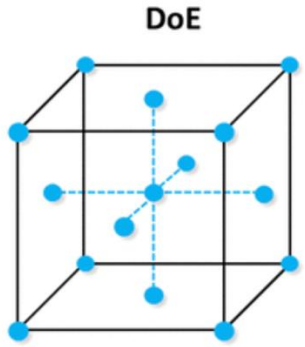
Chem. Rev., **2023**, 123, 6

what are some of the pros and cons of these approaches?

OFAT & DOE – pros and cons



	pros	*cons*
	widely taught	limited experimental space coverage
	straightforward	may miss optimal solution
		fails to identify interactions
		inefficient use of resources



	systematic – thorough coverage of experimental space	minimum entry of experiments (~10)
	efficient – minimal resources required	may need to perform experiments that researcher expects to fail

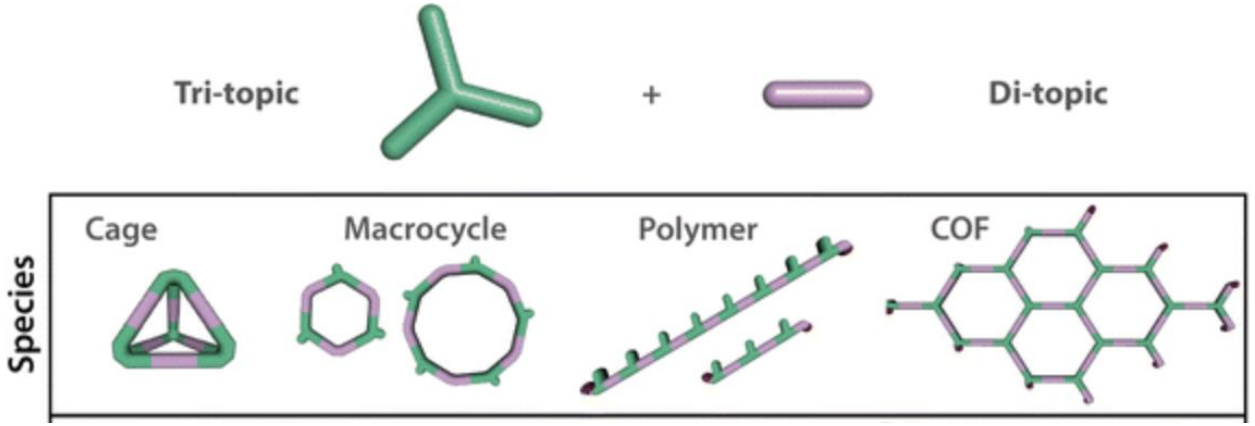
Chem. Rev., **2023**, 123, 6

https://learning.acsgcipr.org/process-design/design-of-experiments/comparing-traditional-approaches-to-experimental-design/

consider chemical space…

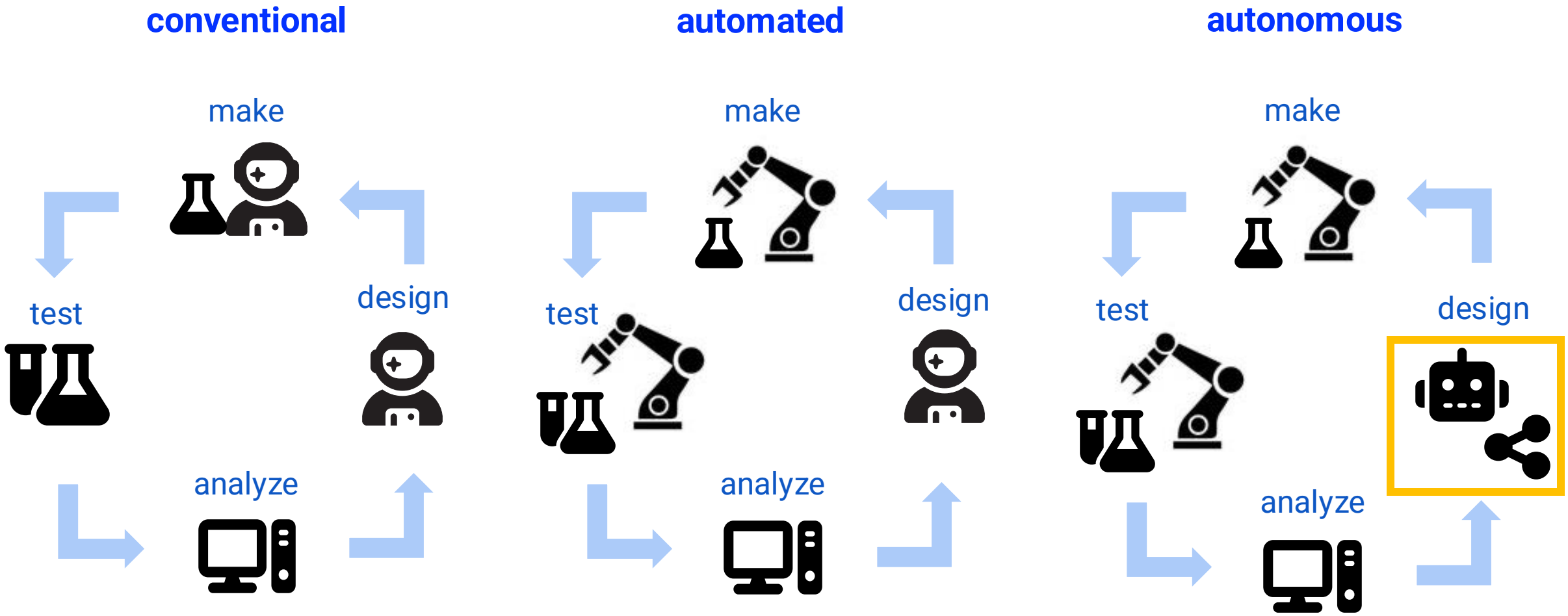
10^{60}

Scale



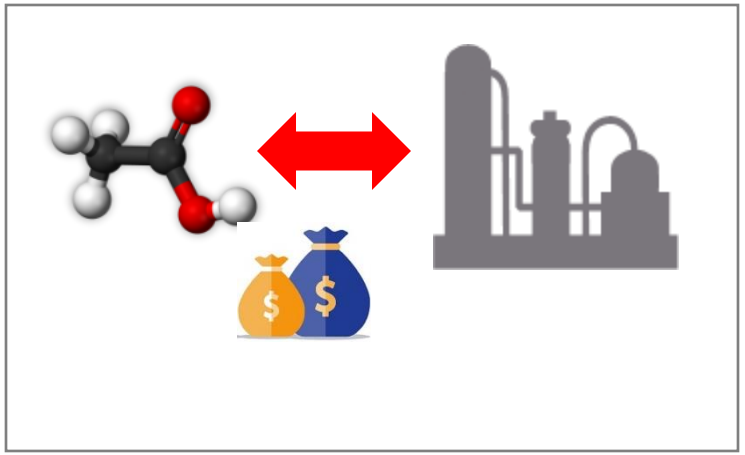
Chem. Sci., **2024**, 15, 6331

Design-make-test-analyze



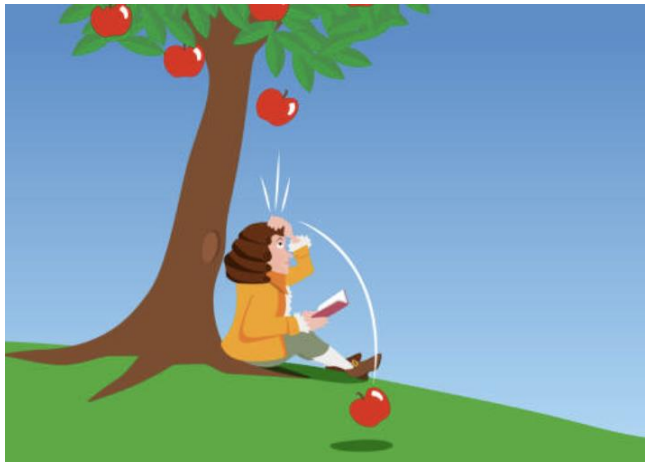
Challenges

Challenge 1



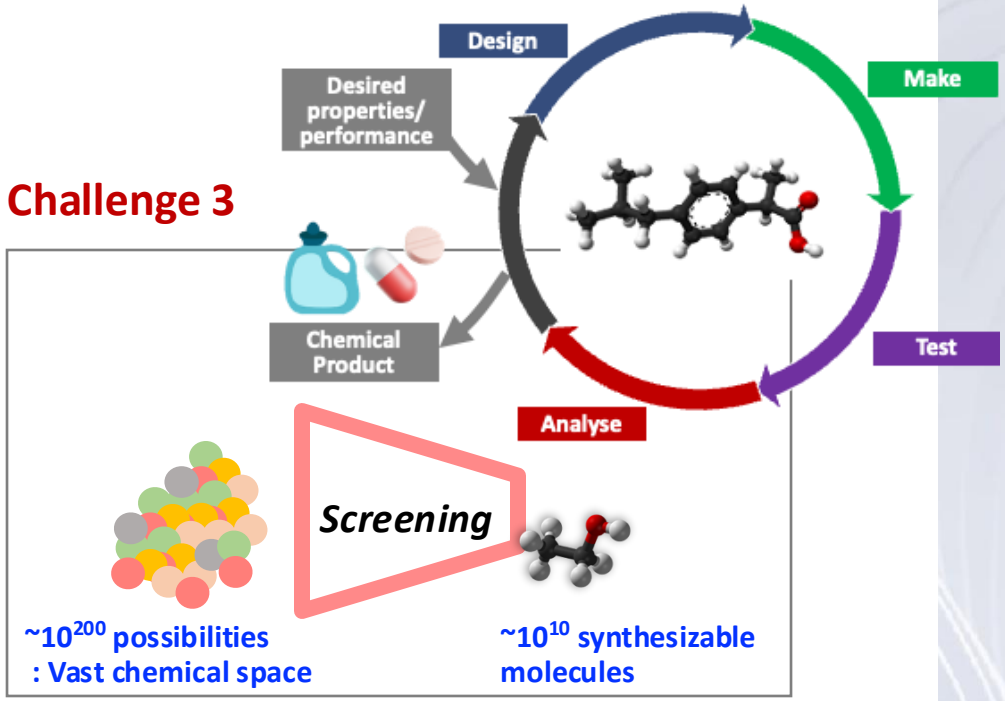
- Overall performance: molecular + system-level decisions
  → high numerical complexity

