Substructure-Aware Program Synthesis for Automated Chemical Reaction Network Discovery

Author:
Adam Piotr Szymaniak | A.P.Szymaniak@student.tudelft.nl

Responsible Professor:
Sebastijan Dumančić

Supervisor:
Reuben Gardos Reid

Background

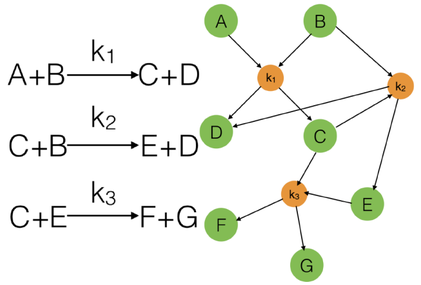

Figure 1: Chemical Reaction Network (CRN) represented as a list of reactions on the left, and as a directed bipartite graph on the right. [Source: networkpages.nl/kinetics-from-networks]

- Nearly every reported **Chemical Reaction Network (CRN)** is incomplete due to empirical data limitations.
- **Program Synthesis** refers to the automatic generation of programs from high-level specifications or I/O examples and a context-free grammar (CFG) that defines the search space.

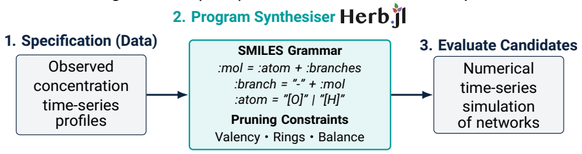

Figure 2: The Program Synthesis Framework applied to CRN Discovery

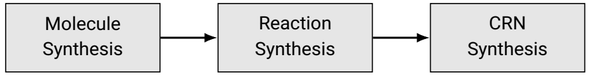

Figure 3: Sequential CRN Synthesis Pipeline

- **The baseline synthesiser** constructs candidate molecules atom-by-atom from scratch. It also lacks awareness of the structural context of known molecules during reaction and network synthesis.

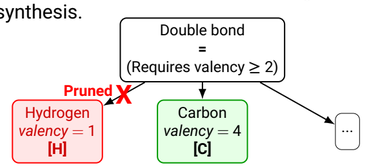

Figure 4: Partial Abstract Syntax Tree (AST) demonstrating an active state transition and structural search space pruning

Research Questions

a. How does the introduction of fragments from known molecules as high-level building blocks impact the molecule synthesiser's ability to construct complex structures?
b. How does molecular similarity scoring impact the number of candidate reactions and networks generated prior to target discovery?

Methods

Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS)

- **BRICS fragments** are extracted from known molecules and introduced into the molecular grammar.

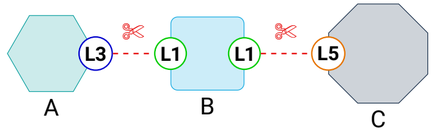

Figure 5: BRICS decomposition of two strategic bonds. The bond between fragment A and B is replaced with connection points L_3 and L_1. The bond between fragment B and C is replaced with connection points L_1 and L_5

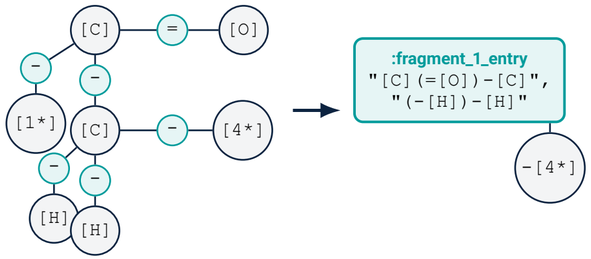

Figure 6: Representation of a BRICS fragment in the baseline molecular grammar on the left, and the proposed compressed grammar on the right.

Chemical Similarity Guidance

- **Collecting Fingerprints:** Molecules are translated into a binary feature-vector representation, known as a (Morgan2) fingerprint, that represents the presence or absence of structural features.
- **Guiding the Synthesis:** The synthesiser prioritises reactions and networks with high Tanimoto similarity (Equation 1) to the known molecules.

$$Tanimoto(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

Experimental Setup

- The framework was evaluated on reaction rebalancing tasks from curated SynRXN datasets and an example incomplete esterification CRN (see Figure 7).

$$CH_2O_2 + C_2H_6O \longrightarrow C_3H_6O_2 + H_2O$$
$$C_3H_6O_2 + CH_4O \longrightarrow C_2H_4O_2 + C_2H_6O$$

Figure 7: The target esterification Chemical Reaction Network. The synthesiser must identify the missing species (H_2O, CH_2O_2, CH_4O) and reconstruct both reactions.

