香港中文大學
The Chinese University of Hong Kong

20251103 Journal Club

Lewen Wang

communications chemistry

Explore content ⌄ About the journal ⌄ Publish with us ⌄

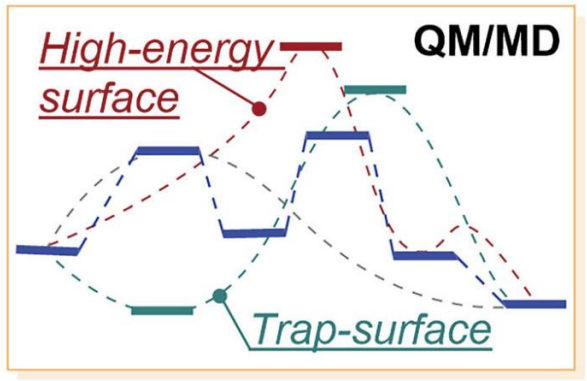
nature › communications chemistry › articles › article

Article | Open access | Published: 24 August 2025

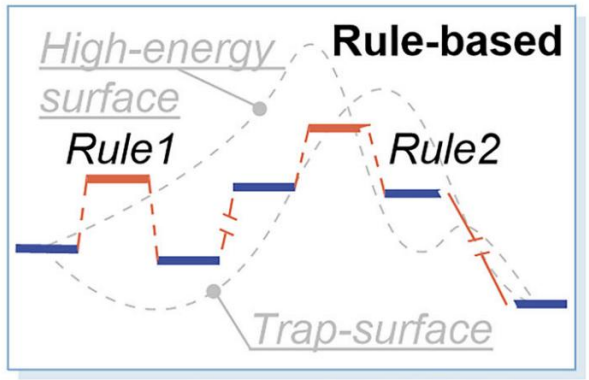
Large language model guided automated reaction pathway exploration

Ruzhao Chen, Yubang Liu, Zhe Chen, Yinwu Li, Fuyi Yang, Jiaxin Lin & Zhuofeng Ke ✉

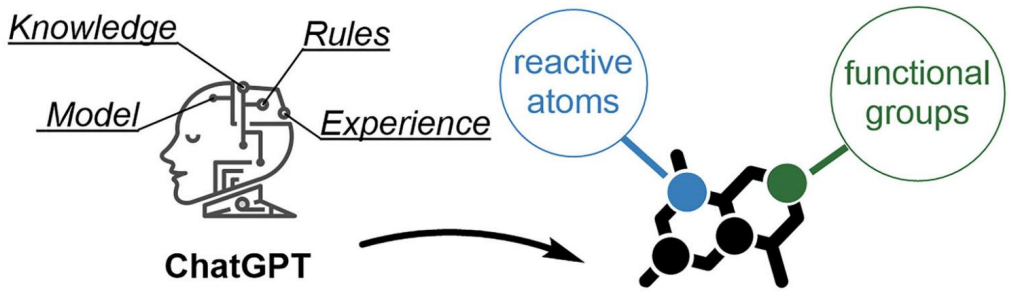
Communications Chemistry **8**, Article number: 255 (2025) | Cite this article

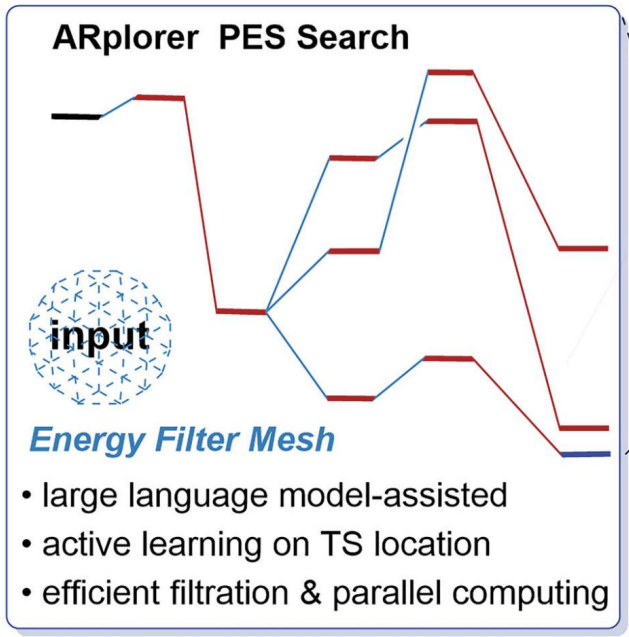


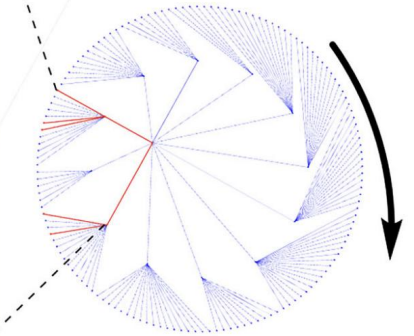
- multistep & various pathways
- time-consuming unbiased search
- unreasonable processes



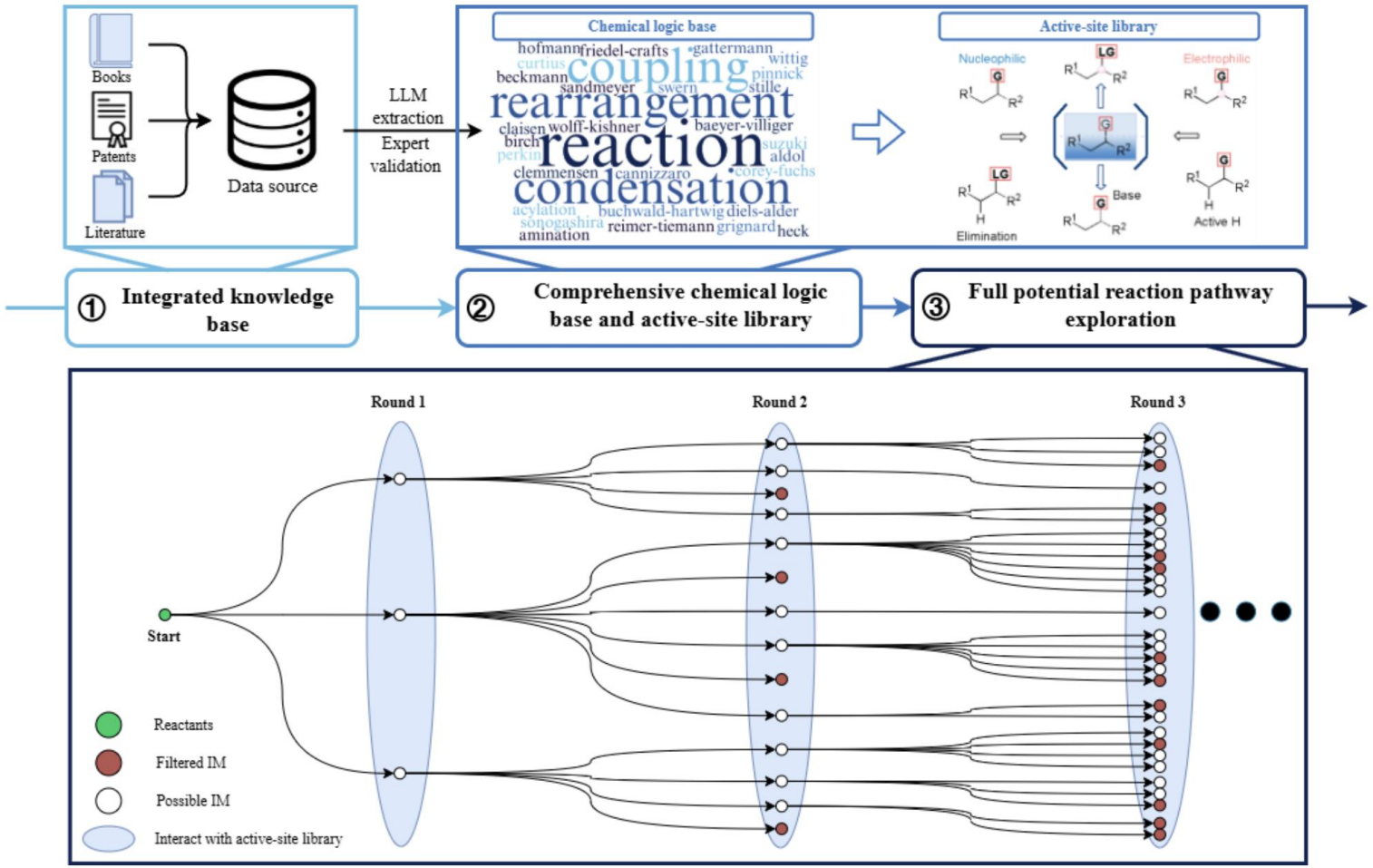
- bias search & avoid exhausted search
- not cover multistep search
- mostly not cover TS search







