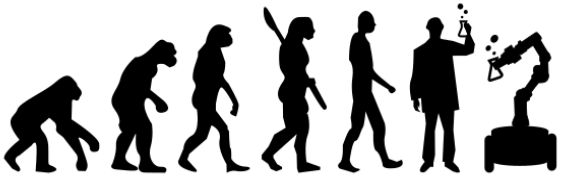
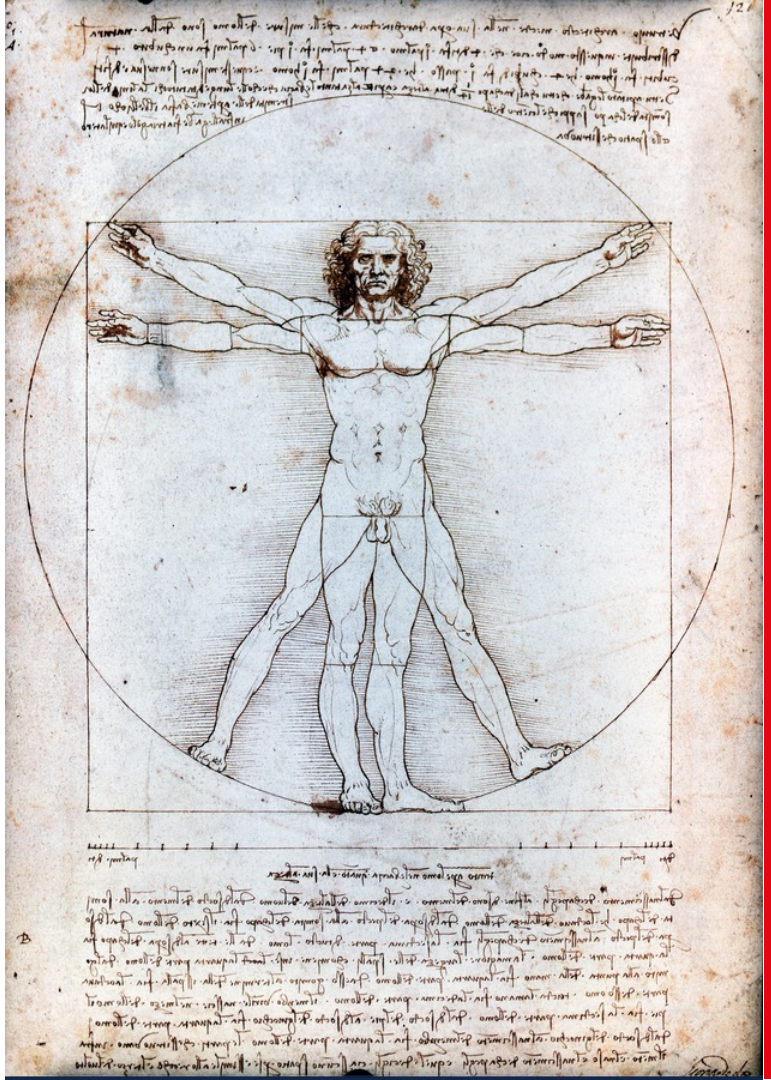


Automated and data-driven chemistry

Lecture 4 first part :
Automation in chemistry II – Self driving lab & closing the loop

