Categorical Chemistry: The architecture of a chemical computer

Wilmer Leal

Gatas Lab
Department of Computer Science
University of Florida, USA

Joint work with:
Benjamin Merlin Bumpus, Eugenio Llanos, & James Fairbanks.

2024-09-10

$$\begin{tikzcd}
G \arrow[r] \arrow[d] \arrow[dr, phantom, "\Downarrow \tau"] & A \arrow[d] \\
A \arrow[r] & S
\end{tikzcd}$$

Goal: Automation in computer aided synthesis planning

Computer-aided chemical discovery cycle[1]:

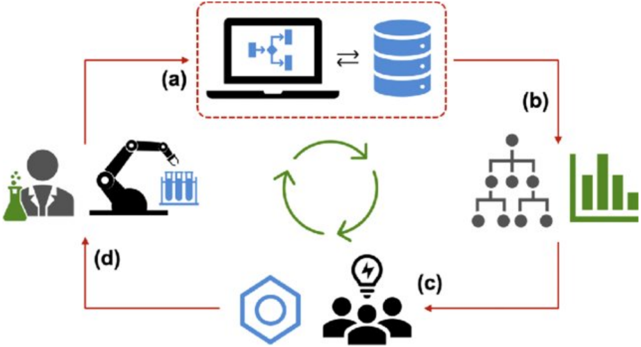


Figure: : (a) the Open Reaction Database; (b) machine learning and cheminformatics; (c) human or automated interpretation and material design; (d) manual or robotic chemical synthesis.

[1] Steven M. Kearnes, et.al. *The Open Reaction Database.* J. Am. Chem. Soc. 2021

What does organic chemistry compute?

Reachability and connectivity (network level): Can I synthesize the desired products with the materials and technology available?

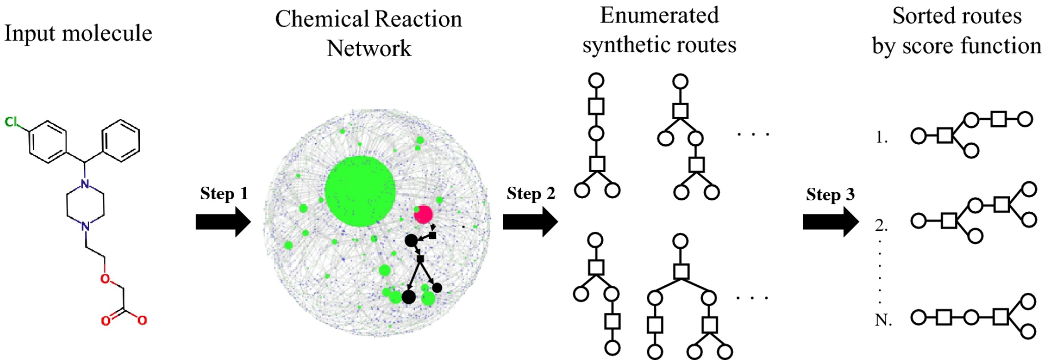

Input molecule
Cl
N
N
O
O
O
Step 1
Chemical Reaction Network
Step 2
Enumerated synthetic routes
. . .
. . .
Step 3
Sorted routes by score function
1.
2.
N.

Figure: Taken from: Shibukawa, R., Ishida, S., Yoshizoe, K. et al. *CompRet: a comprehensive recommendation framework for chemical synthesis planning with algorithmic enumeration.* J Cheminform 12, 52 (2020).

What does organic chemistry compute?

Reachability and connectivity (Molecular structure level): Can I synthesize the desired products with the available materials and technology *while tracking changes in molecular structure (and thus gain information about the reaction mechanism)*?

