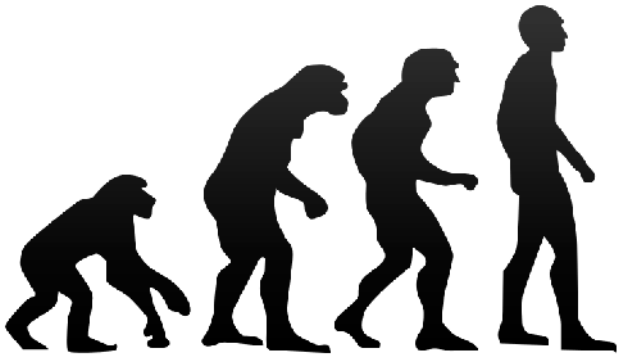
Part 2: Analysing Chemical Spaces

in collaboration with
Jakob L. Andersen, Christoph Flamm, Peter Stadler

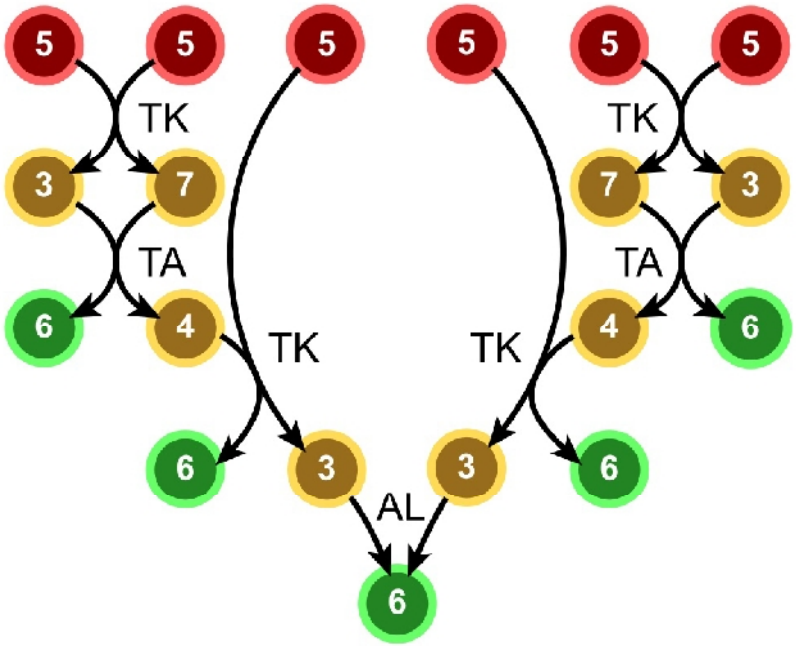
March 19, 2014

Cheminformatics

Bioinformatics



Chemical transformation motif (CTM)



A CTM is a subnetwork with the following properties:

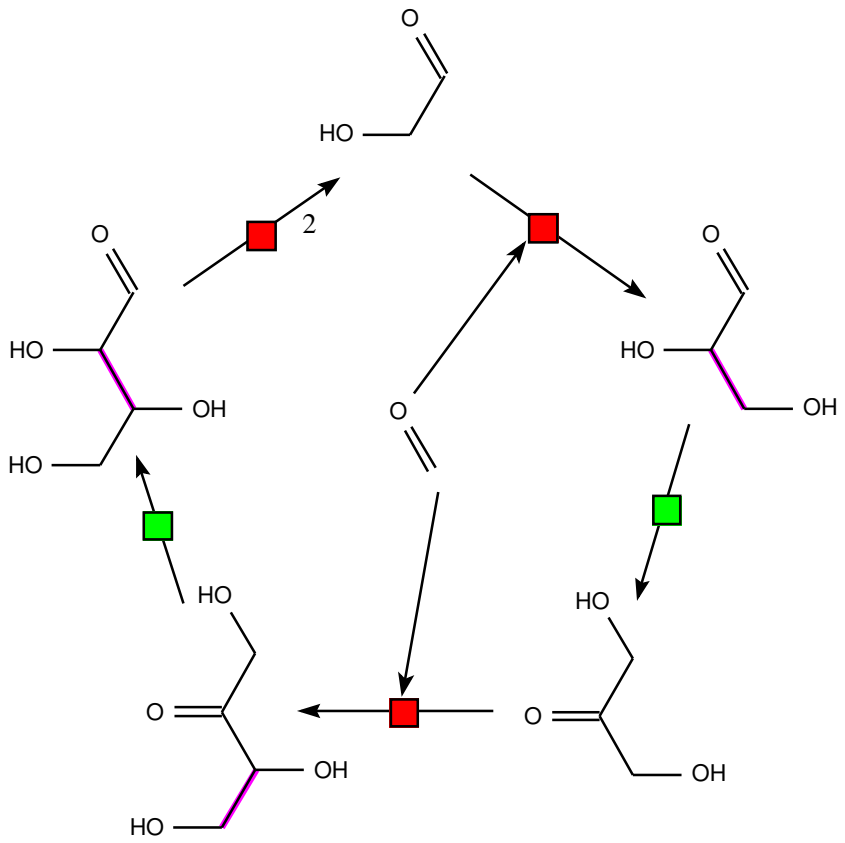
1. Defined boundery to the "outside".
2. Stoichiometrically balanced.
3. Maximize flow between motif boundaries.
4. Must be optimal (e.g. minimal number of reactions)

Figure adapted from Noor, E et al (2011) Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy, *J Mol Cell* **39**:809-820 DOI:10.1016/j.molcel.2010.08.031

Another CTM example: Autocatalysis

$$F + X \longrightarrow 2X + W$$

The formose reaction follows exactly the abstract pattern.



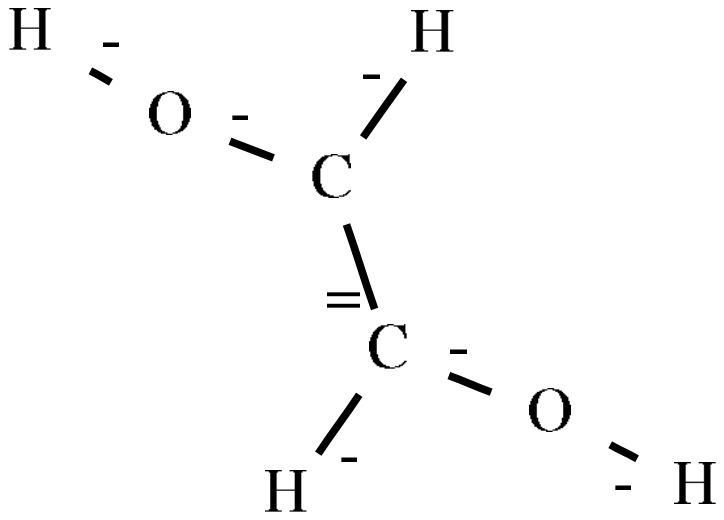
Butlerov AM (1861) Einiges über die chemische Structur der Körper, *Zeitschrift für Chemie* **4**:549-560

Molecule Encoding

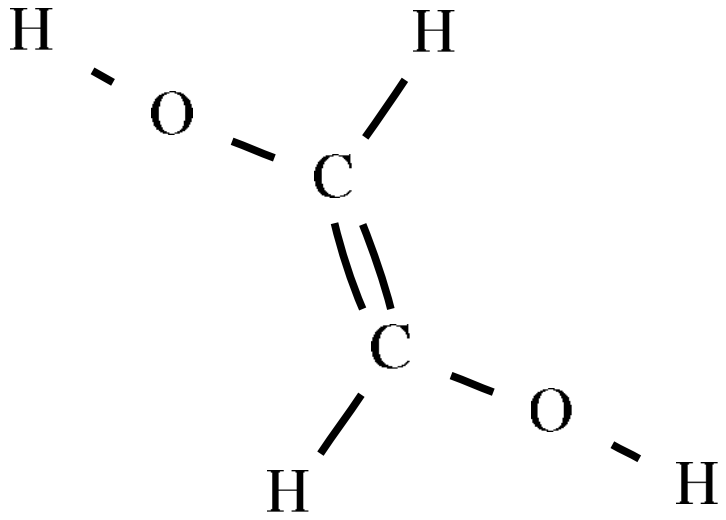
A molecule is an undirected labelled graph.

Vertex label ≡ atom type (e.g., "C" or "O-")

Edge label ≡ bond type (e..g, "-", "=" or "#")



(a) Visualization of encoding



(b) Prettified visualization

(c) Open Babel visualization

Figure: 1,2-ethenediol

Reaction Patterns – Graph Transformation Rules

A reaction pattern is a graph transformation rule, in the Double Pushout Formalism: $p = (L \leftarrow K \rightarrow R)$.