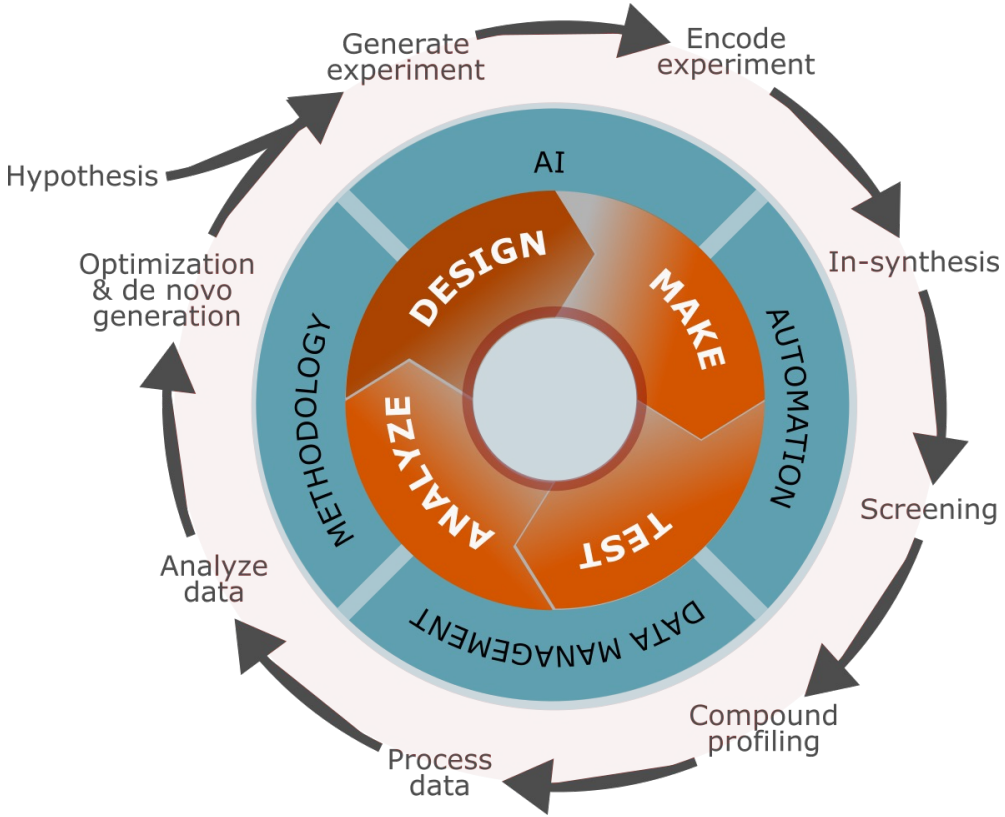
Autumn 2025

Pascal Miéville
Stefano di Leone
Edy Mariano
Jean-Charles Cousty



Intro – SDL & DMTA

Intro – The DMTA/L cycle

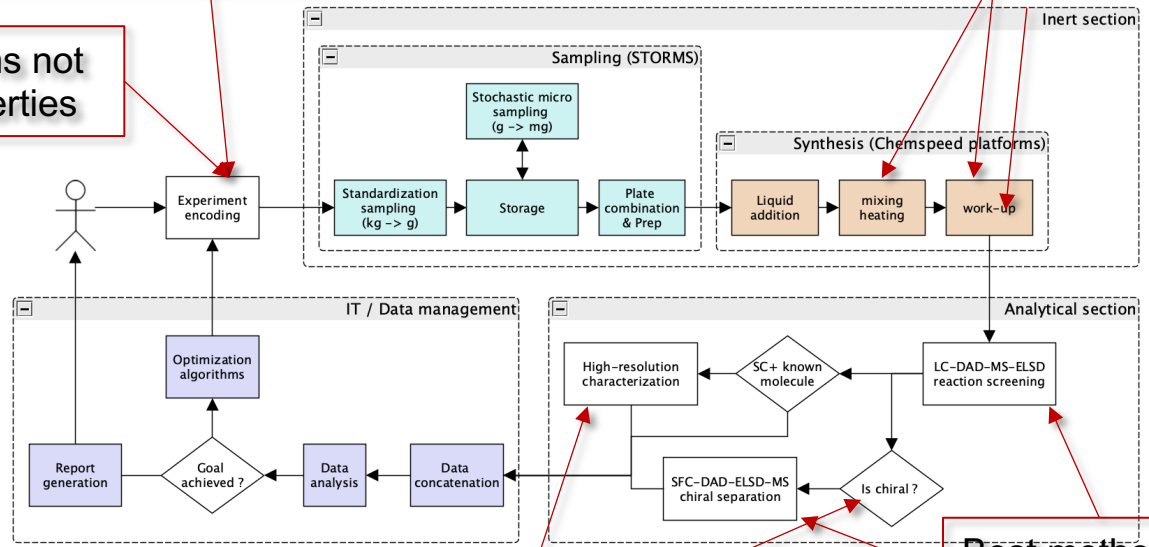


Intro – Why HTE cannot work and requires SDL ?

Reactivity to chemistry events :
- precipitation
- emulsion

Encoding work for 300 reactions/day

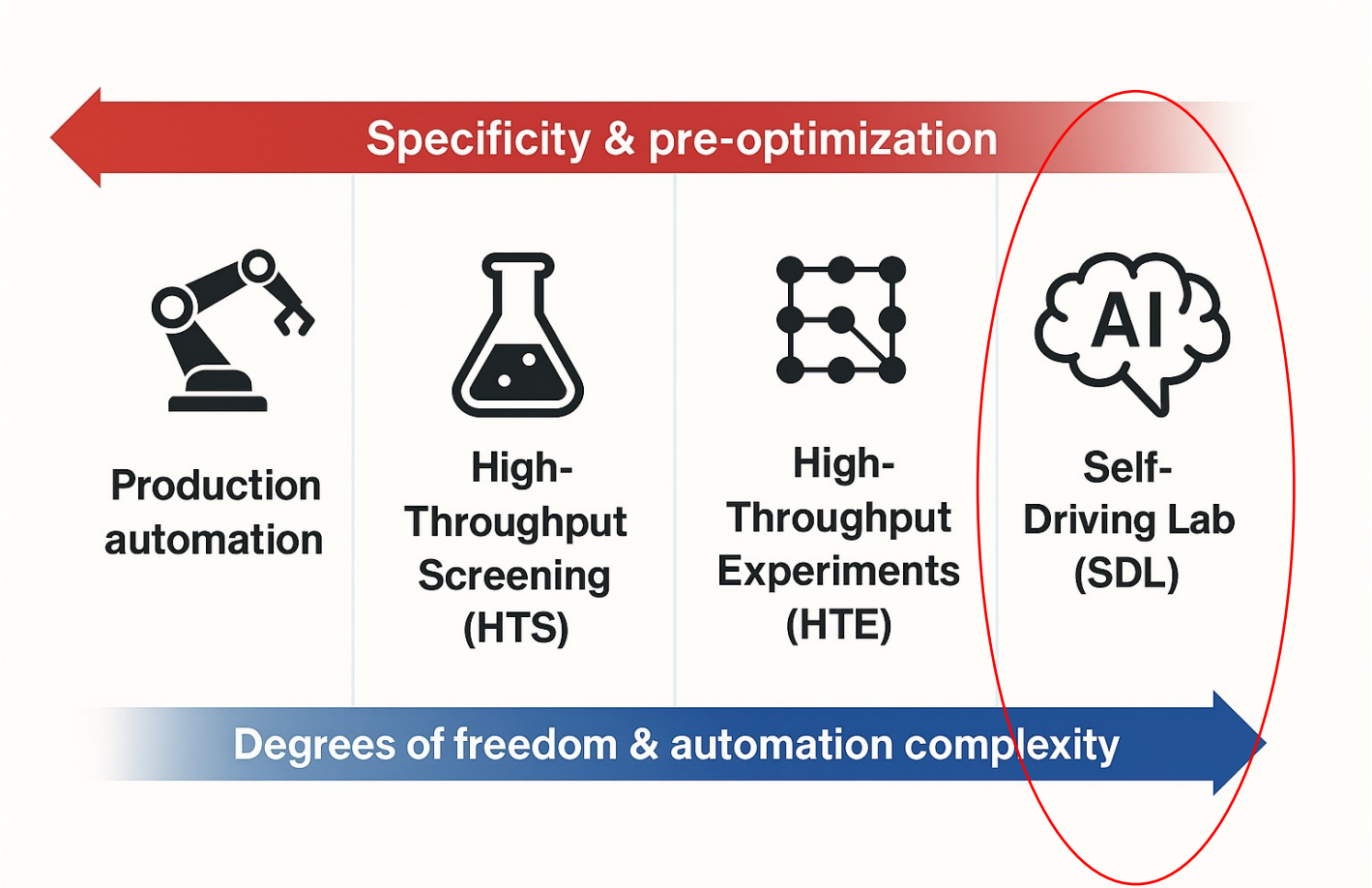
Generative algorithms not aware of setup properties



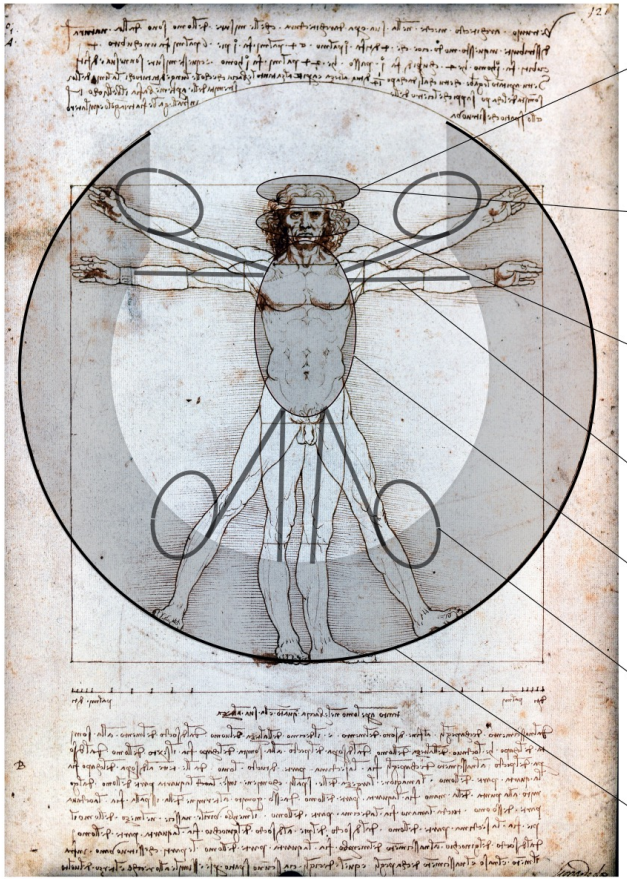
Is the compound chiral ?
Is the compound already known ?