Challenge 2



- Unknown physical mechanism
- Uncertainty in property prediction models

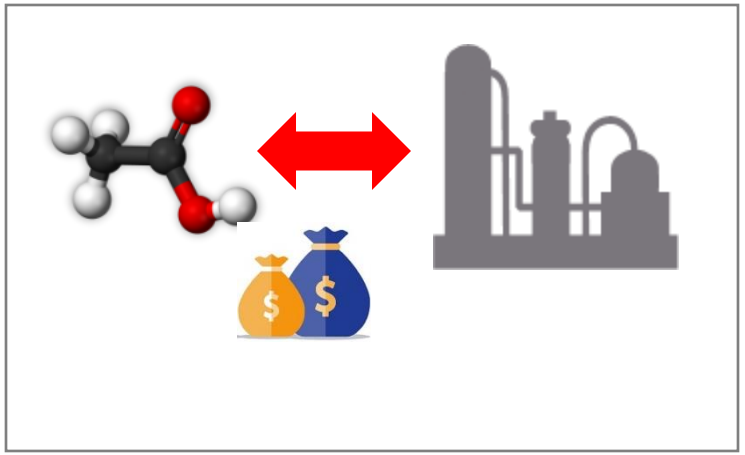
Challenge 3



- Developing a novel chemical involves more than just design
- Synthesizability of molecules
  → Non-smooth & nonlinear function

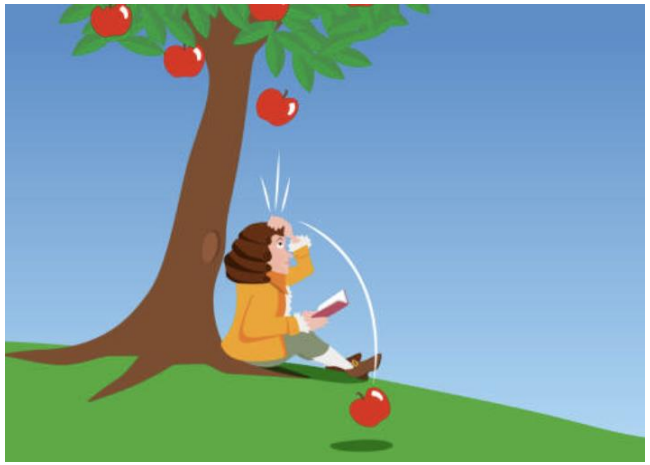
Challenges

Challenge 1



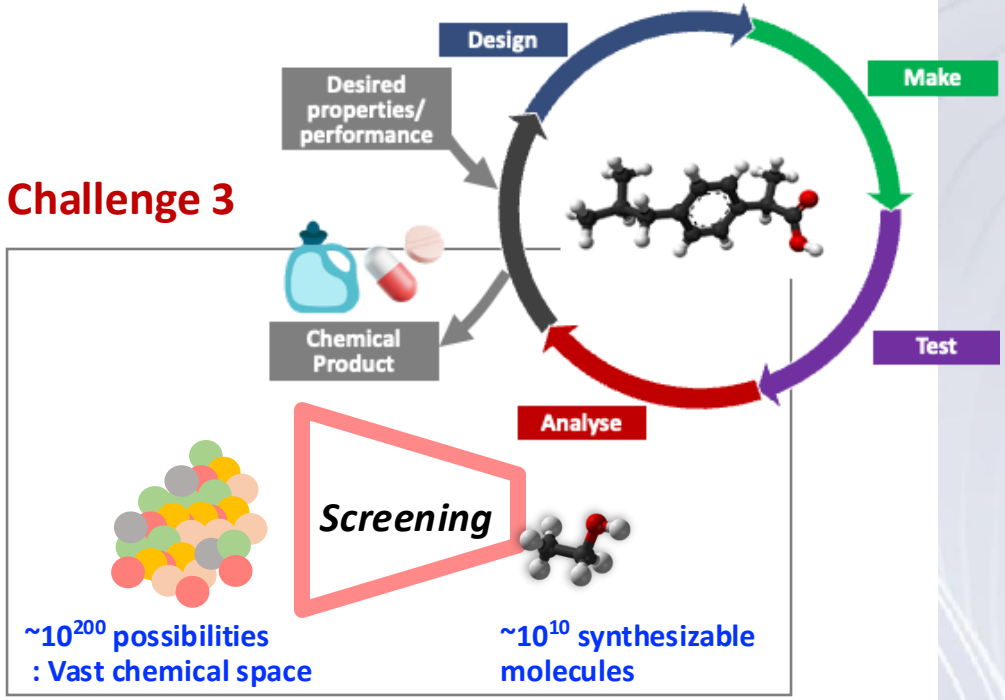
- Overall performance: molecular + system-level decisions
  → high numerical complexity

Challenge 2



- Unknown physical mechanism
- Uncertainty in property prediction models

Challenge 3



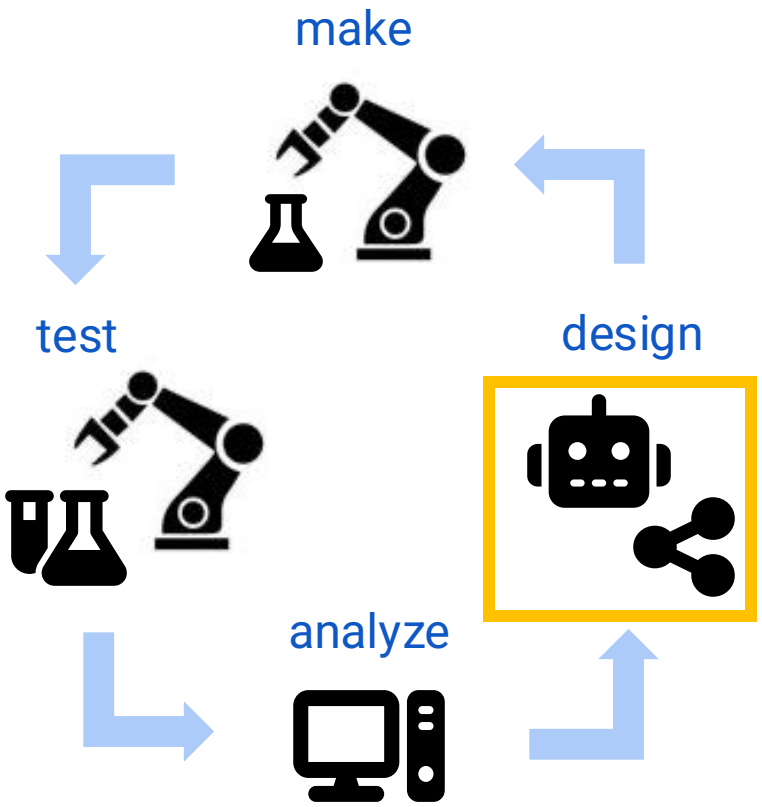
- Developing a novel chemical involves more than just design
- Synthesizability of molecules
  → Non-smooth & nonlinear function

- **Developing new techniques and tools for the design of better products and processes**

