香港中文大學
The Chinese University of Hong Kong

Auto-QChem: an automated workflow for the generation and storage of DFT calculations for organic molecules

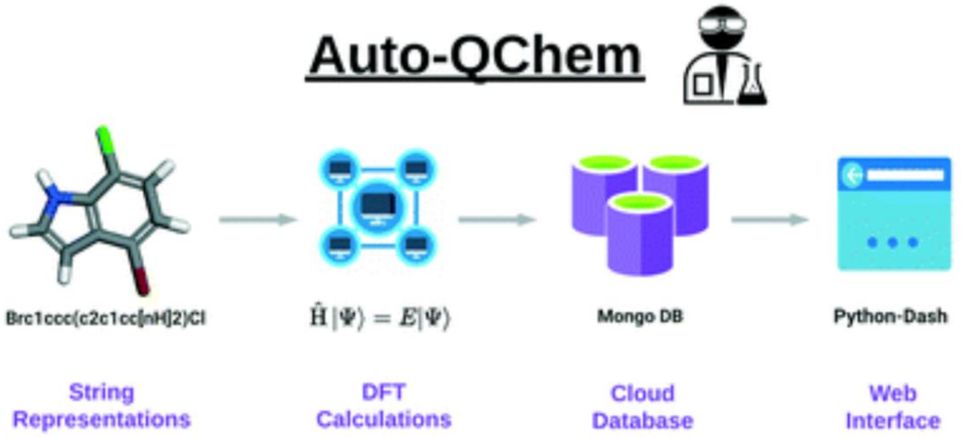
Jingyi Liu

29/09/2025

React. Chem. Eng. **2022**, 7, 1276–1284.

Outline

1. **Background**
2. **Auto-Qchem workflow**
3. **Applications**
4. **Discussion**



Research Background

Research Background

- The application of machine learning models in organic chemistry requires effective representations of chemical structures.

- ML models trained with chemical descriptors are more interpretable than molecular fingerprints and other representations.

- High throughput DFT presents a significant barrier to experimental chemists.

- Previous DFT automation packages are mainly designed for material science rather than small organic molecules.

Auto-Qchem

- An automatic, high-throughput and end-to-end DFT calculation workflow
- Computes chemical descriptors for organic molecules

[1] https://github.com/PrincetonUniversity/auto-qchem;
[2] https://princetonuniversity.github.io/auto-qchem
[3] https://autoqchem.org

Auto-Qchem workflow

❖ Auto-Qchem

➢ Features

1. Generate user-specified input.
2. HPC interface.
3. Automatic information extraction from DFT results.
4. Convenient storage and data accessing.