Results

Reaction Rebalancing

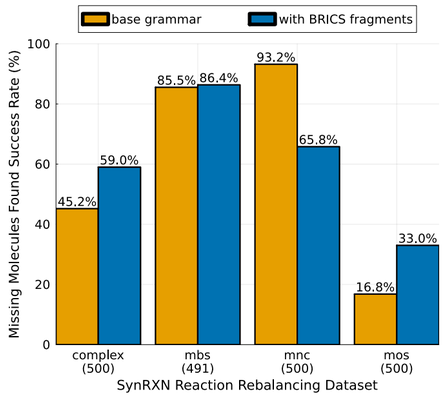

Figure 8: Success rates of synthesising all missing molecules within the first 250 candidates for reaction rebalancing datasets.

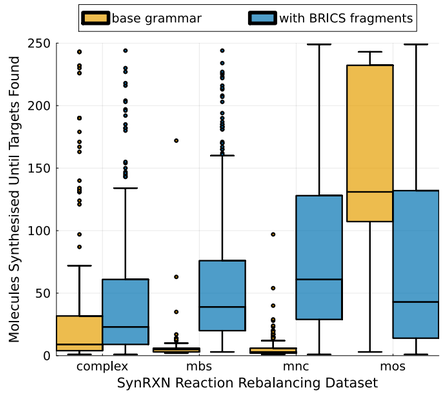

Figure 9: Distributions of the number of molecules synthesised before finding all missing molecules across successfully solved subproblems.

Example Esterification Network Discovery

Table 1: Number of candidate reactions synthesised before finding the target reactions within the first 250 candidate molecules, and the number of networks synthesised before identifying the example esterification network with early synthesiser halting.

Max Stage	Similarity Guidance	Synthesised Until Target	
		Reactions	Network
Reactions	None	22196	N/A
	Tanimoto w/ Morgan2	19631	N/A
Networks	None	144	435
	Tanimoto w/ Morgan2	118	26

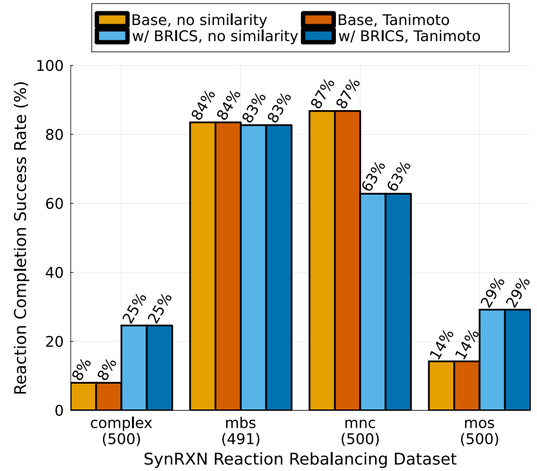

Figure 10: Reaction rebalancing success rates.

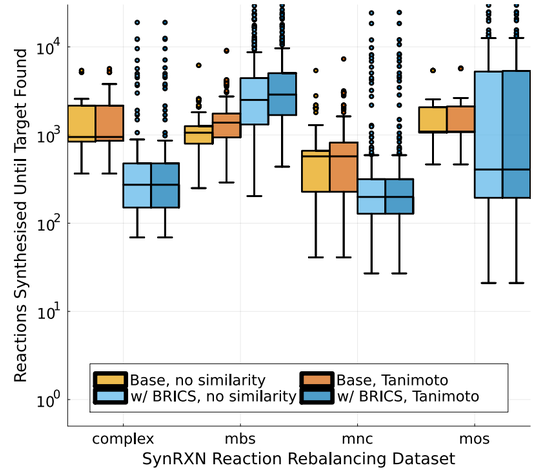

Figure 11: Distributions of the number of reactions synthesised until the target reaction is found across successful runs.

Conclusions and Future Work

- **BRICS Fragmentation** accelerates complex molecule generation by encoding large substructures as single rules.
- **Hybrid Approaches:** Because BRICS delays the discovery of small intermediates, future iterations should split the candidate pool between fragment-enhanced and baseline methods.
- **Molecular Similarity Guidance** reduces the number of candidates explored in the example esterification CRN synthesis.
- **Targeted Search:** Similarity guidance does not help in reaction rebalancing. Future iterations should guide the search using remaining atom counts and chemical plausibility filters.