Mechanistic Modelling  Statistical modelling
Machine learning techniques

DataDrivenDesign-make-test-analyze

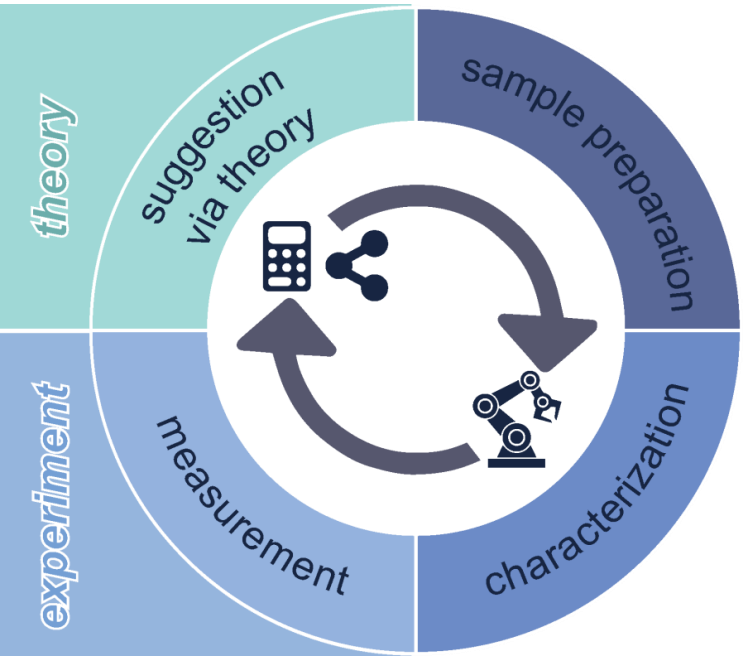
autonomous

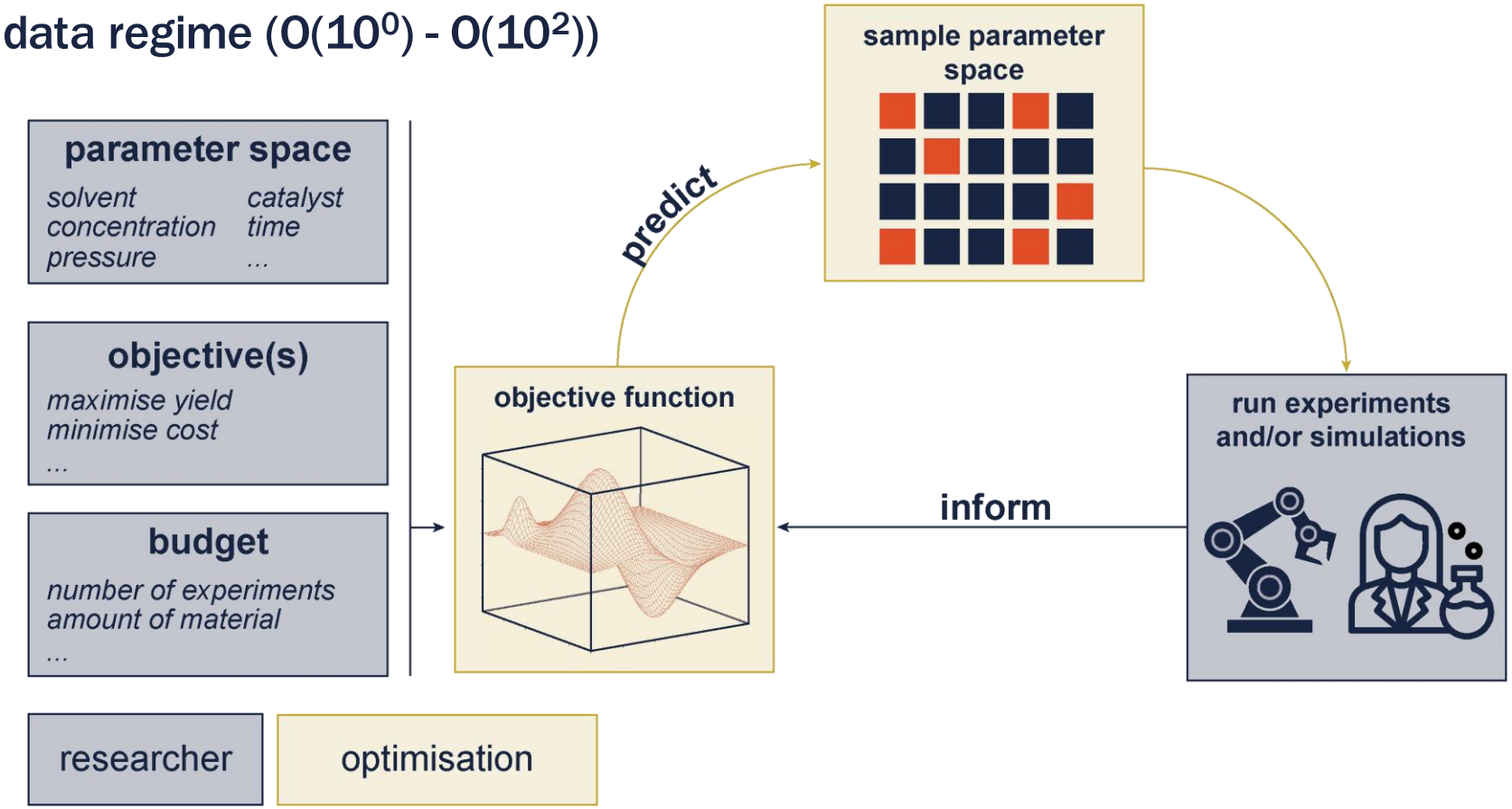


main considerations	*what does our solution need?*
1. we don't know what our objective function looks like	1. we need to be able to sample
2. we are often operating under budgetary and resource constraints	2. resource- and budget—aware
3. we often don't have a lot of starting data, if any at all	3. need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

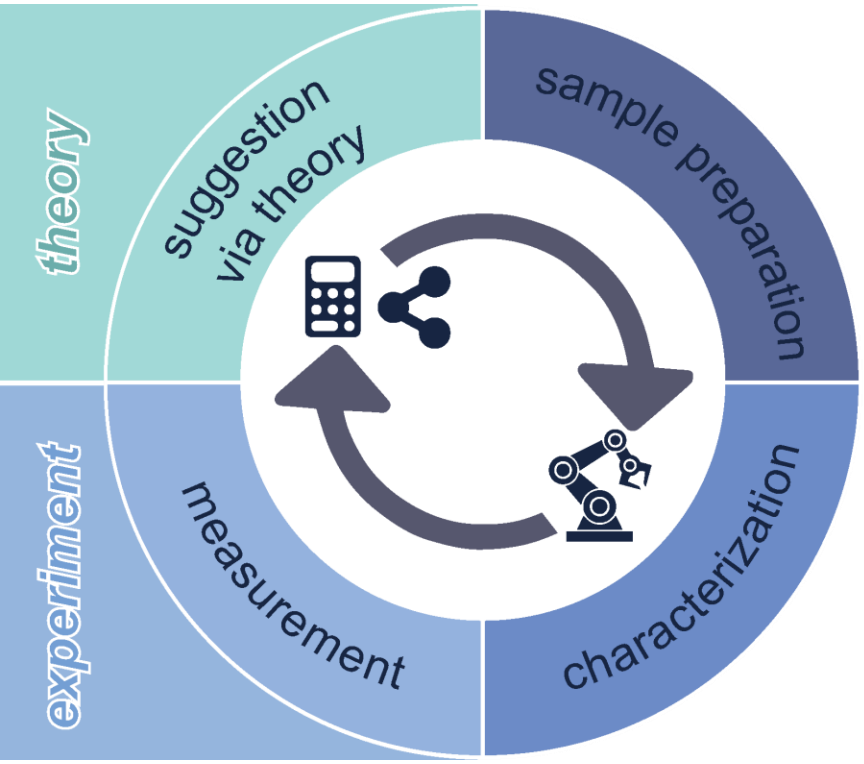
Bayesian optimization for chemistry

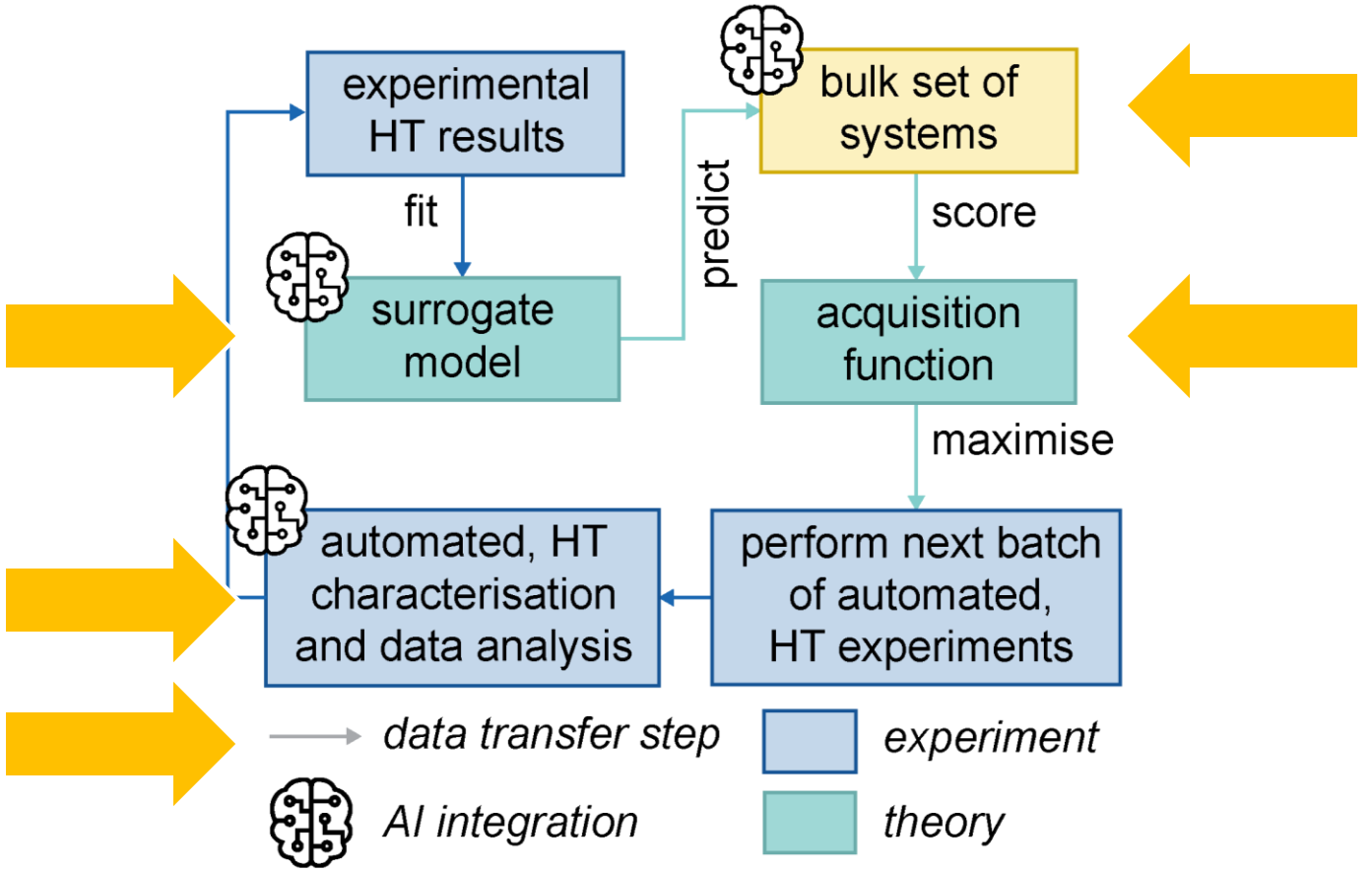
- ✓ resource- and budget-aware
- ✓ balances exploitation and exploration
- ✓ sparse data regime ($O(10^0)$ - $O(10^2)$)