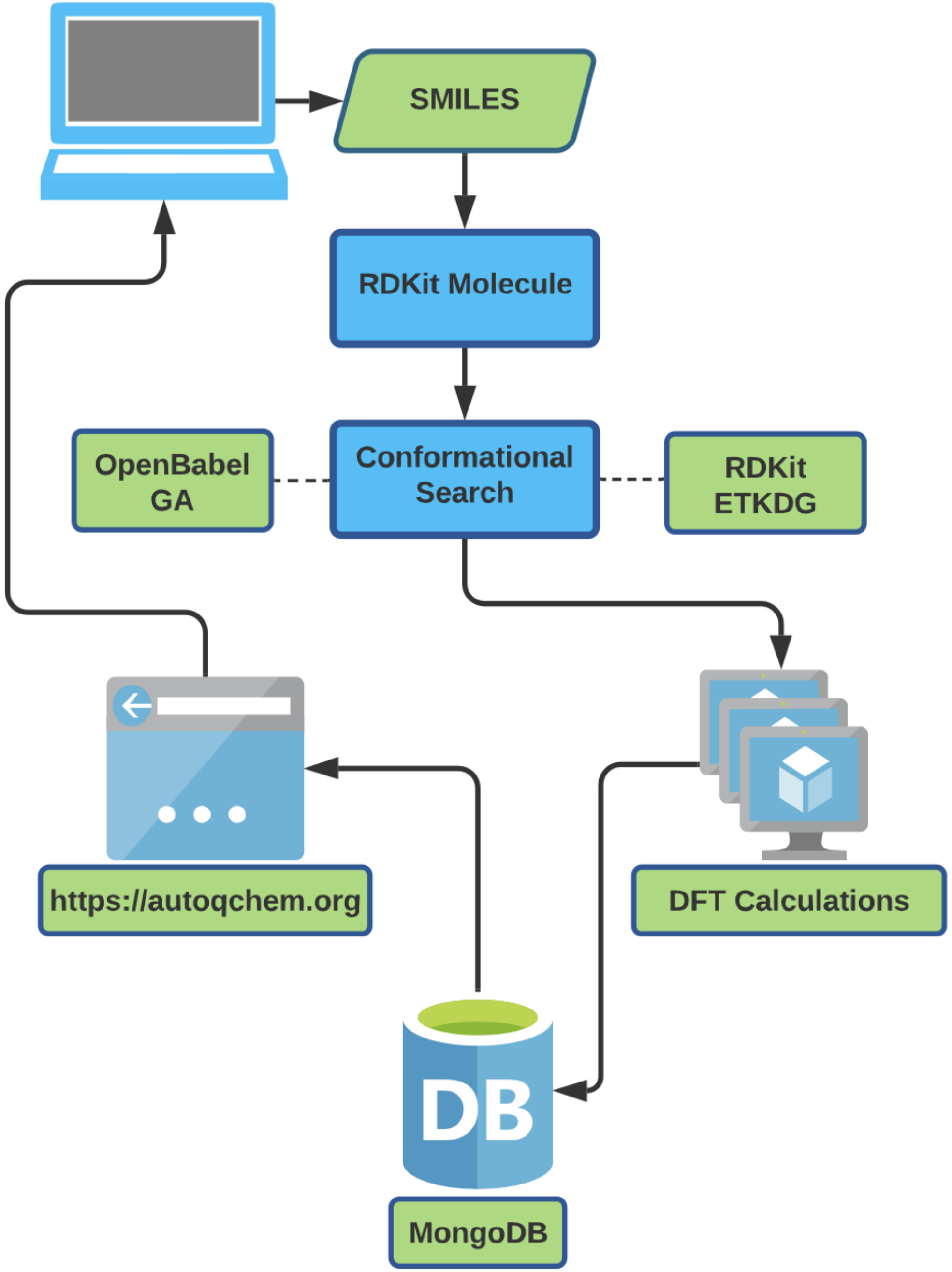
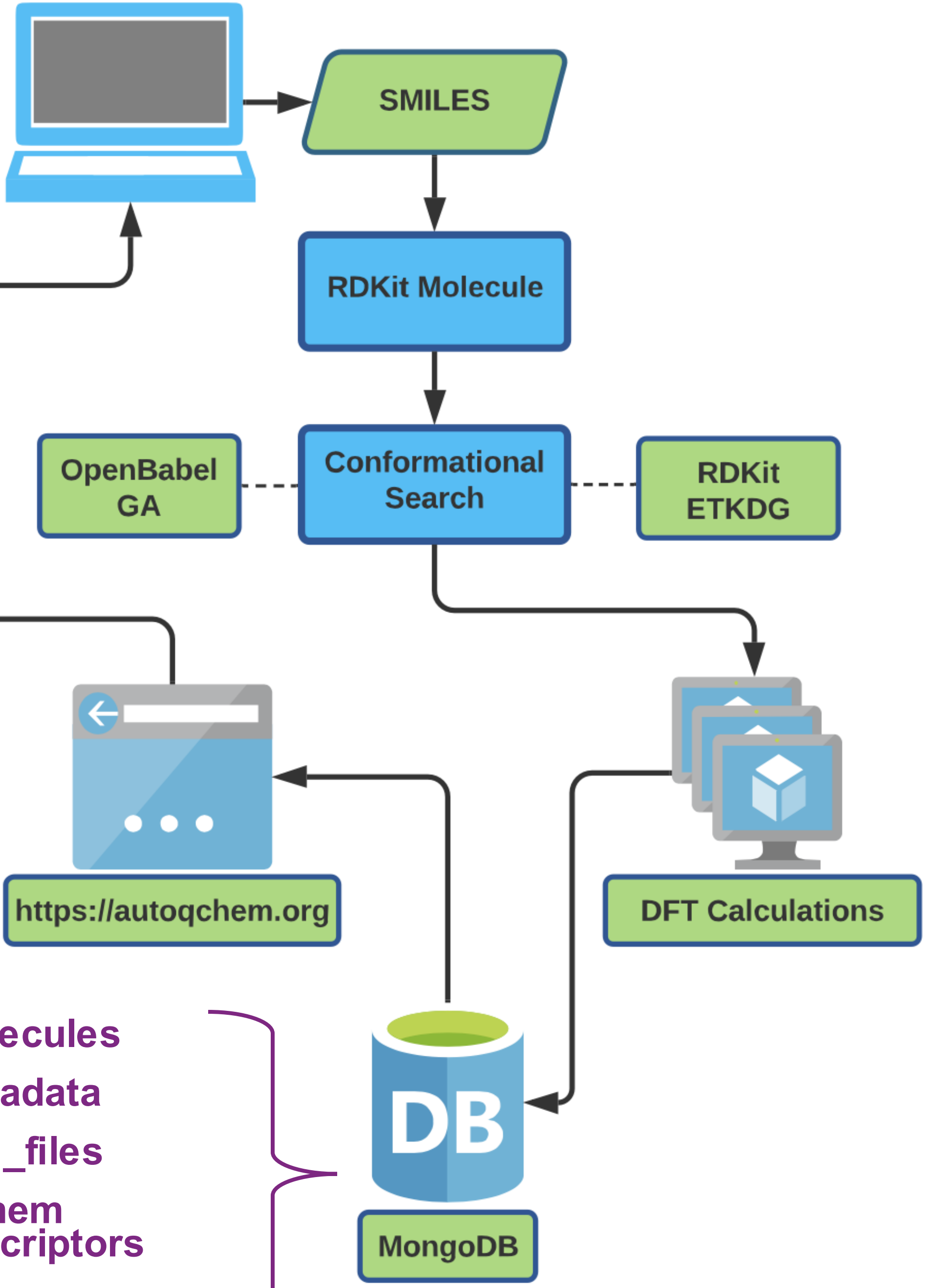


Fig. 1 Computational workflow of Auto-QChem.

❖ Auto-Qchem



- ❖ Functional, basis, ..
- ❖ Geometry optimization
- ❖ Frequency and thermochemistry
- ❖ Excited state TDDFT