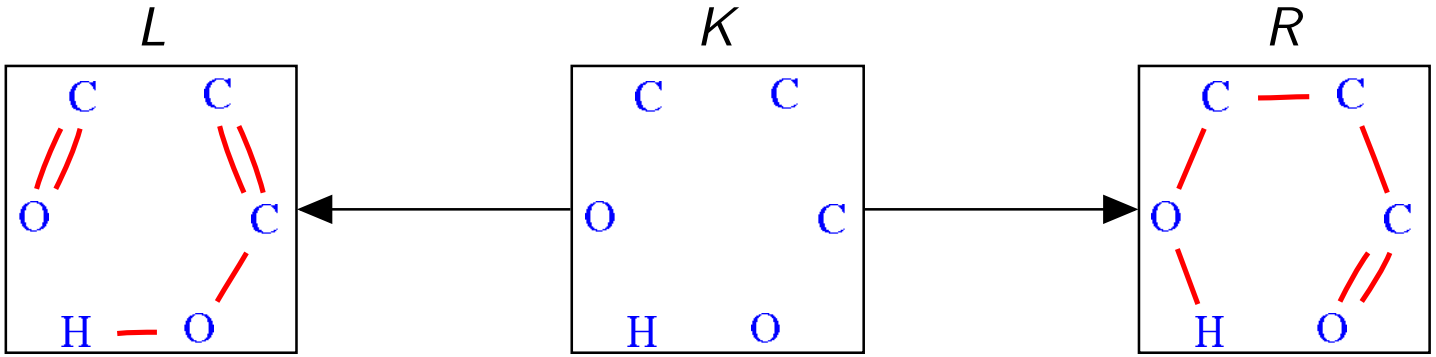
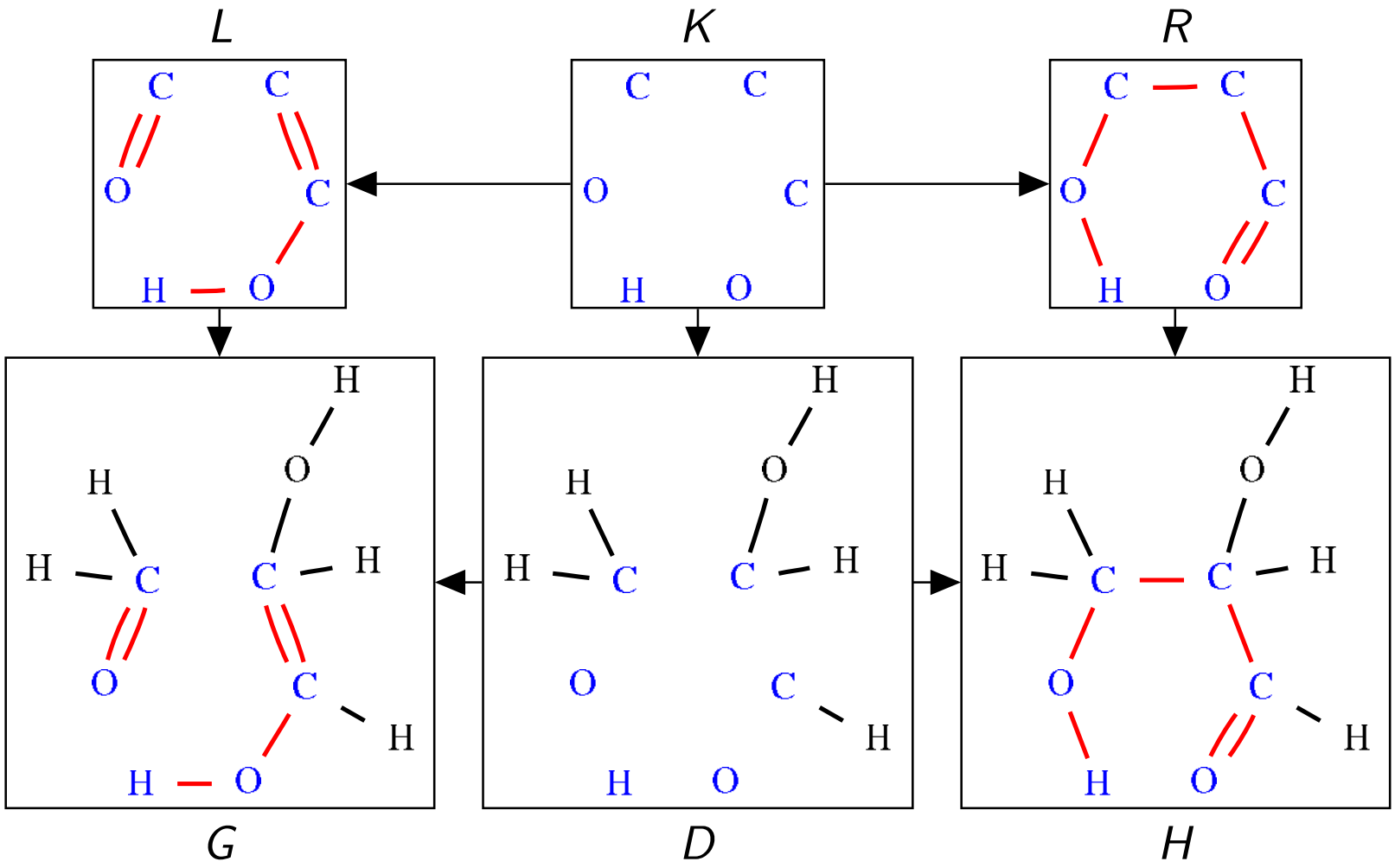


Figure: Transformation rule for aldol addition

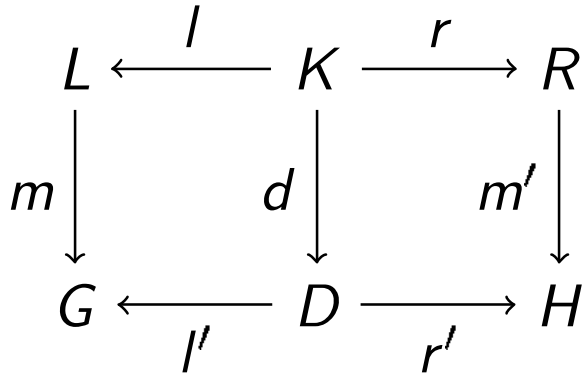
(As for the graphs: the rules are not restricted to chemistry.)

Reactions – Application of Transformation Rules

1,2-ethenediol + formaldehyde $\xrightarrow{\text{aldol addition}}$ glyceraldehyde



Graph Transformation – Double Pushout Approach



Transformation rule: $p = (L, K, R, l, r)$

Specific derivation with p (and m): $G \overset{p,m}{\Longrightarrow} H$

Graph Transformation – Double Pushout Approach

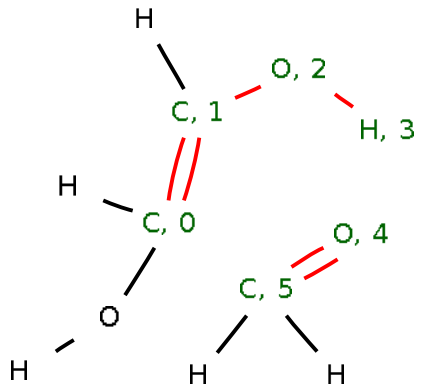
$$\begin{array}{ccccc} L & \xleftarrow{l} & K & \xrightarrow{r} & R \\ {\scriptstyle m}\downarrow & & {\scriptstyle d}\downarrow & & {\scriptstyle m'}\downarrow \\ G & \xleftarrow[l']{} & D & \xrightarrow[r']{} & H \end{array}$$

Transformation rule: $p = (L, K, R, l, r)$

Specific derivation with p (and m): $G \overset{p,m}{\Longrightarrow} H$

If G has 2 connected components:

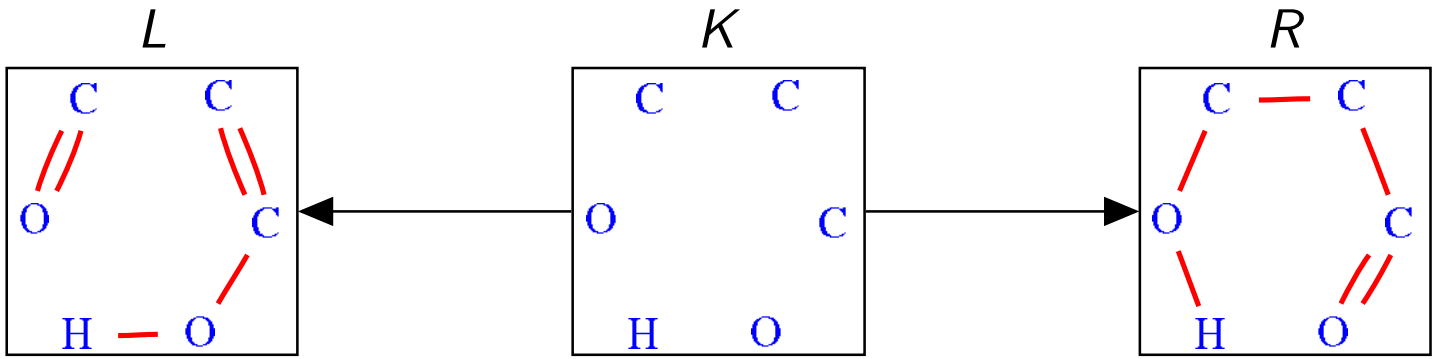
$$G = \{g_1, g_2\} \qquad \{g_1, g_2\} \overset{p,m}{\Longrightarrow} H$$