ARplorer comprehensive reaction steps



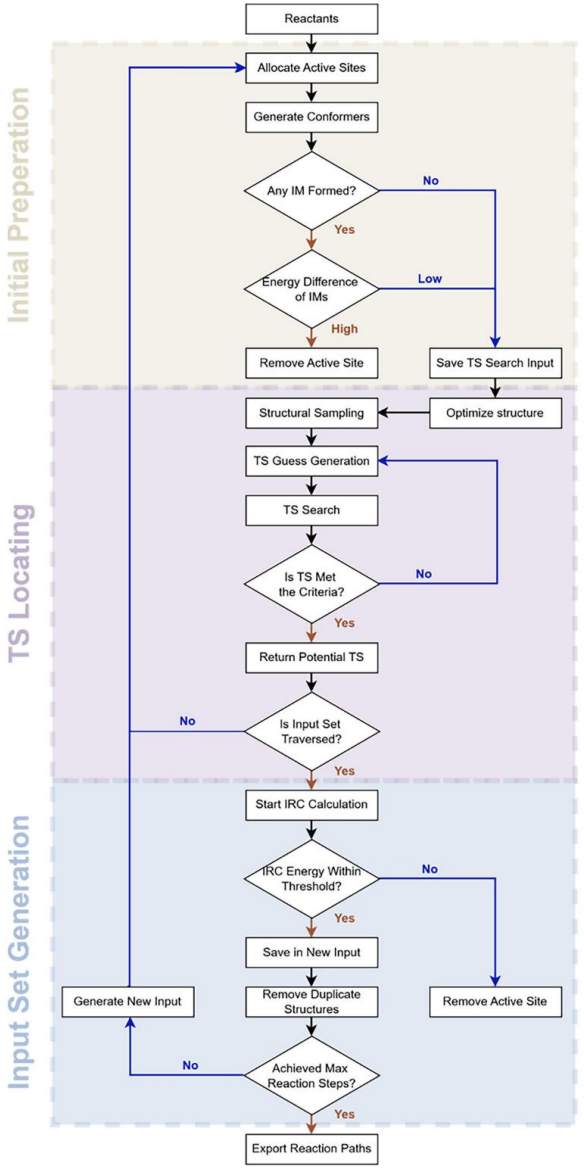
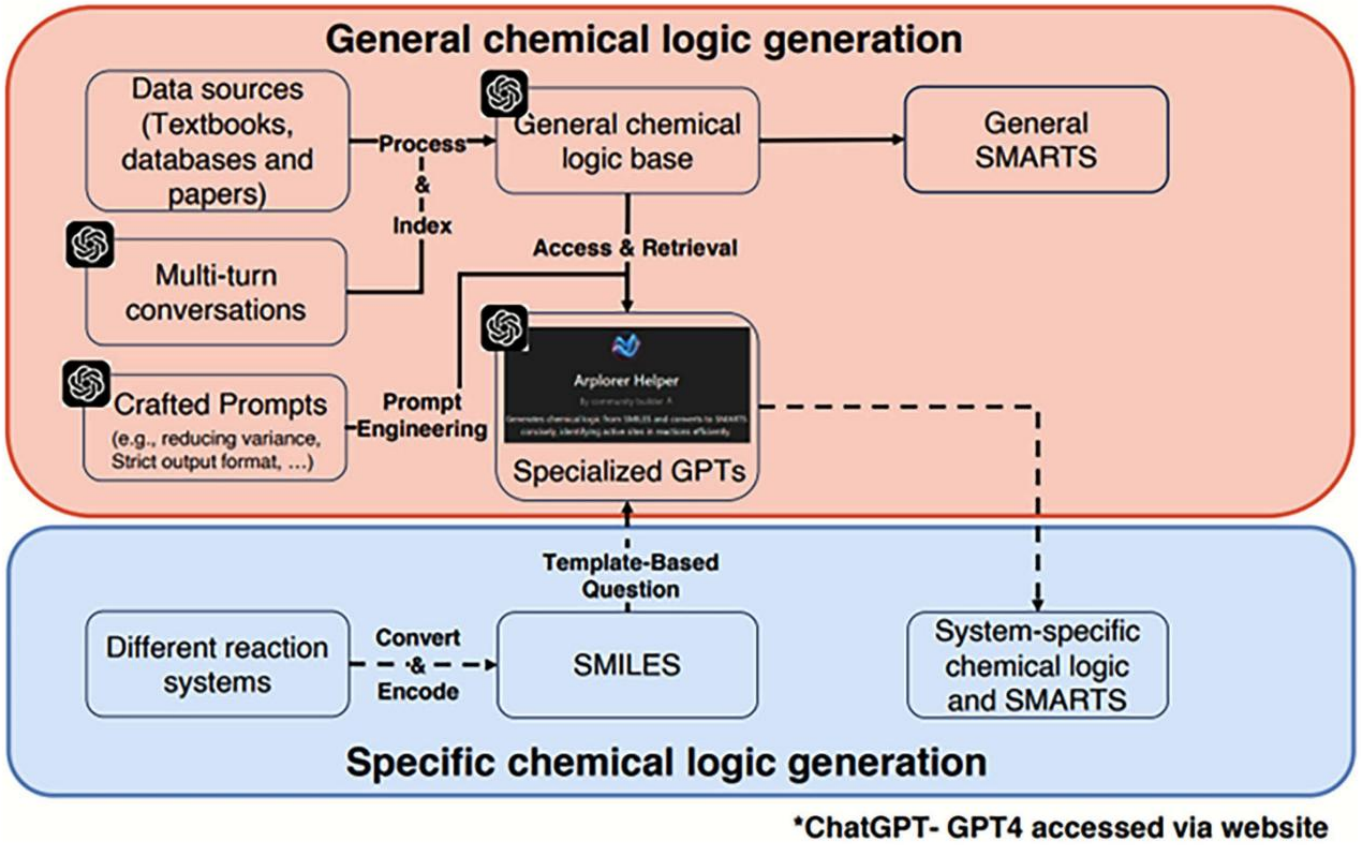
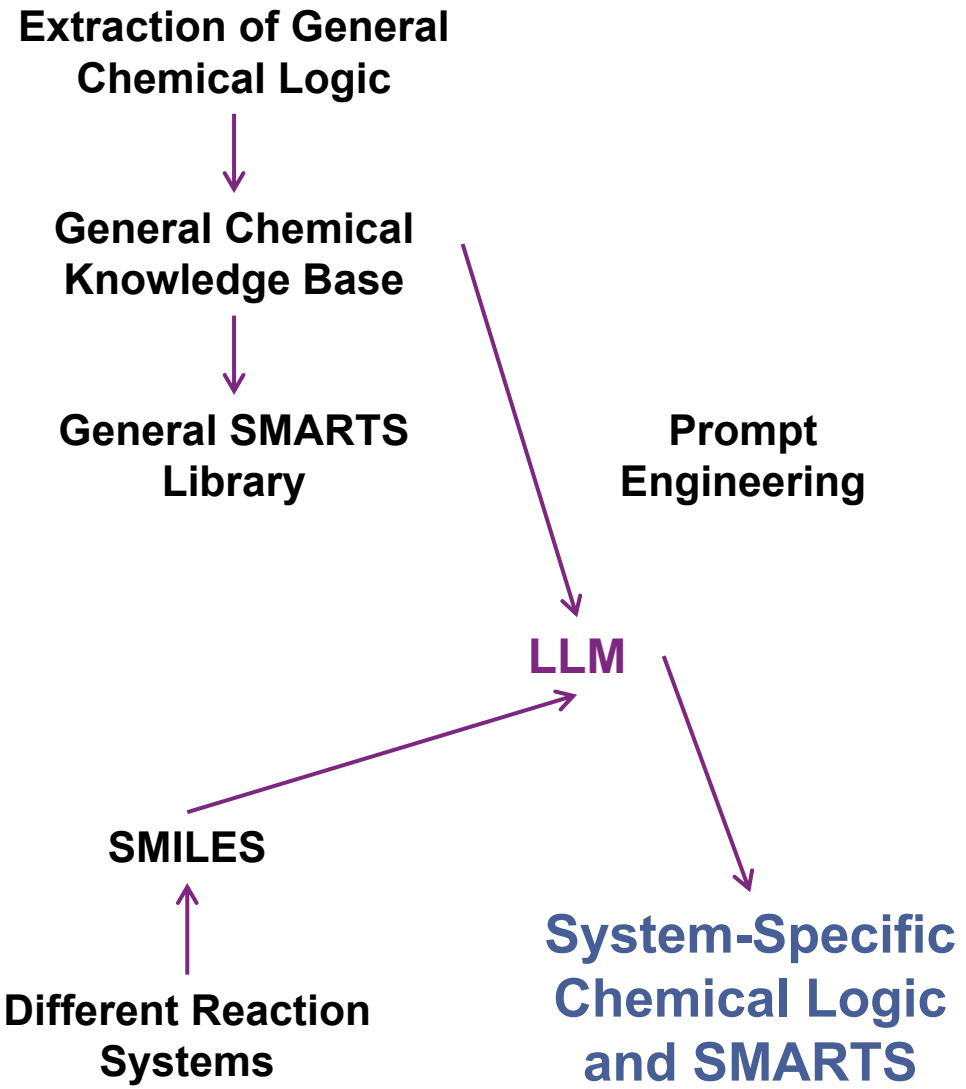
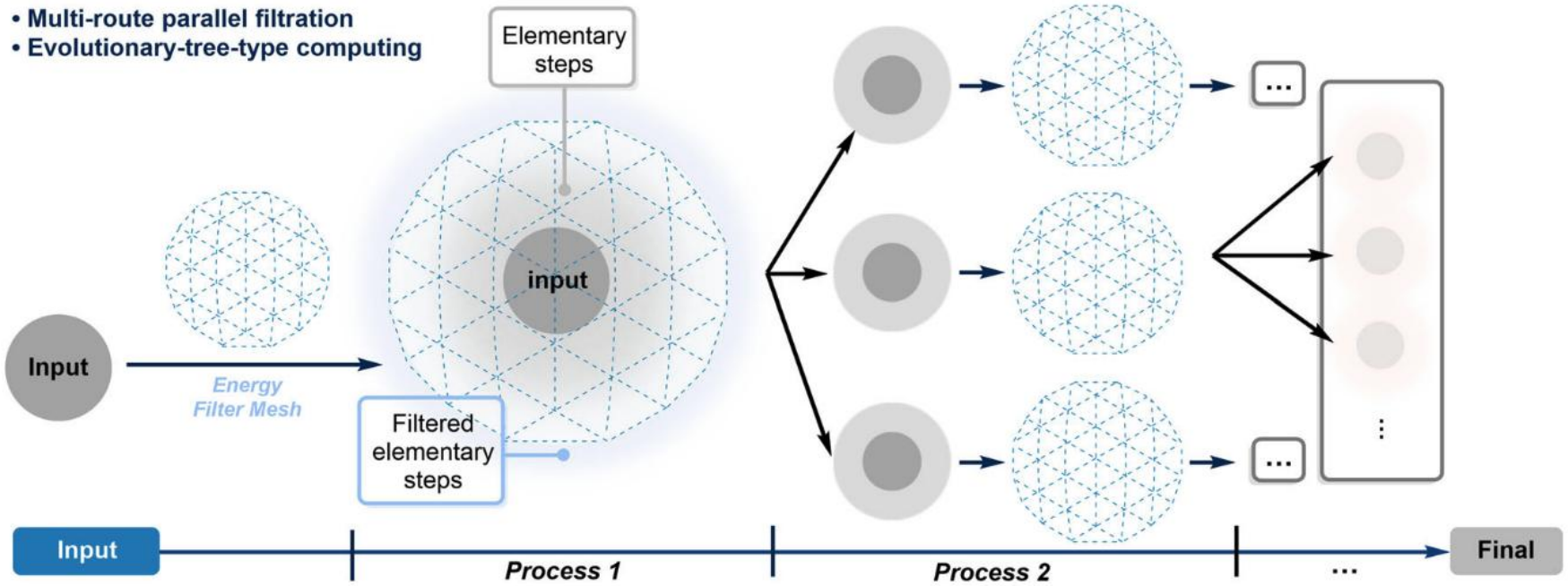