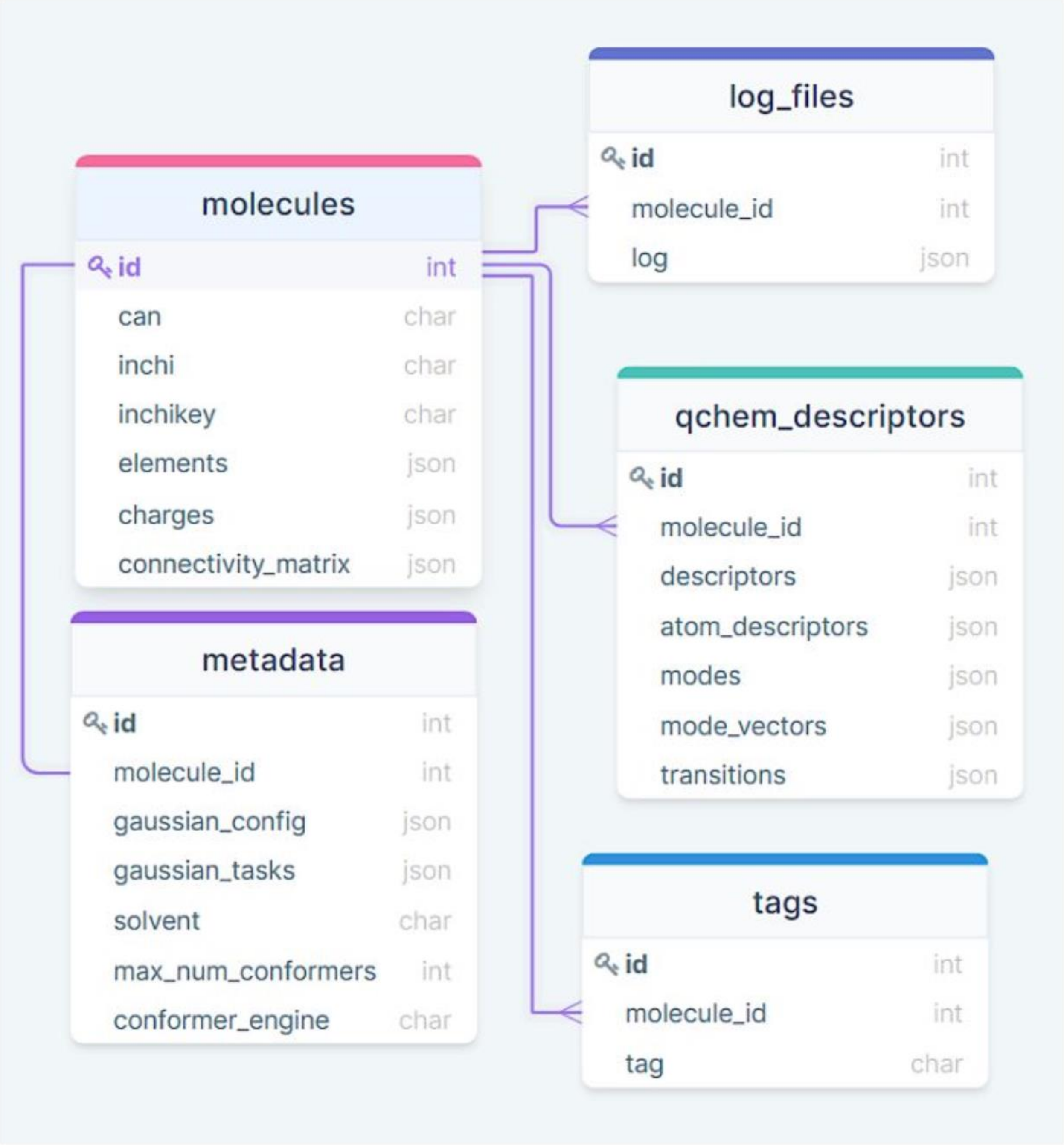
- ❖ Molecules
- ❖ Matadata
- ❖ Log_files
- ❖ Qchem descriptors
- ❖ tags

Fig. 1 Computational workflow of Auto-QChem.

❖ Auto-Qchem

➢ Features

- Avoid duplicate calculations



log_files

id	int
molecule_id	int
log	json

molecules

id	int
can	char
inchi	char
inchikey	char
elements	json
charges	json
connectivity_matrix	json

qchem_descriptors

id	int
molecule_id	int
descriptors	json
atom_descriptors	json
modes	json
mode_vectors	json
transitions	json

metadata

id	int
molecule_id	int
gaussian_config	json
gaussian_tasks	json
solvent	char
max_num_conformers	int
conformer_engine	char

tags

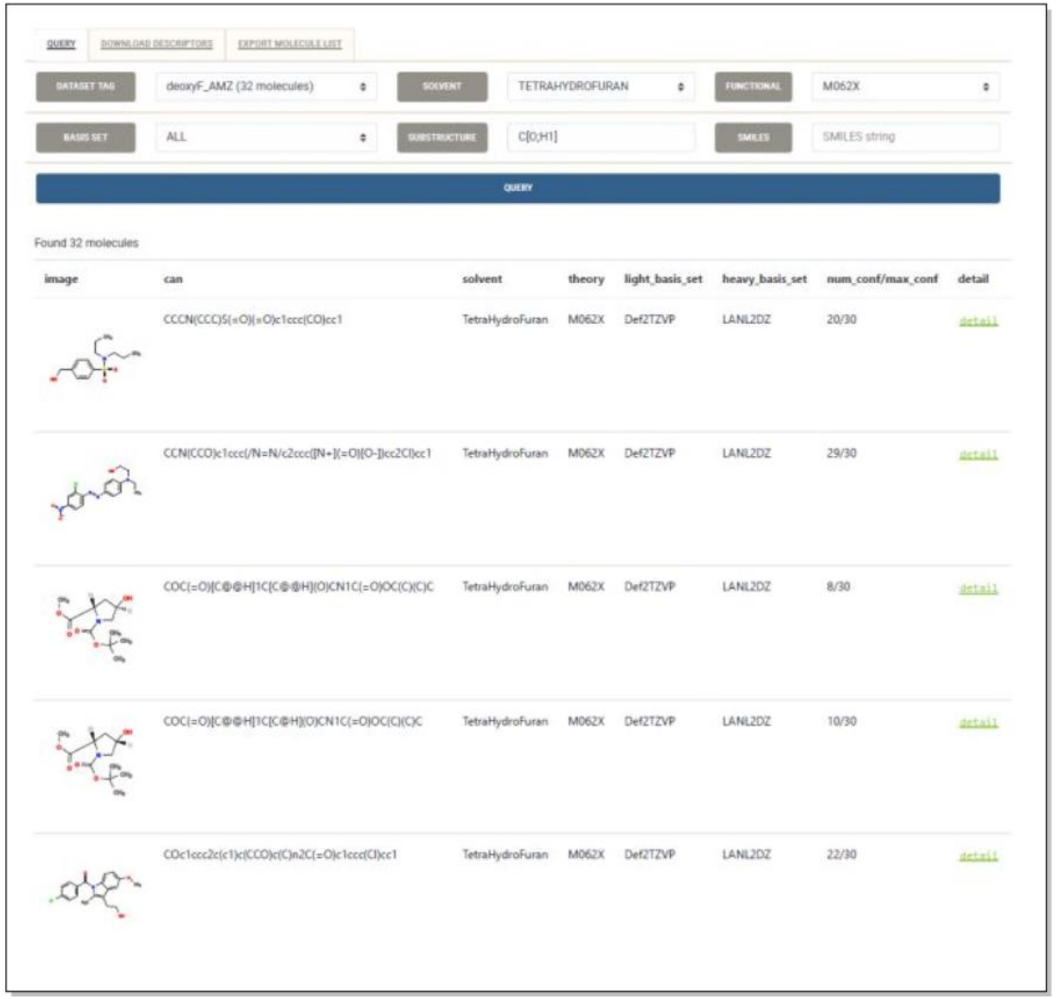
id	int
molecule_id	int
tag	char

Fig. 2 Collection schema of Auto-QChem database.