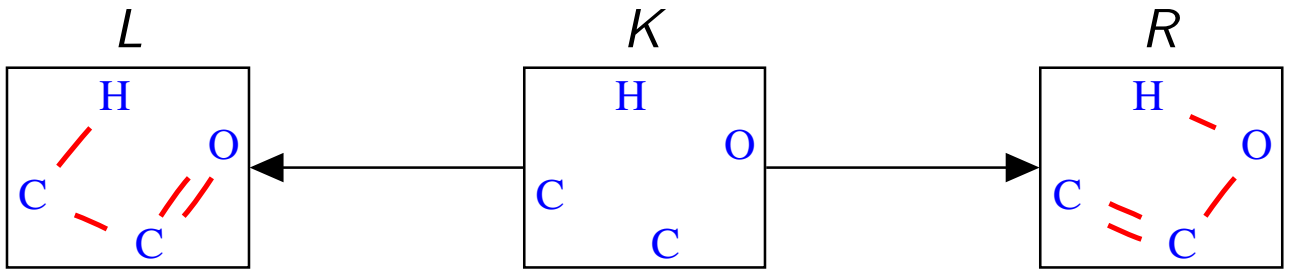
G

The Formose Chemistry

4 reaction patterns:

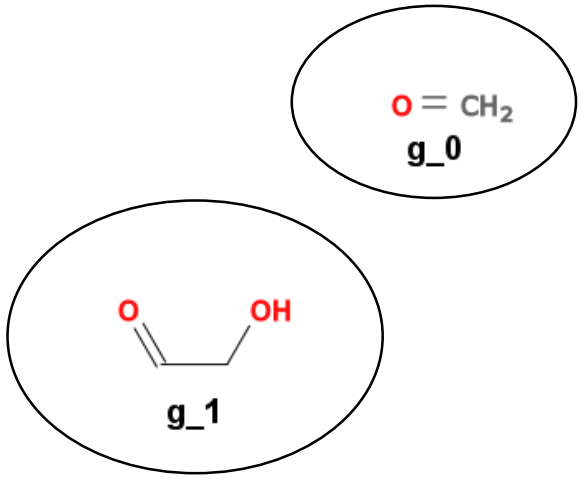


(a) Aldol addition (r_2). (Reverse aldol addition (r_3))



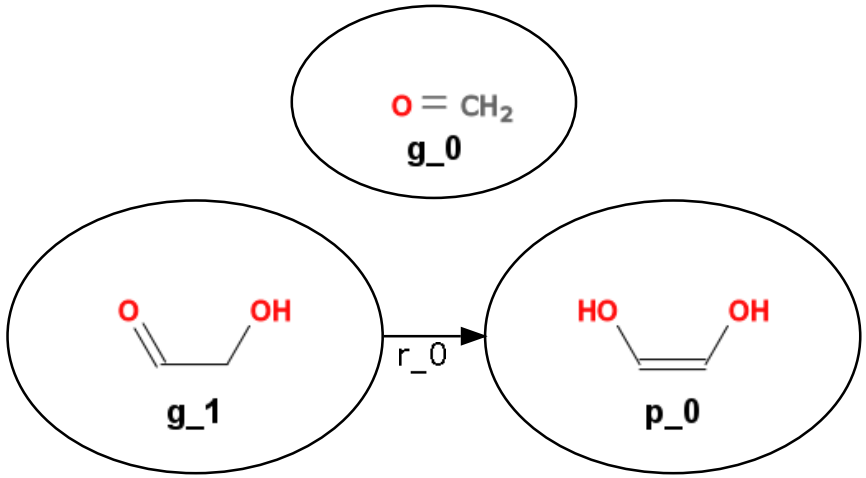
(b) Keto-to-enol (r_0). (Enol-to-keto (r_1))