Bayesian optimization for chemistry





chemistry-specific considerations

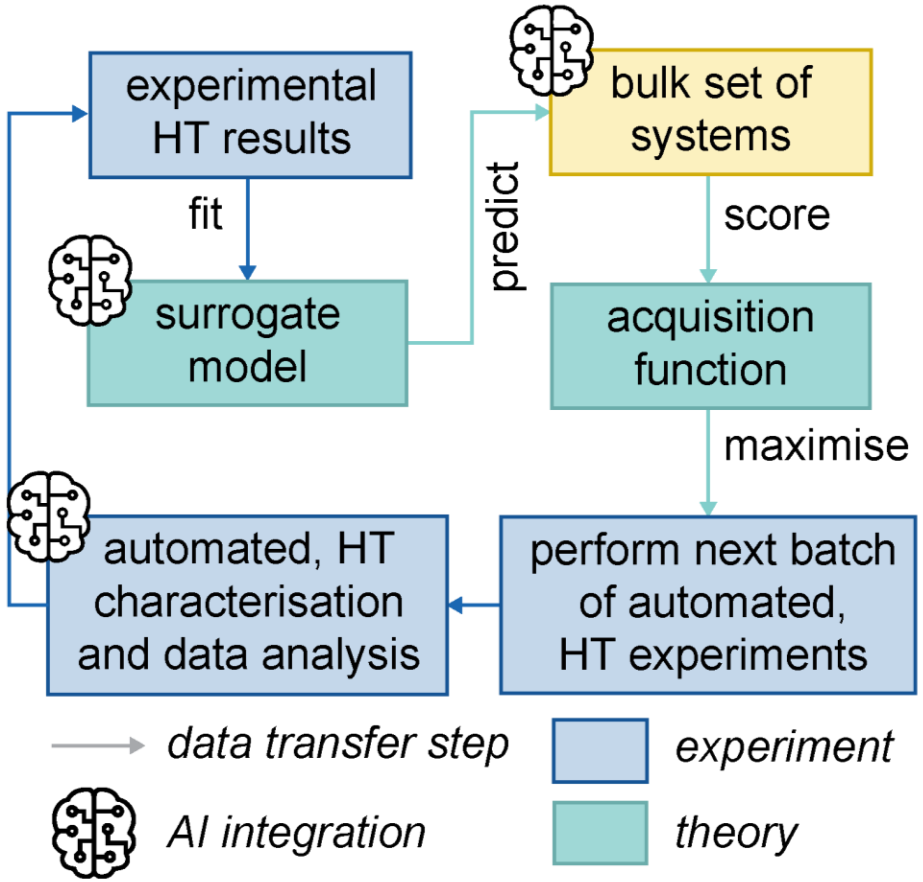
chemistry-specific considerations and problem formulations

objective

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

information streams

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



design space

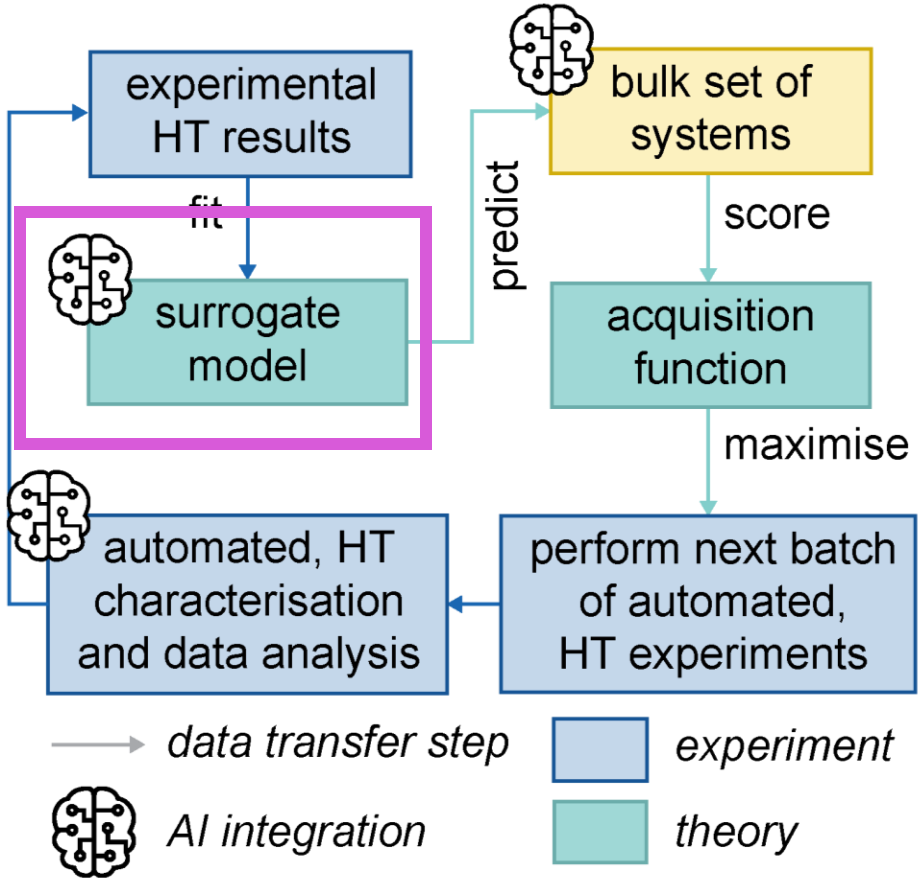
- integer, continuous, categorical, chemical?
- chemical representation / encoding

stopping criteria

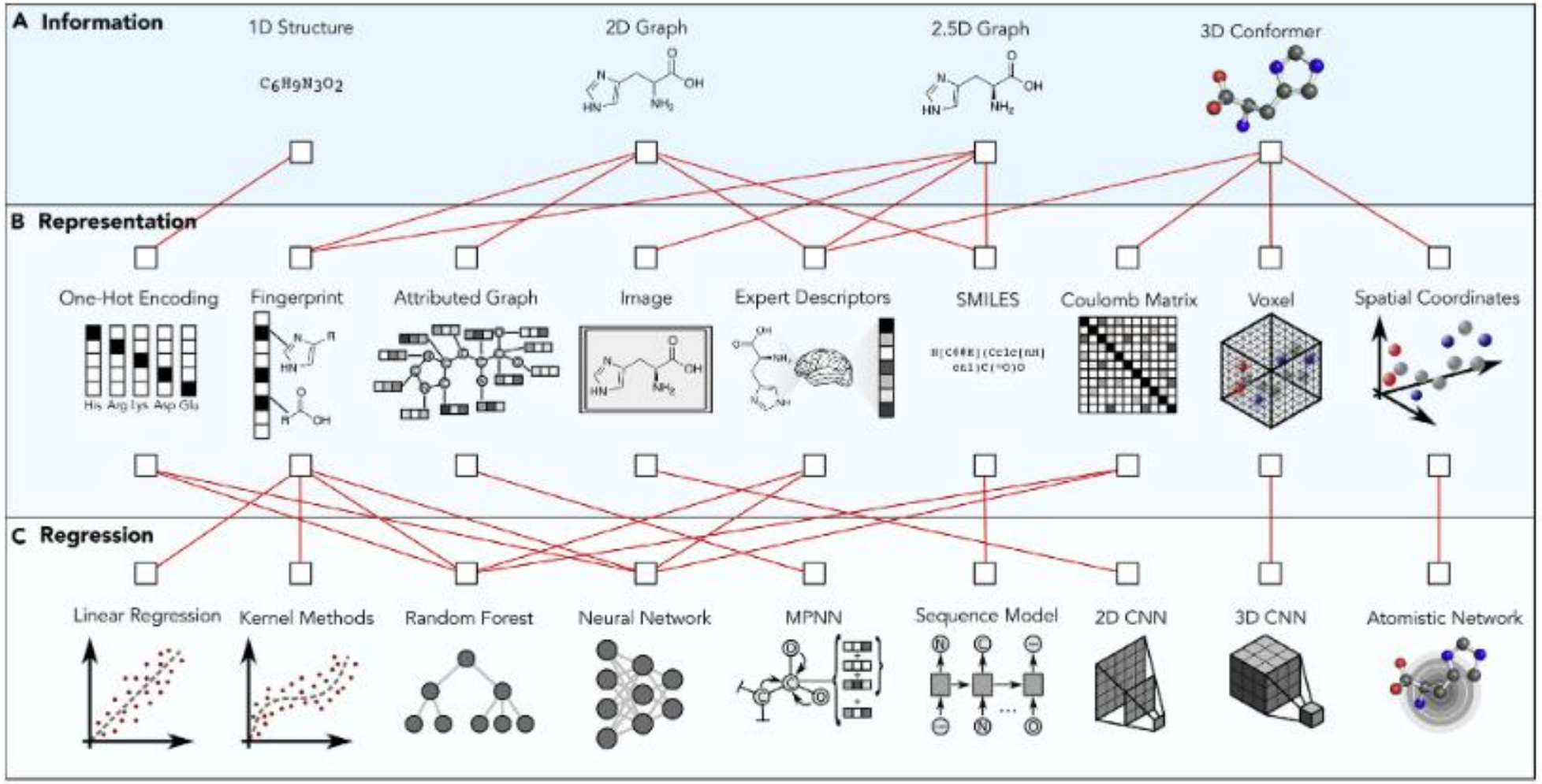
- precursor amounts
- number of experiments
- total time

chemical representations & surrogate model selection

chemical representation



Overview of chemical representations

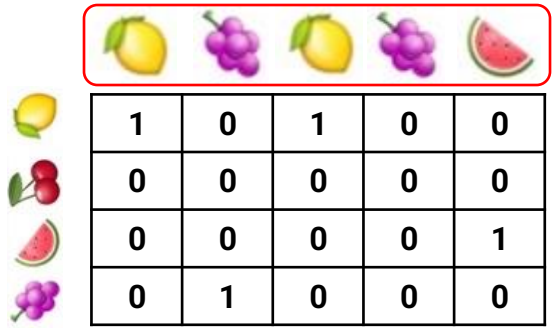


Chem., **2020**, 6, 1204

One-Hot Encoding

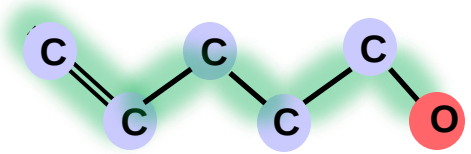
- The conversion of categorical information into a format that may be fed into machine learning.

Example 1: Simple



1	0	1	0	0
0	0	0	0	0
0	0	0	0	1
0	1	0	0	0

Example 2: molecule



4-Penten-1-ol
SMILES: OCCCC=C

	O	C	C	C	C	=	C	…
C	0	1	1	1	1	0	1	0
N	0	0	0	0	0	0	0	0
O	1	0	0	0	0	0	0	0
…	0	0	0	0	0	0	0	0
=	0	0	0	0	0	1	0	0

pros

1. resulting representation prevents ordinal misinterpretation & prevents ranking issues
2. works best for low-dimensional cases
   1. (*e.g.* small number of categories)

cons

1. curse of dimensionality
2. may increase data sparsity

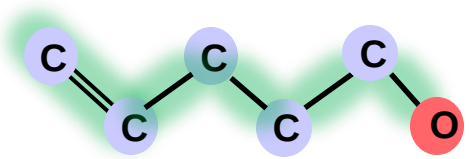
what are some of the pros & cons?

