Database-Guided Program Synthesis of Chemical Reaction Networks

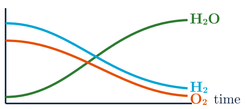
Where reaction-database knowledge is most effective in reducing search

Tymon Jastrzemski — TU Delft, EEMCS · CSE3000 Research Project 2026
Responsible Professor: Sebastijan Dumančić · Supervisor: Reuben Gardos Reid

1 Background

- Chemichal reactions evolve as species **concentrations** rise and fall
- A **CRN** (chemical reaction network) fully determines that evolution
- We see the curves; we must recover the **network** that made them

Observed: concentrations c(t)



recover ⟹ inverse problem

Reaction network

$$2\,H_2 + O_2 \rightarrow 2\,H_2O$$

a **CRN** = species + reactions + rates

Recovering the network from data is done by hand and requires expert knowledge.

2 The Problem & the Gap

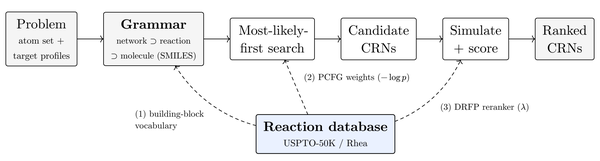
- Framed as **program synthesis**: a grammar defines valid networks; a top-down search **enumerates and simulates** candidates (Wijers 2025)
- The candidate space grows with network size the search space **explodes**
- The bottleneck is **how many candidates** it tries before it reaches the target
- Hard constraints (valence, mass balance) prune the chemically **impossible**, but they cannot say which candidates are **plausible** and likely to occur in nature

The gap: guide the search toward plausible reactions, exactly what a reaction database records.

3 Research Question

To what extent, and at which point in a top-down CRN program-synthesis pipeline, does knowledge from a reaction database reduce the number of candidates explored before the target network is found?

Chemical reactions are catalogued in **reaction databases** (USPTO-50K, Rhea). We inject it at three points:



Q1 Integration point. Which point most reduces candidates-to-target — the PCFG's production-rule probabilities, the building-block vocabulary, or output reranking?

Q2 Content vs. size. Is the reduction driven by the database's frequency content, or just a larger vocabulary?

Q3 Corpus mismatch. How does corpus–target **mismatch** affect the reduction — and whether the target stays reachable?

4 Two Databases

- **USPTO-50K** — organic patent reactions (carbon chemistry)
- **Rhea** — biochemical, inorganic & small-molecule reactions

Benchmark	Chemistry	Matched corpus
Water	inorganic	Rhea
Methane	inorganic	Rhea
Esterification	organic	USPTO
Diels–Alder	organic	neither

5 Building-Block Vocabulary

The molecule pool fixes which species a candidate can contain.

Corpus: USPTO-50K · Rhea
count how often each molecule appears

↓ take the top-N most frequent

Molecule pool = building blocks
a missing species → target unreachable

6 Result

Q: Does database content decide which targets are found? (Q1, Q3)

Benchmark	atom	USPTO	Rhea	∪ sum	∪ rank
Water	6	6	6	5	5
Methane	–	–	**1**	**1**	**1**
Esterification	–	1077	–	–	**421**
Diels–Alder	–	–	–	–	–

candidates-to-target (lower = better) · – = not reached within budget

Each corpus unlocks only its chemistry: USPTO → esterification (idx 1077), Rhea → methane (idx 1). A **rank-normalised** merge recovers both (esterification idx 421); a naive sum **dilutes** and loses esterification.

7 PCFG Over the Grammar

Probabilistic context free grammar (PCFG) puts a probability on grammar production rules. Estimated from corpus count. Search uses cost −log p, so most-likely-first enumeration visits **high-probability** derivations first. We weight two rule groups.

Stoichiometry (surface) prior — coarser

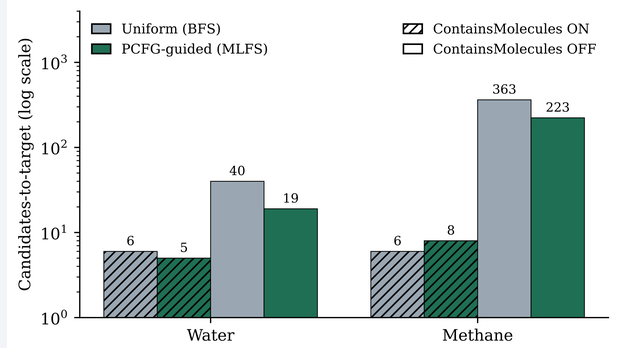
Weights the reaction grammar's **molecule-list rule**, how many molecules sit on a reaction side, by how often each side-size occurs in the corpus.