Figure S15 Overview of the ARplorer workflow


General chemical logic generation
Data sources (Textbooks, databases and papers)
Process & Index
General chemical logic base
General SMARTS
Multi-turn conversations
Access & Retrieval
Crafted Prompts
(e.g., reducing variance, Strict output format, …)
Prompt Engineering
Arplorer Helper
Specialized GPTs
Template-Based Question
Different reaction systems
Convert & Encode
SMILES
System-specific chemical logic and SMARTS
Specific chemical logic generation
*ChatGPT- GPT4 accessed via website


Extraction of General Chemical Logic
General Chemical Knowledge Base
General SMARTS Library
Prompt Engineering
LLM
SMILES
Different Reaction Systems
System-Specific Chemical Logic and SMARTS



System-Specific Reaction Logic

Reaction System Input → **Reactive Site Pair Identification** → **Conformational Search** → **Transition State Search** → **IRC Analysis** → **Intermediate Analysis**

$$U = \sum_{i,j \in bonds} a\left(r_{ij} - r_{ij_{bond}}\right)^2 + \sum_{i<j} \frac{b}{r_{ij}^n}$$

Bond-stretching potential

$$\sum_{i,j \in bonds} a(r_{ij} - r_{ij}^{bond})^2$$

Non-bonded repulsion term

$$\sum_{i<j} \frac{b}{r_{ij}^n}$$

r_{ij}	actual distance between atoms i and j	b	repulsion strength constant
r_{ij}^{bond}	ideal bond length for that atom pair	r_{ij}	interatomic distance
a	force constant (controls bond stiffness)	n	repulsion exponent