Best method choice ?
Best solvent for sample prep ?

Intro - lab automation strategies



Intro – the embodied self-driving lab



1. Counscious layer
High level cognition
Strategic decisions
Mid to long term plans

2. Memory layer
Short and long terms information storage
Can be supported by external memories (Libraries, Cloud...)

3. External communication layer
Communication with humans
Communication with other laboratories
Communication with generative IA

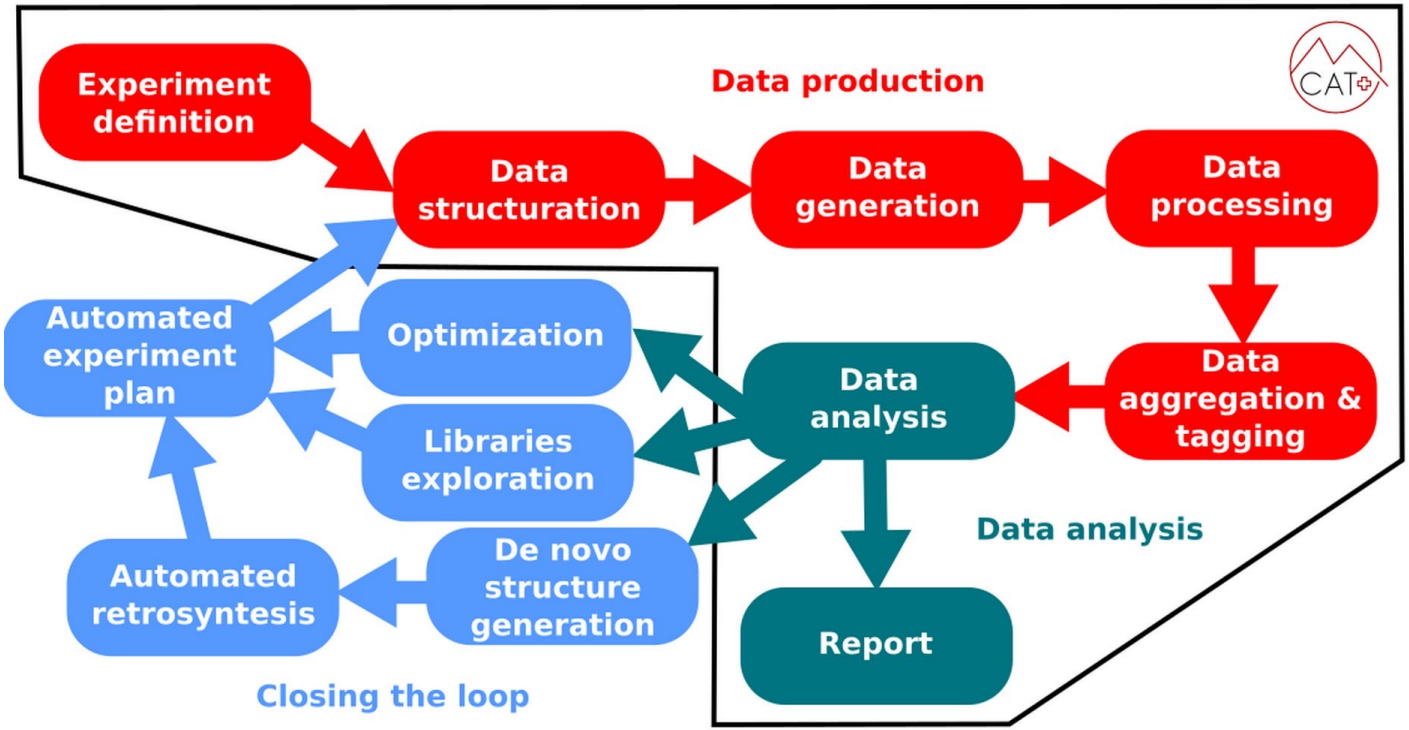
4. Internal communication layer
Axons and synapses
Networks and APIs

5. Autonomic layer
Continuous regulatory tasks (temperature, pressure, waste managment)

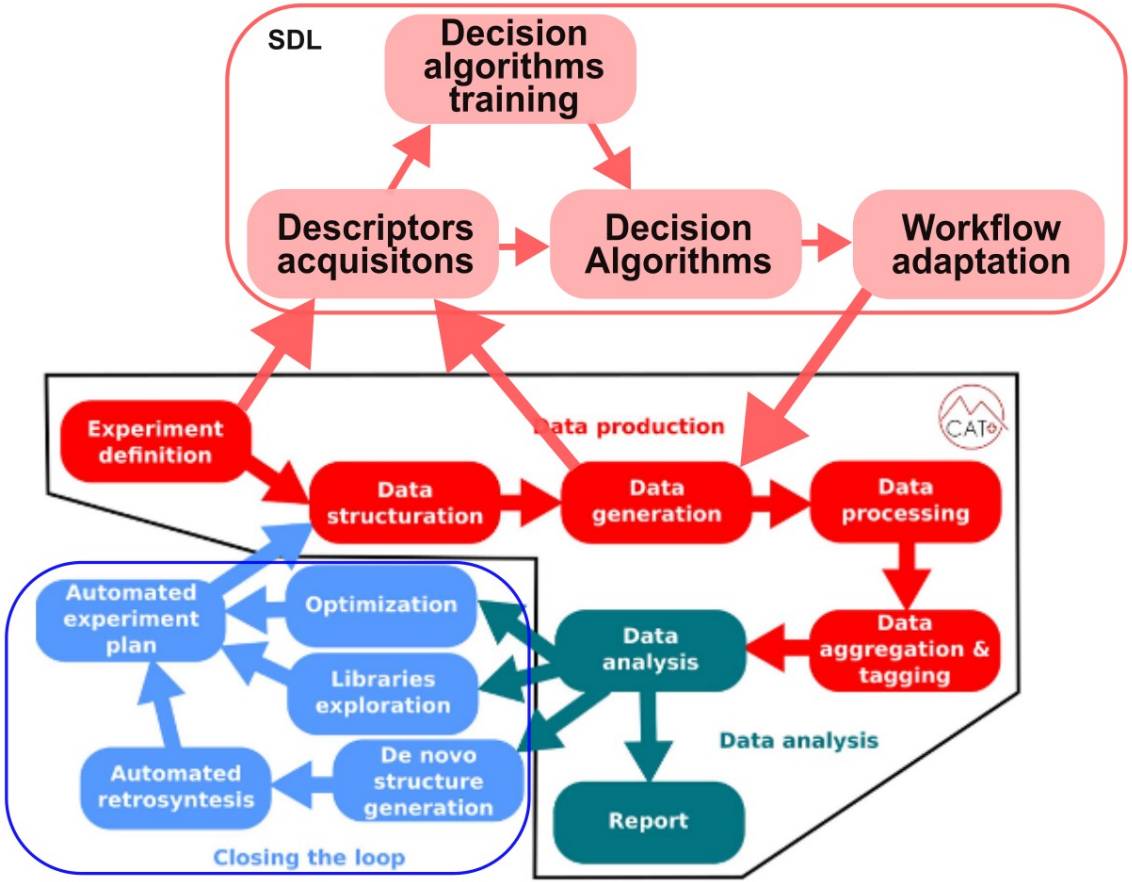
6. Fast reflexive modules layer
Combination of sensors and decision algorithm able to inform layer 1 and to adapt actuator (layer 6) behaviors