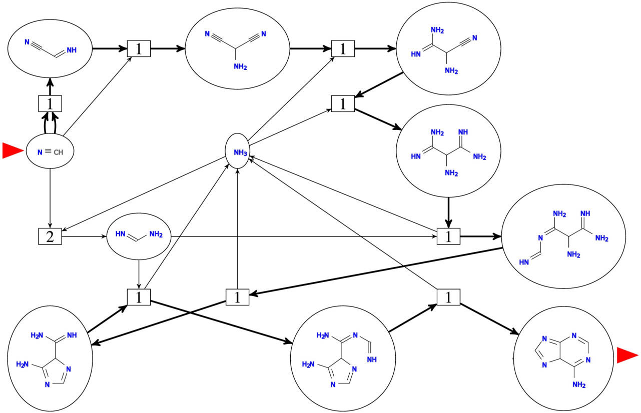
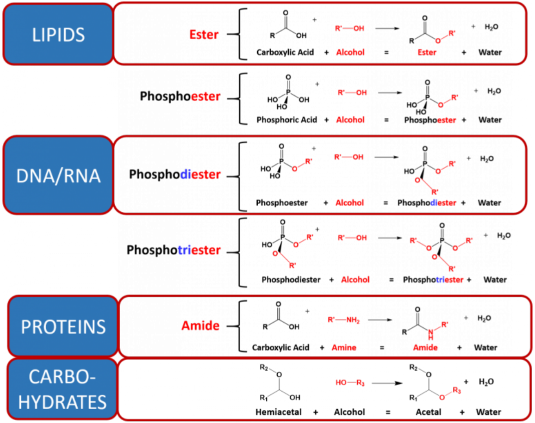


Figure: Is there a synthesis path from hydrogen cyanide (HCN) to adenine? Taken from: Müller, S., Flamm, C. & Stadler, P.F. *What makes a reaction network "chemical"?*. J Cheminform 14, 63 (2022)

The abstractions where chemists compute these problems

Reactivity prediction: How the above structures allow us to deduce the similarity relationships of substances and reactions that predict whether a reaction will take place.



LIPIDS
Ester
Carboxylic Acid + Alcohol = Ester + Water
Phosphoester
Phosphoric Acid + Alcohol = Phosphoester + Water
DNA/RNA
Phosphodiester
Phosphoester + Alcohol = Phosphodiester + Water
Phosphotriester
Phosphodiester + Alcohol = Phosphotriester + Water
PROTEINS
Amide
Carboxylic Acid + Amine = Amide + Water
CARBO-HYDRATES
Hemiacetal + Alcohol = Acetal + Water
R'—OH
R'—NH$_2$
HO—R$_3$
H$_2$O

Figure: Dehydration Synthesis Reactions Involved in Macromolecule Formation. The major organic reactions required for the biosynthesis of lipids, nucleic acids (DNA/RNA), proteins, and carbohydrates.

The abstractions where chemists compute these problems

By chemical computer we mean a formal structure that can compute and solve the above questions. This computer has at least three components (and their interactions):

Level	**Represent**
Reaction Network	Substances connected by reactions
Molecular structure and reactions	Augmented reaction network
Similarity level	Substance and reaction similarity

Arquitecture of a chemical computer

Each component or level is a formal (categorical) structure with rich semantics to compute the problems described before (and others):

Level	Category
Reaction Network	`Petri`
Molecular structure and reactions	`Mol`, `Rxn`
Similarity level	Topological spaces

The capabilities of the computer are extended by enriching the structure of each of its components. For instance, reachability and connectivity, at the network level, can be investigated via closure operators defined on *Petri*:

Arquitecture of a chemical computer: Building the Network level

Necesitamos un modelo categórico de la red que permita codificar y computar:

- Synthetic routes: Paths in the network.
- Reachability: Compute closure of *subgraphs*.
- Reachability at the molecular level: The choice of network model ought to interact well with the *augmented network*.