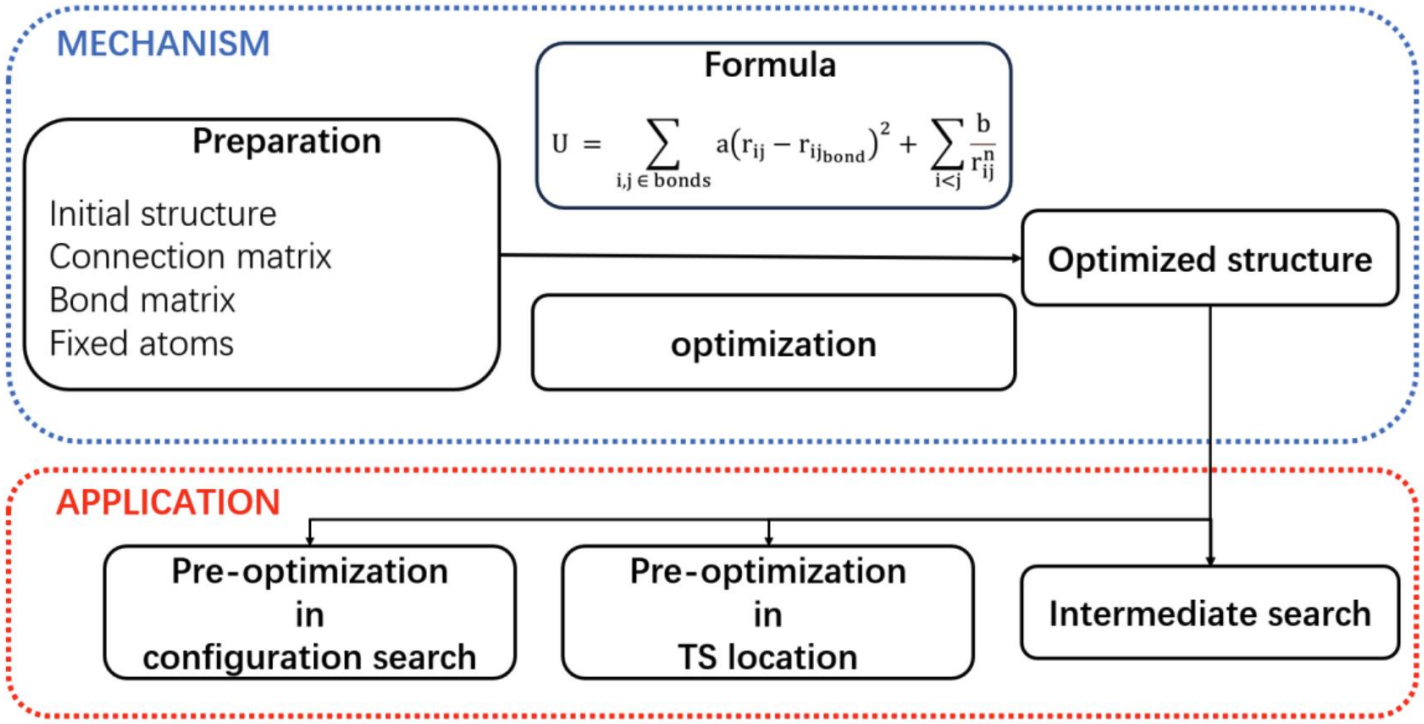
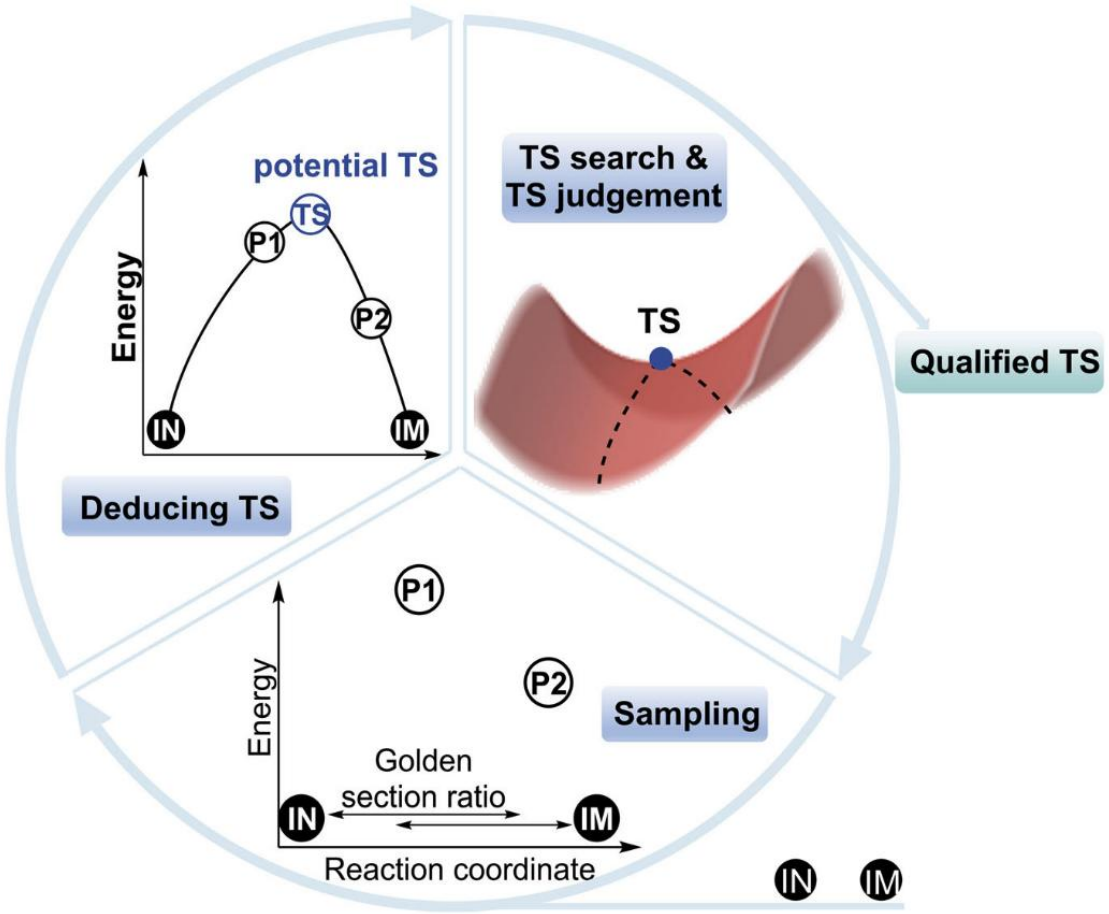


Figure S8. Functions of customized potential U.

$$U = \sum_{i,j \in bonds} a\left(r_{ij} - r_{ij_{bond}}\right)^2 + \sum_{i<j} \frac{b}{r_{ij}^n}$$



TS Locations Sampling

1. For a given IN, determine the corresponding intermediate (IM) using function U.

2. Select three probe points in the golden section ratio region.

3. Fit a quadratic curve and estimate the potential saddle point from the extremum.

4. If no qualified TS is obtained, use the extremum as a new probe point and repeat step 2.

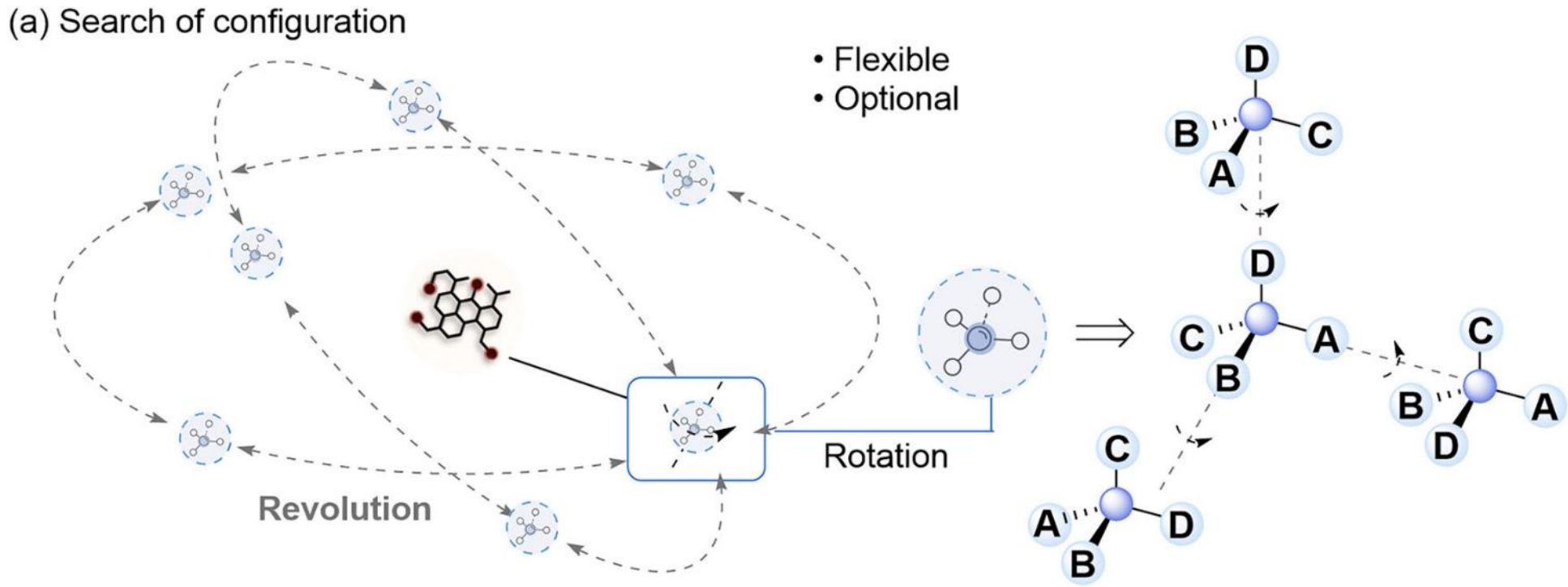
2-Steps Validation

- If the active atom pair shows a top-10 vibration weight in the negative-mode eigenvector.
- One unique virtual frequency in frequency analysis.