7. Optimized actuators and sensors layer
Efficient hardware requiring minimal operational load from top layers
Adaptative robotics

Intro – e.g. of HTE data lifecycle



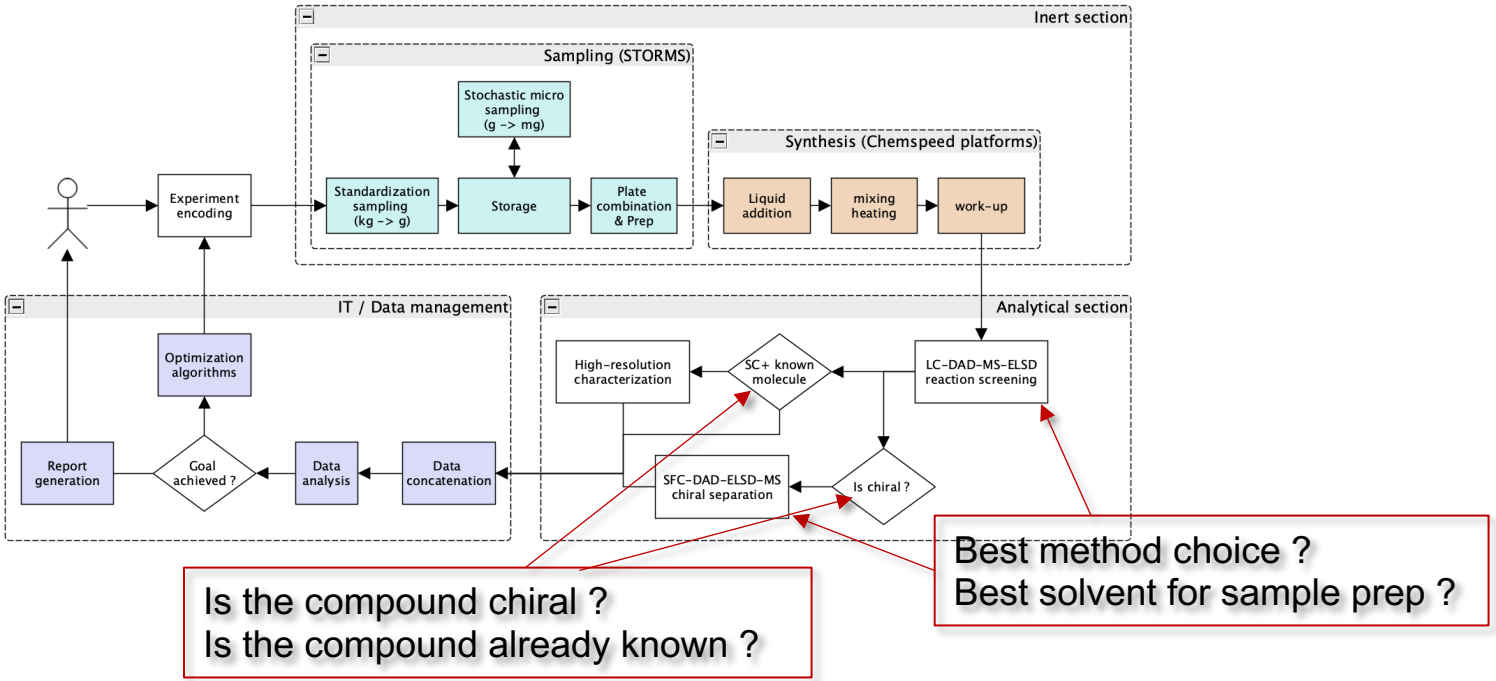
Intro – the expended SDL data lifecycle



SDL Algorithms

SDL Algorithms

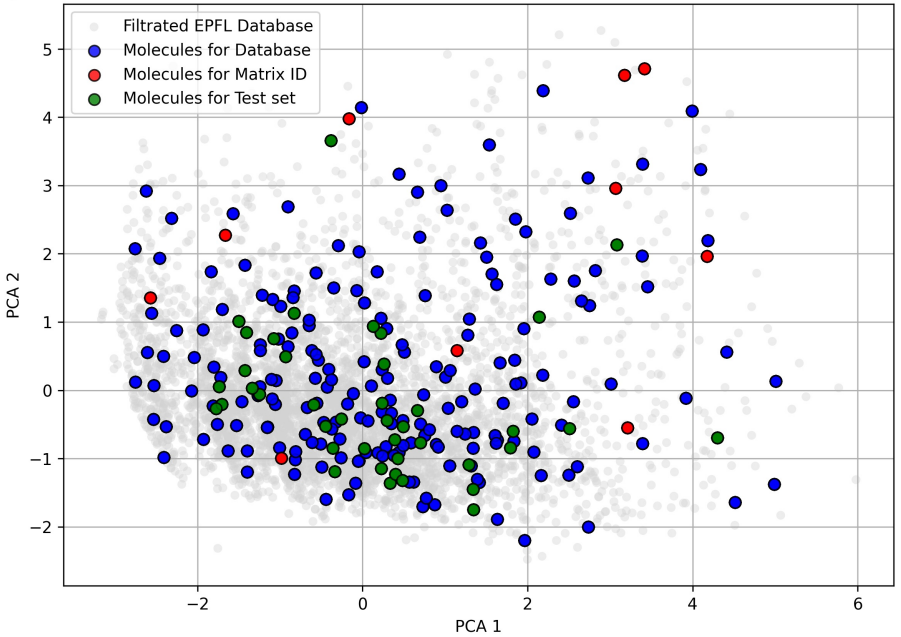
What kind of SDL algorithms can we imagine ?