❖ Auto-Qchem

➢ **Features:** Queries and data retrieval

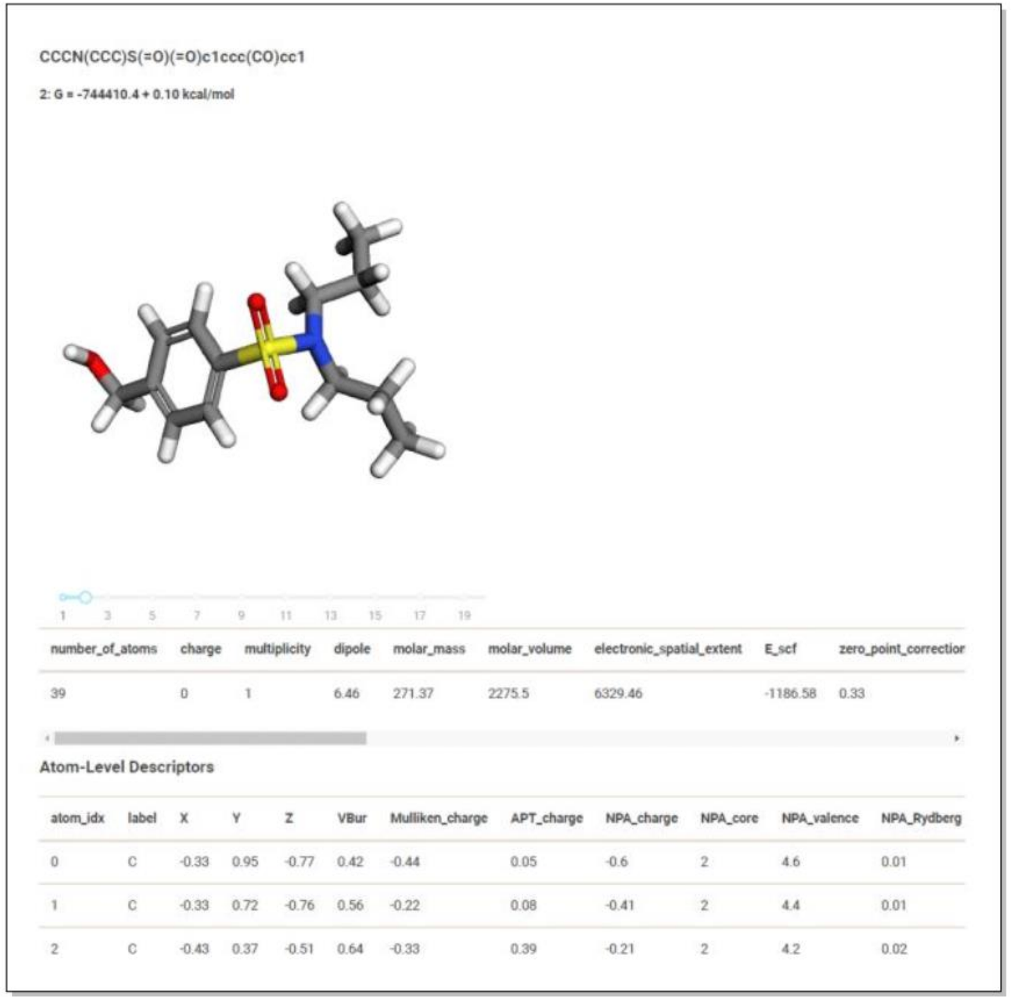
➢ 1692 molecules in total



DATASET TAG: deoxyF_AMZ (32 molecules)
SOLVENT: TETRAHYDROFURAN
FUNCTIONAL: M062X
BASIS SET: ALL
SUBSTRUCTURE: C[0;H1]
SMILES: SMILES string

Found 32 molecules

image	can	solvent	theory	light_basis_set	heavy_basis_set	num_conf/max_conf	detail
	CCCN(CCC)S(=O)(=O)c1ccc(CO)cc1	TetraHydroFuran	M062X	Def2TZVP	LANL2DZ	20/30	detail
	CCN(CCO)c1ccc(/N=N/c2ccc([N+](=O)[O-])cc2Cl)cc1	TetraHydroFuran	M062X	Def2TZVP	LANL2DZ	29/30	detail
	COC(=O)[C@@H]1C[C@@H](O)CN1C(=O)OC(C)(C)C	TetraHydroFuran	M062X	Def2TZVP	LANL2DZ	8/30	detail
	COC(=O)[C@@H]1C[C@H](O)CN1C(=O)OC(C)(C)C	TetraHydroFuran	M062X	Def2TZVP	LANL2DZ	10/30	detail
	COc1ccc2c(c1)c(CCO)c(C)n2C(=O)c1ccc(Cl)cc1	TetraHydroFuran	M062X	Def2TZVP	LANL2DZ	22/30	detail



CCCN(CCC)S(=O)(=O)c1ccc(CO)cc1

2: G = -744410.4 + 0.10 kcal/mol

number_of_atoms	charge	multiplicity	dipole	molar_mass	molar_volume	electronic_spatial_extent	E_scf	zero_point_correctior
39	0	1	6.46	271.37	2275.5	6329.46	-1186.58	0.33