RMG database

Table S2. Benchmark of active learning sampling on TS location

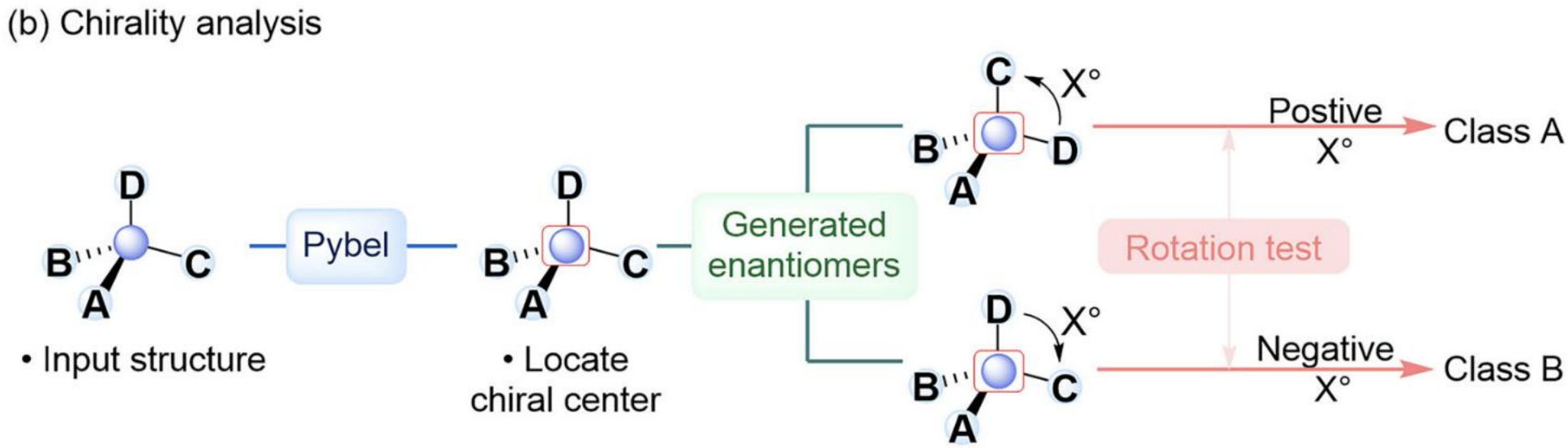
Entry	Reaction	Success (√) or failure (×)
1	[1+2] Cycloaddition	√
2	1,2-Insertion (CO)	√
3	1,2-Insertion (Carbene)	×
4	1,2-Elimination (NH_3)	√
5	1,2-Shift (C)	√
6	1,2-Shift (S)	√
7	1,2-Interchange	√
8	1,3-Insertion (ROR)	√
9	1,3-Insertion (RSR)	√
10	1,3-Insertion (CO_2)	×
11	1,3-Elimination (NH_3)	√
12	1,3-Sigmatropic Rearrangement	√
13	[2+2] Cycloaddition	√
14	6-Membered Central C-C Shift	√
15	Baeyer-Villiger Reaction Step1 (**Cat**)	√
16	Baeyer-Villiger Reaction Step2	√
17	Baeyer-Villiger Reaction Step2 (**Cat**)	√
18	Br atom Abstraction	√
19	Cl atom Abstraction	√
20	Cyclic-Formation (Ether)	√
21	Cyclic-Formation (Thioether)	√
22	Scission (Cyclopentadiene)	√
23	Diels-Alder Addition	√
24	Diels-Alder Addition (Aromatic)	√
25	F Atom Abstraction	√
26	Elimination (Peroxy Radical)	√
27	H Atom Abstraction	√
28	Intra-[2+2] Cycloaddition	√
29	Intra-5-Membered Conjugated Addition (C=C-C=C)	√
30	Intra-Diels-Alder (Monocyclic)	√
31	Concerted Intra-Diels-Alder (Monocyclic 1,2-Shift)	×
32	Intra-Ene Reaction	√
33	Intra-Halogen Migration	√
34	Intra-H Migration	√
35	Intra-OH Migration	√
36	Intra-Retro-Diels-Alder Reaction (Bicyclic)	√
37	Intra-RH-Addition (Endocyclic)	√
38	Intra-RH-Addition (Exocyclic)	√
39	Intra-R-Addition (Endocyclic)	√
40	Intra-R-Addition (Exocyclic)	√
41	Intra-R-Addition (ExoTetCyclic)	√
42	Intra-R-Addition (Exo Scission)	×
43	Intra-Substitution Cyclization (CS)	√
44	Intra-Substitution Isomerization (CS)	√
45	Intra-Substitution Cyclization (S)	√
46	Intra-Substitution Isomerization (S)	√
47	Ketoenol	√
48	Korcek Reaction Step1	√
49	Korcek Reaction Step2	×
50	Korcek Reaction Step1 (Cat)	√
51	Retroene	√
52	R-Addition (CO)	√
53	R-Addition (CS)	√
54	R-Addition (MultipleBond)	√
55	Singlet-Carbene Intra-Disproportionation	√
56	Substitution (S)	√
57	Substitution (O)	√
58	XY-Addition (MultipleBond)	√
59	XY-Elimination (Hydroxyl)	×
	Total number: 59; Succeed cases: 53; **Successful Rate: 90%**	



Focuses only on conformations associated with the active sites.

The portion with active site B rotatedwhile simultaneously revolving around the part with active site A.

The conformations are extracted at a certain frequency.