SDL Algorithms

Example 1: Automated selection of most appropriate LC method

- **20'000 small molecules** available at EPFL light grey dots
- Molecular descriptors-based classification **LogP, MW, H_donor, H_acceptor, TPSA** (extracted from Pubchem)
- **200 selected for transformer training** (K-mean sampling algorithm, blue dots)
- **40 selected for transformer testing** (K-mean sampling algorithm, green dots)
- **10 for method matrix ID (gluttony approach**, max distance between points, cheap and easily accessible, red dots)

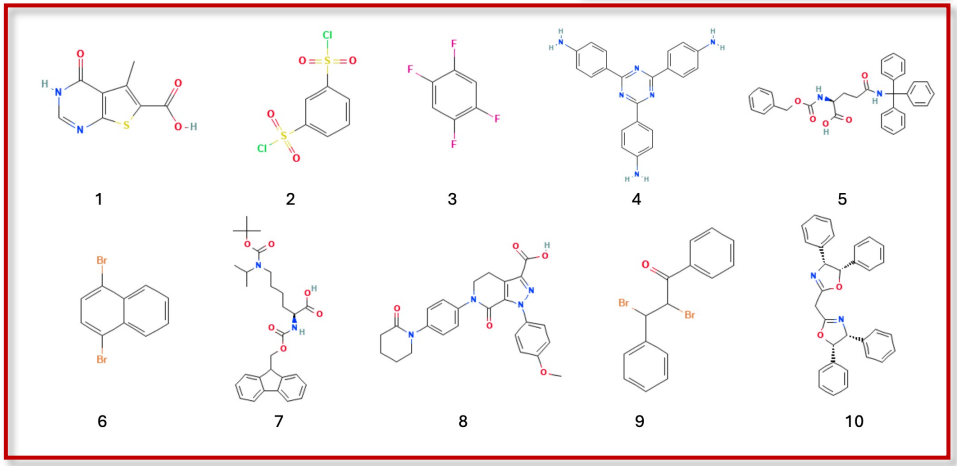


Principal Component Analysis (PCA) of Five Molecular Descriptors: LogP, H-bond Donors, H-bond Acceptors, Molecular Weight (MW), and Topological Polar Surface Area (TPSA)

SDL Algorithms

Example 1: Automated selection of most appropriate LC method

#	Column	Column description	Method
1	Bluebird	C18 and hydrophilic (OH) encapping	1
2	PolarTec	C18 with embedded polar group	1
3	PFP	Pentafluorophenylpropyl multi-encapping	1
4	RP18	C18 multi-endcapping	1
5	Sphinx	Propylphenyl and C18 encapping (1:1)	1
6	RP18	C18 multi-encapping	2
7	Sphinx	Propylphenyl and C18 encapping (1:1)	2



#1 - Bluebird, method 1			
molecule	RT	A_{10}	W_{10}/S
1	1.932	1.610	0.262
2	3.453	0.109	0.959
3	3.124	2.697	0.081
4	5.424	2.081	0.722
5	6.396	0.045	0.408
6	2.314	1.691	0.984
7	1.181	1.936	0.825
8	4.126	0.051	0.614
9	6.449	2.980	0.271
10	3.692	0.139	0.277

Automated and data-driven chemistry

E. Venancio, R. Guichonnet, V.-T. Nguyen, J.-C. Cousty, C. Portmann, P. Miéville, in preparation

SDL Algorithms

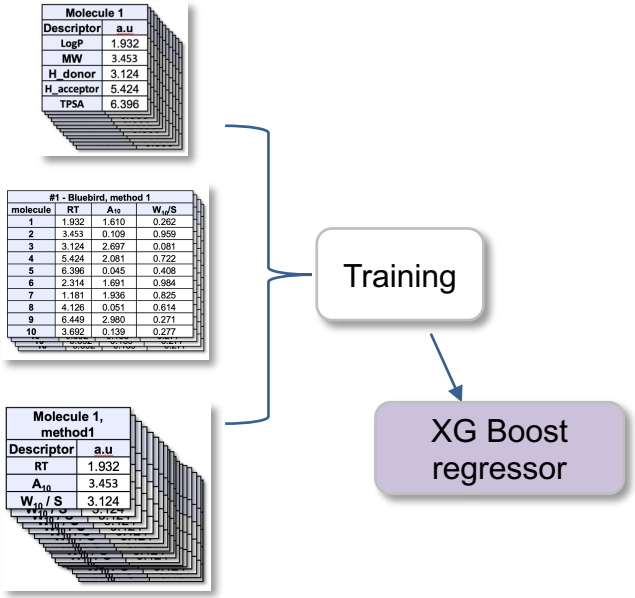
Example 1: Automated selection of most appropriate LC method

Transformer training :

Molecular Descriptors (LogP, MW, H_donor, H_acceptor, TPSA) 200 vectors of 5 dimensions each

Method & column couple represented by their 3x10 matrix ID 7 vectors of 30 dimensions each

Separation Properties (RT, A_{10}, W_{10} / S) for all the molecules with all methods 7 x 200 = 1400 vectors of 3 dimensions each

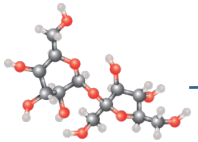


E. Venancio, R. Guichonnet, V.-T. Nguyen, J.-C. Cousty, C. Portmann, P. Miéville, in preparation

SDL Algorithms

Example 1: Automated selection of most appropriate LC method

1 Reaction output prediction



DB Query

DFT calc

New molecule 1	
Descriptor	**a.u.**
LogP	1.932
MW	3.453
H_donor	3.124
H_accept	5.424
TPSA	6.396

2 Separation properties prediction per method

New molecule 1	
Descriptor	**a.u.**
LogP	1.932
MW	3.453
H_donor	3.124
H_accept	5.424
TPSA	6.396

XG Boost regressor

New molecule 1 Method 1	
Descriptor	**a.u.**
RT	2.152
A_{10}	0.453
W_{10} / S	0.124