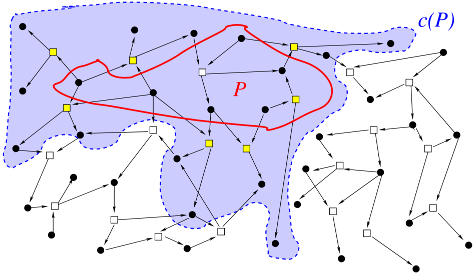


Figure: Bärbel M. R. Stadler and Peter F. Stadler. *Reachability, Connectivity, and Proximity in Chemical Spaces.* MATCH Communications in Mathematical and in Computer Chemistry (2018)

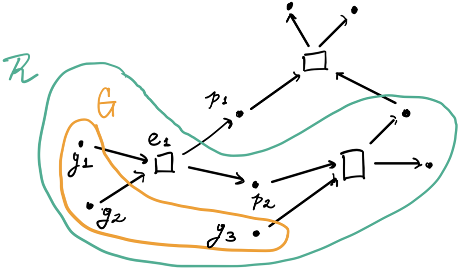
Set theoretically, reachability can be defined on Petri nets

Relative closure functions $C_R^K(G)$ and $R[G]$

- Define a reference set $R \subseteq X$ of compounds (interesting/possible in my lab, etc.).
- Then, define the closure relative to R, denoted as $C_R(-)$ on 2^X, defined by

$$G \mapsto C_R(G) = \bigcup \{V : (U, V) \in E, U \subseteq G \ , V \subseteq R\}$$

- $C_R(G)$ is interpreted as the set of compounds that can be produced from G by the help of a single reaction within R.



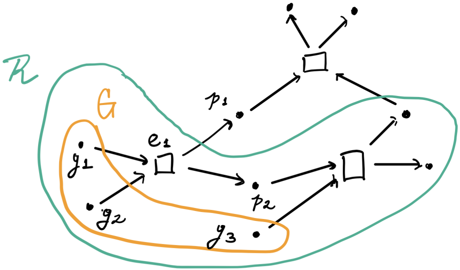
Set theoretically, reachability can be defined on Petri nets

Relative closure functions $C_R^K(G)$ and $R[G]$

- Define a reference set $R \subseteq X$ of compounds (interesting/possible in my lab, etc.).
- Then, define the closure relative to R, denoted as $C_R(-)$ on 2^X, defined by

$$G \mapsto C_R(G) = \bigcup\{V : (U, V) \in E, U \subseteq G\ , V \subseteq R\}$$

- $C_R(G)$ is interpreted as the set of compounds that can be produced from G by the help of a single reaction within R.



Set theoretically, reachability can be defined on Petri nets

$R[G]$: the set of all compounds eventually reachable from G within R

- We can iterate relative to R:

  $$C_R^K(P) = C_R(P \cup C_R(P) \cup C_R^2 \cup ... \cup C_R^{k-1}(P))$$

  for $k \geq 2$.

- $P' \subseteq P \subseteq R$ implies $C_R(P') \subseteq C_R(P)$, so, $C_R^{(j)}(P) \subseteq C_R^K(P)$ for $j \leq K$. Recursively, they obtain the following expression

  $$C_R^{(K)}(P) = C_R(P \cup C_R^{(K-1)}(P))$$

- Let's define $R[G] = C_R^{(\infty)}$.

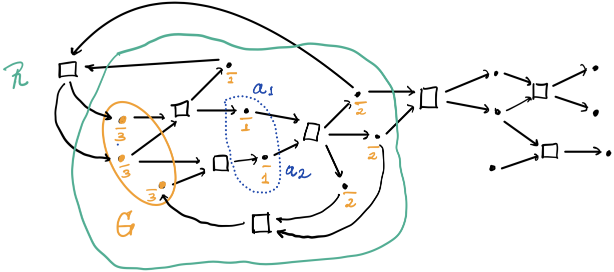
Set theoretically, reachability can be defined on Petri nets

Generalized Reach $\succ$ and Separation Relations $\mid$ and their relation

- $G \succ R$ over 2^X is expressed as follows:

$$G \succ R \leftrightarrow (R[G] = R \vee R = \emptyset)$$

G is related to R if from G all compounds in R can be synthesized (including the initial compounds in G), and we say that G reaches R.