1. Recognition and collection of chiral carbon atoms using Pybel;

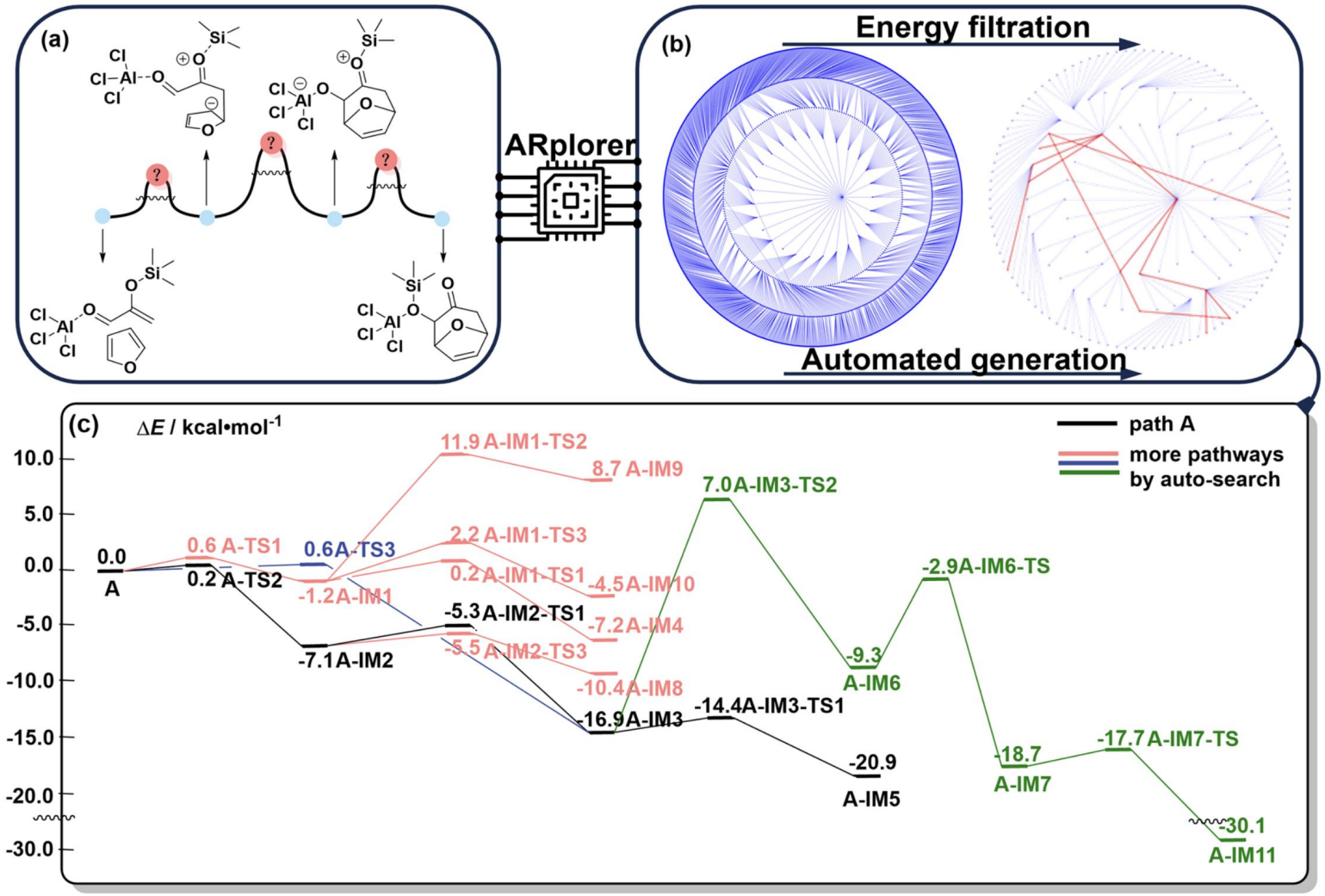
2. Generate enantiomers through rotation and the use of function U;

3. Classifie the elementary step accurately by rotating the structure to maintain the same atomic connections but considering the different spatial relationship of the chiral atom and its connected atoms.

Table S3-1. Comparison between ARplorer and other programs (Claisen rearrangement of CH_2=CH−O−CH_2−CH=CH_2)

Program	Total reactions	Outcome reaction	Outcome yield	Total number of TSs	Total number IMs	Costed Time (h)	Calculation level
ARplorer	427	156	36.5%	156	178	5	xTB (GFN-2)
ADCR	24949	221	0.8%	221	282	18.9	xTB (GFN-2)
GRRM	-	200	-	200	197	168	B3LYP /6-31G
Kinbot	220	64	29.1%	34	-	72	B3LYP /6-31G
LASP	1575	303	19.2%	303	2182	8	MLP

In ARplorer and ADCR, the reaction is specified to proceed in two reaction steps. In other programs, reaction steps could not be specified, and these programs were terminated at a given time. In ADCR, four of the twelve atoms were frozen. The site with "-" means this data could not be found.



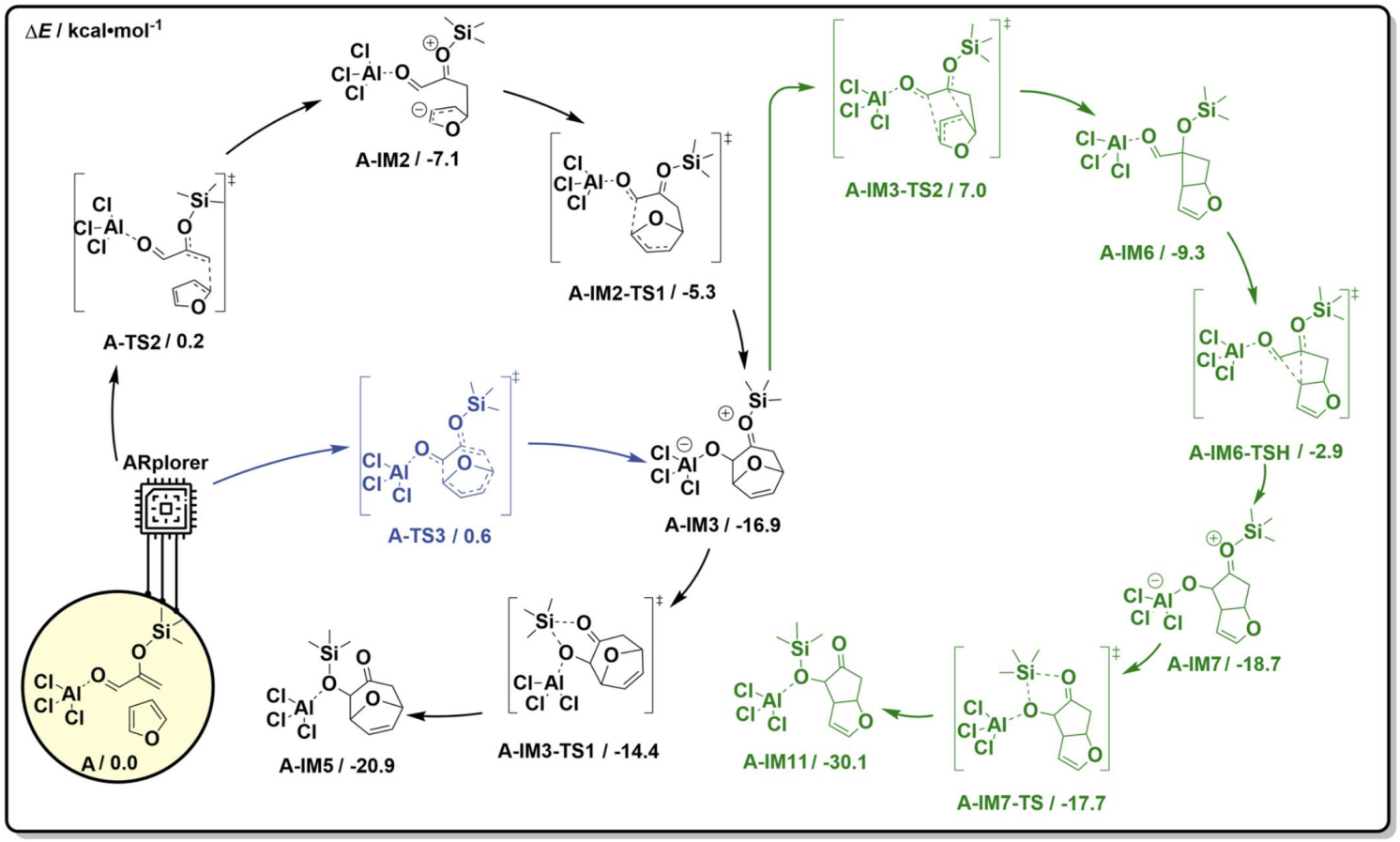
1. Reproduced the literature-reported reaction pathway (path A);

2. Discovered additional plausible pathways;

3. Identified a previously unreported concerted [4+3] pathway;

4. Discovery of a new side reaction.

~1.5 days, generating 874 reaction steps, on 24-core Xeon GOLD 6342 CPU



- Although this product cannot form directly due to Woodward–Hoffmann constraints, it undergoes a 1,2-migration leading to A–IM7 and a final stable product A–IM11.

- This new pathway suggests a thermodynamically controlled alternative product channel that was missed by manual DFT studies.