E. Venancio, R. Guichonnet, V.-T. Nguyen, J.-C. Cousty, C. Portmann, P. Miéville, in preparation

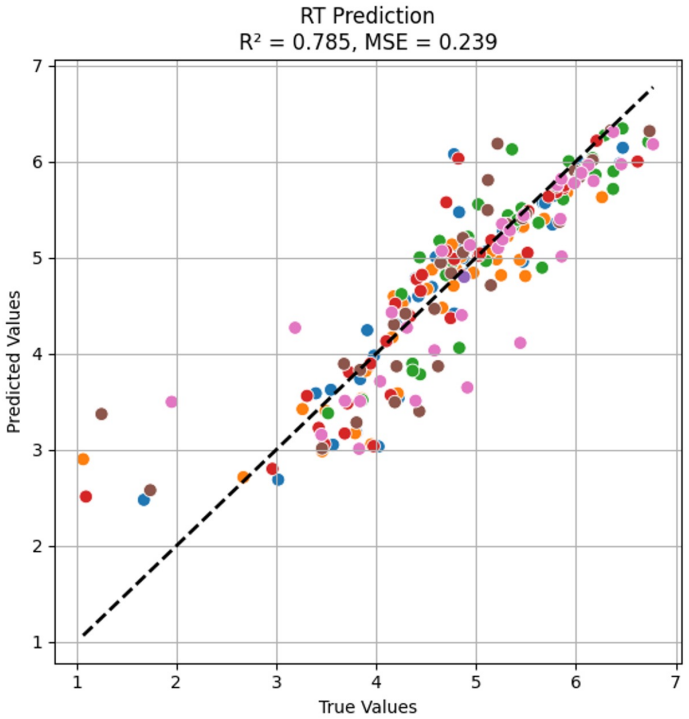
SDL Algorithms

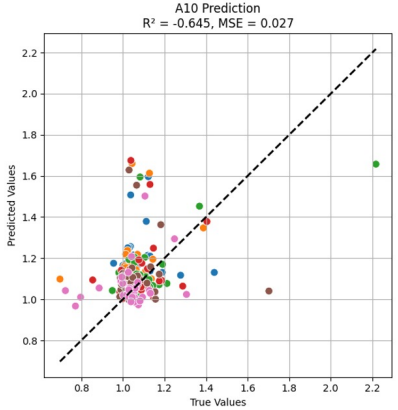
Example 1: Automated selection of most appropriate LC method

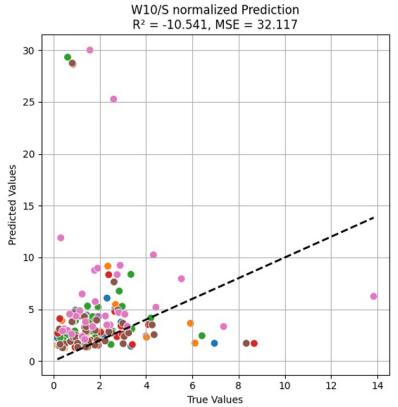
XG boost regressor :

predictions for **40 test molecules (spots)** with **7 methods (colors)**

logP	0.61
MW	0.32
H donor	0.43
H acceptor	0.38
TPSA	0.46







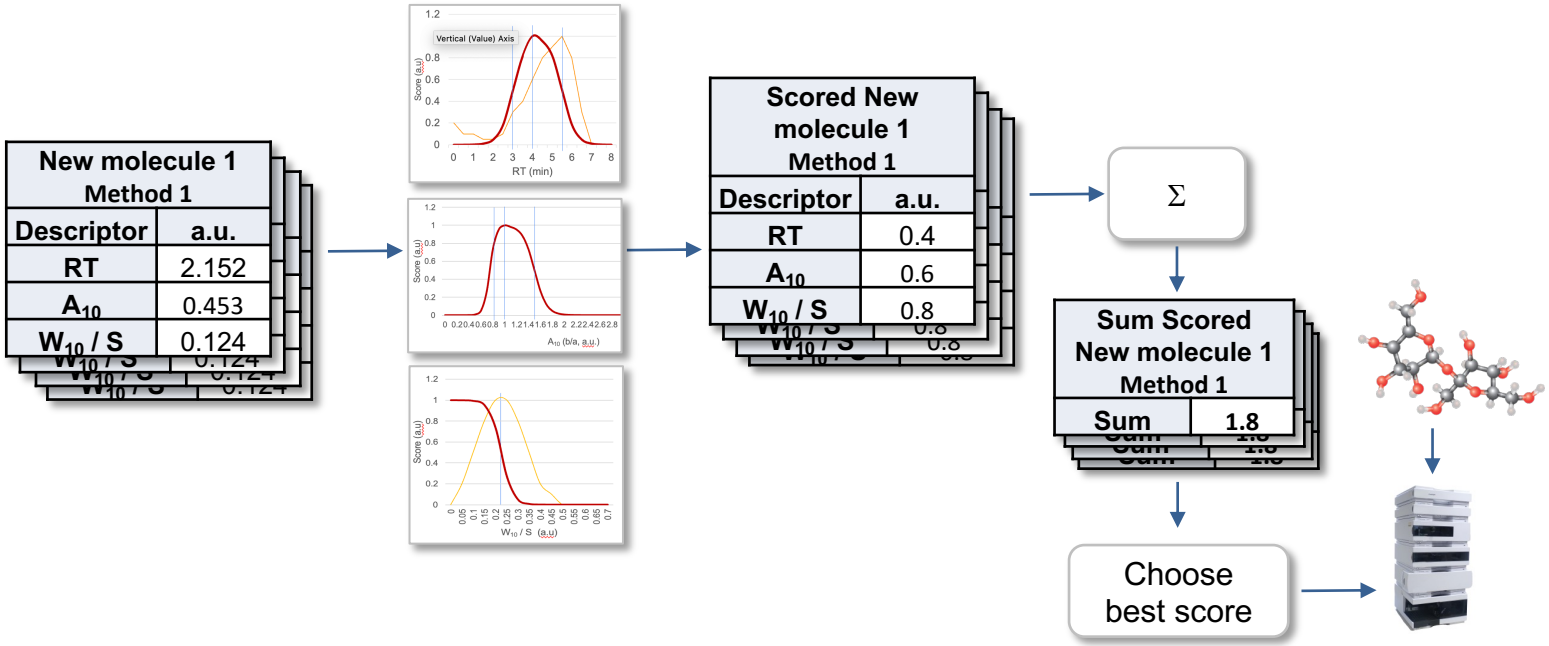
Automated and data-driven chemistry

J. Bradshaw, A. Zhang Babak, M. David E. Graff, M. H. S. Segler, C. W. Coley, *ACS Cent. Sci.* 2025, 11, 4, 539–549, doi.org/10.1021/acscentsci.5c00055
E. Venancio, R. Guichonnet, V.-T. Nguyen, J.-C. Cousty, C. Portmann, P. Miéville, in preparation

SDL Algorithms

Example 1: Automated selection of most appropriate LC method

Scoring



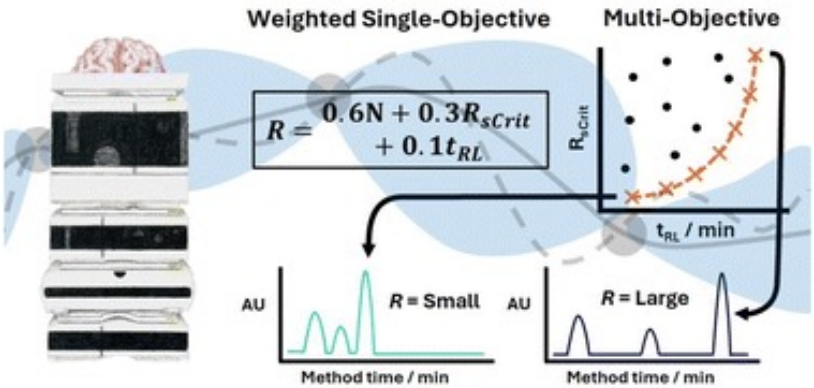
E. Venancio, R. Guichonnet, V.-T. Nguyen, J.-C. Cousty, C. Portmann, P. Miéville, in preparation