Set theoretically, reachability can be defined on Petri nets

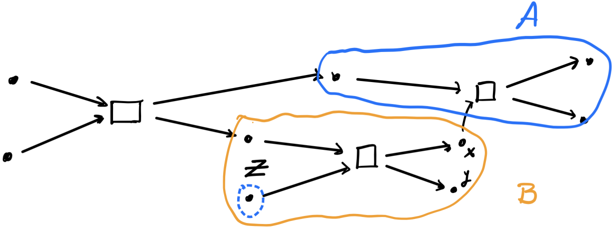
Generalized Reach $\succ$ and Separation Relations $|$ and their relation

- $A|B$ (A is separable of B) on 2^X if

$$(A \cup B)[B] \cap A = \emptyset$$

.

- Consider the following example



Interpretation: If $A|B$, then from reactions involving substances from B, it is not possible to produce any substance from A.

Set theoretically, reachability can be defined on Petri nets

Testing separation using rechability:

$$(A, B) \between (P, Q) \iff P \cup Q \subseteq A \cup B, P \subseteq B, Q \cap A \neq \emptyset$$

We can proof that:

Lemma: If $(A, B) \between (P, Q)$ then:

1. $P \succ Q$ then $A \nmid B$, where $A \nmid B \iff (A \cup B)[B] \neq \emptyset$
2. $A|B$ then $P \nsucc\ Q$, where $P \nsucc\ Q \iff Q[P] \neq Q$

Set theoretically, reachability can be defined on Petri nets

Galois connection induced by relation $\between$

$$\begin{tikzcd} 2^{2^X \times 2^X} \arrow[r, bend right, "\beta"] & 2^{2^X \times 2^X} \arrow[l, bend right, "\alpha"'] \end{tikzcd}$$

Where β is defined as
$C \mapsto \beta(C) = \{(P, Q) \in 2^X \times 2^X : (A, B) \between (P, Q), \forall (A, B) \in C\}$, for all $C \in 2^{2^X \times 2^X}$.

Set theoretically, reachability can be defined on Petri nets

Galois connection induced by relation $\between$

$$\begin{tikzcd} 2^{2^X \times 2^X} \arrow[r, bend right, "\beta"] & 2^{2^X \times 2^X} \arrow[l, bend right, "\alpha"'] \end{tikzcd}$$

Where β is defined as
$C \mapsto \beta(C) = \{(P, Q) \in 2^X \times 2^X : (A, B) \between (P, Q), \forall (A, B) \in C\}$, for all $C \in 2^{2^X \times 2^X}$.

Interpretation: Let $C \in 2^{2^X \times 2^X}$ such that $A|B$ for all $(A, B) \in C$. If we consider any pair $(A, B) \in C$ by lemma 4.2, we conclude that $P \not\succ Q$ for all $(P, Q) \in \beta(C)$. This implies that all pairs $(P, Q) \in \beta(C)$ are non-reachable, hence β maps separable things to non-reachable ones.

Set theoretically, reachability can be defined on Petri nets

Galois connection induced by relation $\between$

$$\begin{tikzcd}
2^{2^X \times 2^X} \arrow[r, bend right, "\beta"] & 2^{2^X \times 2^X} \arrow[l, bend right, "\alpha"']
\end{tikzcd}$$

Where α is defined as
$D \mapsto \alpha(D) = \{(A, B) \in 2^X \times 2^X : (A, B) \between (P, Q), \forall (P, Q) \in D\}$, for all $D \in 2^{2^X \times 2^X}$.

Set theoretically, reachability can be defined on Petri nets

Galois connection induced by relation $\between$

$$\begin{tikzcd}
2^{2^X \times 2^X} \arrow[r, bend right, "\beta"'] & 2^{2^X \times 2^X} \arrow[l, bend right, "\alpha"']
\end{tikzcd}$$

Where α is defined as
$D \mapsto \alpha(D) = \{(A, B) \in 2^X \times 2^X : (A, B) \between (P, Q), \forall (P, Q) \in D\}$, for all $D \in 2^{2^X \times 2^X}$.

Interpretation: Let $D \in 2^{2^X \times 2^X}$ such that $P \succ Q$ for all $(P, Q) \in D$. If we consider any pair $(P, Q) \in D$ by lemma 4.1, we conclude that $A \not| B$ for all $(A, B) \in \alpha(D)$. This implies that all pairs $(A, B) \in \alpha(D)$ are non-separable, hence α maps reachable things to non-separable ones.