Text-based (or String-based) Representation

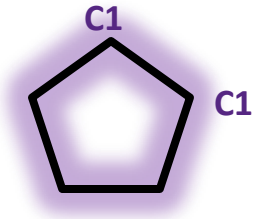
❑ SMILES

- Simplified Molecular-Input Line-Entry system
- It is a line notation for describing the structure of chemical species
- A series of characters that represents atoms and bonds.

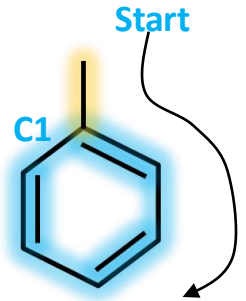
→ Normally, hydrogen atoms are not explicitly shown.



4-Penten-1-ol
SMILES: OCCCC=C



Cyclopentane
SMILES: C1CCCC1



Toluene

Do it yourself !

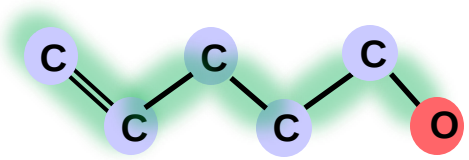
- Some vocab:
  = : double bond, #: triple bond
  () : branches
  Digits: ring

Text-based (or String-based) Representation

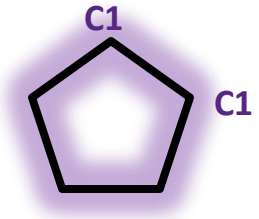
❑ SMILES

- Simplified Molecular-Input Line-Entry system
- It is a line notation for describing the structure of chemical species
- A series of characters that represents atoms and bonds.

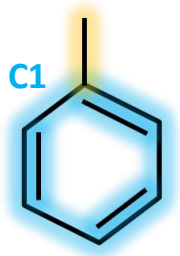
→ Normally, hydrogen atoms are not explicitly shown.



4-Penten-1-ol
SMILES: OCCCC=C



Cyclopentane
SMILES: C1CCCC1



Toluene
SMILES: CC1=CC=CC=C1 (Pubchem)
Cc1ccccc1
c1(C)ccccc1

- Some vocab:
  = : double bond, #: triple bond
  () : branches
  Digits: ring

What can you notice here?

Text-based (or String-based) Representation

❑ SMILES

- It is necessary to select a "unique SMILES" for a molecu
  - ➔ "Canonicalization"

similar molecules can have very different SMILES representation → fail to capture structural similarity!

Canonical SMILES

PubChem Cyclopentane (Compound)

2.1.3 InChIKey

RGSFGYAAUTVSQA-UHFFFAOYSA-N

Computed by InChI 1.0.6 (PubChem release 2021.10.14)

PubChem

2.1.4 Canonical SMILES

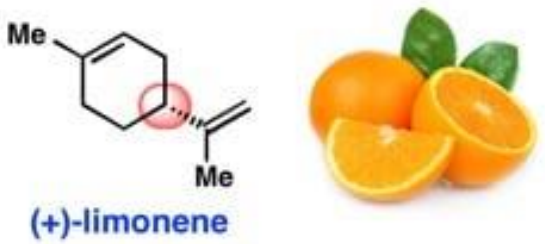
C1CCCC1

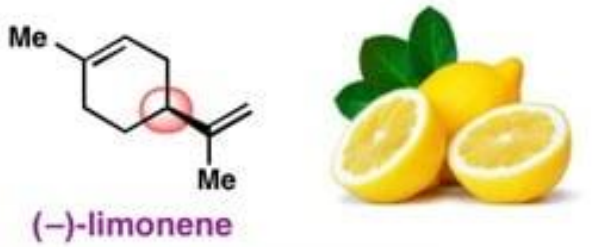
Computed by OEChem 2.3.0 (PubChem release 2021.10.14)

PubChem

2.2 Molecular Formula

C_5H_{10}

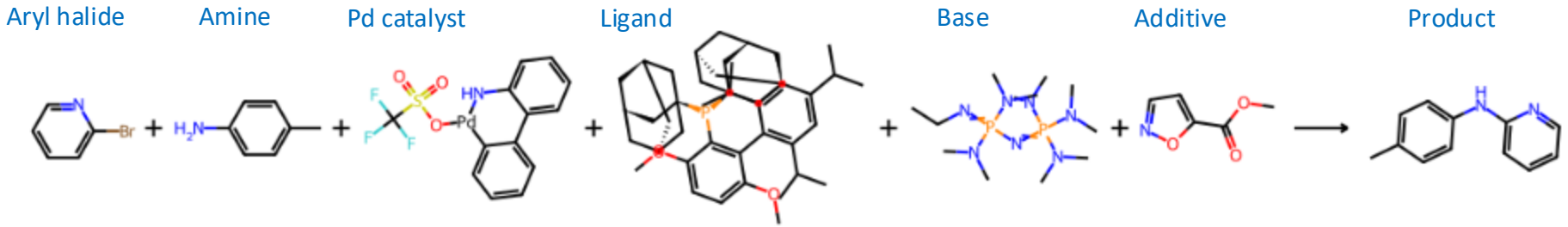




Text-based (or String-based) Representation

Reaction Representation

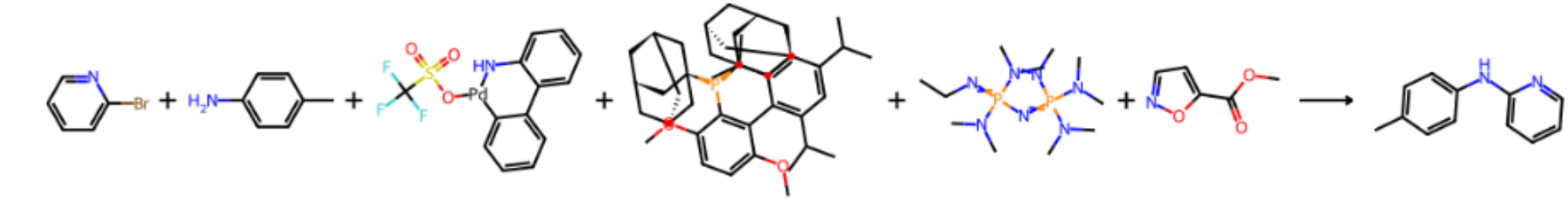
- The SMILES format used for describing molecules has been extended to so-called Reaction SMILES
- Each molecule in the reactants, agents, and products is represented by a SMILES string, and disconnected structures are separated by a period; this includes the individual molecules, ions and ligands, which are listed in an arbitrary manner. Reactants, agents, and products are separated by either the '>' or '≫' symbol



Text-based (or String-based) Representation

❑ Reaction Representation

- The SMILES format used for describing molecules has been extended to so-called Reaction SMILES
- Each molecule in the reactants, agents, and products is represented by a SMILES string, and disconnected structures are separated by a period; this includes the individual molecules, ions and ligands, which are listed in an arbitrary manner. Reactants, agents, and products are separated by either the '>' or '≫' symbol

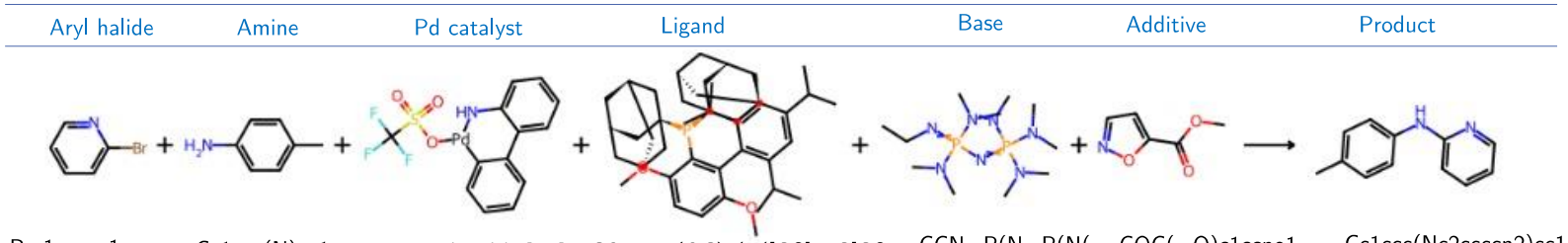


Aryl halide	Amine	Pd catalyst	Ligand	Base	Additive	Product
Brc1ccccn1	Cc1ccc(N)cc1	O=S(=O)(O[Pd]1c2ccccc2-c2ccccc2N~1)C(F)(F)F	COc1ccc(OC)c(P([C@]23C[C@H]4C[C@H](C[C@H](C4)C2)C3)[C@]23C[C@H]4C[C@H](C[C@H](C4)C2)C3)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C	CCN=P(N=P(N(C)C)(N(C)C)N(C)C)(N(C)C)N(C)C	COC(=O)c1ccno1	Cc1ccc(Nc2ccccn2)cc1

Sargent Centre for Process Systems Engineering

Text-based (or String-based) Representation

❑ Reaction Representation



Dose the order of precursor matter?
Does the SMILES have physical meaning?