Reaction Network for Formose



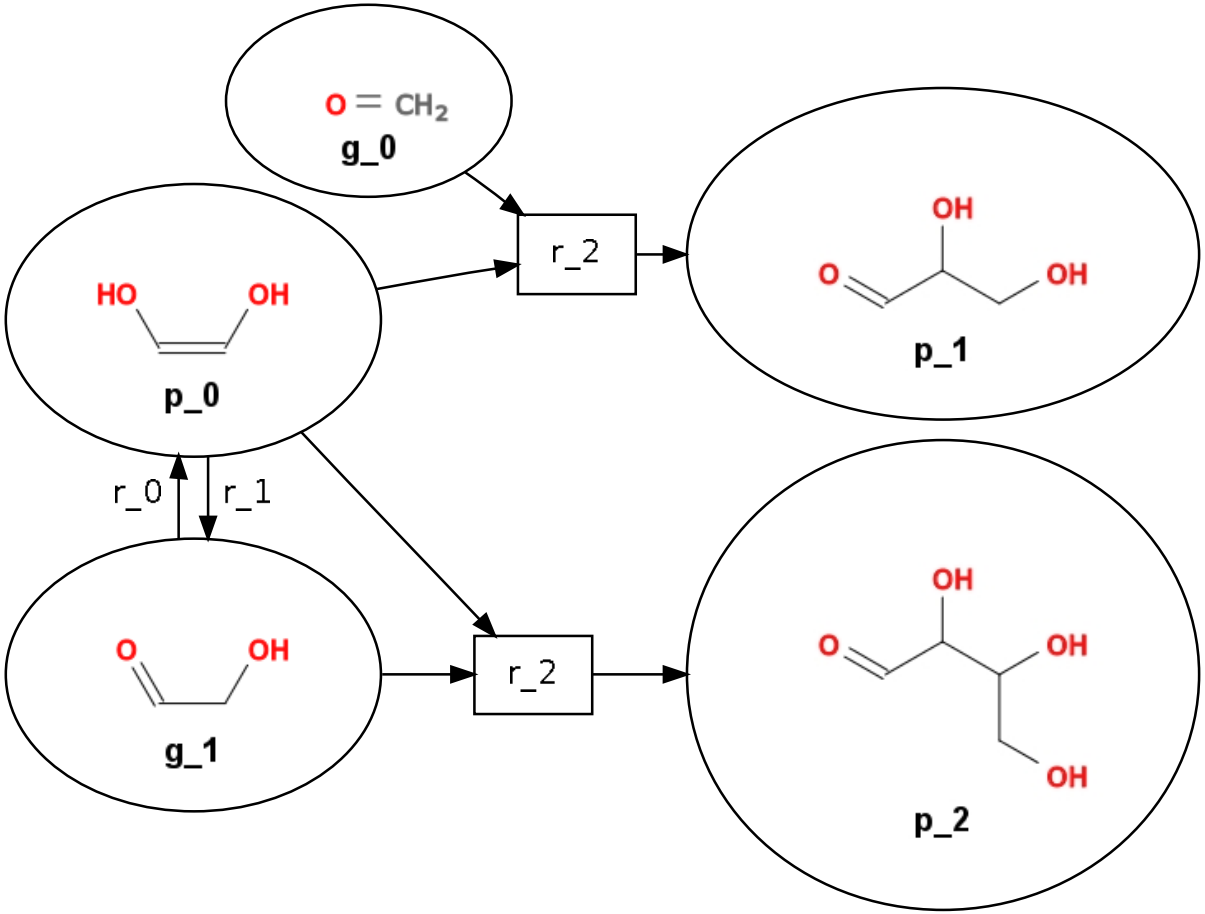
Initial State

Reaction Network for Formose



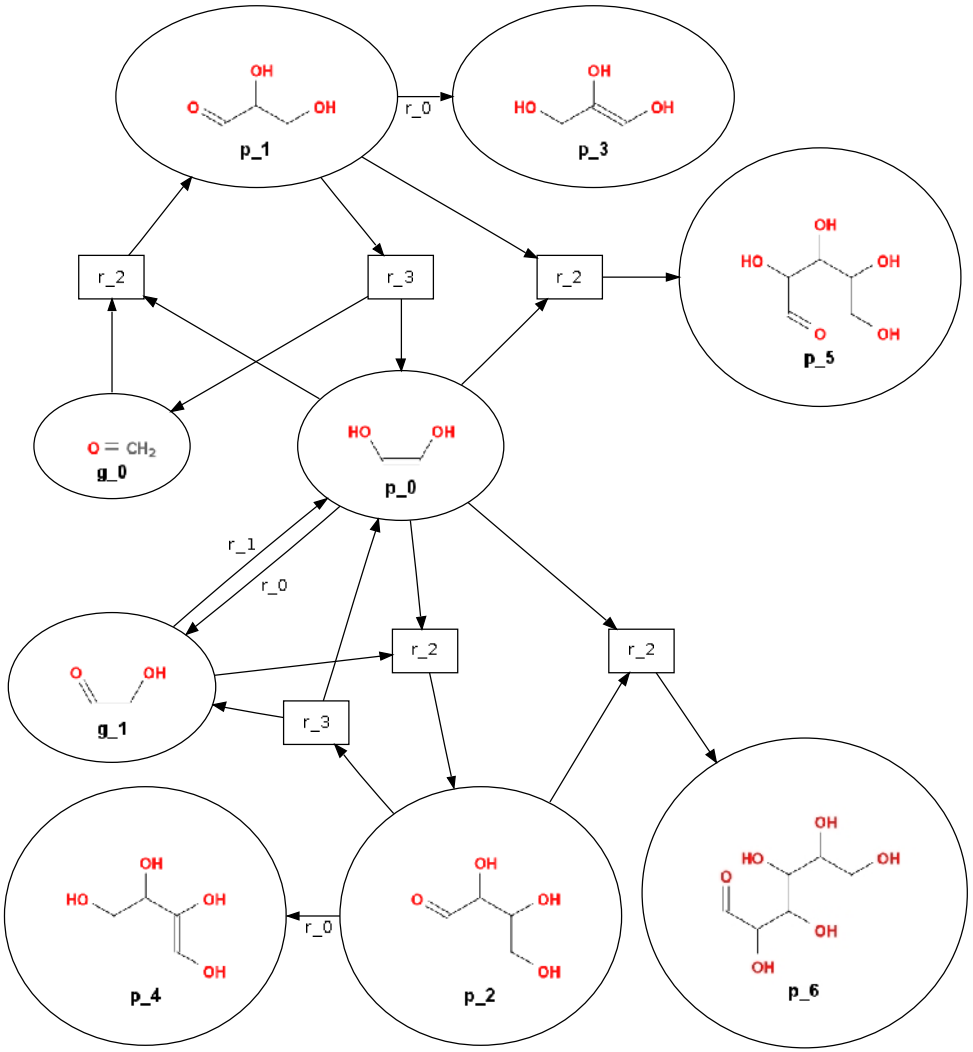
Generation 1

Reaction Network for Formose



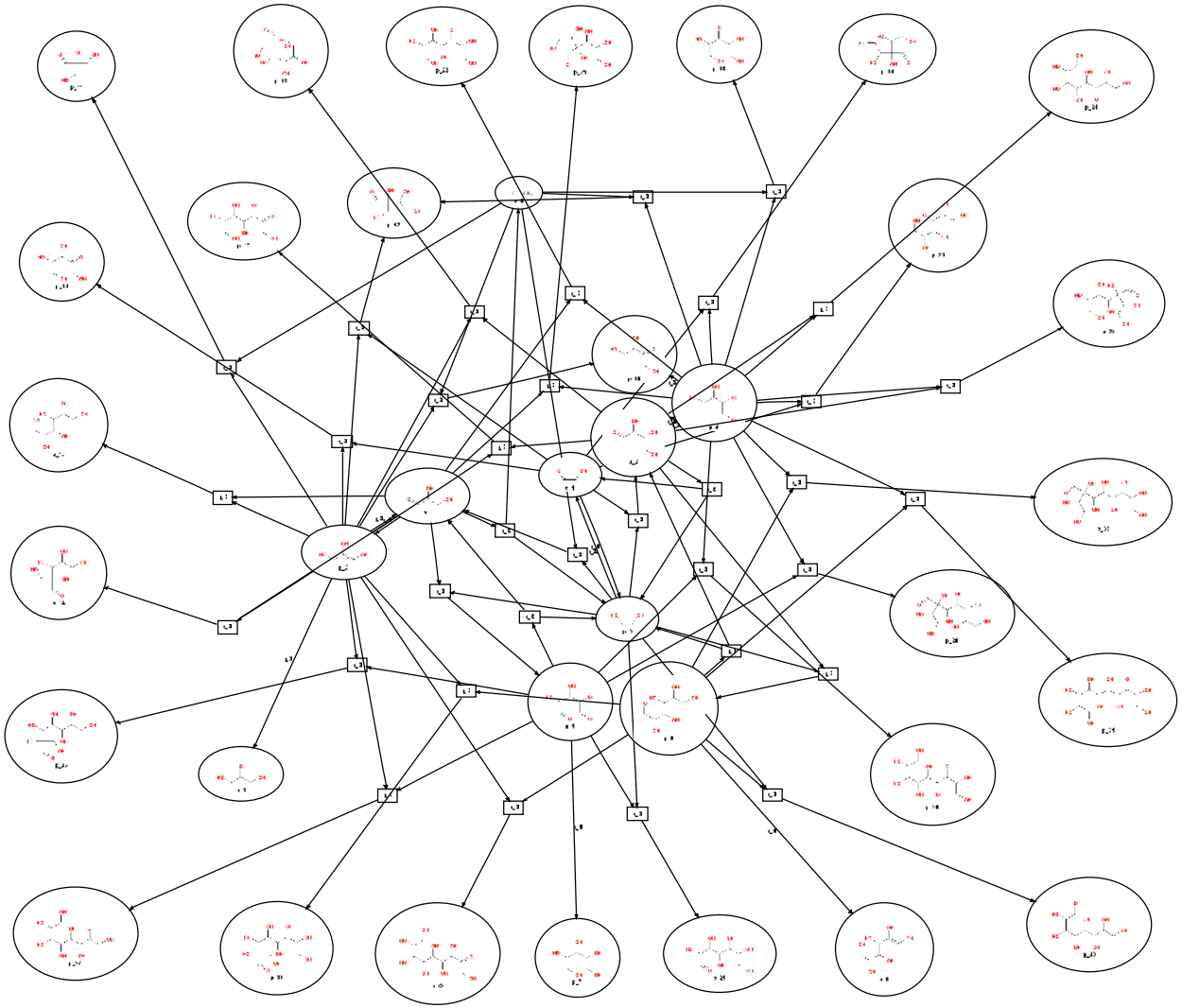
Generation 2

Reaction Network for Formose



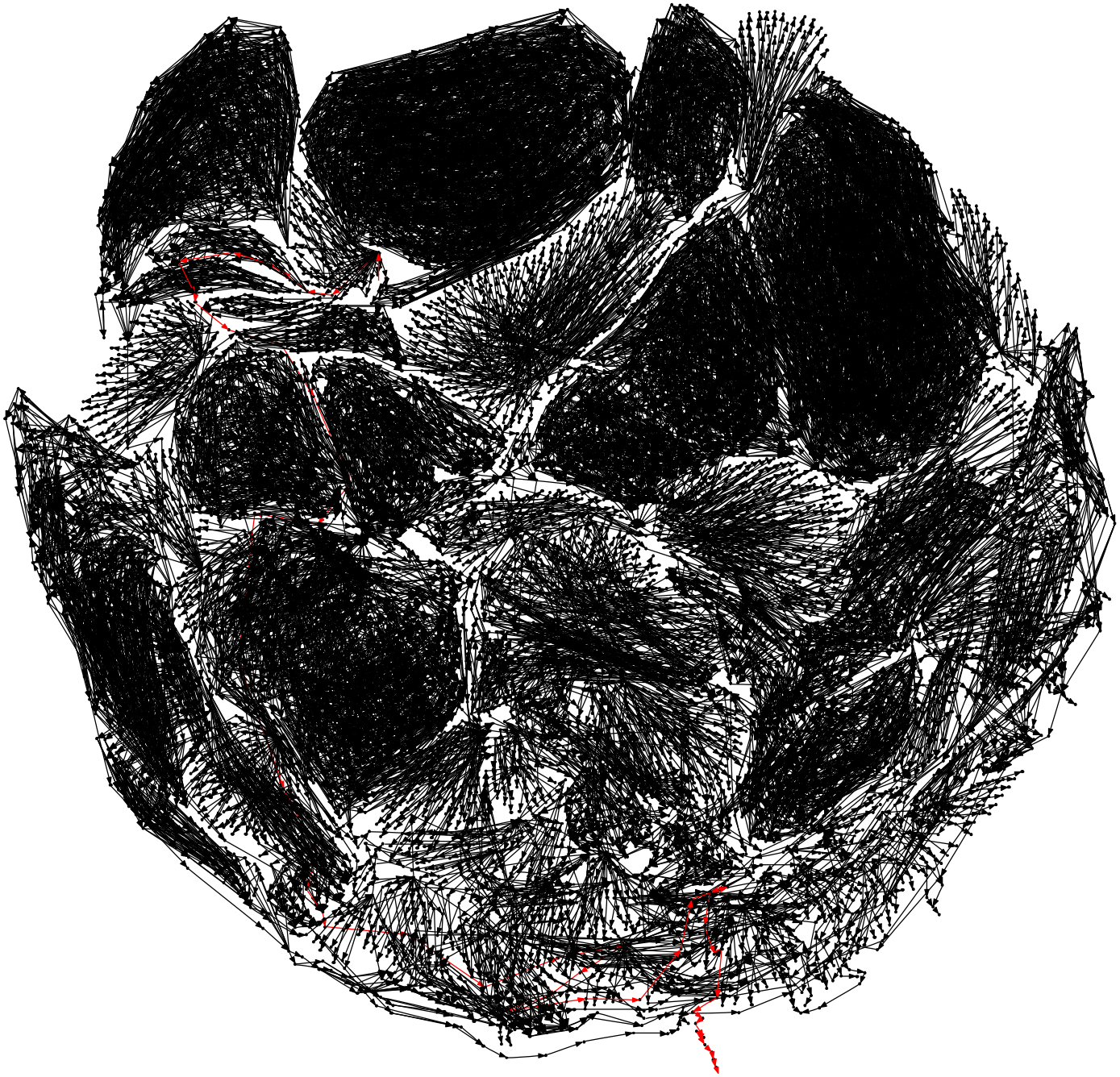
Generation 3

Reaction Network for Formose



Generation 4

Another Example of a Search Space (Catalan)



Two Important Questions Concerning CTMs

1. How to define/find/enumerate CTMs in arbitrary reaction networks?
   - Generic reaction network generator (graph grammar).
   - Reformulate question as Network Integer Flow Problem.
   - Solve with Integer Linear Programming (ILP)

2. Can a CTM be realized in different chemistry?
   - Inverse problem leading from an abstract reaction mechanism to a molecular implementation of the CTM.
   - Use Satisfiability Modulo Theories (SMT) to attack this problem.