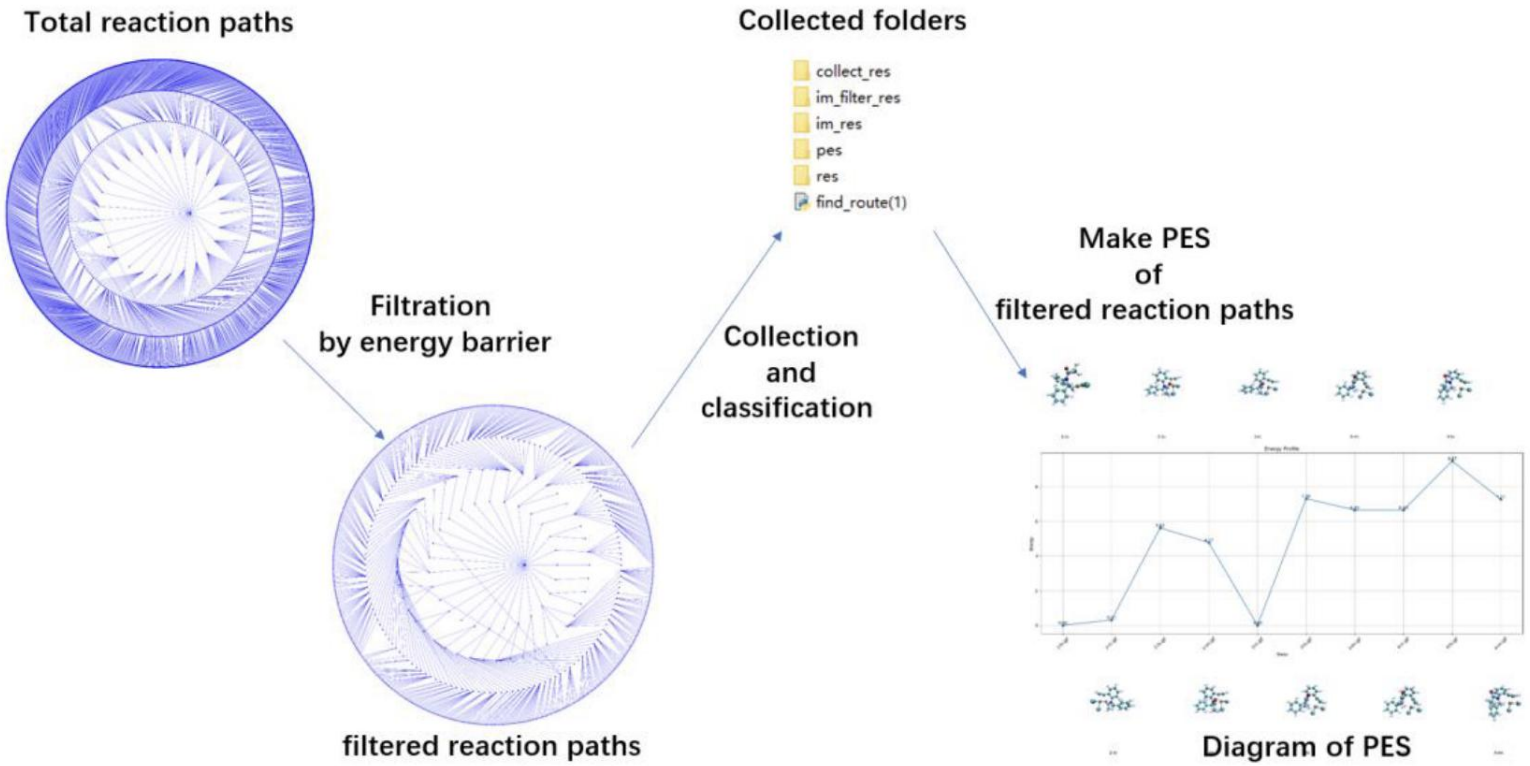
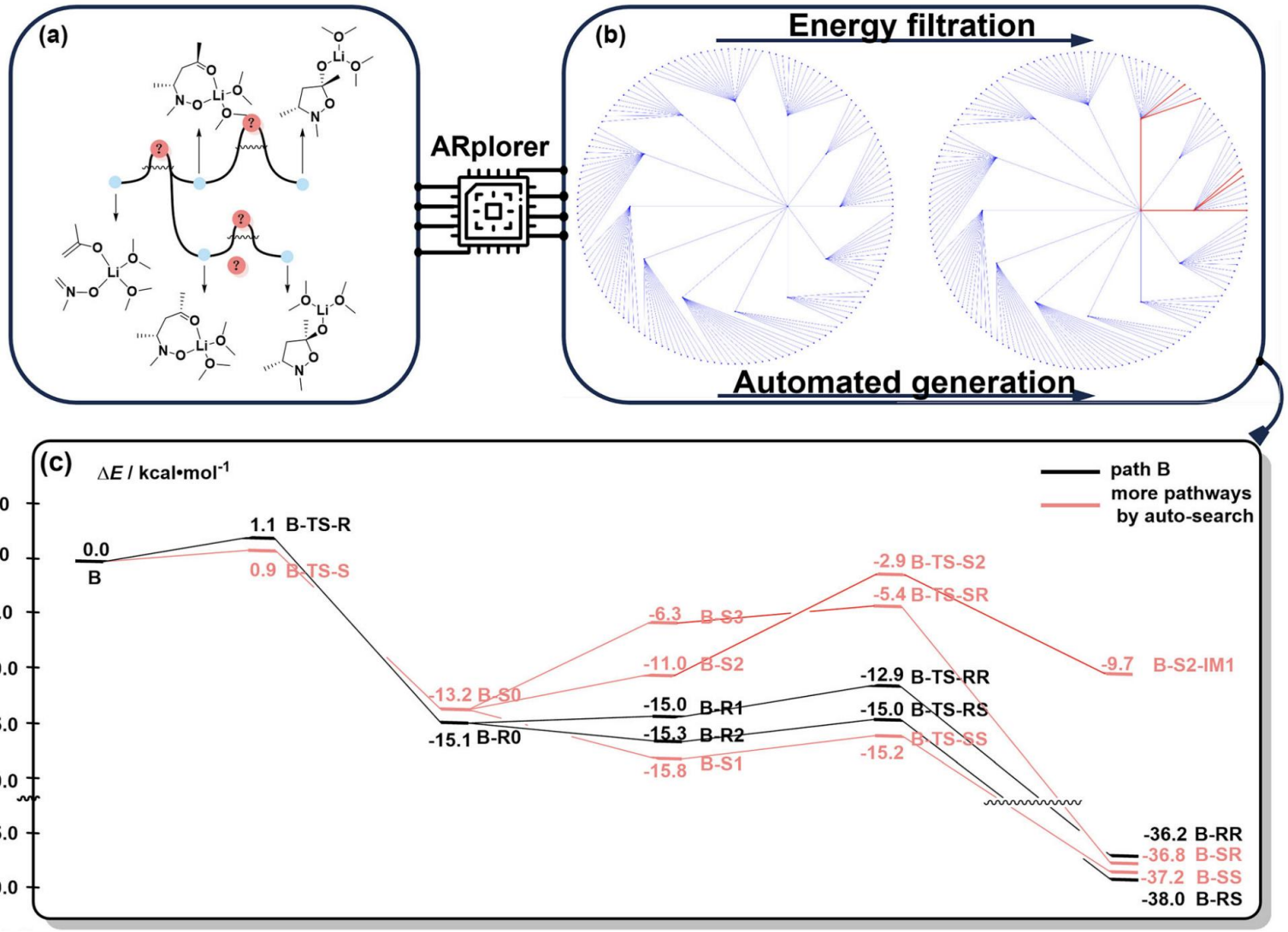


Figure S9. Workflow of postprocessing in the program.