- Each reactant, solvent, additive, catalyst is separated by a "."
- Sometime, for the substrate which do not contribute atoms to the product is separated by >
- A product is separated by ≫ (precursors ≫ products) OR (reactants > substrates ≫ products)

- **Reaction SMILES:**

Brc1ccccn1.Cc1ccc(N)cc1.O=S(=O)(O[Pd]1c2ccccc2-c2ccccc2N~1)C(F)(F)F.COc1ccc(OC)c(P([C@]23C[C@H]4C[C@H](C[C@H](C4)C2)C3)[C@]23C[C@H]4C[C@H](C[C@H](C4)C2)C3)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C.CCN=P(N=P(N(C)C)(N(C)C)N(C)C)(N(C)C)N(C)C.COC(=O)c1ccno1≫Cc1ccc(Nc2ccccn2)cc1

OR

Brc1ccccn1.Cc1ccc(N)cc1>O=S(=O)(O[Pd]1c2ccccc2-c2ccccc2N~1)C(F)(F)F.COc1ccc(OC)c(P([C@]23C[C@H]4C[C@H](C[C@H](C4)C2)C3)[C@]23C[C@H]4C[C@H](C[C@H](C4)C2)C3)c1-c1c(C(C)C)cc(C(C)C)cc1C(C)C.CCN=P(N=P(N(C)C)(N(C)C)N(C)C)(N(C)C)N(C)C.COC(=O)c1ccno1≫Cc1ccc(Nc2ccccn2)cc1

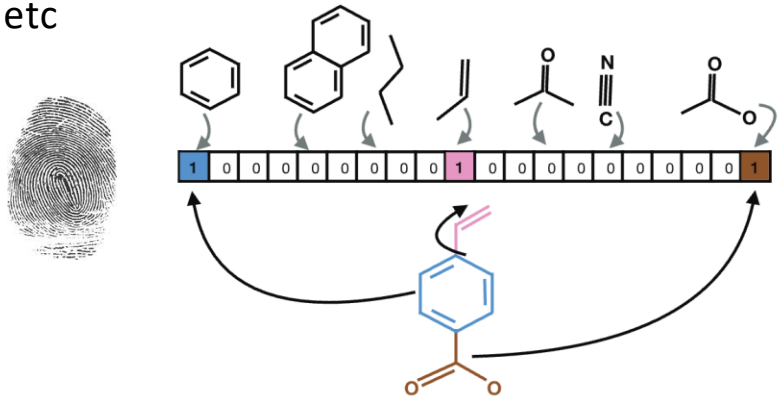
Molecular Fingerprint

- A vectorised representation of molecules capturing precise details of atomic configurations within a molecule.

❑ 1) Structural keys

- In structural keys, the structure of a molecule is encoded into a binary bit string (that is, a sequence of 0's and 1's), each bit of which corresponds to a "pre-defined" structural feature (e.g., substructure or fragment).
- It encode the presence (1) or absence (0) of certain substructures in a compound (one-hot-encoding)
- It is important to understand that structural keys cannot encode structural features that are not pre-defined in the fragment library.
- Examples are the **MACCS keys** and PubChem Fingerprints.

- **MACCSkeys (**Molecular ACCess System)
  - The predefined keys are implemented in popular open-source cheminformatics software packages, including RDKit, OpenBabel, CDK, etc



Substructure = Key
(pre-defined)

A size of feature vector is determined and fixed in a priori.

J Chem Inf Comput Sci., **2002,** 42, 1273

Molecular Fingerprint

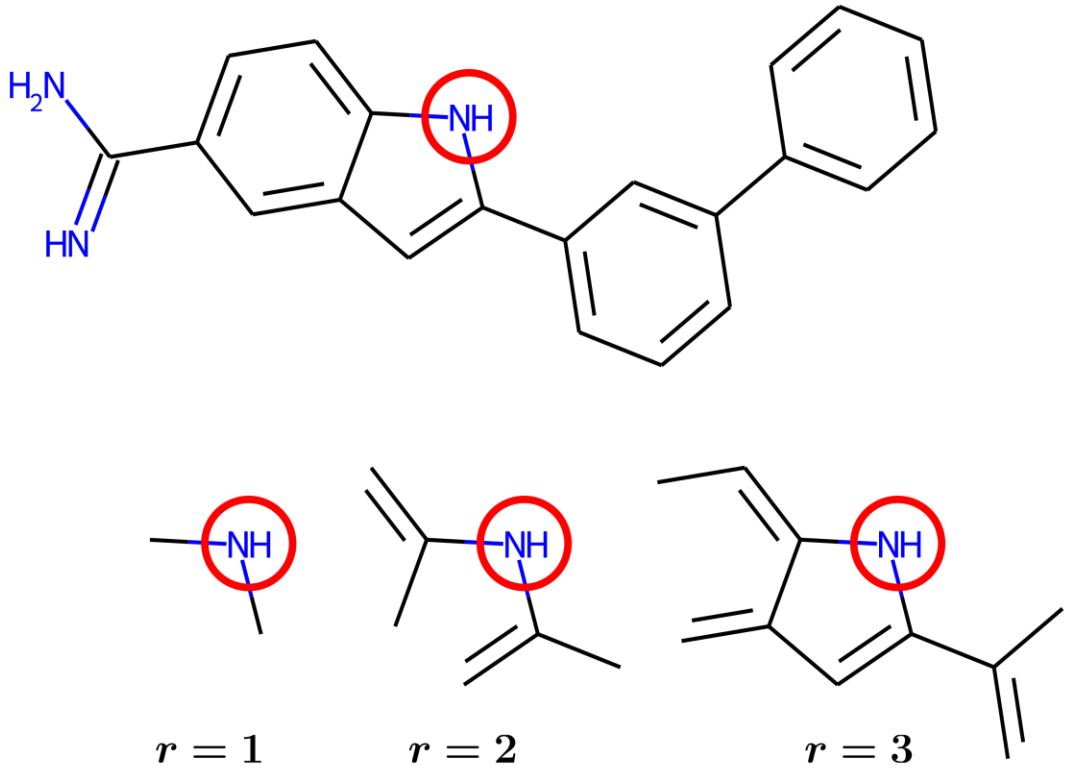
❑ 2) Hashed keys

- An alternative to structural keys.
- Hashed fingerprints do not require a pre-defined fragment library.
- Instead, they are generated by enumerating through the molecule all possible fragments that are not bigger than a certain size and then converting these fragments into numeric values using a "hash" function.
- These numeric values can be used to indicate bit positions in the hashed fingerprints.
- Hashing them into values within a fixed range inevitably results in "bit collisions", in which different fragments are converted into the same numeric value (and the same bit position). Because of this, there is no one-to-one correspondence between fragments and fingerprint bits (contrary to structural keys).
- Hashed fingerprints may be further classified into topological or path-based fingerprints and circular fingerprints, according to the way by which the fragments are enumerated.
- Let's focus on circular fingerprint – Morgan fingerprint

J Chem Inf Comput Sci., **2002,** 42, 1273

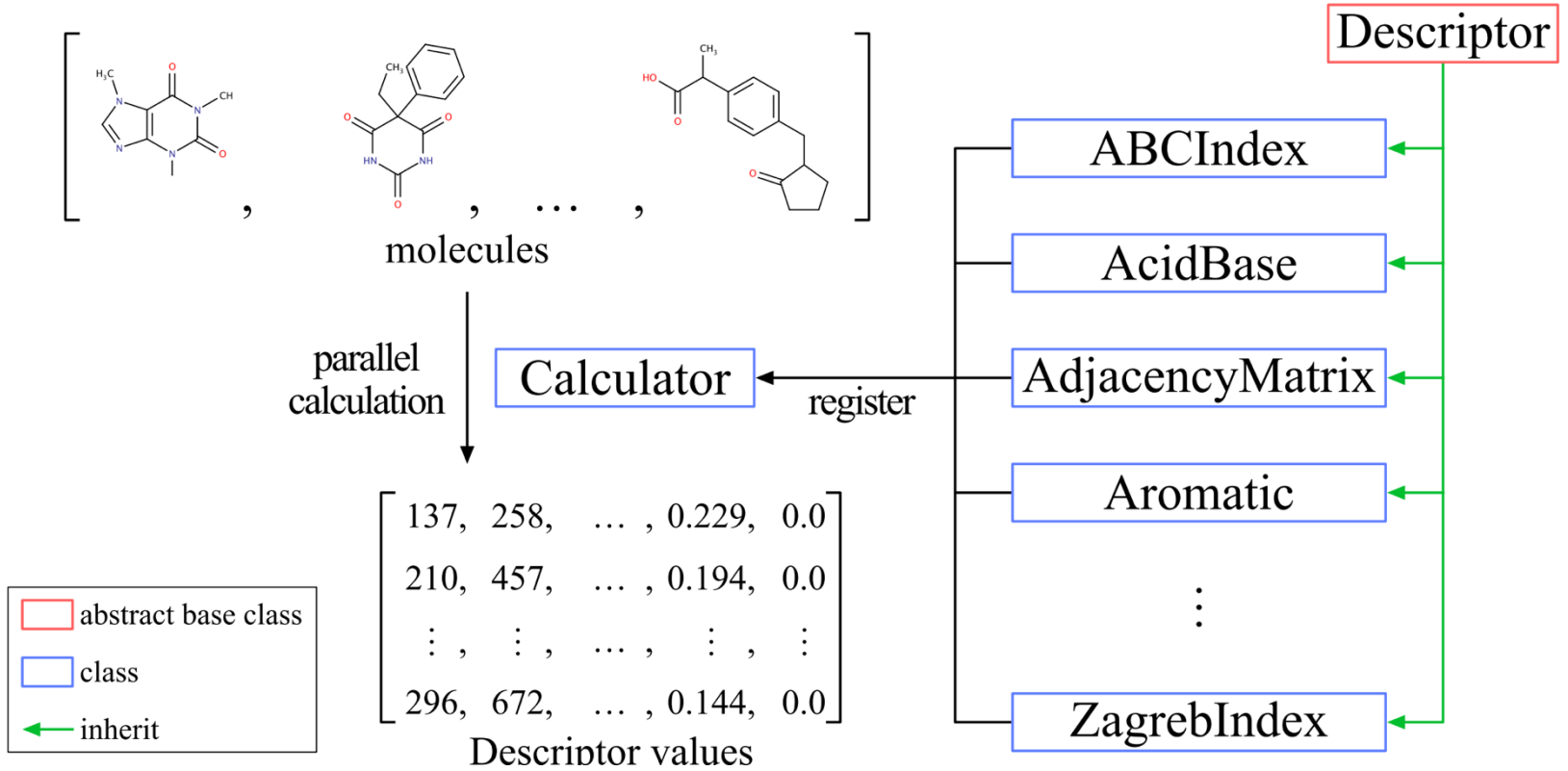
Morgan fingerprints

- family of fingerprints based on the Morgan algorithm
- bits correspond to the circular environments of each atom in a molecule