Atom-Level Descriptors

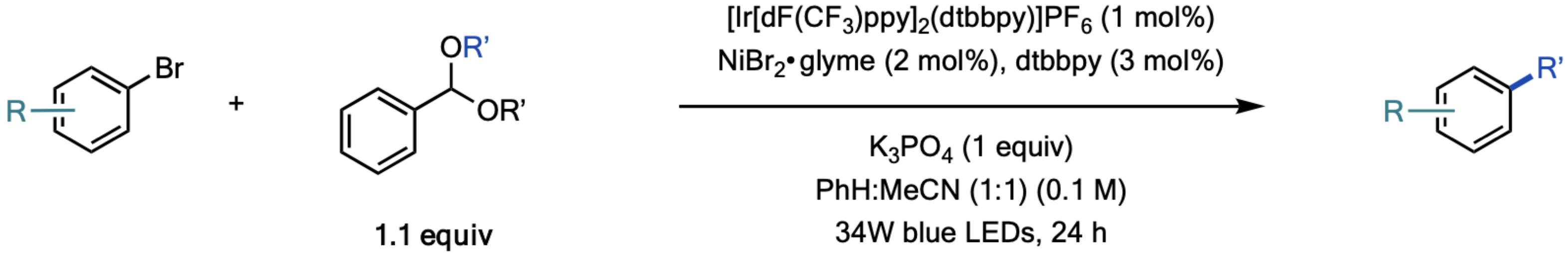
atom_idx	label	X	Y	Z	VBur	Mulliken_charge	APT_charge	NPA_charge	NPA_core	NPA_valence	NPA_Rydberg
0	C	-0.33	0.95	-0.77	0.42	-0.44	0.05	-0.6	2	4.6	0.01
1	C	-0.33	0.72	-0.76	0.56	-0.22	0.08	-0.41	2	4.4	0.01
2	C	-0.43	0.37	-0.51	0.64	-0.33	0.39	-0.21	2	4.2	0.02

Fig. 3 Query view (left) and the molecule view (right) of the web interface. The molecule view is a snapshot while viewing the second lowest energy conformation in 3D.

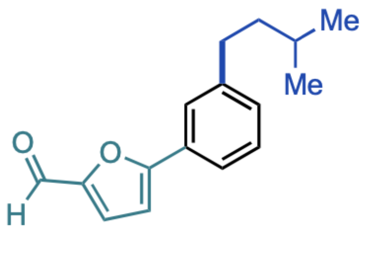
❖ Applications

1. Substrate scope design in Ni/ photoredox methodology development

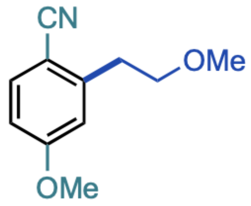
- Ni/ photoredox catalyzed alkylation reaction of aryl halides using acetals as alcohol-derived aliphatic radical sources.
- Data set: 2683 aryl bromides, with 168 electronic and steric features,
- 95 of 168 features were employed for clustering.



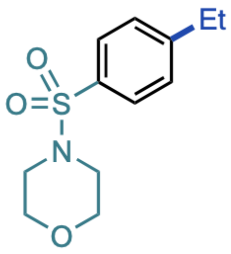
R
Br
+
OR'
OR'
1.1 equiv
[Ir[dF(CF$_3$)ppy]$_2$(dtbbpy)]PF$_6$ (1 mol%)
NiBr$_2$•glyme (2 mol%), dtbbpy (3 mol%)
K$_3$PO$_4$ (1 equiv)
PhH:MeCN (1:1) (0.1 M)
34W blue LEDs, 24 h
R
R'