SDL Algorithms

Example 1: Automated selection of most appropriate LC method

Selected method self optimization

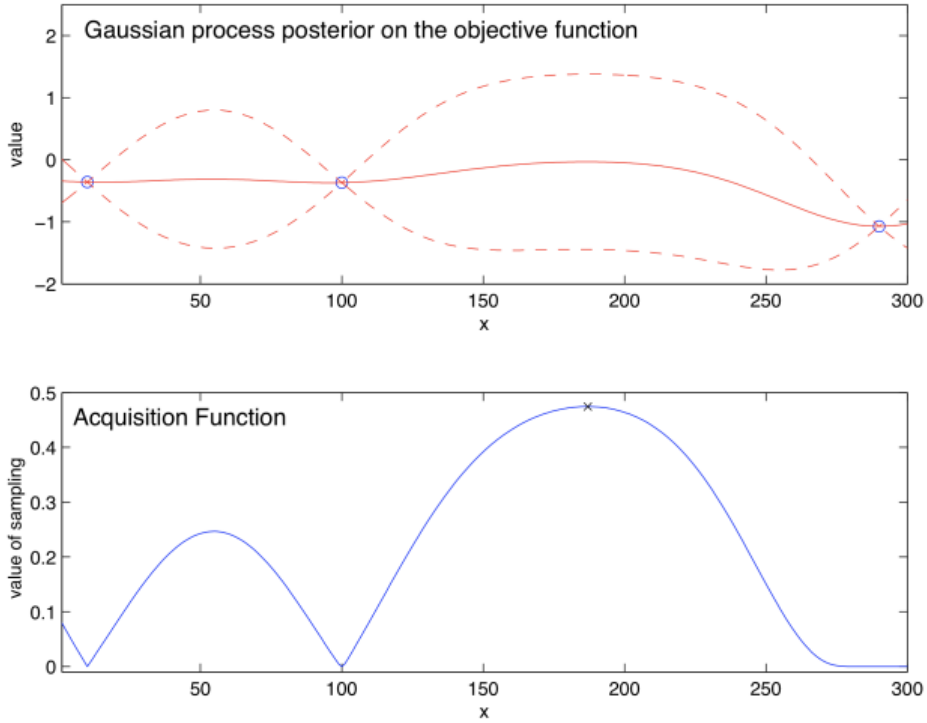
Operator-free HPLC automated method development guided by Bayesian optimization



T. M. Dixon et al., Digital Discovery, 2024, **3**, 1591-1601 10.1039/D4DD00062E

SDL Algorithms

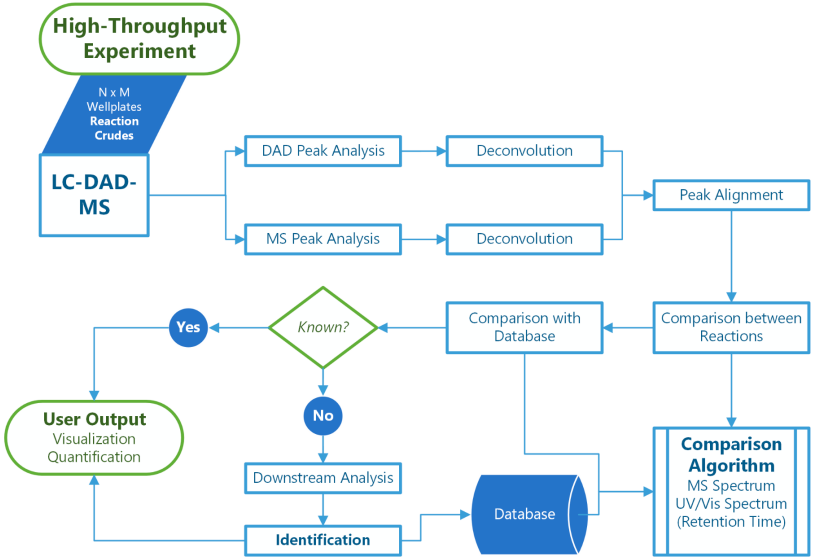
Short explanation of BO

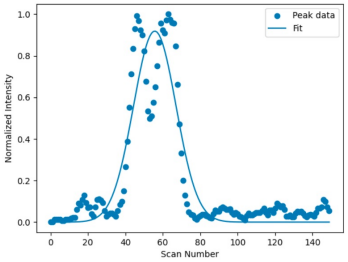


Peter I. Frazier, **A Tutorial on Bayesian Optimization,** arXiv:1807.02811

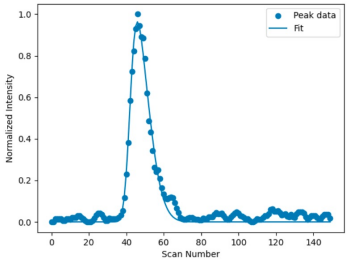
SDL Algorithms

Example 2: New compound automatic identification

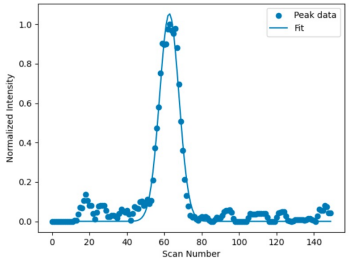




(a) Two overlapping peaks.



(b) Peak one.

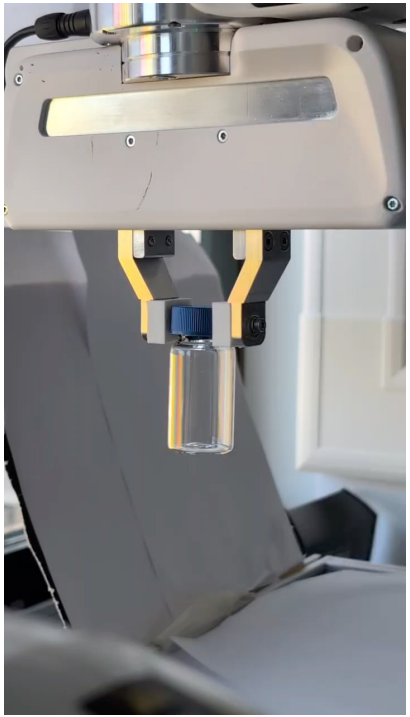


(c) Peak two.