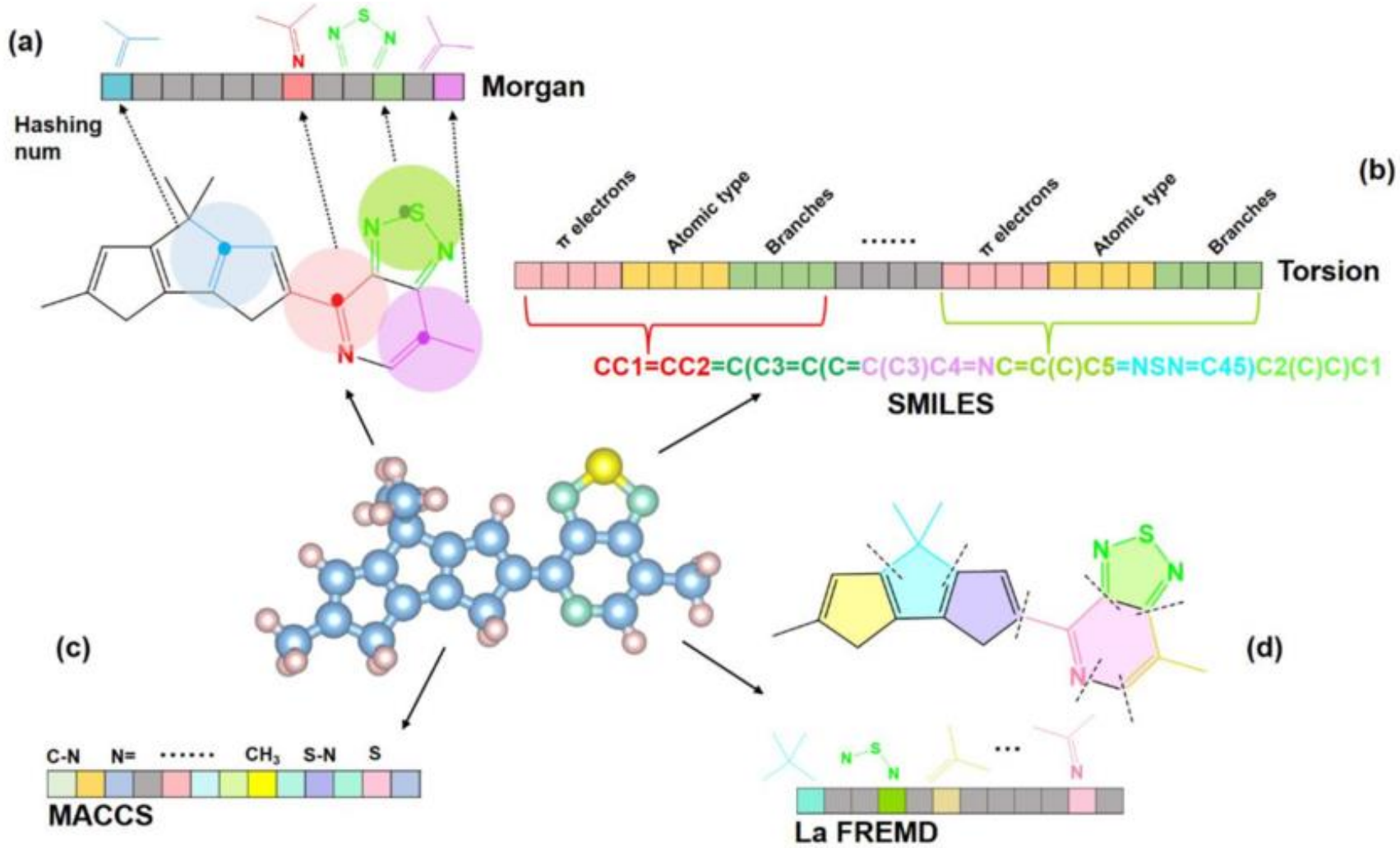
Mordred descriptors

a numerical vector of several molecular properties known to be important in structure-property relationships



J Cheminf., **2018,** 10, 4

representation comparison



ACS Appl Mater. Inter., **2023**, 15, 17

packages to support chemical representations

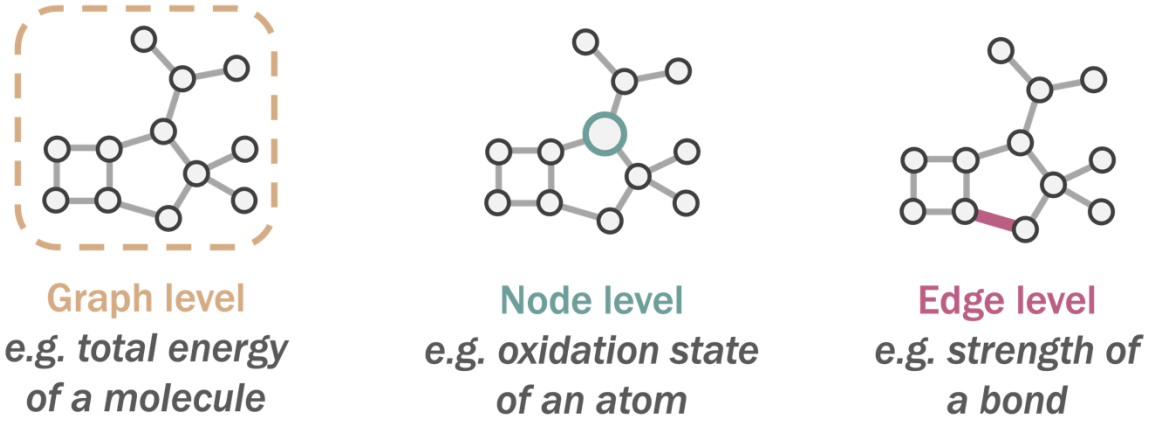
RDKit

MORFEUS

molecular features for machine learning

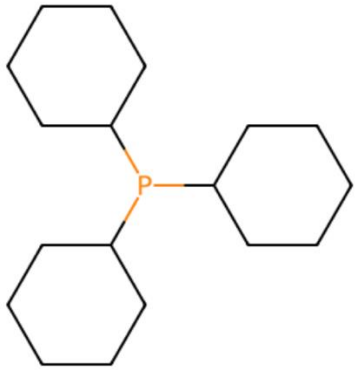
MORDRED

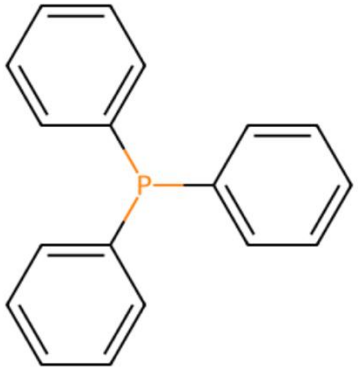
DScribe



- learned embeddings (contrastive learning)
- custom fingerprints (MOFid)

what does it look like in practice?

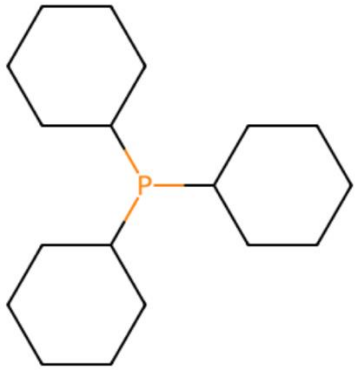


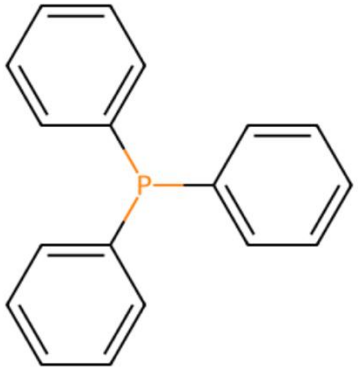


```python
from rdkit import Chem
from rdkit.Chem import Draw

# first, generate the rdkit molecule object from the smiles
mol_L6 = Chem.MolFromSmiles('P(C1=CC=CC=C1)(C2=CC=CC=C2)C3=CC=CC=C3')
Draw.MolToImage(mol_L6)

mol_L5 = Chem.MolFromSmiles('C3CCC(P(C1CCCCC1)C2CCCCC2)CC3')
Draw.MolToImage(mol_L5)
```

snappify.com

what does it look like in practice?

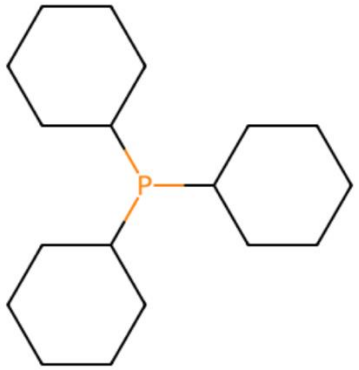


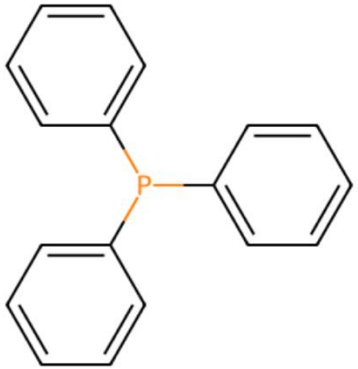


```python
from rdkit import Chem
from rdkit.Chem import Draw, AllChem

# first, generate the rdkit molecule object from the smiles
smiles_list = [
  'C3CCC(P(C1CCCCC1)C2CCCCC2)CC3',
  'P(C1=CC=CC=C1)(C2=CC=CC=C2)C3=CC=CC=C3',
  ]
rdmols = [Chem.MolFromSmiles(smiles) for smiles in smiles_list]

from rdkit.Chem import AllChem
fpgen = AllChem.GetMorganGenerator(radius=3)
morgan_fingerprints = [
  fpgen.GetSparseCountFingerprint(mol) for mol in rdmols
  ]
```

snappify.com

what does it look like in practice?






```python
from mordred import Calculator, descriptors

# create descriptor calculator with all descriptors
calc_2D = Calculator(descriptors, ignore_3D=True)
calc_3D = Calculator(descriptors, ignore_3D=False)

df_2D = calc_2D.pandas(rdmols)
df_3D = calc_3D.pandas(rdmols)
```

snappify.com

BO implementations

Implementing BO for chemistry

Table 2 A collection of open-source Python software libraries for Bayesian optimisation. Surrogate models include Gaussian process (GP), random forest (RF) and tree of Parzen estimators (TPE)