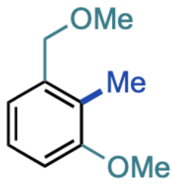
A (R' = i-Amyl)
22% yield



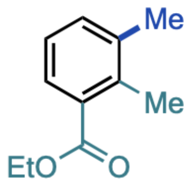
D (R' = CH_2CH_2OMe)
98% yield



E (R' = Et)
87% yield



K (R' = Me)
0% yield

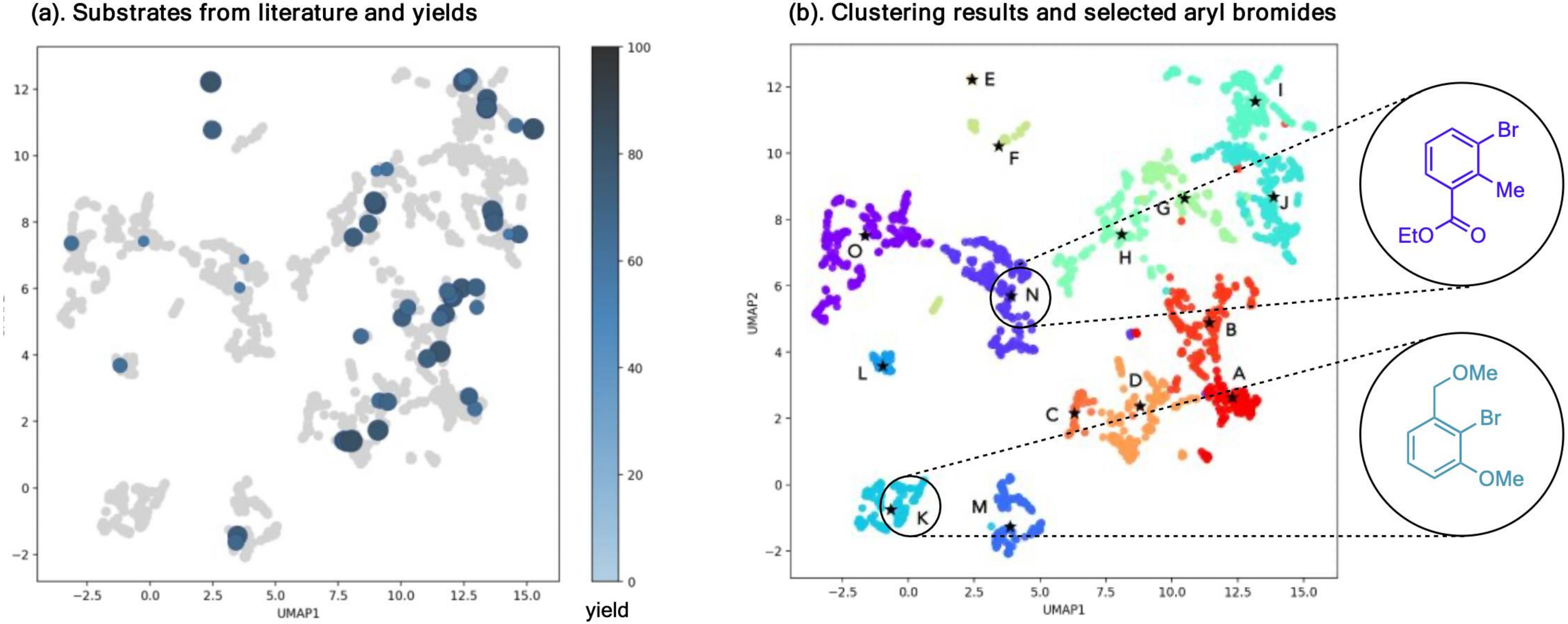


N (R' = Me)
70% yield

❖ Applications

1. Substrate scope design in Ni/ photoredox methodology development

- Electronegativity of the aryl bromides was highly correlated with yield.
- Generalized additive model (GAM) trained on electronegativity outperformed ML models trained on literatures



(a). Substrates from literature and yields

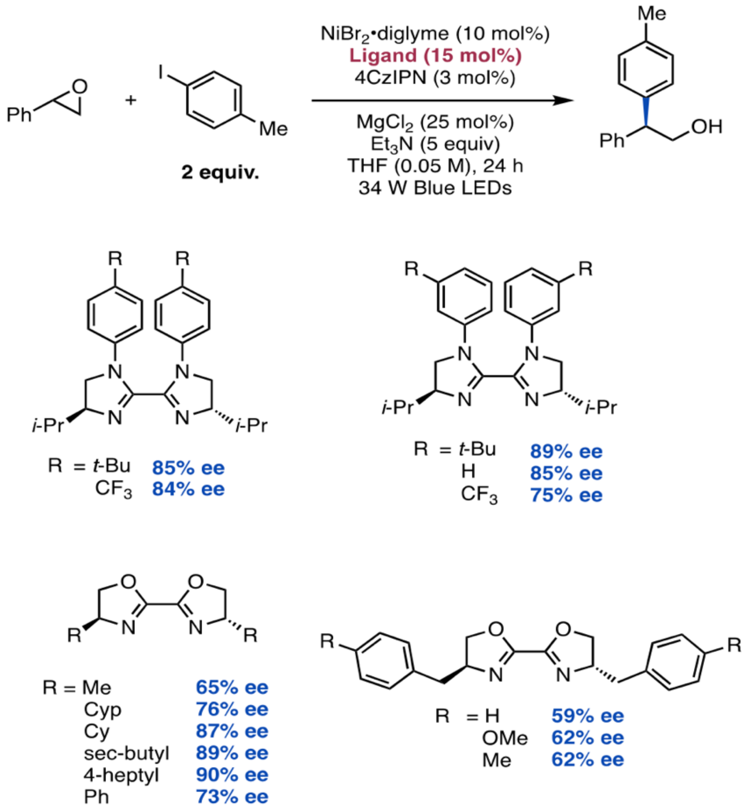
(b). Clustering results and selected aryl bromides

❖ Applications

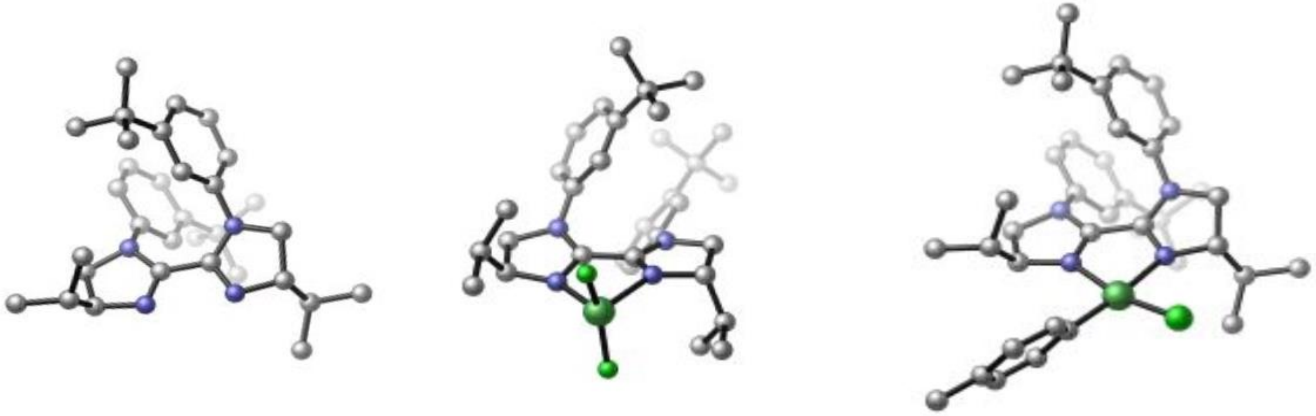
2. Ligand parametrization and enantioselectivity prediction in nickel catalysis

- Previous used ligand: Bioxazoline (BiOx)
  - Good enantioselectivity, low yield
- Optimal ligand: chiral biimidazoline (BiIm) : time-consuming to discover

(a). Model reaction system and representative examples of ligands tested.