Fagerberg, Flamm, Merkle, Peters: *Exploring Chemistry Using SMT*. CP 2012: 900-915

Flows in Derivation Graphs

Idea: use (a generalization of) network flows as a model for chemical pathways, and find interesting flows

Problem: derivation graphs are hypergraphs

Solution: use integer linear programming

Examples of chemically interesting questions:

- Product Optimization: Given k ribulose-5-phosphate maximize production of fructose-6-phosphate.
  How can this maximum be realized?
- Autocatalysis: $(\exists v \in V : 0 < \mathit{inFlow}(v) < \mathit{outFlow}(v))$
  Does this exist in the chemical space of sugar chemistry?
  What is the mechanism?

Both problems are NP-complete

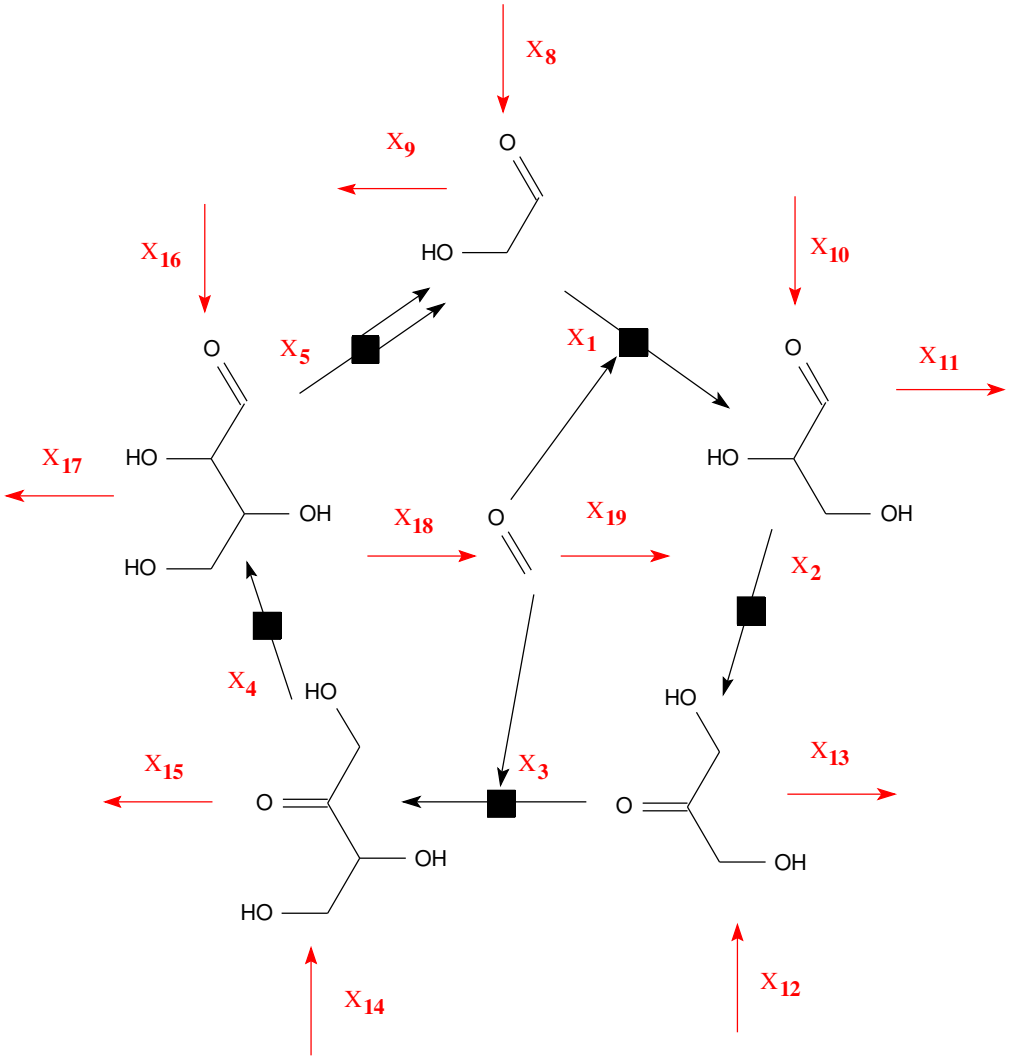
Andersen, Flamm, Merkle, Stadler (2012): Maximizing output and recognizing autocatalysis in chemical reaction networks is NP-complete. *Journal of Systems Chemistry*, **3**:1