Package	Models	Features	License	Ref.
General purpose				
Ax[a]	GP, others	Modular framework built on BoTorch	MIT	
Bayesianopt[b]	GP	Parallel optimisation	MIT	29
BayesOpt[c]	GP	Single objective	MIT	30
BoTorch[d]	GP, others	Multi-objective optimisation	MIT	31
COMBO[e]	GP	Multi-objective optimisation	MIT	32
Dragonfly[f]	GP	Multi-fidelity optimisation	Apache	33
GPyOpt[g]	GP	Parallel optimisation	BSD	34
Hyperopt[h]	TPE	Serial/parallel optimisation	BSD	35
NEXTorch[i]	GP, others	Modular framework built on BoTorch	MT	36
Optuna[j]	RF	Hyperparameter tuning	MIT	37
Skopt[k]	RF, GP	Batch optimisation	BSD	38
SMAC3[l]	GP, RF	Hyperparameter tuning	BSD	39
GPax[m]	GP	Multi-task/fidelity	MIT	40 and 41
Physical science domain				
Atlas[n]	GP	Mixed-parameter optimisation for self-driving labs	MIT	42
BOSS[o]	GP	Crystal structure optimisation	Apache	43
Edbo[p]	GP	Tailored chemical synthesis descriptors	MIT	5
GAUCHE[q]	GP	Tailored molecular representations	MIT	44
NUBO[r]	GP	Transparent BO to personalise problem	BSD	45
Olympus[s]	GP, TPE, BNN	Benchmarking and noisy optimisation	MIT	46
Phoenics[t]	BNN	Bayesian kernel density estimation	Apache	47
Summit[u]	GP, RF	Multi-task optimisation for chemical reactions	MIT	48

Digital Discovery, **2024**, 3, 1086

(this is not an exhaustive list!)

chemistry-specific BO packages and tools

- What makes something chemistry-specific?
  - Chemical/Reaction encoding – one-hot encoding, Mordred, ECFP
- Packages that can readily implement this
  - GAUCHE
  - BayBE
  - BOFire

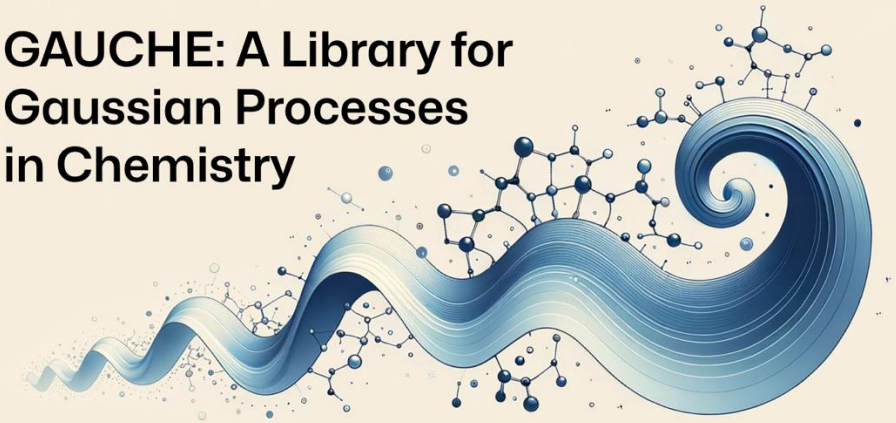
but what if I don't want to code?

***open-source* GUIs for BO!**

BayBE
Bayesian Back End

BoFire

GAUCHE: A Library for Gaussian Processes in Chemistry



chemistry-specific BO packages and tools

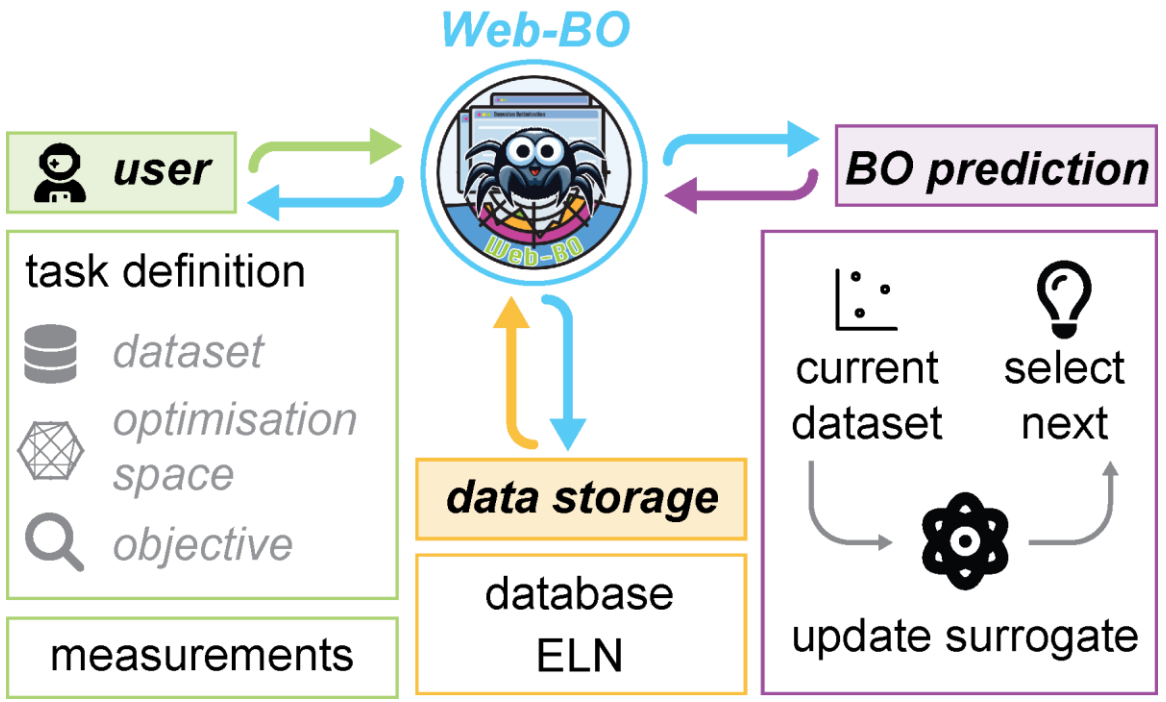
- What makes something chemistry-specific?
  - Chemical/Reaction encoding – one-hot encoding, Mordred, ECFP
- Packages that can readily implement this
  - GAUCHE
  - BayBE
  - BOFire

but what if I don't want to code?

***open-source* GUIs for BO!**

Web-BO

Improving accessibility of data-driven optimisation for chemical tasks via a graphical user interface



Faraday Discuss., **2024**

Interactive Web-BO task:

Let’s compare representations for reaction optimization

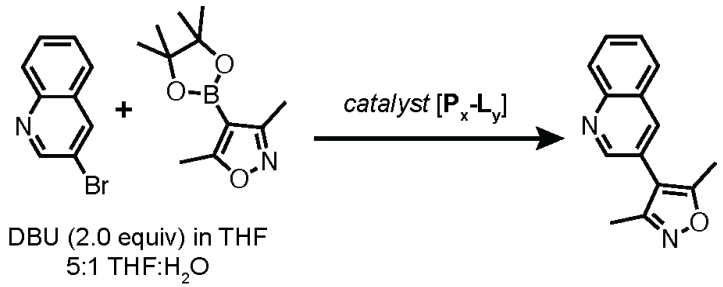
Emulators as tools for benchmarking

- Allow us to explore varying algorithm formulations without performing experiments
- Packages that offer emulators supported/maintained for recent versions of PyTorch
  - SUMMIT
  - EDBO
- **Example cases**
  - Suzuki-Miyura Cross Coupling optimisation

Sargent Centre for Process Systems Engineering

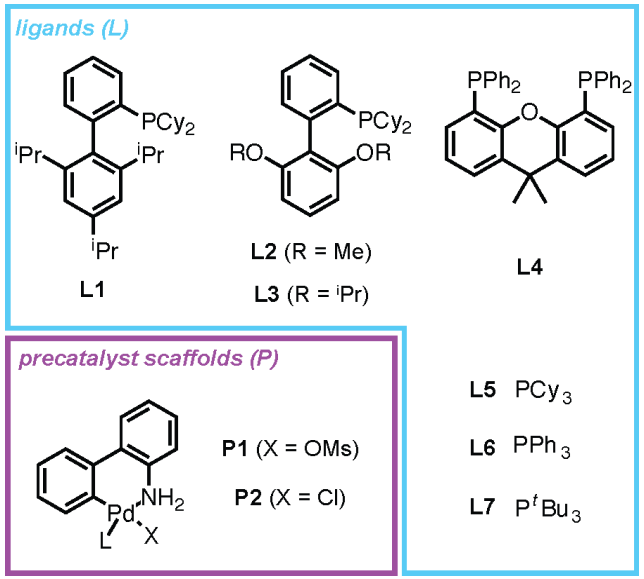
Emulators as tools for benchmarking

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



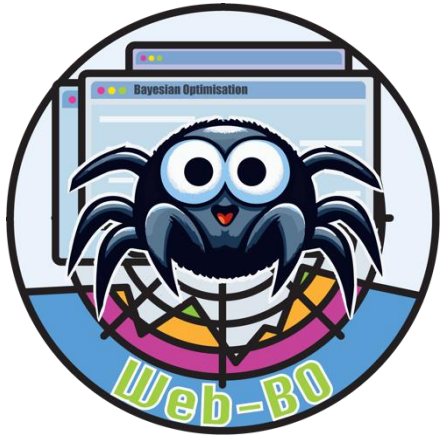
”experiments” are supported by an emulator developed by Lapkin, et al.

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.)

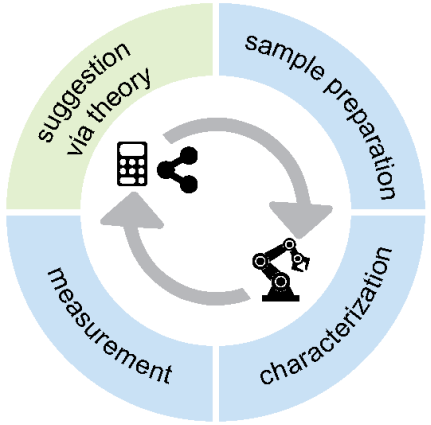


let’s see how different chemical representations compare for our catalysts

Let's code!



suprashare



tutorial notebooks

break