(b). Ligand environment models.



free ligand

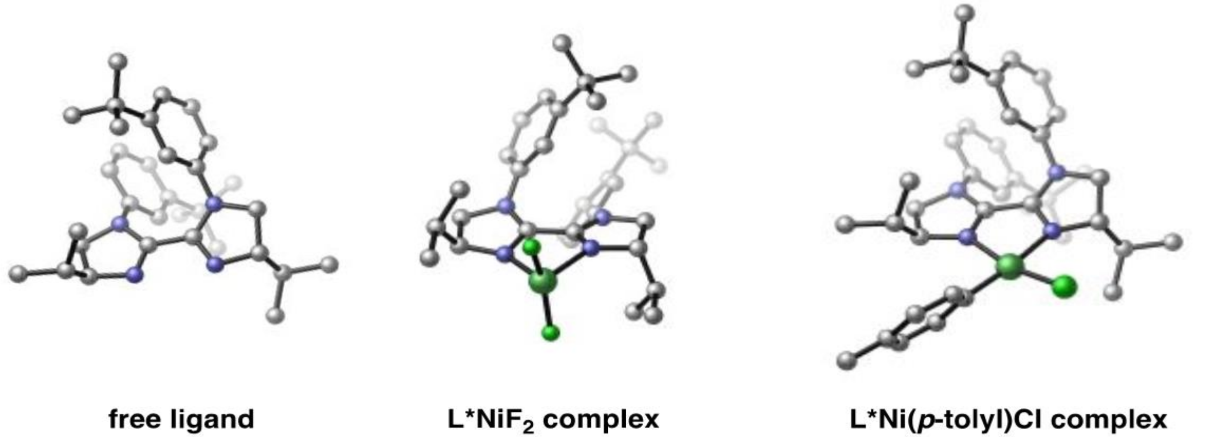
L*NiF$_2$ complex

L*Ni(*p*-tolyl)Cl complex

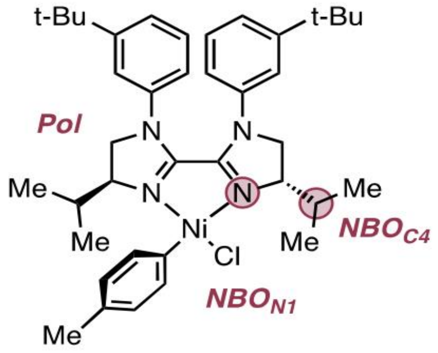
❖ Applications

2. Ligand parametrization and enantioselectivity prediction in nickel catalysis

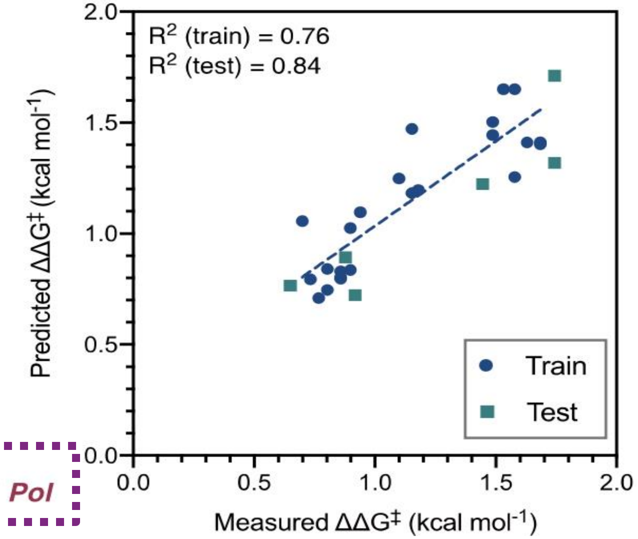
(b). Ligand environment models.



(c). Regression modeling for L*Ni(*p*-tolyl)Cl with DFT-derived features.



$\Delta\Delta G^{\ddagger}$ = **-0.58** ***NBO$_{N1}$*** + **0.46** ***NBO$_{C4}$*** **- 0.27** ***Pol***



- Hypothesis: computed features depend on ligand environment.
- Manually generated conformers ← Auto-Qchem cannot handle transition metal complexes
- Conclusion: electronic, rather than steric attributes of BiIm ligands govern the enantioselectivity of this reaction