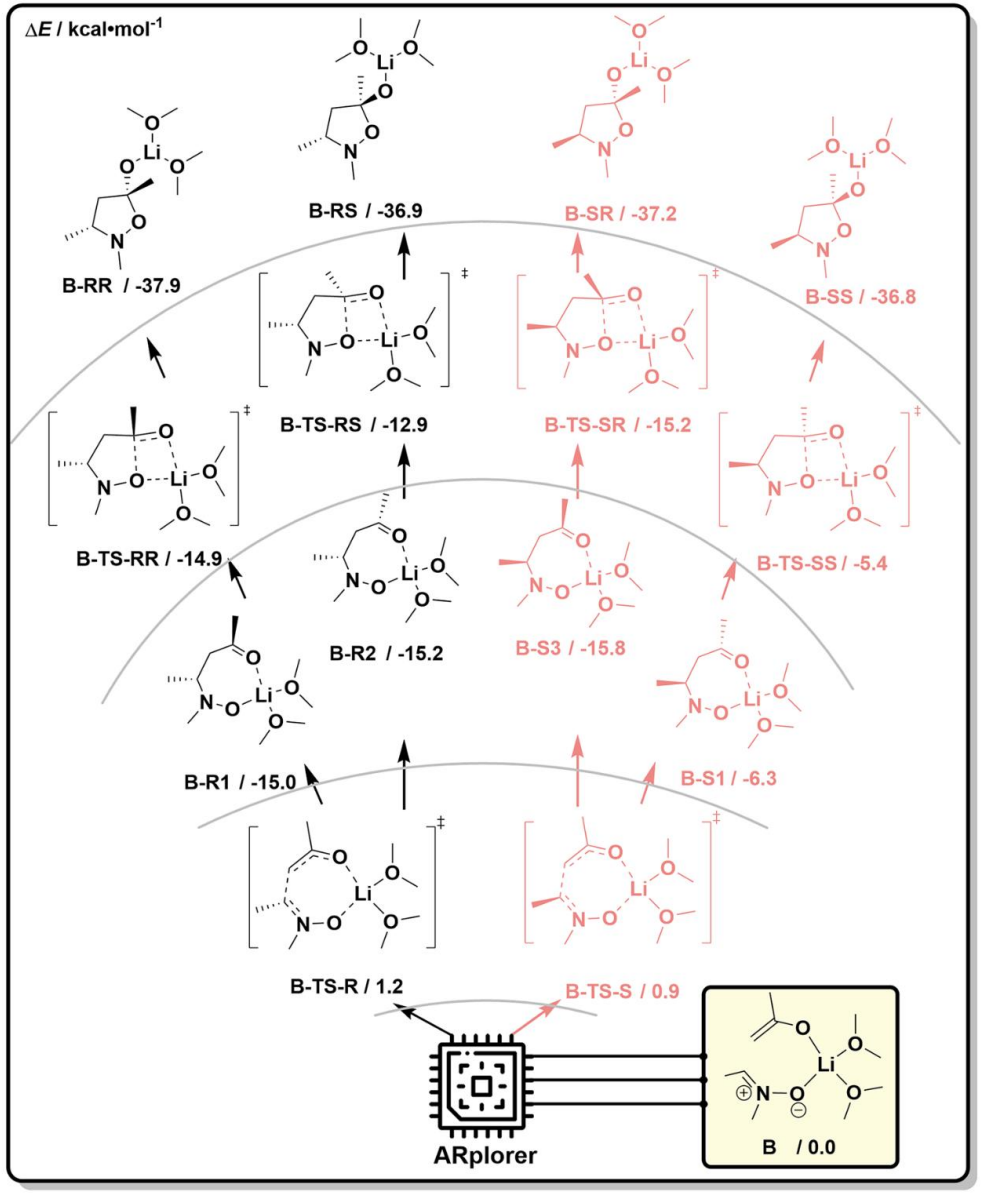


~6 hours, generating 156 reaction steps, on 24-core Xeon GOLD 6342 CPU

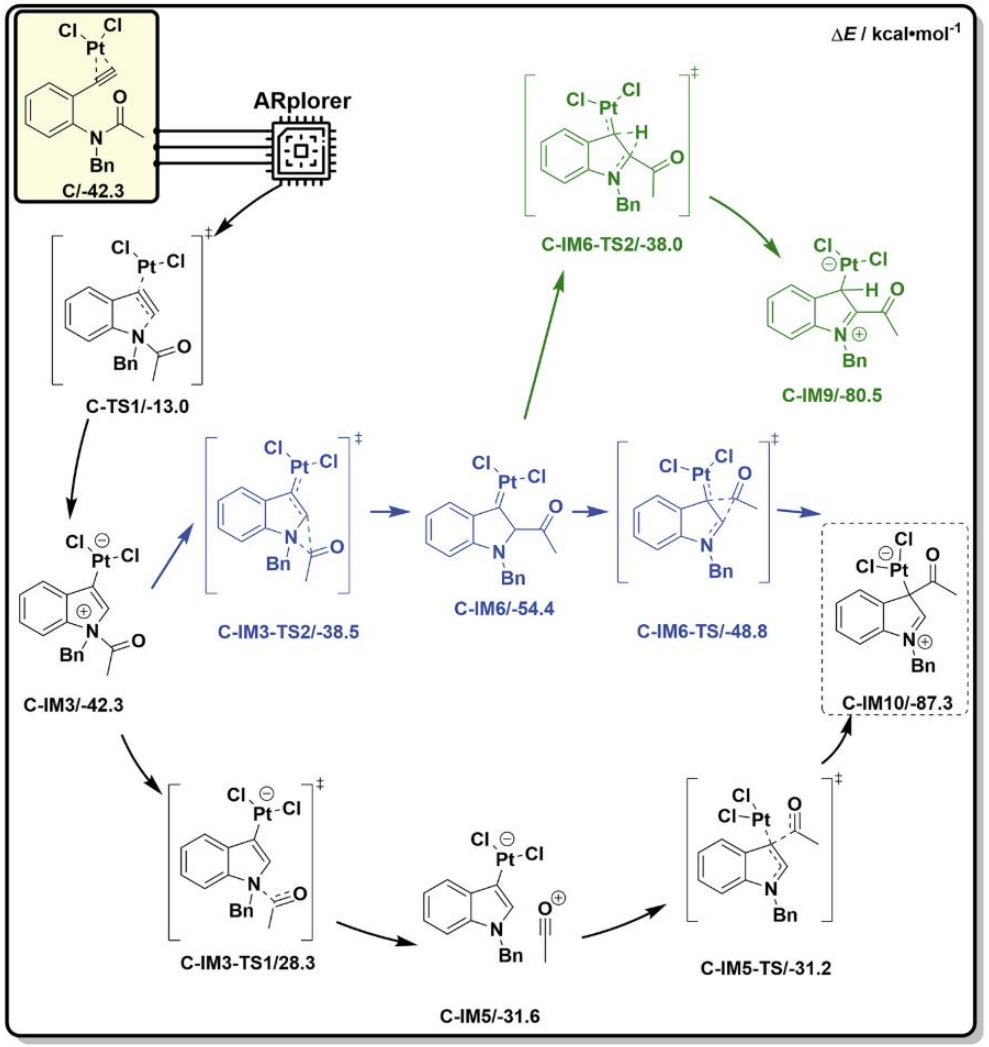
1. Successfully reproduced the literature-reported asymmetric reaction pathway (path B);

2. Mapped all stereoselective pathways leading to B-SS, B-SR, B-RS, and B-RR products;

3. The automated exploration network illustrates the full set of transition states and intermediates obtained.



ARplorer systematically and efficiently mapped all stereoselective pathways in a remarkably short timeframe. The exploration covered all enantio- and diastereoselective pathways towards B-SS, B-SR, B-RS,and B-RR products. While further refinement with high-level theory and consideration of solvation effects were not yet incorporated in this preliminary exploration with ARplorer



~4.5 days, generating 2693 reaction steps, on 24-core Xeon GOLD 6342 CPU

1. Successfully replicated the known pathway via the acylium intermediate;

2. Identified an alternative intramolecular 1,2-acyl migration pathway;

3. Uncovered a previously unreported, kinetically favored carbene pathway, missed by manual DFT calculations.

Limitations

- Reaction logic libraries should be further expanded for some elementary reactions and special synthons
- The computational complexity still scales exponentially with the number of steps in the reaction pathway
- For efficiency, solvent effects and high-level electronic energy corrections were not included during automated exploration.

Thank You