Network Flow Problem as ILP Formulation

$$\text{minimize} \quad 0 \quad \text{s.t}:$$

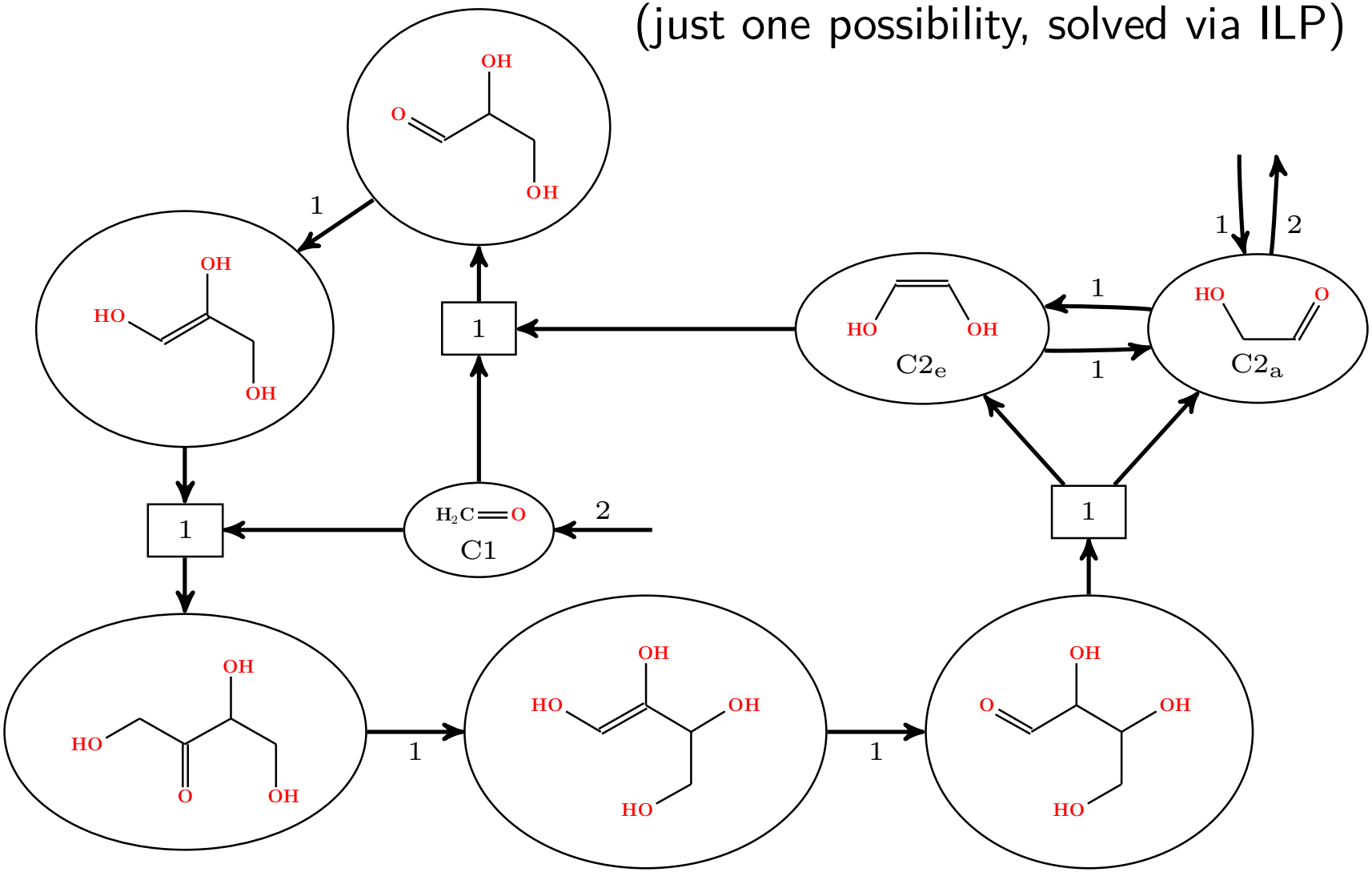
$$x_8 + 2x_5 - x_9 - x_1 = 0$$
$$x_{10} + x_1 - x_{11} - x_2 = 0$$
$$x_{12} + x_2 - x_{13} - x_3 = 0$$
$$x_{14} + x_3 - x_{15} - x_4 = 0$$
$$x_{16} + x_4 - x_{17} - x_5 = 0$$
$$x_{18} - x_{19} - x_1 - x_3 = 0$$

$$\forall i \in [1; 19] \qquad x_i \geq 0$$
$$x_{18} > 0$$
$$x_8 = 1$$
$$x_9 = 2$$



The Formose Reaction — $\mathrm{C2_a} + 2\,\mathrm{C1} \longrightarrow 2\,\mathrm{C2_a}$

(just one possibility, solved via ILP)



The Formose Reaction

Number of *different* autocatalytic cycles

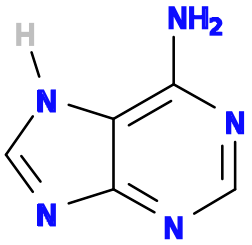
Network: *all* molecules with at most 9 carbon atoms

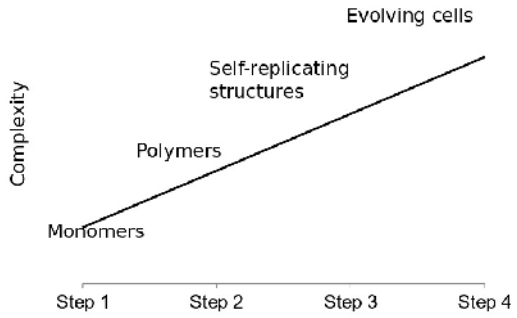
	Maximum #C					
Reactions used	4	5	6	7	8	9	Sum
6	0	0	1	1	1	2	5
7	0	0	0	0	0	2	2
8	1	5	7	17	37	68	135
9	0	0	12	12	37	69	130
10	0	12	50	274	849	—	$\geq$ 1185
11	0	5	41	190	738	—	$\geq$ 974
							$\geq$ 2431

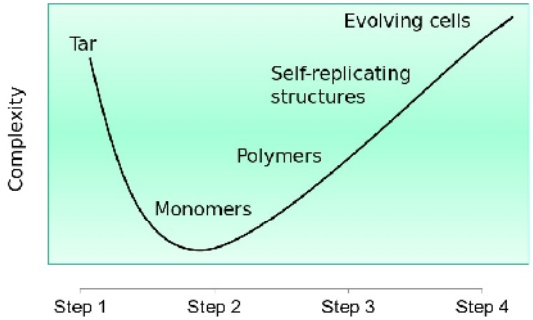
Other systems analysed:
TCA cycle, PPP, Borate inhibited Formose Reaction, non-oxidative glycolysis pathway, ...

The HCN Chemistry

- Initial Molecules: HCN, amonium (and water)
- Subspaces of many different polymers
- Found outside Earth, e.g., on Titan
- Possibly an early Earth chemistry
- Synthesis of adenine





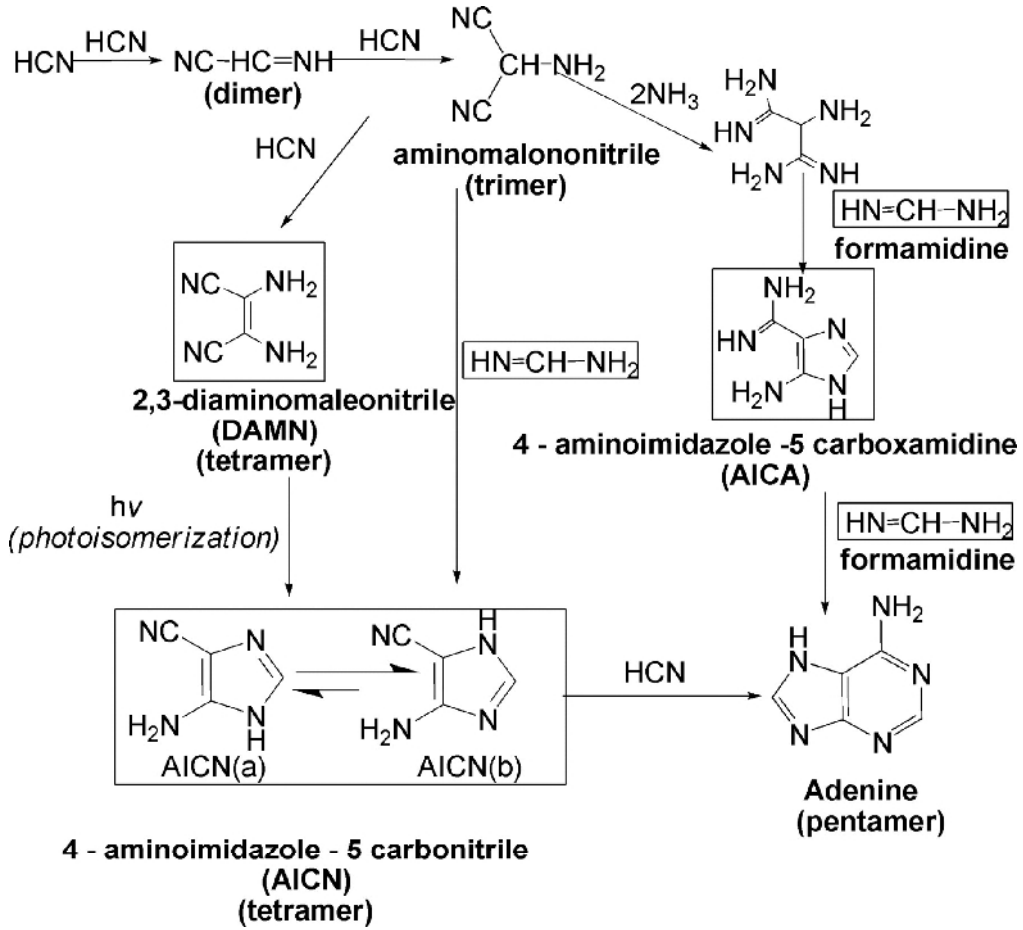


The HCN Chemistry – Graph Grammar Model

Input Molecules: HCN, amonium (and water)

Input Rules: 23 rules

Adenine synthesis:



Roy et al, PNAS 2007

Lab Experiments

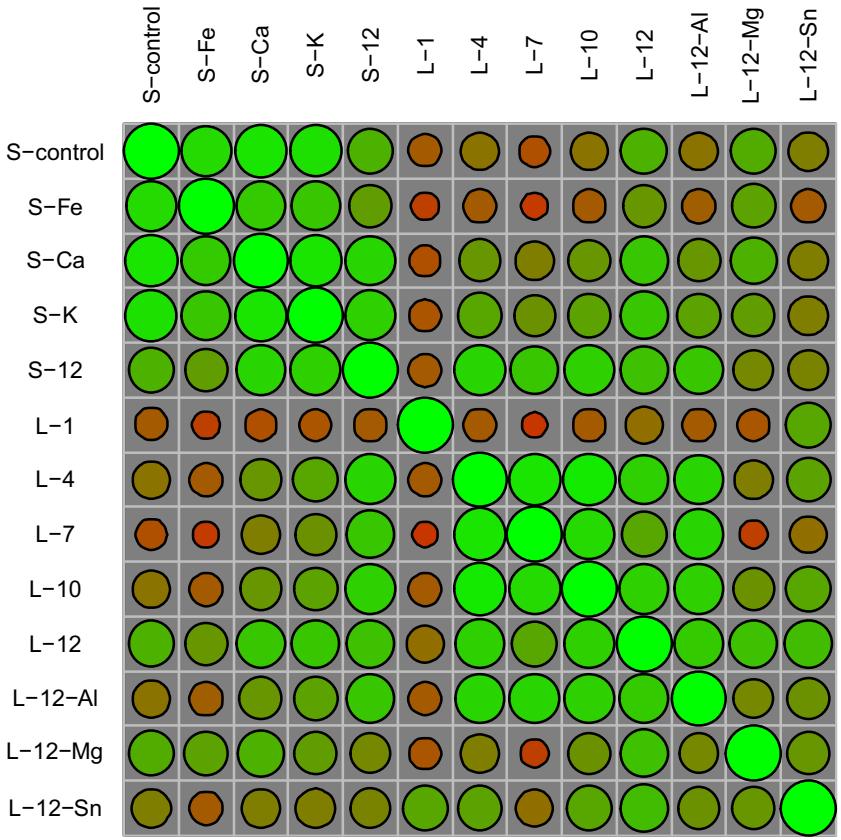
1. Polymerization of HCN
2. Hydrolysis under different conditions
3. Analysis by liquid chromatography and mass spectrometry
   Scans both with and without fragmentation

Variations:

- Long term hydrolysis, difference in pH and added salts
- Short term hydrolysis, difference in added salts

13 different conditions in total

Mass Spectrometry



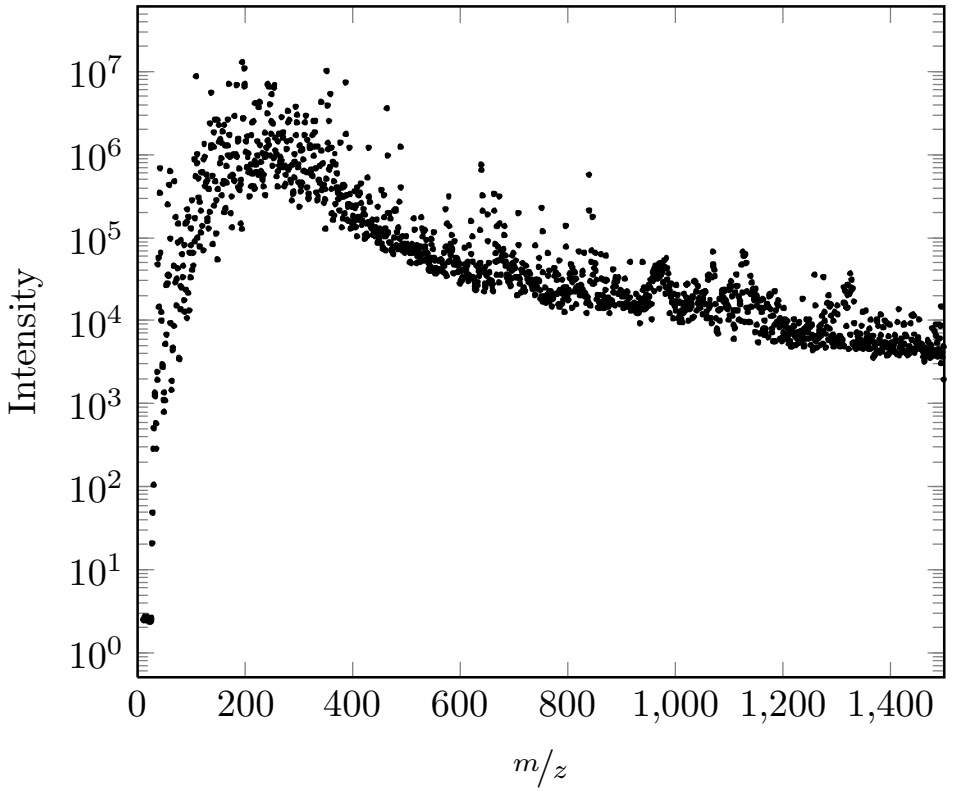


Figure: Left: Correlation Matrix under different reaction conditions; Right: maximal observed intensity for each m/z-value observed in any of the scans.

General Exploration

Let $r = \{r_1, r_2, \ldots, r_{23}\}$ be the set of transformation rules.

Exploration Strategy: breadth-first expansion?

$$\texttt{addActive}[\{\mathrm{HCN}, \mathrm{amonium}, \mathrm{water}\}] \rightarrow \texttt{repeat}(r)$$

Not feasible: 3 steps gives 996 molecules and 2050 reactions

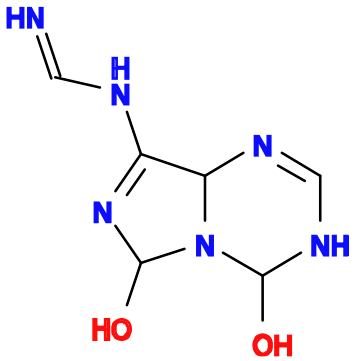
Pruning:

- Limit growth of molecules by weight (get rid of size doubling)
- Boltzmann factors, among isomers
- Bias expansion towards high-intensity molecules

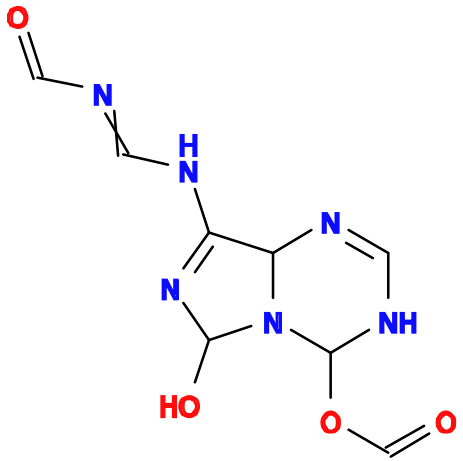
Hydrolysis Strategy

$$
\begin{aligned}
&\text{polymerization} \rightarrow \texttt{filterSubset}[\text{false}] \\
&\rightarrow \texttt{addActive}[\{\text{HCN}, \text{amonium}, \text{water}\}] \\
&\rightarrow \texttt{repeat}(\\
&\qquad \texttt{eductPredicate}[P_{weight}](r) \\
&\qquad \rightarrow \texttt{filterUniverse}[\text{isomerBoltzmann} > p_{\text{isomer}}] \\
&\qquad \rightarrow \texttt{sortUniverse}[\lambda\, g_1, g_2 \rightarrow \text{intensity}(g_1) < \text{intensity}(g_2)] \\
&\qquad \rightarrow \texttt{takeUniverse}[100] \\
&\qquad \rightarrow \texttt{addPassive}[\{\text{HCN}, \text{amonium}, \text{water}, \ldots\}] \\
&\quad)
\end{aligned}
$$

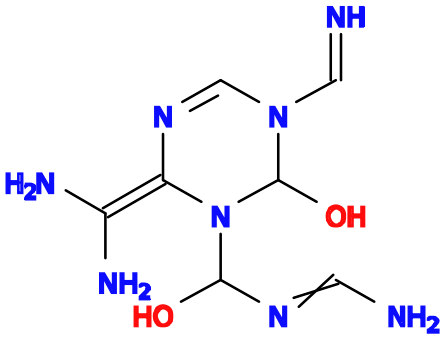
Surviving Molecules (a Small Sample)



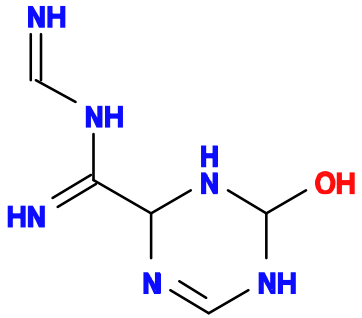
Weight: 198, intensity: $1.02 \cdot 10^7$