Atom-bag (molecular-formula) prior — finer

Weights the **molecule grammar's productions** by how often a molecule of that **atomic composition** (its formula, ignoring how the atoms bond) appears in the corpus.

8 Result Stoichiometry Prior

Q: Does a reaction-stoichiometry prior reach the target sooner? (Q1)



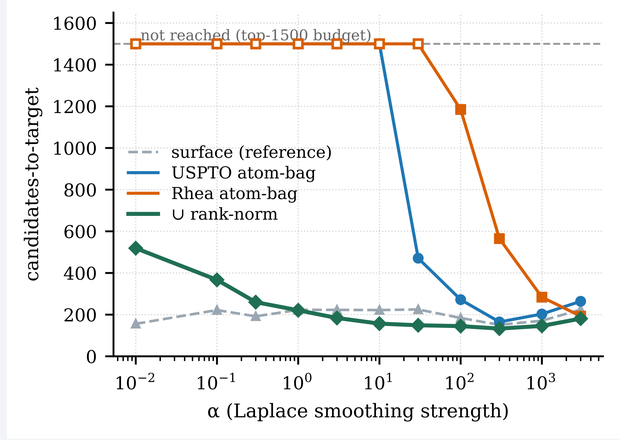
With the **ContainsMolecules** constraint **off**, the coarse **reaction-stoichiometry (surface)** prior reaches the target sooner: water 2.11× (40→19), methane 1.63× (363→223). With the constraint **on**, counts are already tiny (6) — so the prior's ordering gain shows clearest when the constraint is off.

Constraint + prior are complementary: the constraint prunes the invalid candidates, the prior orders valid ones.

The stoichiometry prior orders the reachable set (≤ 2.11×) — complementary to the hard constraint.

9 Result Atom-Bag Prior

Q: Does a finer molecular-formula prior help further? (Q1, Q3)



Pushing to a finer **molecular-formula (atom-bag)** prior **backfires**: it erases water's gain and makes methane unreachable on **both** single corpora — even Rhea, where O_2 is abundant. Results return only as the prior is smoothed toward uniform (α→∞); only the rank-normalised union finds methane across the range (idx 221).

A finer formula prior backfires.

10 DRFP Reranker

Post-process the finished list: reward candidates that **resemble known reactions**. We use Tanimoto similarity between their DRFP reaction fingerprint and the corpus.

Benchmark	none	USPTO	Rhea	∪ rank
Water	3	5	**1**	**1**
Methane	11	9	**7**	**7**
Esterification	12	**10**	14	**10**
Diels–Alder	13	15	15	15

match → helps mismatch → hurts

Reranking only reorders the finished list, it helps when the corpus matches, hurts when it does not.

11 Where Knowledge Helps Most

All three points are governed by the same **corpus-to-chemistry match**. Answering each sub-question:

- **Q1 which point?** The **vocabulary** changes if the target is found; the **PCFG** makes the target appear sooner; the **reranker** only reorders the finished list.
- **Q2 — content or size? Content.** Simple union of two corpora dialutes the pool.
- **Q3 — corpus mismatch?** A **matched** corpus helps at every point and a **mismatched** one hurts; rank-normalised merging of USPTO + Rhea recovers both.

12 Conclusion

You have to match the corpus to the chemistry.

- Building-block vocabulary can find targets the baseline cannot
- PCFG can make target networks appear sooner in enumeration
- A matched corpus helps; a mismatched one hurts
- Rank-normalised merging is what lets two corpora cooperate

13 Limitations & Future Work

Limitations

- Small benchmark set
- Corpus bias: USPTO is organic, Rhea inorganic, there is chemistry leaving outside both
- Vocabulary can **lose** a target (Diels–Alder, in neither corpus);

Future work

- More, chemically diverse benchmarks
- Weight the reaction- and network-grammar rules too
- Merge more than two corpora; broaden the database reach