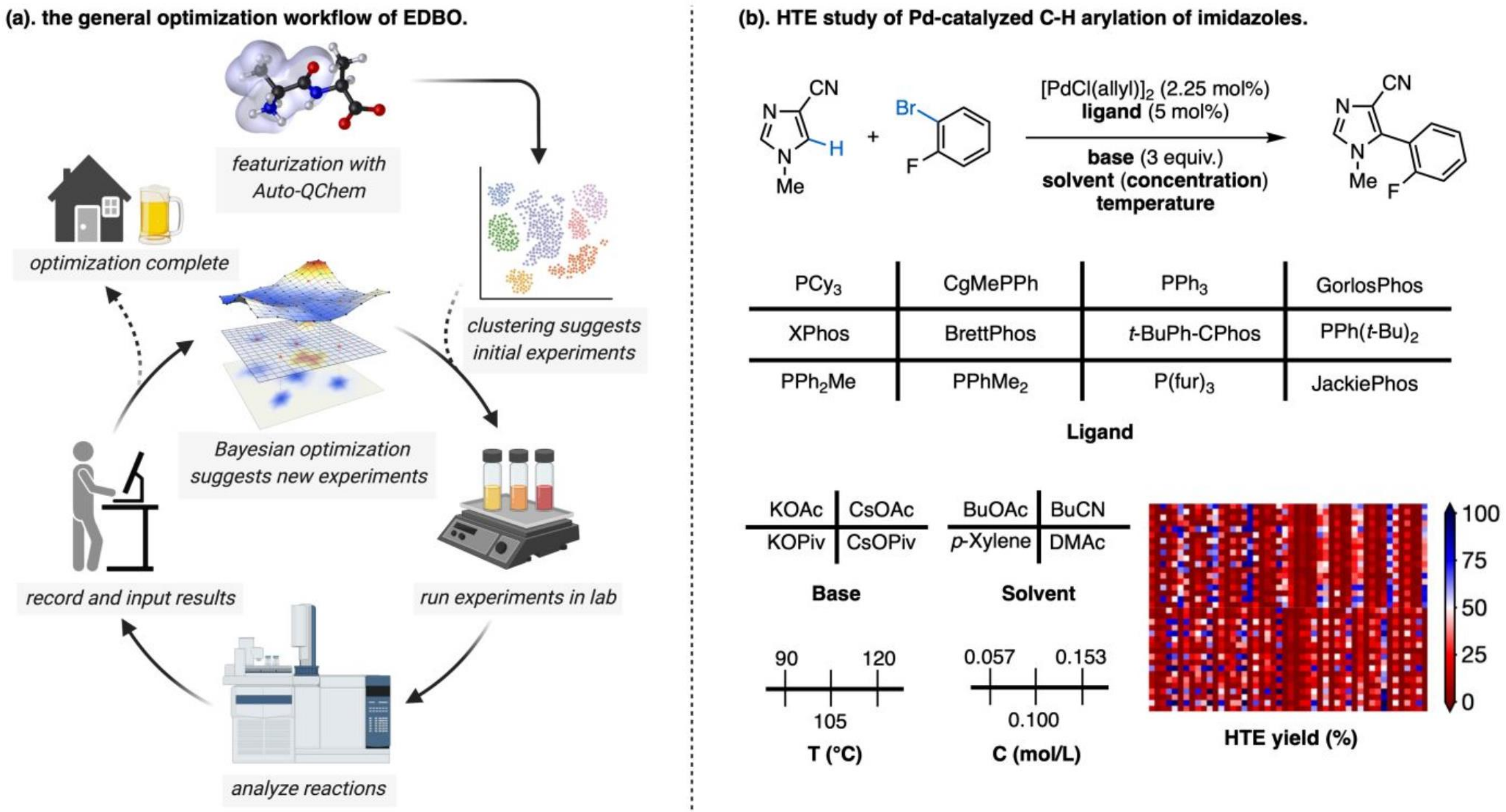
❖ Applications

3. Reaction condition optimization via Bayesian optimization

- Traditional reaction optimization is slow and resource-intensive, relying on trial-and-error or costly methods, with limited exploration of chemical space.
- Bayesian optimization, a sequential design algorithm for global optimization of black-box functions, in efficient reaction condition optimization.
- EDBO (Experimental Design via Bayesian Optimization)

❖ Applications

➢ 3. Reaction condition optimization via Bayesian optimization



(a). the general optimization workflow of EDBO.

featurization with Auto-QChem
clustering suggests initial experiments
run experiments in lab
analyze reactions
record and input results
Bayesian optimization suggests new experiments
optimization complete

(b). HTE study of Pd-catalyzed C-H arylation of imidazoles.

PCy_3	CgMePPh	PPh_3	GorlosPhos
XPhos	BrettPhos	*t*-BuPh-CPhos	PPh(*t*-Bu)$_2$
PPh_2Me	$PPhMe_2$	P(fur)$_3$	JackiePhos

Ligand

KOAc	CsOAc
KOPiv	CsOPiv

Base

BuOAc	BuCN
p-Xylene	DMAc

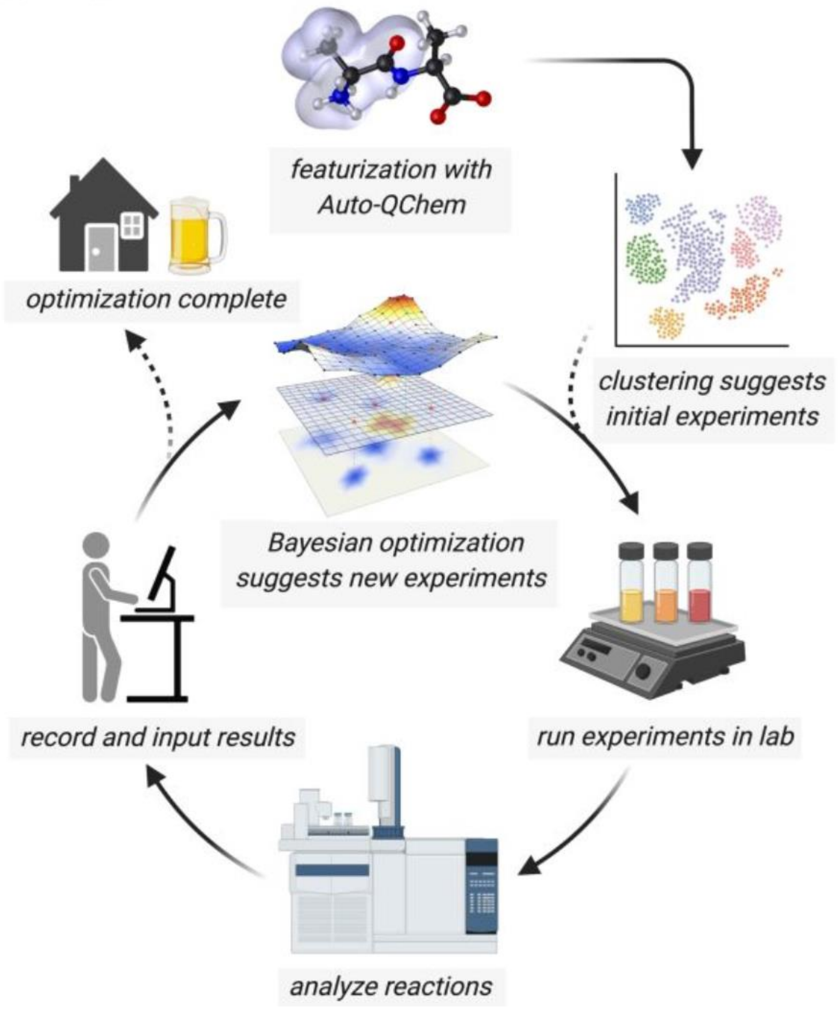
Solvent

Fig. 6 Use case 3: reaction condition optimization via Bayesian optimization.

❖ Applications

3. Reaction condition optimization via Bayesian optimization

(a). the general optimization workflow of EDBO.


featurization with Auto-QChem
clustering suggests initial experiments
run experiments in lab
analyze reactions
record and input results
Bayesian optimization suggests new experiments
optimization complete

EDBO (simulated 50 times) achieved a higher average performance within the first 15 experiments even with random initialization and found conditions with >99% yield 100% of the time!

Discussion

❖ Limits

- Cannot generate accurate conformers for transition metal complexes and molecules with non-canonical bonds.
- Lack of supports for other cluster schedulers

❖ Future perspective

- External packages support
- Data insufficiency

Questions? Comments?

Thank You