Weight: 254, intensity: $6.49 \cdot 10^6$



Weight: 242, intensity: $5.71 \cdot 10^6$



Weight: 170, intensity: $6.09 \cdot 10^6$

(Just some more) Adenine Pathways

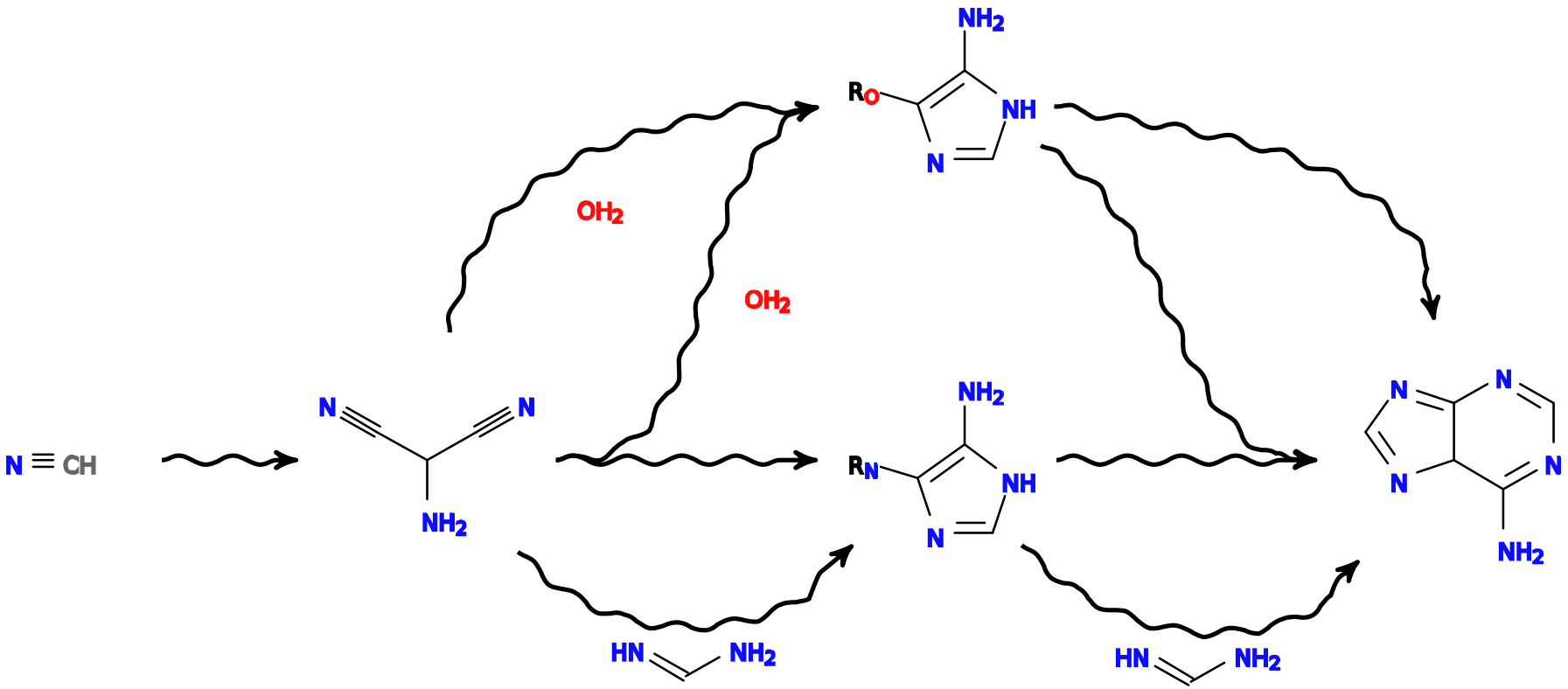
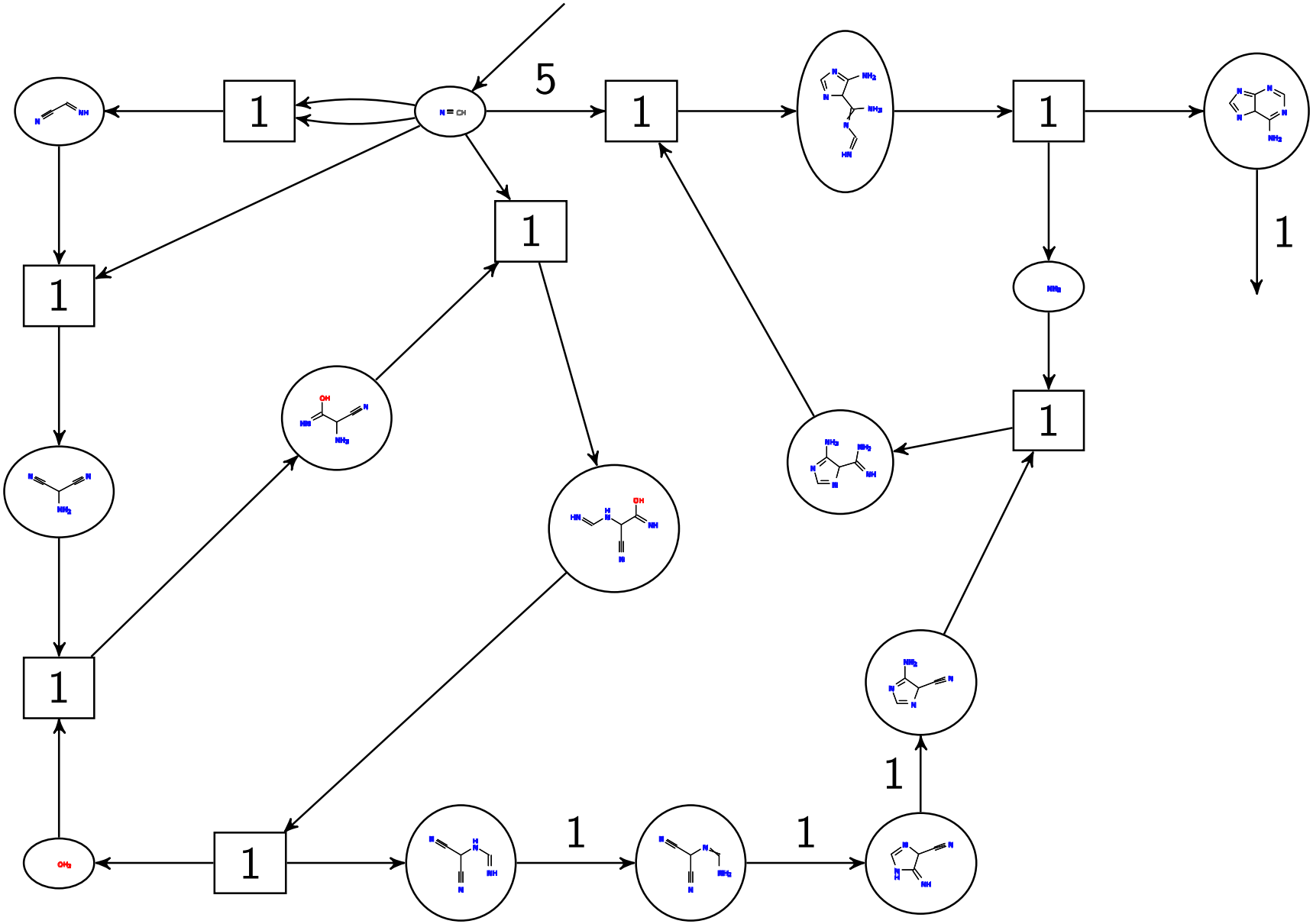


Figure: Schematic of the merge of different pathways to create adenine.

Maximizing the Sum of MS Data Intensities



WIRED.CO.UK

FOLLOW

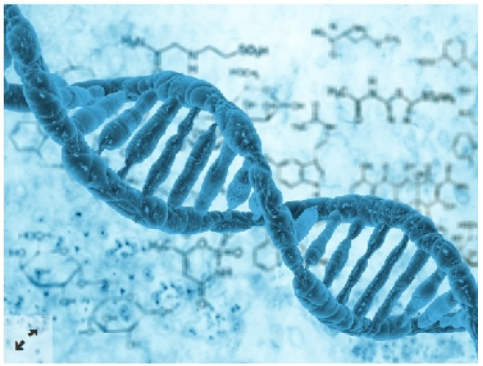
NEWS Topics / SCIENCE DNA LIFE CHEMISTRY

Unravelling the origins of life with mathematical chemistry

SCIENCE / 06 SEPTEMBER 13 / by KADHIM SHUBBER

How life began is one of the most compelling questions humanity has ever asked. Atoms and molecules, driven by nothing more than unthinking chemical processes, somehow became the complex reproductive organisms that we see roaming the Earth today -- somehow, they became us.

Those tiny baby steps at the start of life, when some unknown molecule somehow became self-replicating, for example, hold the key to understanding how life began and how likely it is to have sprouted throughout the Universe.



Andersen, Flamm, Hanczyc, Merkle*, Stadler (2013) : Navigating the Chemical Space of HCN Polymerization and Hydrolysis: Guiding Graph Grammars by Mass Spectrometry Data, *Entropy* **15**:4066-4083

Summary

- Generative graph grammars are a powerful approach to study chemical transformation spaces
- Systematic approach to explore Chemical Spaces
- Recognizing Chemical Transformation Motifs in arbitrary reaction networks