P. Sandor Kelley, L. Choudhury et al. In preparation

SDL Algorithms

Example 3: Optimized robotics using vision and ML



Combined with LLM to perform automated classification of powders

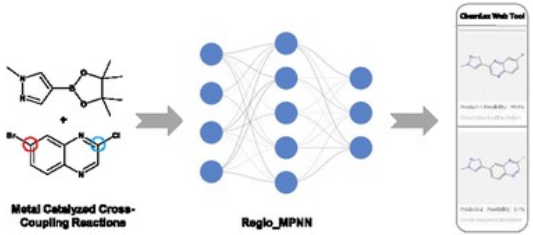
Jasper Tan, Edy Mariano et al. in preparation

Closing the loop Algorithms

Closing the loop Algorithms

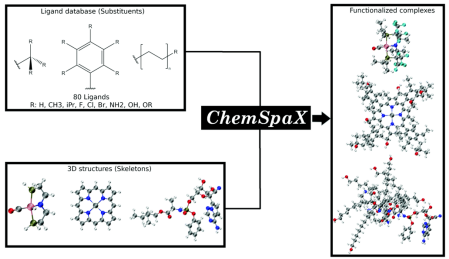
Automated exploration and proposal of new possible molecules

Regio-MPNN: predicting regioselectivity for general metal-catalyzed cross-coupling …



Baochen Li et al., Digital Discovery, 2024
10.1039/d4dd00244j

ChemSpaX: exploration of chemical space by automated functionalization of molecular scaffold

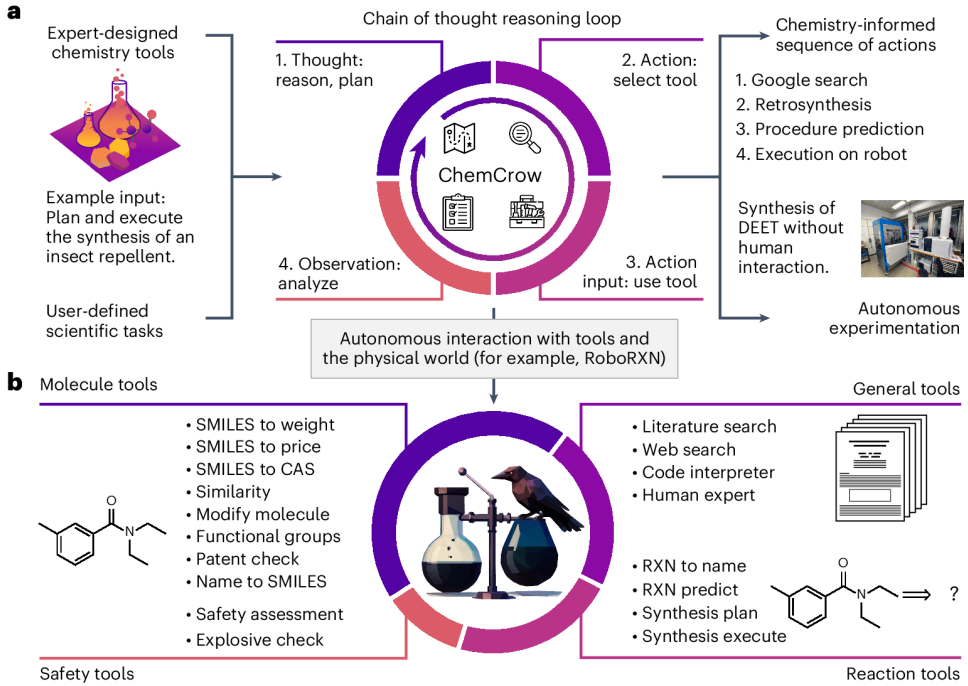


A. V. Kalikadien et al., Digital Discovery, 2022,**1**, 8-25
10.1039/d1dd00017a

Closing the loop Algorithms

New automatically generated synthetic pathways

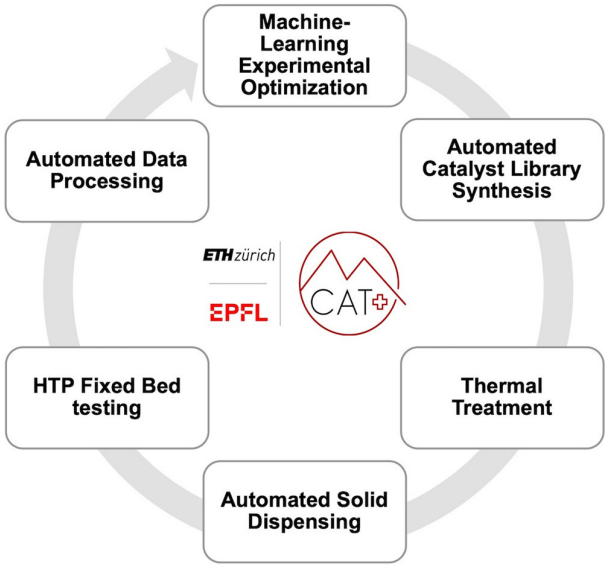
Augmenting large language models with chemistry tools

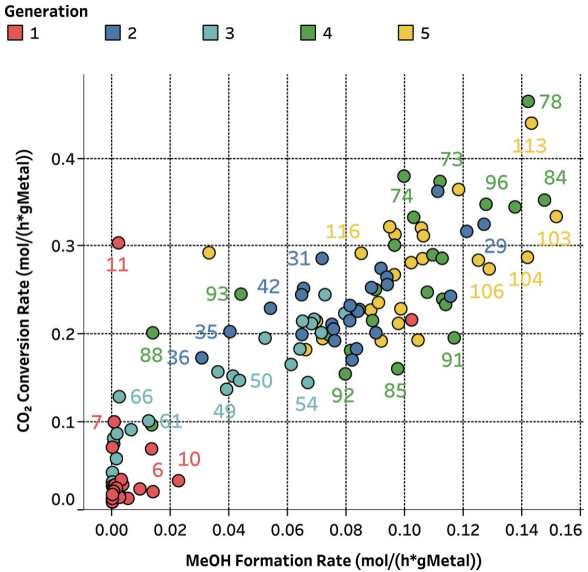


M. Bran, A., Cox, S., Schilter, O. *et al.* Augmenting large language models with chemistry tools. *Nat Mach Intell* **6**, 525–535 (2024). 10.1038/s42256-024-00832-8

Closing the loop Algorithms

Bayesian optimization of new formulations or reaction conditions





Automated and high-throughput synthesis and testing of **144 catalysts in 6 weeks**

Significant improvement of the catalyst's performances **in 5 generations**

Similar to **ca. 10 years of catalytic development**

Ramirez et al., Chem Catalysis 4, 100888, February 15, 2024

Next session

Today
Automation in chemistry III - Workflows analysis