Building the level of molecular structure and reactions: formalization of Stadler's intermediate level of abstraction for chemistry

In this model, substances are molecular graphs and reactions are spans that account for molecular changes.
Problems:

- Computationally implemented but not formalized: What are graphs/molecular structures? DPOs in which category?
- Chemists use different representations/abstractions for mol structures: what is the best choice of model?
- Disconnected from network level: set-theoretic formulation is difficult to integrate with network and similarity levels.

Building the level of molecular structure and reactions:

Proposal:

- We formalize Chemical Structure Theory instead as a category $\texttt{ChemStructTh}$ of syntactic models!
- Definition of molecular structures as functors

$$M : \texttt{StructMod} \to \texttt{Set}$$

  on a scheme $\texttt{StructMod} \in \texttt{ChemStructTh}$.
- Let's call the these collection of functors $\texttt{Subst} = \texttt{Set}^{\texttt{StrThMod}}$.
- There are two possible choices of morphisms to get a category:
  - Syntactic relations between substances. Using natural transformations between these functors.
  - Reactivity. rxns modeled as rewritings of molecular structures: $F \leftarrow G \to H \in \texttt{Span}(\texttt{Subs})$

Building the level of molecular structure and reactions:

Why reactions as spans of molecular structures?

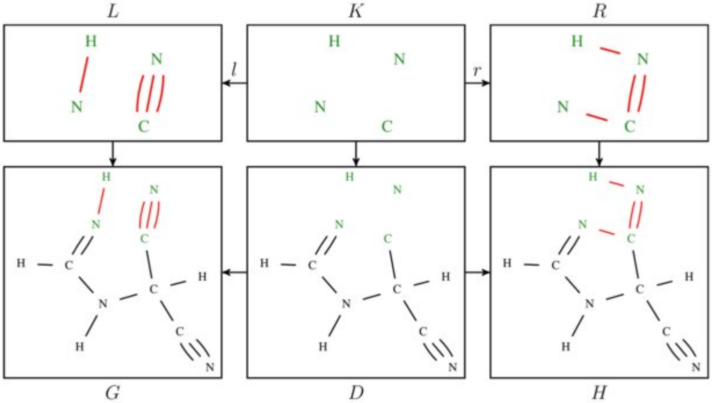
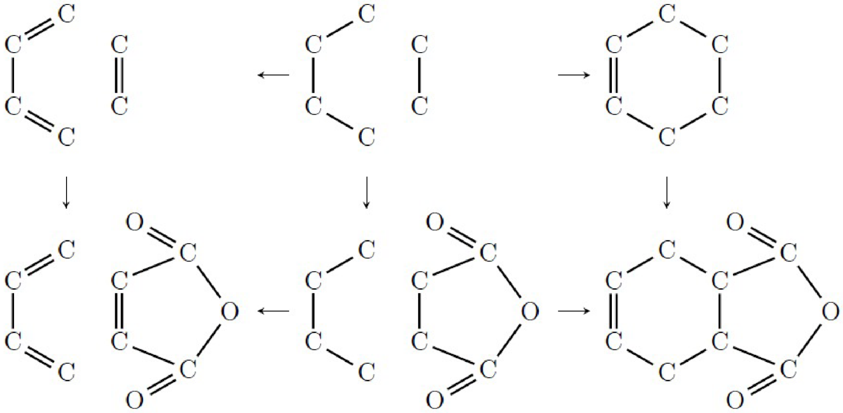


Figure: Illustration of a chemical reaction using the Double Pushout approach. The chemical transformation of complete molecules (i.e., the application of the graph grammar rule as defined in the first row) is represented as the graph derivation $G \Rightarrow H$ in the second row.

Building the level of molecular structure and reactions: a better model!

The previous proposal is not entirely satisfactory:

- It collapses inputs and outputs of reactions into single "molecular structures".



Building the level of molecular structure and reactions: a better model!

The previous proposal is not entirely satisfactory:

- It collapses inputs and outputs of reactions into single "molecular structures".
- We want something with this or more details: