An open source computational workflow for the discovery of autocatalytic networks in abiotic reactions: ESI

Aayush Arya, Jessica Ray, Siddhant Sharma, Romulo Cruz, Alejandro Lozano, Harrison Smith, Jakob Lykke Andersen, Huan Chen, Markus Meringer, and Henderson James Cleaves II

Contents

List of Tables **2**

List of Figures **2**

1 SI File Descriptions **6**

2 Selected SI Figures **7**

3 Detailed Methods **19**
3.1 Reaction Network Generation 19
3.2 Rule Selection 20
3.3 Post-Generation Filtering 20
3.4 Minimizing the Combinatorics for Specific Rules 21
3.5 Treatment of Tautomers 21
3.6 Comparison with literature analyses 22
3.7 Experimental Degradation of Glucose 23
3.7.1 Sample Preparation 23
3.7.2 General Precautions 23
3.7.3 FT-ICR MS Instrument Specifications 23
3.7.4 Ionization 23
3.7.5 Mass Calibration 24
3.7.6 Data Cleansing 24
3.8 Cheminformatic Descriptor Computation 24
3.9 Thermochemical Calculations 35
3.10 Detection of Autocatalytic Motifs 36
3.11 Network Data Visualization 39

4 SI Tables **41**

Bibliography **46**

List of Tables

SI1 The number of products and reactions produced per generation as a function of computation time. *The 'number of edges' presented here is equal to the number that should appear in the species representation of the network. 41

SI2 A list of all reaction rules encoded in our workflow. 42

SI3 Counts of reaction rules implemented per generation in the generated network. . . . 45

List of Figures

SI1 High resolution negative-ESI mode FT-ICR-MS spectra of carbonaceous meteorite organics and laboratory abiotic synthesis simulants. **A:** an H_2O-methanol extract of the Allende meteorite, **B:** a Miller-Urey experiment (Levy et al. 2000), **C:** the reaction of aqueous NH_2CN (Levy et al. 2000), and **D:** a model formose reaction (Kebukawa et al. 2013; Omran et al. 2020). 7

SI2 **A:** The 'Benzilic Acid Rearrangement rule applied to cyclohexan-1,2-dione (the graph G) as implemented in MØD. The blue highlighted bonds in L are broken and the green highlighted bonds in R are created in the course of the reaction; bonds in black do not change. The graph H is the product of the reaction (1-hydroxycyclopentanecarboxylic acid), while D is an intermediary graph in the transformation algorithm. This reaction normally proceeds through the action of a base (e.g., KOH or NaOH), the availability of such bases in the medium is assumed. Asterisks represent generic wildcards that could represent any species. **B:** The same reaction shown as a fragment of a reaction network, formally a directed multi-hypergraph (Andersen et al. 2012). Reactions (the rectangular node in (**3B**)) between species store the information about which reaction rule was used to generate the products. 8

SI3 Comparison of the resolved mass peaks in the experimental ADG reaction (blue, from D-glucose heated in water, see ESI Detailed Methods Section 3.7) with the modeled reaction network output (pink). There are several peaks which do not correspond between the model and experimental data, which suggests either limitations of the model or limitations of the analytical technique. For example, the *in silico* ADG model could only practically be performed for five generations using desktop computational resources, thus the model likely under samples the chemical space the real-world reaction explores, the reaction rules used may not be not complete and thus cannot accurately match every detected mass, and FT-ICR-MS instruments are often less sensitive to lower mass species. The flexible nature of the presented workflow allows users to add or remove reaction rules based on their own chemical intuition or experimental data. 9

SI4 **A:** Estimated in-degree as a function of Exact Mass in the computed ADG products, **B:** Estimated out-degree as a function of Exact Mass in the computed ADG products, **C:** Estimated in-degree as a function of Number of StereoIsomers in the computed ADG products, **D:** Estimated out-degree as a function of Number of StereoIsomers in the computed ADG products. 10

SI5 The in-degree and out-degree of each node plotted against each other. Clearly, most species have higher out-degree than they do, in-degree. This is likely a bias in our methods and rule structure than any inherent property of real chemical systems. As we can see, there is no clear difference between the species matched with the Yang and Montgomery (1996) test set (red crosses) and the remaining species (blue circles). 11

SI6 Network distributions of select reaction rules applied in the network evolution. The highlighted edges are those that correspond to the reaction rules mentioned underneath. Here node size is proportional to the generation in which they were produced. 12

SI7 The abundance of resolved mass peaks in experimental degradation (blue) contrasted with the model (pink). Note that the frequency with which products of a certain mass occur in our model does not correspond to its abundance in real chemistry. Clearly, there are model predictions that are not recovered in our own experiment. However, we caution that a major part of this effect could be due to the reduced sensitivity of the instrument at lower masses. Still, we do see peaks in the mass spectra that aren't predicted by our model. First, since our *in silico* degradation could be performed only up to five generations, we haven't exhausted the chemical space that can be explored. Even then, it is possible that our selection of reaction rules is not complete and cannot span the entire chemical space. The flexible nature of our workflow allows the user to add any reaction rules of their own based on their chemical intuition. The sample shown here is dextrose wet (see Section 3.7 for details). 13

SI8 The compounds identified in Yang and Montgomery (1996) matched by the computed ADG network as a function of generation. This table was generated using the `mols2grid` library https://github.com/cbouy/mols2grid. An interactive viewer of the matched molecular structures can be accessed at https://reaction-space-explorer.github.io/match-viz/ 14

SI9 The effect of experimental conditions on the product suite of glucose degradation can be seen by comparing the van Krevelen diagrams of the dry (top panels) and wet samples (bottom panels). 9A and 9C show the comparison of all measured FT-ICR-MS signals vs the ADG CRNR, while 9B and 9D show the same comparison, but with the experimental data limited to signals $\geq$ 200 amu. Both the wet and dry samples are characterized by a principal signal density along a 45° slope in the O/C - H/C plane, while the dry sample (9A) is somewhat more scattered and includes a major signal cluster along the H/C = 2 value. Despite that, in the 150-200 m/z regime, the products of both samples exhibit similar atomic ratios. Some of the computed products that fall in these regions are shown in SI10. 15

SI10 Molecules produced in the computed ADG network closest to the region of aromatic compounds in the Van Krevelen plot in Figure **4C.** 16

SI11 Estimated Number of Stereoisomers as a function of Exact Mass in the computed ADG products. 16

SI12 **A:** Histogram showing the variance of $\Delta_{rxn}G$ vs the total number of reactions in our workflow as a function of the pH. **B:** Histogram showing the variance of $\Delta_{rxn}G$ vs the number of "most frequent" reactions in our study as a function of the pH. . . 17

SI13 Comparison of the CRNR products and their matches in the HMDB (Wishart et al. 2018), KEGG (Kanehisa and Goto 2000) and ECMDB (Sajed et al. 2016) databases as a function of generation. 18

SI14 **Top**: Estimated Total Surface Area (TSA) as a function of mass and generation in the computed ADG products. **Bottom**: Estimated Polar Surface Area (PSA) as a function of mass and generation in the computed ADG products (Stanton et al. 2002) 26

SI15 **Top**: Estimated Topological Polar Surface Area (TPSA) as a function of mass and generation in the computed ADG products. TPSA is defined as the sum of solvent accessible surface areas of atoms with absolute value of partial charges greater than or equal 0.2 (Stanton et al. 2002). **Bottom**: Estimated Relative Polar Surface Area (RPSA) as a function of mass and generation in the computed ADG products. RPSA is defined as a value of Topological Polar Surface Area divided by Total Surface Area (Stanton et al. 2002) 27

SI16 **Top**: Estimated Molecular Complexity as a function of mass and generation in the computed ADG products (von Korff and Sander 2019). **Bottom**: Estimated Molecular Flexibility (from 0, completely rigid, to 1, completely flexible) as a function of mass and generation in the computed ADG products (Kier 1989). 28

SI17 **Top**: The generational diversity of HybRatio as a function of generation in the ADG CRNR. HybRatio is defined as the ratio of sp^3 hybridized carbon atoms in a molecule over the sum of sp^2 and sp^3 hybridized carbon atoms ($N_{sp^3}/(N_{sp^2}+N_{sp^3})$). This was computed using CDK (Willighagen et al. 2017). **Bottom**: Estimated Aqueous Solubility (ESOL) in mol/L as a function of mass and generation in the computed ADG products. ESOL is the estimated aqueous solubility of a molecule inferred from its structure (Delaney 2004). Higher values indicate higher aqueous solubility. ADG products tend towards lower aqueous solubility as a function of generation. 29

SI18 Principal component analysis (PCA) of the cheminformatic descriptor variables colored by modularity class computed using the methods described in (Blondel et al. 2008) and visualized using Gephi (Bastian et al. 2009). 31

SI19 PCA loading plot showing how strongly each descriptor influences component identification. IC0, SIC0 and CIC0 are information theoretic indices with neighborhood order equal to zero. IC represents the Information Content of a molecule graph, SIC is related to the Structural Information Content and CIC is the complementary information content that is defined as the difference between the maximum possible complexity of a molecule graph and its Information Content (IC). A more detailed definition of these descriptors can be found in (Basak et al. 2000). There is a positive correlation among the modularity class and cLogP/ΔG(pH = 10) and ΔG(pH = 10)/Shape Index. There is also a positive correlation among the Number of Stereoisomers and the Number of Oxygen Atoms, Polar Surface Area and Number of H-Acceptors or H-Donors. In the vertical direction the observation apparently clusters in relation to Molecular Polarity and Exact Mass (PC2 Component). The PC1 component is more related to the Molecular Complexity descriptor and leads to the formation of sub clusters related principally to the number of carbon atoms and the number of rotatable bonds or the flexibility of the molecule. There is also a positive correlation among Molecular Complexity, IC0 and O/C Ratio. 32

SI20 Representative structures of the molecules in each cluster. A general trend of decreasing flexibility from left to right is evident, as is increasing polarity from top to bottom. 33

SI21 The molecules in the ADG CRNR can be grouped into five natural modularity classes using Gephi (Bastian et al. 2009) with a resolution of $\sim$ 2.8. Modularity analysis was carried out with different resolution values and usually all gave four large groups (a-d), as well as other small subgroups. Modularity class e here contains several subgroups that do not cluster at this level of resolution (Lambiotte et al. 2008) (In this plot node size is proportional to in-degree). 34

SI22 Examples of compounds present in modularity class e. 35

SI23 Autocatalytic cycle definitions proposed by Gánti (1975); Peretó (2012); Andersen et al. (2021) and Blokhuis et al. (2020). Circles represent molecules and squares reactions. Autocatalytically produced molecules are highlighted in yellow. Brown circles are external molecules to the cycle (e.g., feeder molecules or output molecules). Purple circles represent an "exclusive autocatalytic node" (see Andersen et al. 2021), which takes into account the edges as flows allowing for a net production of that molecule. In Fig. SI23E, the autocatalytic motif used in the imperative programming approach is illustrated. The source node represents the autocatalytic molecule and the target node represents the autocatalytic reaction. In Fig. SI23F, the autocatalytic motif used in the declarative programming approach is illustrated. 37

SI24 **Left:** Examples of Gephi's Species-Reaction representations, which preserve information regarding the molecularity of the reaction. **Right:** Gephi's spline interpolation function used to set the size of the nodes as a function of in-degree or out-degree. 40

1 SI File Descriptions

This model's code is available open-source on GitHub, along with supporting documentation on https://github.com/Reaction-Space-Explorer/reac-space-exp. Additional files such as the output glucose degradation network and downloaded SDF formatted versions of the ECMDB and HMDB are available.

SI File 1

The *declarative* Neo4J queries for searching toplogically autocatalytic cycles (see Section 3.10 for details) can be found here. The code for the *imperative* Ford-Fulkerson approach (written in Python) has also been provided on the same repository (see this link).

These methods have a subroutine to print the candidate cycles found for visual inspection. The imperative method additionally includes $\Delta_r G$ values in cycle evaluation. Using that, we can verify that all spontaneous cycles have negative free energies as established by the algorithm. Relevant instructions have been provided on the repository.

SI File 2

Graphical representation of all the reaction rules encoded in our model as double-pushout diagrams. Available: here.

SI File 3

Forbidden substructures used in the reaction expansion methodology (see Section 3.3 for details). Available: here.

SI File 4

The change in free energy for each reaction, $\Delta_r G^\circ$ using the eQuilibrator API. Available: here.

2 Selected SI Figures

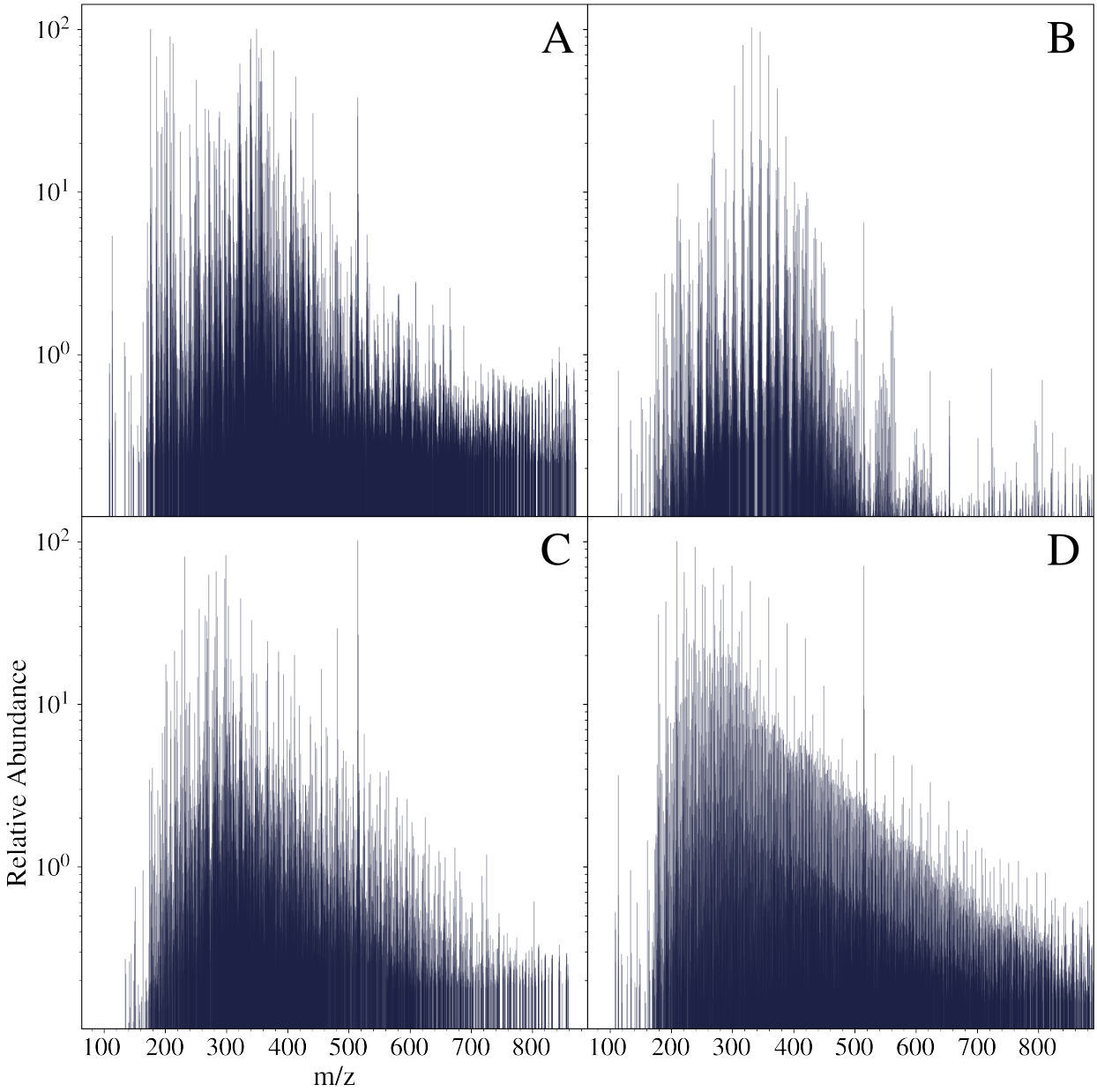

Fig. SI1: High resolution negative-ESI mode FT-ICR-MS spectra of carbonaceous meteorite organics and laboratory abiotic synthesis simulants. **A:** an H_2O-methanol extract of the Allende meteorite, **B:** a Miller-Urey experiment (Levy et al. 2000), **C:** the reaction of aqueous NH_2CN (Levy et al. 2000), and **D:** a model formose reaction (Kebukawa et al. 2013; Omran et al. 2020).

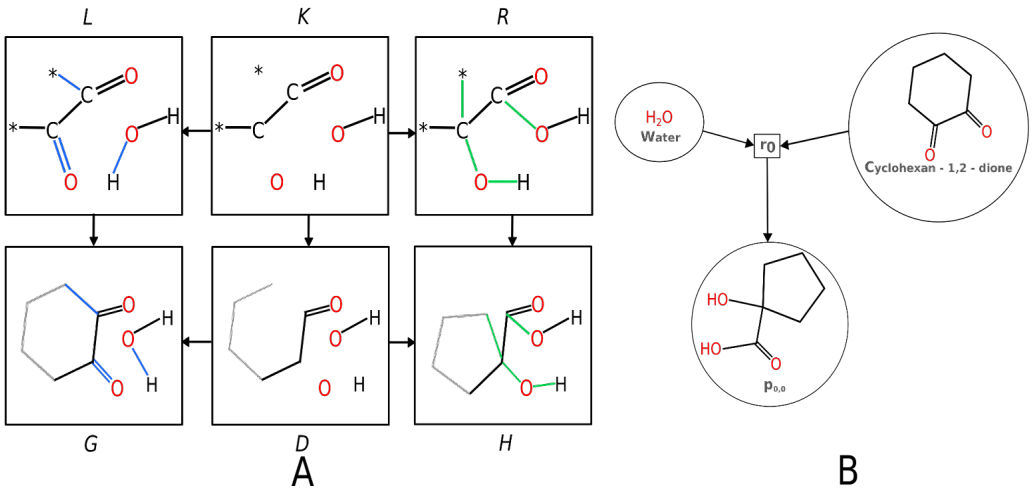

Fig. SI2: **A:** The 'Benzilic Acid Rearrangement rule applied to cyclohexan-1,2-dione (the graph G) as implemented in MØD. The blue highlighted bonds in L are broken and the green highlighted bonds in R are created in the course of the reaction; bonds in black do not change. The graph H is the product of the reaction (1-hydroxycyclopentanecarboxylic acid), while D is an intermediary graph in the transformation algorithm. This reaction normally proceeds through the action of a base (e.g., KOH or NaOH), the availability of such bases in the medium is assumed. Asterisks represent generic wildcards that could represent any species. **B:** The same reaction shown as a fragment of a reaction network, formally a directed multi-hypergraph (Andersen et al. 2012). Reactions (the rectangular node in (**3B**)) between species store the information about which reaction rule was used to generate the products.

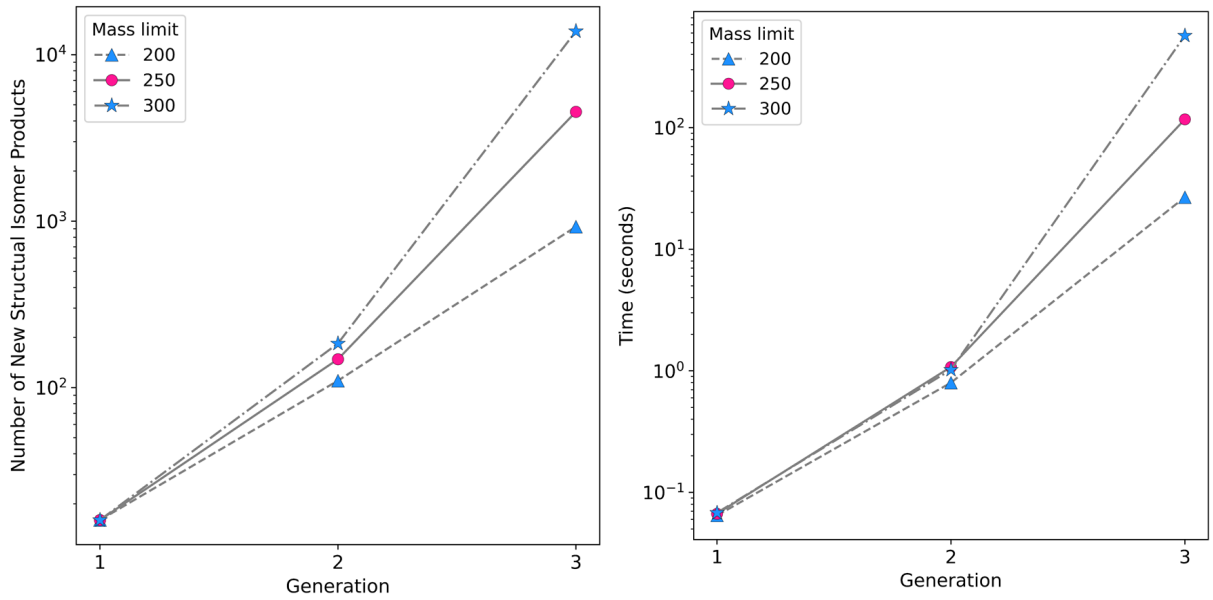

Fig. SI3: Comparison of the resolved mass peaks in the experimental ADG reaction (blue, from D-glucose heated in water, see ESI Detailed Methods Section 3.7) with the modeled reaction network output (pink). There are several peaks which do not correspond between the model and experimental data, which suggests either limitations of the model or limitations of the analytical technique. For example, the *in silico* ADG model could only practically be performed for five generations using desktop computational resources, thus the model likely under samples the chemical space the real-world reaction explores, the reaction rules used may not be not complete and thus cannot accurately match every detected mass, and FT-ICR-MS instruments are often less sensitive to lower mass species. The flexible nature of the presented workflow allows users to add or remove reaction rules based on their own chemical intuition or experimental data.

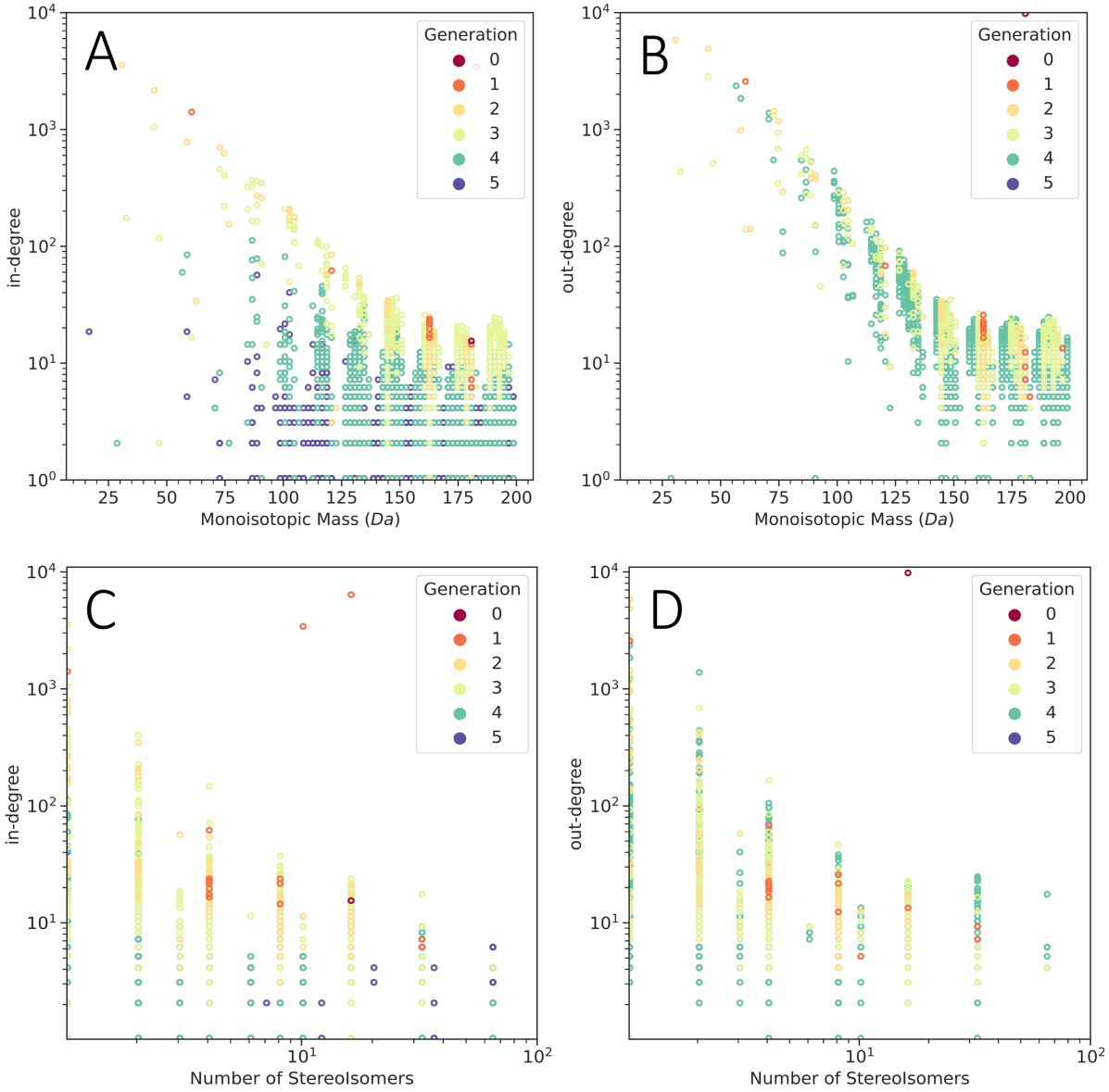

Fig. SI4: **A:** Estimated in-degree as a function of Exact Mass in the computed ADG products, **B:** Estimated out-degree as a function of Exact Mass in the computed ADG products, **C:** Estimated in-degree as a function of Number of StereoIsomers in the computed ADG products, **D:** Estimated out-degree as a function of Number of StereoIsomers in the computed ADG products.

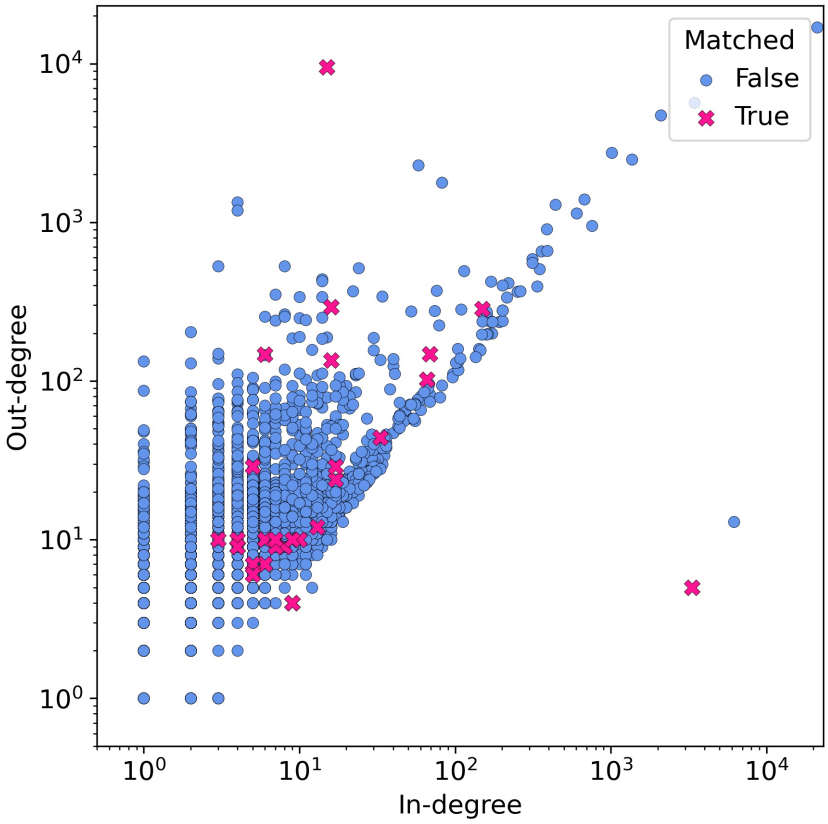

Fig. SI5: The in-degree and out-degree of each node plotted against each other. Clearly, most species have higher out-degree than they do, in-degree. This is likely a bias in our methods and rule structure than any inherent property of real chemical systems. As we can see, there is no clear difference between the species matched with the Yang and Montgomery (1996) test set (red crosses) and the remaining species (blue circles).

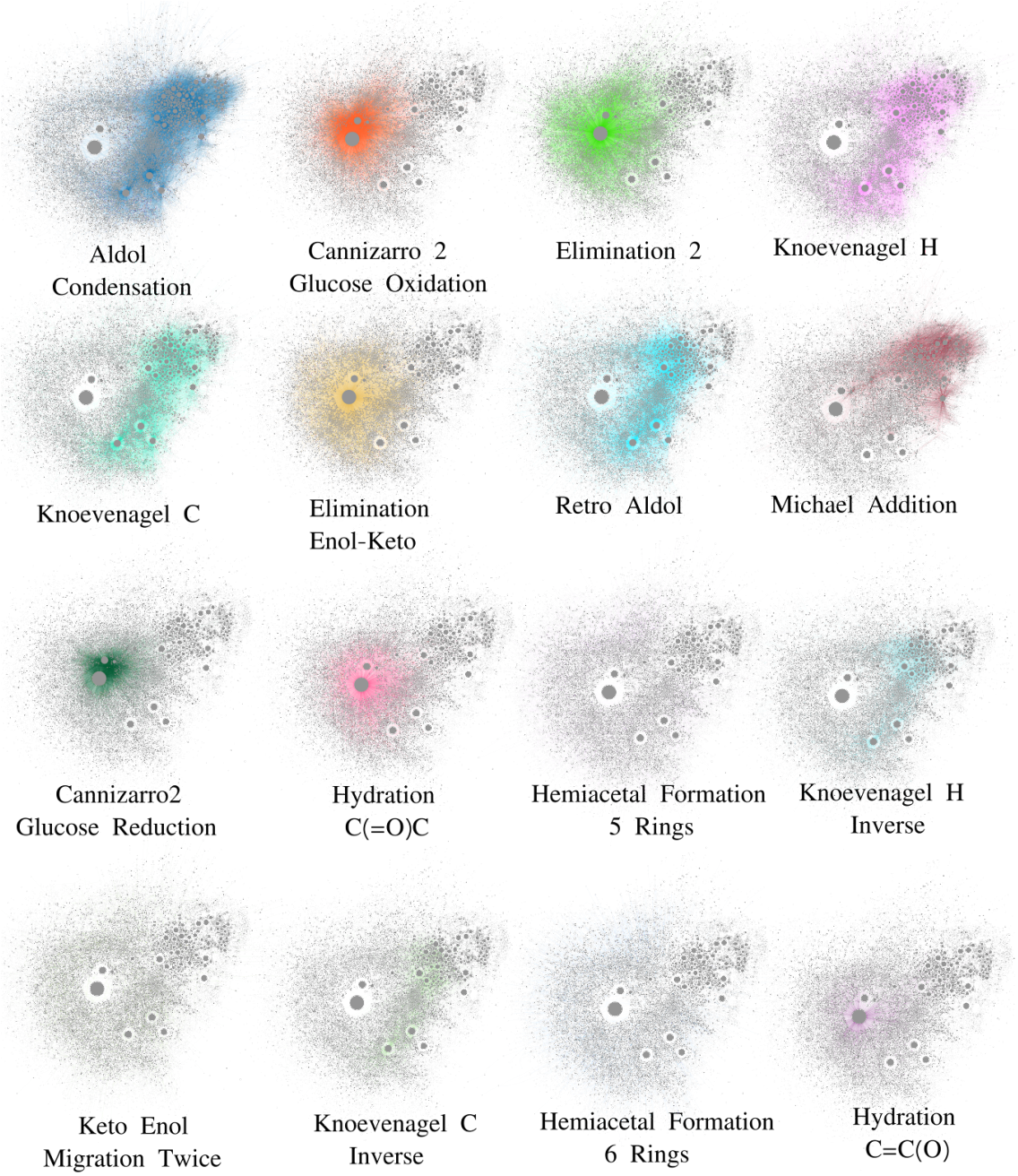

Fig. SI6: Network distributions of select reaction rules applied in the network evolution. The highlighted edges are those that correspond to the reaction rules mentioned underneath. Here node size is proportional to the generation in which they were produced.

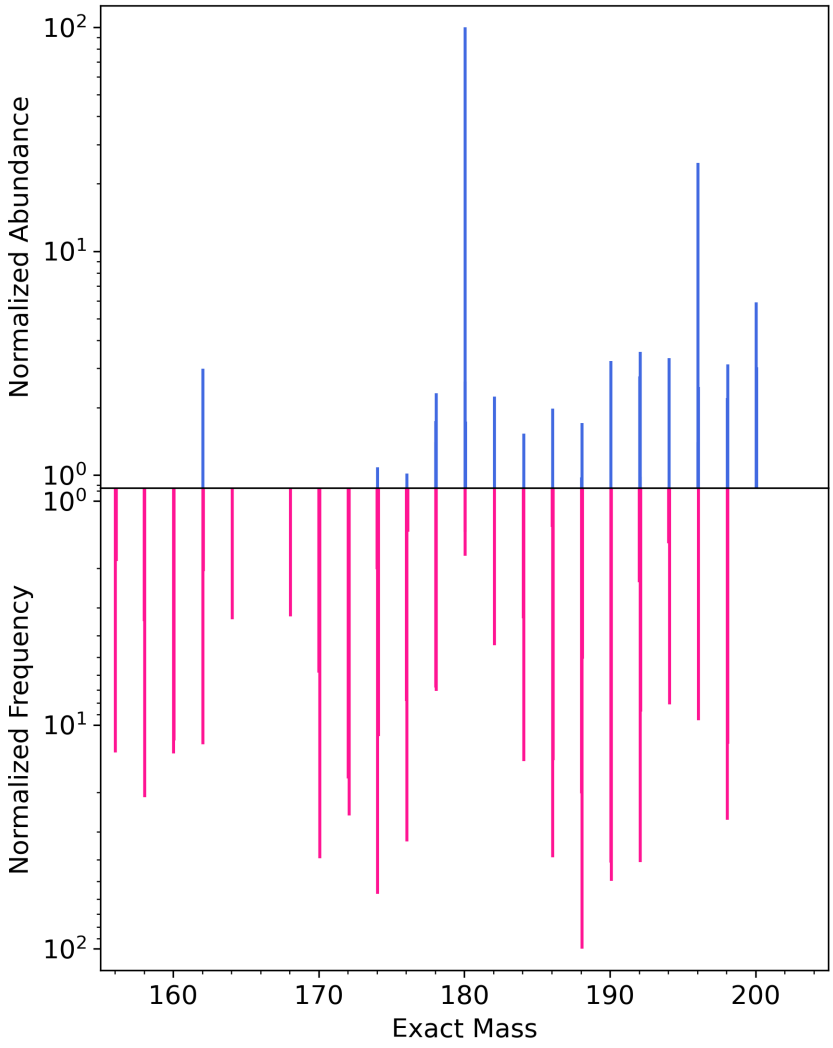

Fig. SI7: The abundance of resolved mass peaks in experimental degradation (blue) contrasted with the model (pink). Note that the frequency with which products of a certain mass occur in our model does not correspond to its abundance in real chemistry. Clearly, there are model predictions that are not recovered in our own experiment. However, we caution that a major part of this effect could be due to the reduced sensitivity of the instrument at lower masses. Still, we do see peaks in the mass spectra that aren't predicted by our model. First, since our *in silico* degradation could be performed only up to five generations, we haven't exhausted the chemical space that can be explored. Even then, it is possible that our selection of reaction rules is not complete and cannot span the entire chemical space. The flexible nature of our workflow allows the user to add any reaction rules of their own based on their chemical intuition.
The sample shown here is dextrose wet (see Section 3.7 for details).

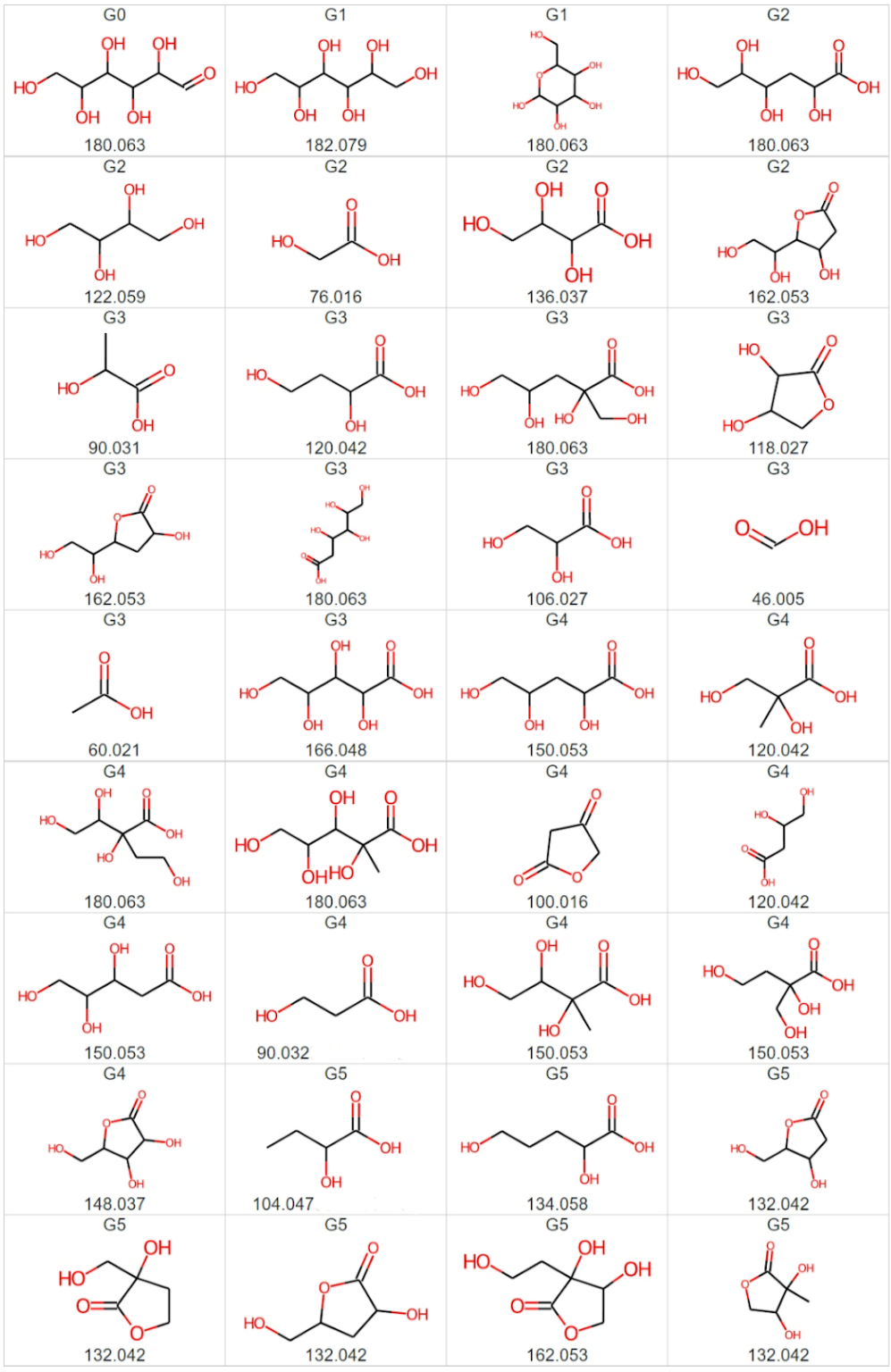

Fig. SI8: The compounds identified in Yang and Montgomery (1996) matched by the computed ADG network as a function of generation. This table was generated using the `mols2grid` library https://github.com/cbouy/mols2grid. An interactive viewer of the matched molecular structures can be accessed at https://reaction-space-explorer.github.io/match-viz/

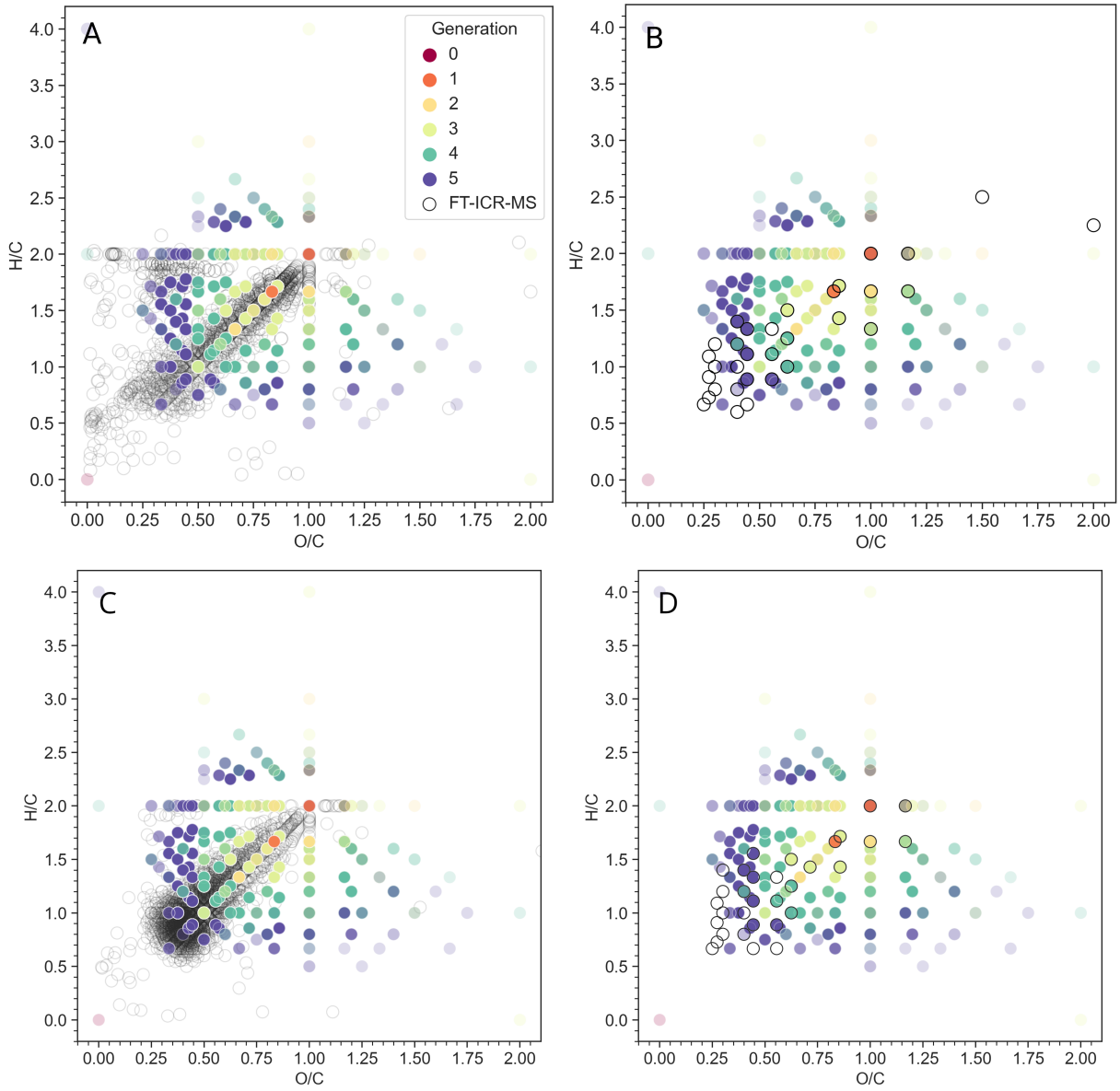

Fig. SI9: The effect of experimental conditions on the product suite of glucose degradation can be seen by comparing the van Krevelen diagrams of the dry (top panels) and wet samples (bottom panels). 9A and 9C show the comparison of all measured FT-ICR-MS signals vs the ADG CRNR, while 9B and 9D show the same comparison, but with the experimental data limited to signals $\geq$ 200 amu. Both the wet and dry samples are characterized by a principal signal density along a 45° slope in the O/C - H/C plane, while the dry sample (9A) is somewhat more scattered and includes a major signal cluster along the H/C = 2 value. Despite that, in the 150-200 m/z regime, the products of both samples exhibit similar atomic ratios. Some of the computed products that fall in these regions are shown in SI10.

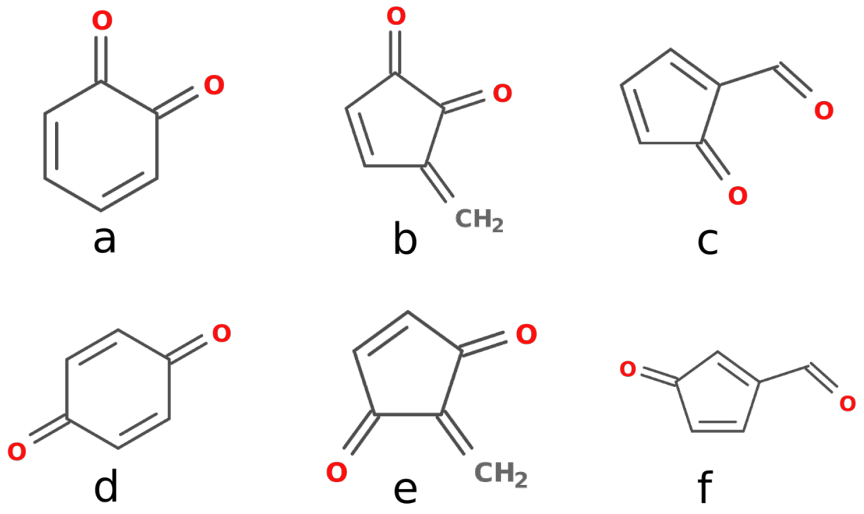

Fig. SI10: Molecules produced in the computed ADG network closest to the region of aromatic compounds in the Van Krevelen plot in Figure **4C.**

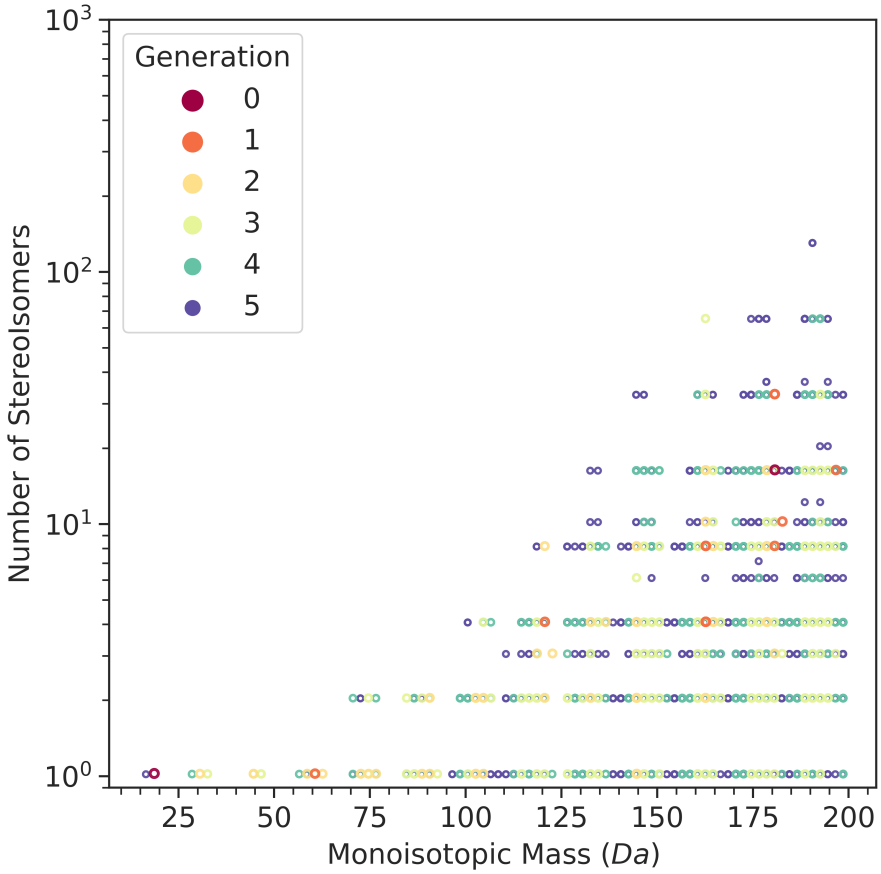

Fig. SI11: Estimated Number of Stereoisomers as a function of Exact Mass in the computed ADG products.

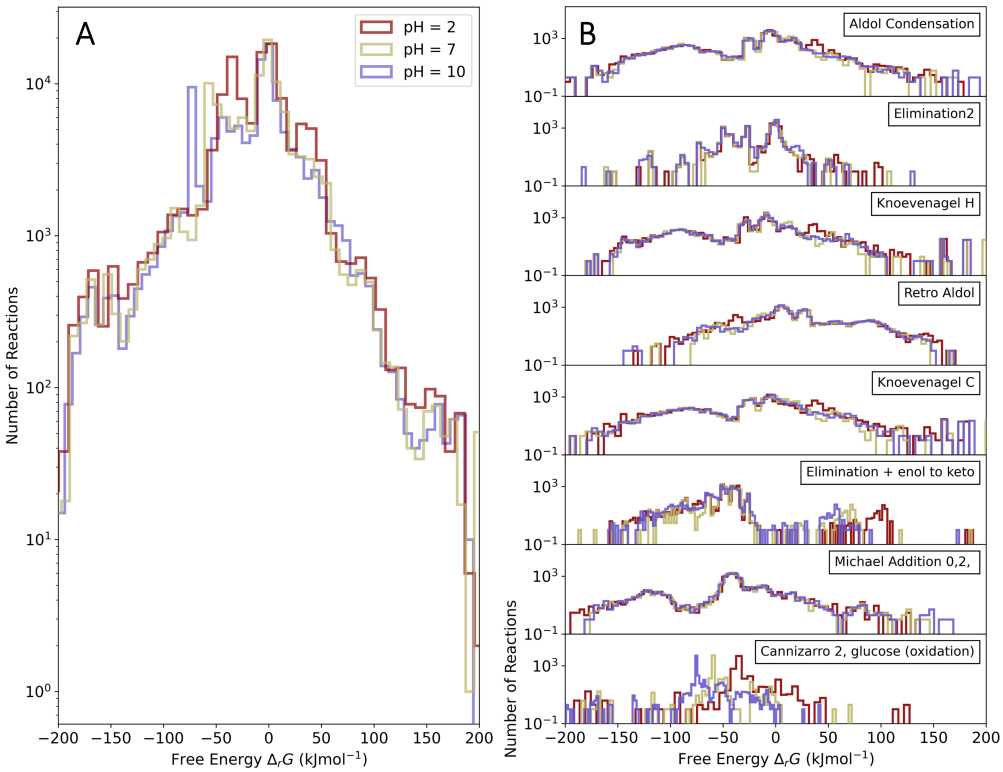

Fig. SI12: **A:** Histogram showing the variance of $\Delta_{rxn}G$ vs the total number of reactions in our workflow as a function of the pH. **B:** Histogram showing the variance of $\Delta_{rxn}G$ vs the number of "most frequent" reactions in our study as a function of the pH.

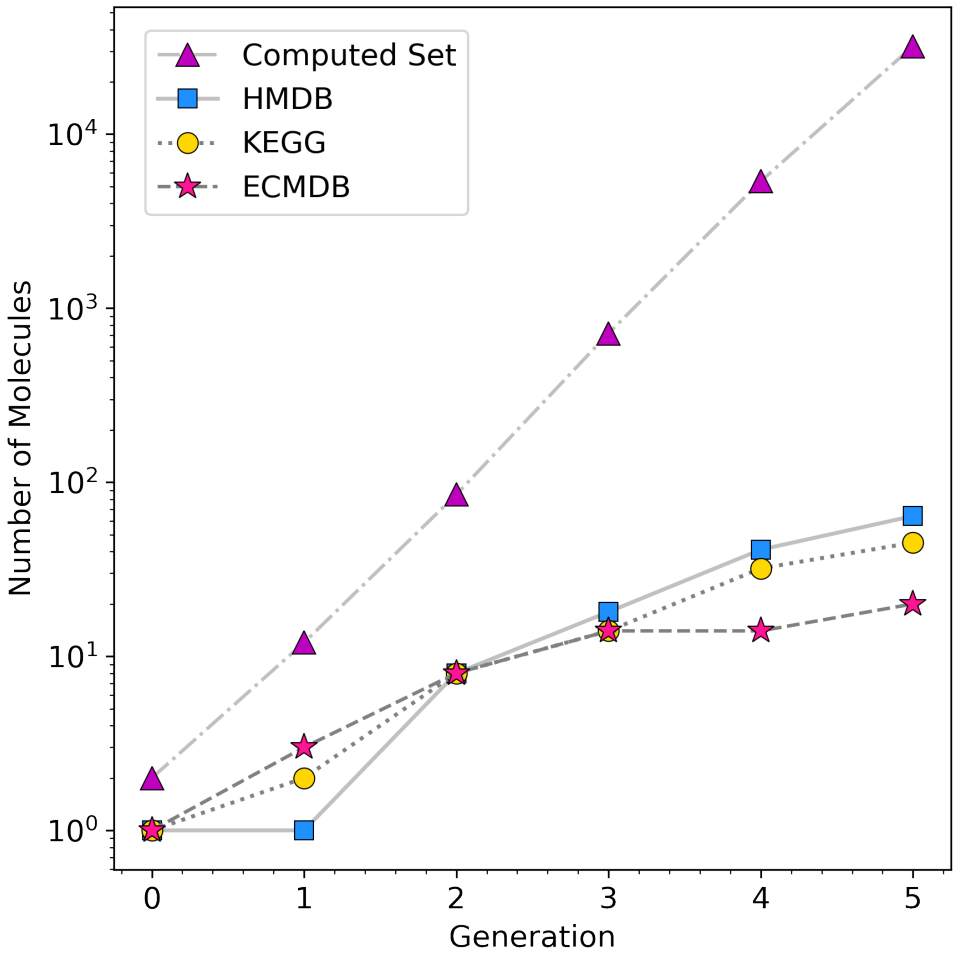

Fig. SI13: Comparison of the CRNR products and their matches in the HMDB (Wishart et al. 2018), KEGG (Kanehisa and Goto 2000) and ECMDB (Sajed et al. 2016) databases as a function of generation.

3 Detailed Methods

3.1 Reaction Network Generation

The core of this workflow is the MØD software package (Andersen et al. 2016), that can generate CRNRs using user-defined reaction rules (see SI File 2). The ADG network explored here was computed using MØD run on a Hewlett-Packard Z820 computer with 128 GB RAM and 16 cores in two Intel® Xeon® E5-2670 2.60 GHz processors, running Ubuntu 20.04 in a virtual machine hosted on a Windows 10 system.

MØD was used due to its availability as an open-source Python package, and since its core library is written in C++, its performance is on par with pure C or C++ based libraries. It can thus be easily pipelined with other Python packages, and its flexible object-oriented framework removes the need for an external library for graph object handling. Alternate tools for reaction network generation such as Reaction Mechanism Generator (Liu et al. 2021), and ReacNetGenerator (Zeng et al. 2020) also exist, however MØD uses state-of-the-art graph canonicalization algorithms which are efficient at dealing with large collections of molecular graphs (Andersen and Merkle 2020), and RMG is not fully customizable, while ReacNetGenerator uses ab initio MD, which may pose scaling problems as discussed in (Sharma et al. 2021).

Reactant molecules were loaded into MØD using SMILES (Weininger 1988), but GML (Himsolt 1997) format can also be used. In MØD, molecules are treated as graphs in which labelled nodes and edges represent atoms and edges, respectively. Reactions are processed by application of reaction rules which provide templates describing which transformations substructures in reactant molecules can undergo (see Fig. SI6). Products of such reactions then become reactants for subsequent iterations of this process, leading to a concatenated chain of reactions (hereafter termed "generations"). The product graphs are then output in an internally canonicalized SMILES format[1]. It should be noted that information allowing reconstruction of the entire network can be exported from MØD to other desired formats and parsed by other software for further analysis such as Graph Tool (Peixoto 2017) or NetworKit (Staudt et al. 2014).

The initial reactants, H_2O and open-chain DL-glucose, were input in SMILES. A set of defined reaction rules (summarized in Table SI2) developed based on literature precedent was then loaded. Stereochemistry is undefined in the formalism employed here, e.g., DL-glyceraldehyde is treated as a single "flattened" isomer (R, S-2,3-dihydroxypropanal), threose and erythrose are treated as equivalent, etc.. Ring-chain isomers (e.g., open chain and ring isomers of glucose, lactones and their associated acids) are treated as unique species since these could serve as unique reactants due to their structurally unique reactive motifs. This ADG CRNR could also have been initiated using the cyclic hemiacetal of glucose (as α-D-glucose interconverts in aqueous solution to give a mixture of ring epimers and open-chain forms (Dworkin and Miller 2000), but would then require one more generation to enter the initial state described here.

Reaction rules are templates for reactions and are formally Double Push-Out (DPO) graph transformation rules (Andersen et al. 2016; Ehrig et al. 2006). A DPO rule consists of three graphs (or molecular patterns): the left graph 'L' represents the reactant patterns to search for, the right graph 'R', the product pattern that will be substituted in place of "L", and the rule

[1] https://jakobandersen.github.io/mod/dataDesc/dataDesc.html

context graph 'K' which is needed along with the two arrows to formally encode which atoms of L map to which in R.

If the reactant pattern L can be found in a molecule, then the corresponding edges of the graph (the chemical bonds of the molecule(s)) can be changed by removing all edges defined by L that are not in K, and adding all of the edges of R (Andersen et al. 2013). The initial step in reaction rule application is thus the search for the reactant pattern L. An exemplary transformation for the benzilic acid rearrangement (Liebig 1838) reaction of cyclohexan-1,2-dione with water is shown in Fig. SI2.

3.2 Rule Selection

This approach does not allow for automated consideration of the effects of pH, temperature, or other kinetic considerations, or specifically consider kinetic effects except as they are latently represented by the selection of reaction rules. Reaction choice thus involves user-defined biases, and considers all user-defined mechanistically plausible reactions (see Table SI2). The implemented reactions were however vetted by their ability to create products that can explain the compounds reported by the comprehensive study of Yang and Montgomery (1996), and selected to be chemically reasonable in terms of their reversibility, for example reactions known to be essentially reversible in basic medium, e.g. ester hydrolysis or decarboxylation, were encoded here as only proceeding in the forward direction.

3.3 Post-Generation Filtering

During application of reaction rules, structures deemed to be chemically unstable were sometimes produced which are unlikely to appear in real chemistry. Apart from this, the exponential growth in the number of species produced with each generation quickly made the computation resource heavy. To ameliorate these issues, each computed product was examined for sterically problematic or "bad" substructures, and a molecular weight (MW) cutoff filter was applied (herein 200 amu). Any species containing a bad substructure or exceeding the prescribed MW limit created during reaction generation was discarded from the network. This helped to both clean bad products and limit the size of the product space input to the next generation, making further generation computation more tractable The forbidden "bad" structures were selected based on literature precedent (e.g., (Weber 2004), as well as those included in MolGen 5.0 (Gugisch et al. 2015), see SI File 3), e.g., three- and four-membered rings were excluded as they are typically unstable due to their large ring strain (Wiberg 1986). Some chemically implausible substructures (e.g., structures with double bonds at bridgehead carbons, which violate Bredt's rule (Bredt 1924) were also specifically forbidden. The prohibition of X=X=X bonding was suppressed specifically for CO2 and isocyanates.

This filtering reduces the computational load for succeeding generations, and since each product must be computed before it is checked, limiting mass growth helps avoid "combinatorial explosions", which allows computation of more iterations than would otherwise be possible.

3.4 Minimizing the Combinatorics for Specific Rules

We quickly found that some rules in our set caused a sharp growth in computing time. For example, due to the large combinatorics involved, computing time for a simple rule encoding the Cannizzaro reaction posed a resource bottleneck despite not adding many products, as it simply showed that any two aldehydes could redox disproportionately (Morooka et al. 2005). After some consideration, we decided this could be an "inferred reaction" to speed processing: glucose (which itself contains an aldehyde group) could be allowed to be the oxidant/reductant for all Cannizzaro reactions (Geissman 1944), as opposed to allowing all possible network-generated aldehydes to play this role.

This "fixing" of a particular reactant aldehyde does not affect the novel products that can be formed, as at this level of analysis the only transformation happening is the oxidation of one aldehyde group and the reduction of another — the full structure of the molecule is irrelevant for the transformation.Thus, it minimized the computing time by reducing the number of possible combinations of aldehydes from being proportional to n^2 to instead being proportional to n, where n is the number of reactants that can undergo the Cannizzaro reaction. This generalization allowed more reaction iterations to be performed while still producing the same novel products. This is a valid approximation as long as the aldehyde that is "fixed", exists in the reaction network, but this approximation excludes the discovery of reaction motifs which depend on specific Cannizzaro redox reactions, e.g., in which glucose is not, or not yet, a member of the reaction network. In the chemistry studied here, glucose is available in the network since the beginning, allowing Cannizzaro to proceed unimpeded.

3.5 Treatment of Tautomers

Tautomers are compounds with identical molecular formulas, which are structural (i.e., constitutional) isomers of compounds that readily interconvert, often by proton migration, among other mechanisms (Kochev et al. 2013). Many organic compounds containing conjugated bonds can be represented by multiple human-interpretable structures, though chemists often prefer one tautomeric representation. Compounds with unique anhydrous crystal structures often interconvert into multiple bonding states in solution, and even relatively simple compounds such as guanine ($C_5H_5N_5O$), with completely defined core bonding, may have as many as 36 valid tautomeric representations (Sabio et al. 1990). MØD treats different tautomeric forms as distinct chemical species; it was thus necessary to suppress the generation of multiple tautomeric forms for the same molecule. Tautomerism is generally acknowledged to be a difficult representational problem in computational chemistry (see (Sayle 2010), (Dhaked et al. 2020)), although resolutions for this issue have been proposed (e.g., (Warr 2010)).

Tautomers may engage in chemical reactions in unique ways, thus their unique representation in generative chemical models is meaningful. However, most chemists would also consider them to be chemically latent structures. Tautomeric structures are capable of undergoing the same reactions because they are interconvertible, and denoting tautomers as unique in this context is meaningful in some ways (e.g., with respect to precise reaction mechanisms or hydrogen bonding), but not in others (e.g., the overall outcome of reactions).

We explored the use of RDKit tautomer classes (Landrum 2013) with an open source wrapper

`MolVS`[2] integrated into RDKit to handle the enumeration of possible tautomers, keeping only one molecule out of a set of tautomers in a given generation. To have a more configurable method than RDKit could provide, the same exercise was also done using the Java library Ambit-Tautomer (Kochev et al. 2013) pipelined with our workflow using the `JPype` python package.[3] Neither of these approaches proved satisfactory for our goals. We finally handled tautomerism at the level of encoded reaction mechanisms, such that reaction schemes would provide tautomeric species that a chemist would recognize as "catalog representations" of tautomers (Taylor et al. 2013), and such that only one unique tautomer of a given compound would be produced anywhere in the network, e.g., enol forms were forced into keto or aldo forms, and reaction mechanisms were built to infer the existence of transient reactive species such as enolates. This simplification of tautomers reduces the number of unique species representations generated, reducing computational cost and assisting in output interpretation.

3.6 Comparison with literature analyses

The predictive validity of the generated ADG network was evaluated by checking if compounds detected in (Yang and Montgomery 1996) using Gas Chromatography-Mass Spectrometry (GC-MS) could be explained by this workflow. To do so, the molecules detected in (Yang and Montgomery 1996) were rendered into sdf format. These reported species (henceforth referred to as the "test set") were then imported into our workflow using RDKit and matched against the compounds generated in our workflow by performing isomorphism tests using the VF2 algorithm in MØD (Cordella et al. 2004). Molecules are isomorphic if all the atoms in one molecule are bonded in exactly the same fashion as another, despite their initial arbitrary labeling.

We did not represent aromaticity explicitly in our reaction rules, aromatic compounds would need to be identified from their Kekulé forms. Since there are multiple Kekulé representations for a single aromatic species, compounds in the test set were kekulized to a canonical representation using RDKit before comparison. Since we could perform only five generations of reaction expansion due to computational limitations, matching for some of the compounds in the test set requiring more than five generations for their production was explored using a modified workflow. To explore our technique's ability to produce molecules identified in (Yang and Montgomery 1996) which were not produced in the first five computed generations, each ADG product was reacted with water, and the same glucose degradation rules were applied. Due to the reversibility of most of our reaction rules, this effectively runs the reactions "backwards" and this process should eventually connect with molecules produced by the "forward" generation method (see Table SI3). These retrospective networks were explored for four generations to see if the species generated matched products of the "forward" glucose network. If experimentally measured molecules were found after n generations of reactions within these test networks, it stands to reason it would take 5+n total generations to construct the corresponding test set species in the "forward" glucose degradation network.

[2] https://molvs.readthedocs.io/en/latest/
[3] https://jpype.readthedocs.io/en/latest/

3.7 Experimental Degradation of Glucose

3.7.1 Sample Preparation

D(+)-glucose ($\geq$ 99.5%, Aldrich) was thermally altered in the dry state, and in aqueous solution in sealed glass ampoules under nitrogen, thermostatted at fixed temperature on dry heating blocks at 150° C for 2 days.

3.7.2 General Precautions

All glassware and ceramics were ashed at 500 °C to remove organic contaminants. Water was from a Millipore MilliQ system and of maximum conductivity of 18 MΩ. All other solvents were High-Pressure Liquid Chromatography (HPLC) grade. Glucose (Reagent Plus Grade, $\geq$ 99.5% purity) was purchased from Sigma Aldrich (St. Louis, MO, USA).

3.7.3 FT-ICR MS Instrument Specifications

A custom-built FT-ICR mass spectrometer (Kaiser et al. 2011a) equipped with a 9.4 Tesla horizontal 220 mm bore diameter superconducting solenoid magnet operated at room temperature, and a modular ICR data station (Predator, Blakney et al. 2011) which facilitated instrument control, data acquisition, and data analysis, located at the National High Magnetic Field Laboratory in Tallahassee, Florida, was used to analyze the samples. Helium gas introduced into the octopole collisionally cooled ions prior to transfer through radio-frequency (RF)-only quadrupoles (total length 127 cm) equipped with an auxiliary RF waveform (Kaiser et al. 2014) into a 7-segment open cylindrical cell (Kaiser et al. 2011b) with capacitively coupled excitation electrodes based on the Tolmachev configuration (Tolmachev et al. 2011). Next, 100–150 individual transients of 5.9 s duration were signal-averaged, apodized with a half-Hanning weight function, and zero-filled once prior to fast Fourier Transformation (FT). Data was collected at maximum memory depth of the data station hardware (16 million samples), apodized with a single-sided Hanning apodization, zero-filled to 16 megasamples (16,777,216 samples or 224). An additional zero fill brought the pre-FT data packet to 32 megasamples. Due to increased complexity at higher m/z, broadband phase correction (Beu et al. 2004) was applied to the mass spectra to increase the resolution of isobaric species as previously described (Xian et al. 2010). The sonicated aqueous extracts were briefly centrifuged to remove fines. Mass spectra were calibrated with custom-built software (MIDAS, Blakney et al. 2011).

3.7.4 Ionization

Extracts were analyzed by negative electrospray ionization (ESI). The samples were further diluted to 500 µg/mL with equal parts (by volume) of MeOH spiked with 0.25 %(by volume) TMAH to ensure efficient deprotonation (Lobodin et al. 2013). Sample solutions were pumped through a microelectrospray source (Emmett et al. 1998) (50 µm i.d. fused silica emitter) at 0.5 µL/min by a syringe pump. The sample and the Pierce ® LTQ Velos ESI negative ion calibration solution (Thermo Fisher Scientific, Waltham, MA, USA) were electrosprayed consecutively by use of a dual needle electrospray source (Hannis and Muddiman 2000). Predator software controlled the duration when each needle was positioned at the entrance of the mass spectrometer. Negative ESI tends to favor the detection of compounds containing functional groups that can readily lose a proton, such as alcohols, carboxylic acids, cyanides, nitric- and sulfonic-acids. Positive ion ESI data were not collected here. Based on previous data regarding Dissolved Organic Matter (DOM)

functional group characteristics, most previously published works on DOM molecular composition have been carried out using negative ion ESI when combined with FT-ICR MS. A previous study on extraterrestrial organic matter in the Murchison meteorite found that positive and negative ESI modes were complementary, and spectra collected in both modes show repetitive patterns (Schmitt-Kopplin et al. 2010).

3.7.5 Mass Calibration

ICR frequencies were converted to ion masses based on the quadrupolar trapping potential approximation (Shi et al. 2000; Grosshans et al. 1991). Each m/z spectrum was first externally calibrated by the Pierce ® LTQ Velos ESI negative ion calibration solution, and internally calibrated based on the "walking" calibration equation (Savory et al. 2011) for several highly abundant homologous series (mass of –CH2– repeating unit 14.01565 Da) confirmed with isotopic fine structure. Mass spectra with m/z from 150 to 750 were exported to peak lists at a signal-to-noise ratio (S/N) > 2. Elemental compositions assignment and data visualization were facilitated using PetroOrg©software.

3.7.6 Data Cleansing

In the raw processed data, several peaks could not be assigned a molecular formula by the spectrometer. These peaks could possibly have been inorganic salt clusters. Therefore, we discarded all data entries with "No Hit" for an assigned formula. An exception was made for the peaks in the m/z 150-200 range which we found to be consistent with containing a ^{13}C isotope within a precision of 4 decimal places. Next, given that our intended sample and modelled chemistry (glucose degradation) involves only C, H and O, we further filtered out data points containing nitrogen and sulphur atoms. This is the final form in which the data presented in Figures **2** and **4** of the main text were utilized.

3.8 Cheminformatic Descriptor Computation

The compounds produced in these networks represent real chemical compounds with real chemical properties, and it is of interest to understand how real chemical reactions explore chemical property space for emergent behaviors (e.g., Boogerd et al. 2005; Jain and Krishna 2001). Energy minimization optimizations of output SMILES were carried out using OpenBabel (O'Boyle et al. 2011). Physical and topological chemoinformatic descriptors were computed using MolGen-QSPR (Kerber et al. 2005), and DataWarrior (Sander et al. 2015). Two-dimensional descriptors were calculated and benchmarked using Mordred (Moriwaki et al. 2018). Computed chemical descriptors as a function of monoisotopic mass for the ADG CRNR from Datawarrior are present in the SI as follows: Total Surface Area (Fig. SI14 Top), Polar Surface Area (Fig. SI14 Bottom, Ertl et al. 2000), Topological Polar Surface Area (Fig. SI15 Top, Stanton et al. 2002), Relative Polar Surface Area (Fig. SI15 Bottom, Stanton et al. 2002), Molecular Complexity (Fig. SI16 Top, von Korff and Sander 2019), Molecular Flexibility (Fig. SI16 Bottom, Kier 1989), and HybRatio [4] (Fig. SI17 Top, Willighagen et al. 2017).

[4] https://cdk.github.io/cdk/1.5/docs/api/org/openscience/cdk/qsar/descriptors/molecular/HybridizationRatioDescriptor.html

cLogP is the calculated logarithm of a molecule's predicted partition coefficient between n-octanol and water, e.g., if a compound is shaken between equal volumes of n-octanol and water, and a thousand times as much of the compound partitions into the octanol phase as partitions into the water phase, then the logP is 3. cLogP is a proxy for the ability of compounds to partition between hydrophobic and hydrophilic phases, and potentially to form new phases altogether. Estimated Aqueous Solubility (ESOL, Delaney 2004) was calculated using `PatWalters/solubility`[5] (see Fig. SI17).Stereoisomers were enumerated initially by using ChemAxon's `stereocalc`[6] commandline function, which also includes the option to check structures for their geometric validity. Final set of stereoisomers were enumerated using the RDKit's `EnumerateStereoisomers` module.[7]

[5] https://github.com/PatWalters/solubility
[6] https://chemaxon.com/
[7] https://www.rdkit.org/docs/source/rdkit.Chem.EnumerateStereoisomers.html

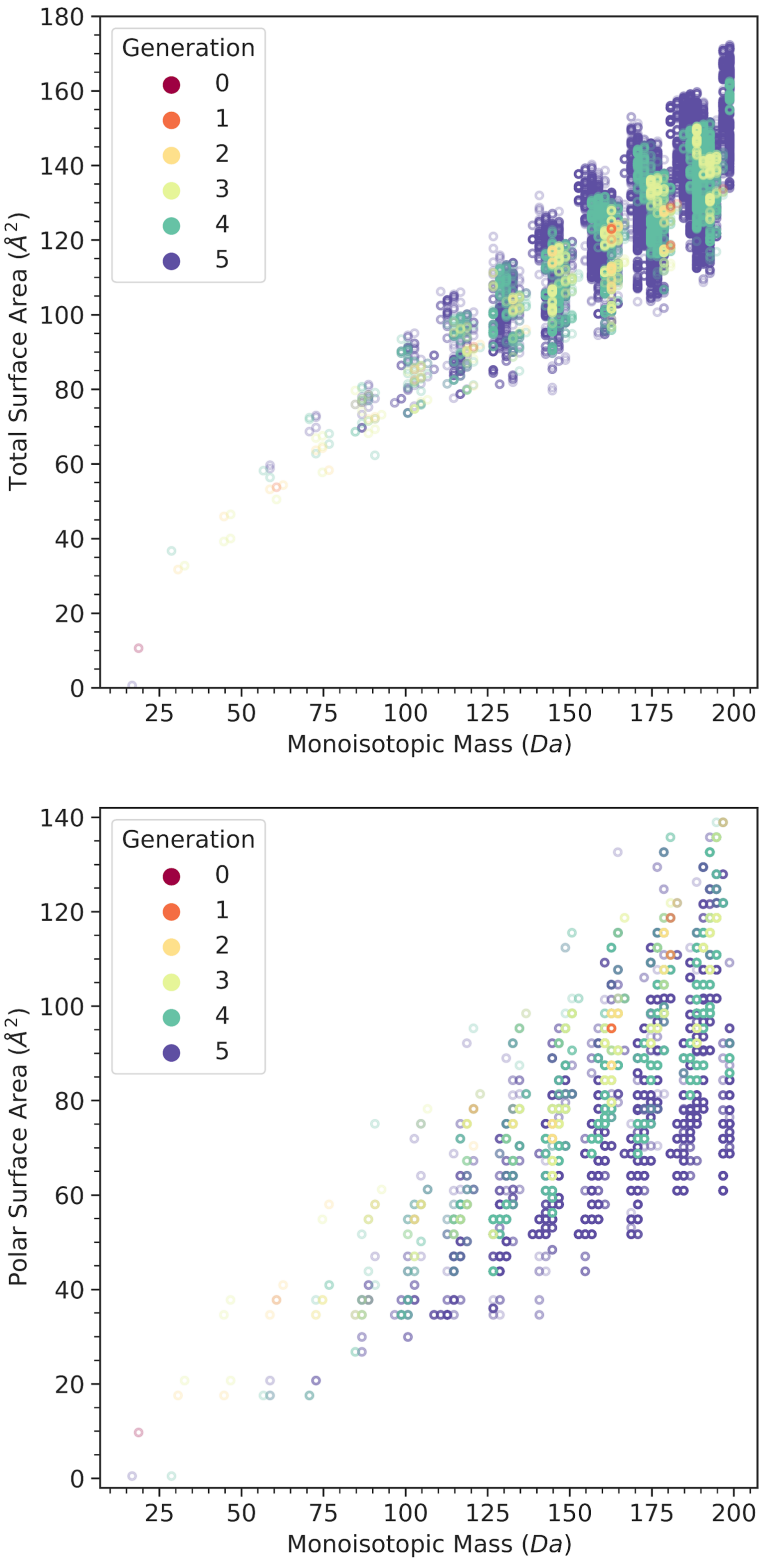

Fig. SI14: **Top**: Estimated Total Surface Area (TSA) as a function of mass and generation in the computed ADG products. **Bottom**: Estimated Polar Surface Area (PSA) as a function of mass and generation in the computed ADG products (Stanton et al. 2002)

.

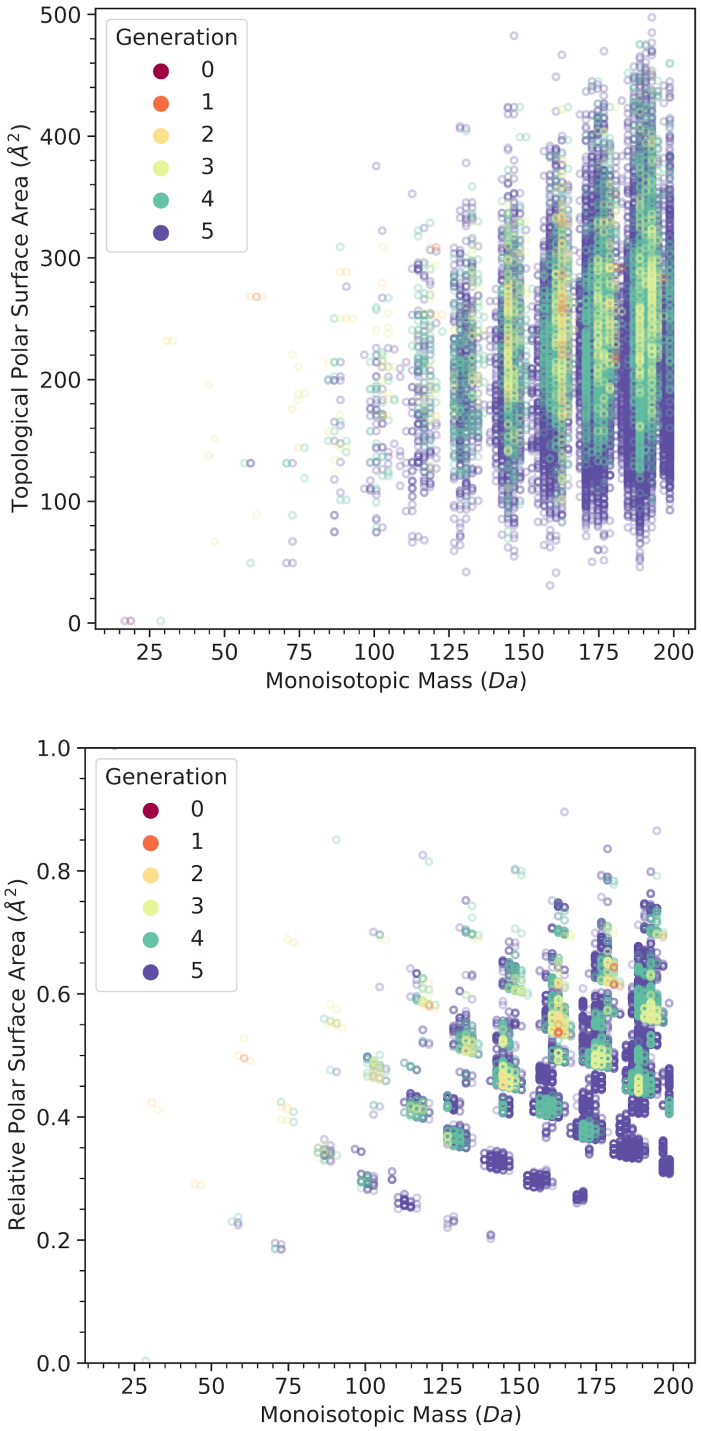

Fig. SI15: **Top**: Estimated Topological Polar Surface Area (TPSA) as a function of mass and generation in the computed ADG products. TPSA is defined as the sum of solvent accessible surface areas of atoms with absolute value of partial charges greater than or equal 0.2 (Stanton et al. 2002). **Bottom**: Estimated Relative Polar Surface Area (RPSA) as a function of mass and generation in the computed ADG products. RPSA is defined as a value of Topological Polar Surface Area divided by Total Surface Area (Stanton et al. 2002)

.

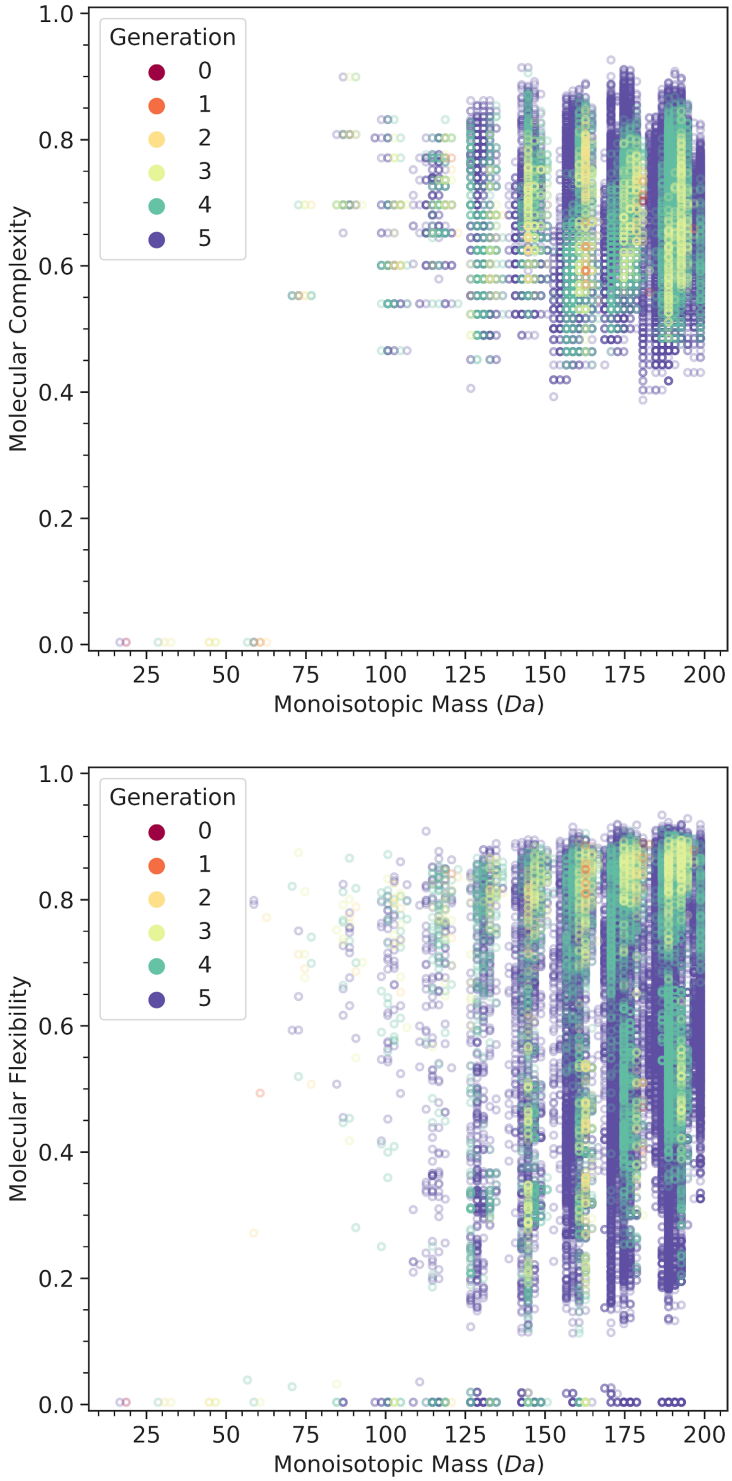

Fig. SI16: **Top**: Estimated Molecular Complexity as a function of mass and generation in the computed ADG products (von Korff and Sander 2019). **Bottom**: Estimated Molecular Flexibility (from 0, completely rigid, to 1, completely flexible) as a function of mass and generation in the computed ADG products (Kier 1989).

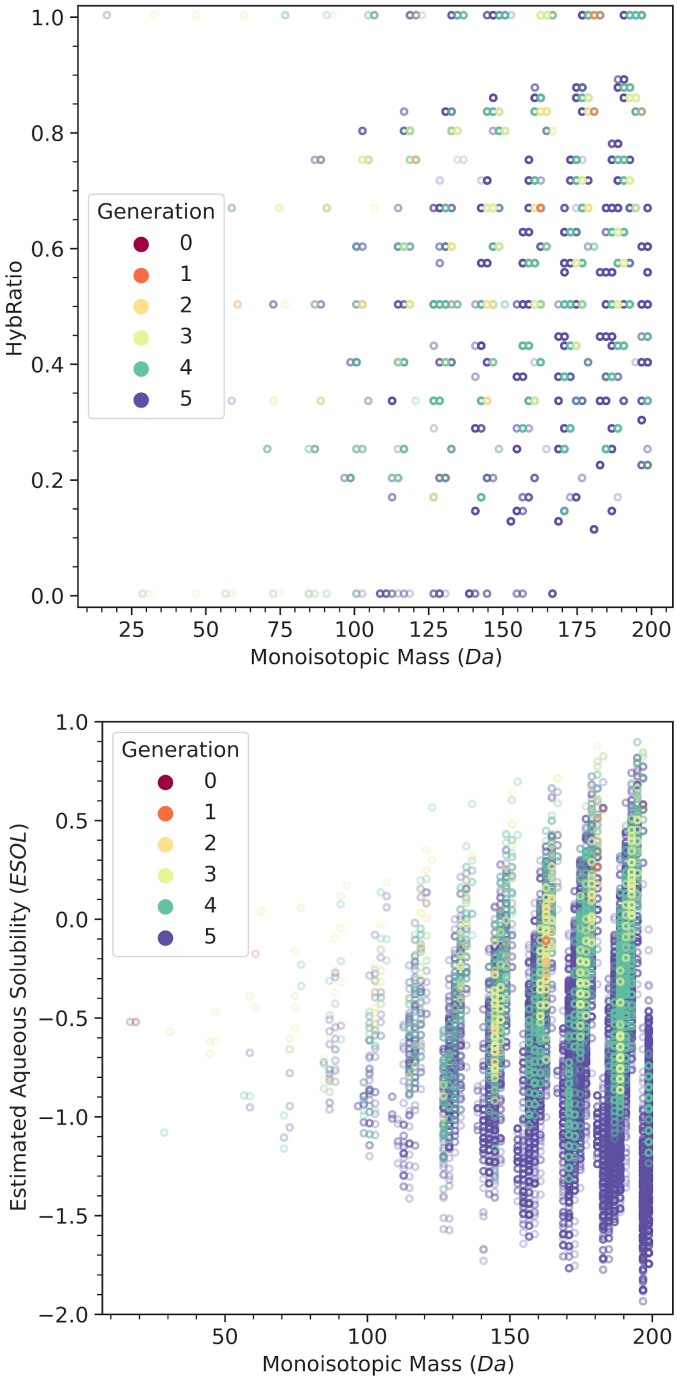

Fig. SI17: **Top**: The generational diversity of HybRatio as a function of generation in the ADG CRNR. HybRatio is defined as the ratio of sp^3 hybridized carbon atoms in a molecule over the sum of sp^2 and sp^3 hybridized carbon atoms ($N_{sp^3}/(N_{sp^2} + N_{sp^3})$). This was computed using CDK (Willighagen et al. 2017). **Bottom**: Estimated Aqueous Solubility (ESOL) in mol/L as a function of mass and generation in the computed ADG products. ESOL is the estimated aqueous solubility of a molecule inferred from its structure (Delaney 2004). Higher values indicate higher aqueous solubility. ADG products tend towards lower aqueous solubility as a function of generation.

Principal Component Analysis (see Figures SI18,SI19) was conducted using the Python library scikit-learn (Pedregosa et al. 2011). Computed descriptors including Monoisotopic Mass, H-Acceptors, H-Donors, Druglikeness, Shape Index, Rotatable Bonds, Acidic Oxygens, Van der Waals-Surface, Van der Waals-Volume, Number of StereoIsomers, Number of Carbons, Number of Hydrogens, Number of Oxygens, H/C, O/C, IC0, SIC0, CIC0 and DGph10, the free energy of formation at pH 10 of the molecules computed using Equilibrator (see below) were used as variables. Modularity is a measure of network structure (Newman 2006). It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). There are dense connections between nodes within modules but sparse connections between nodes in different modules (see Fig. SI20). As can be seen in Figures SI18 and SI19, a strong correlative relationship exists between network properties (modularity classes) and the cheminformatic descriptor of each compound in the network. For example, modularity classes "c" and "d" are characterized by molecules with high flexibility, high number of carbon atoms and overall greater nonpolar character. Modularity classes "a" and "b" are associated with molecules with high complexity, high acidic oxygen content and high polarity. Comparing Fig. SI18 with Fig. **3A**, it is evident that modularity class "d" is mostly composed of molecules with high clogP produced in generation 5. Modularity class "e" is not visibly clustered but rather distributed across the PCA space (see Figures SI18, SI21 and SI22).

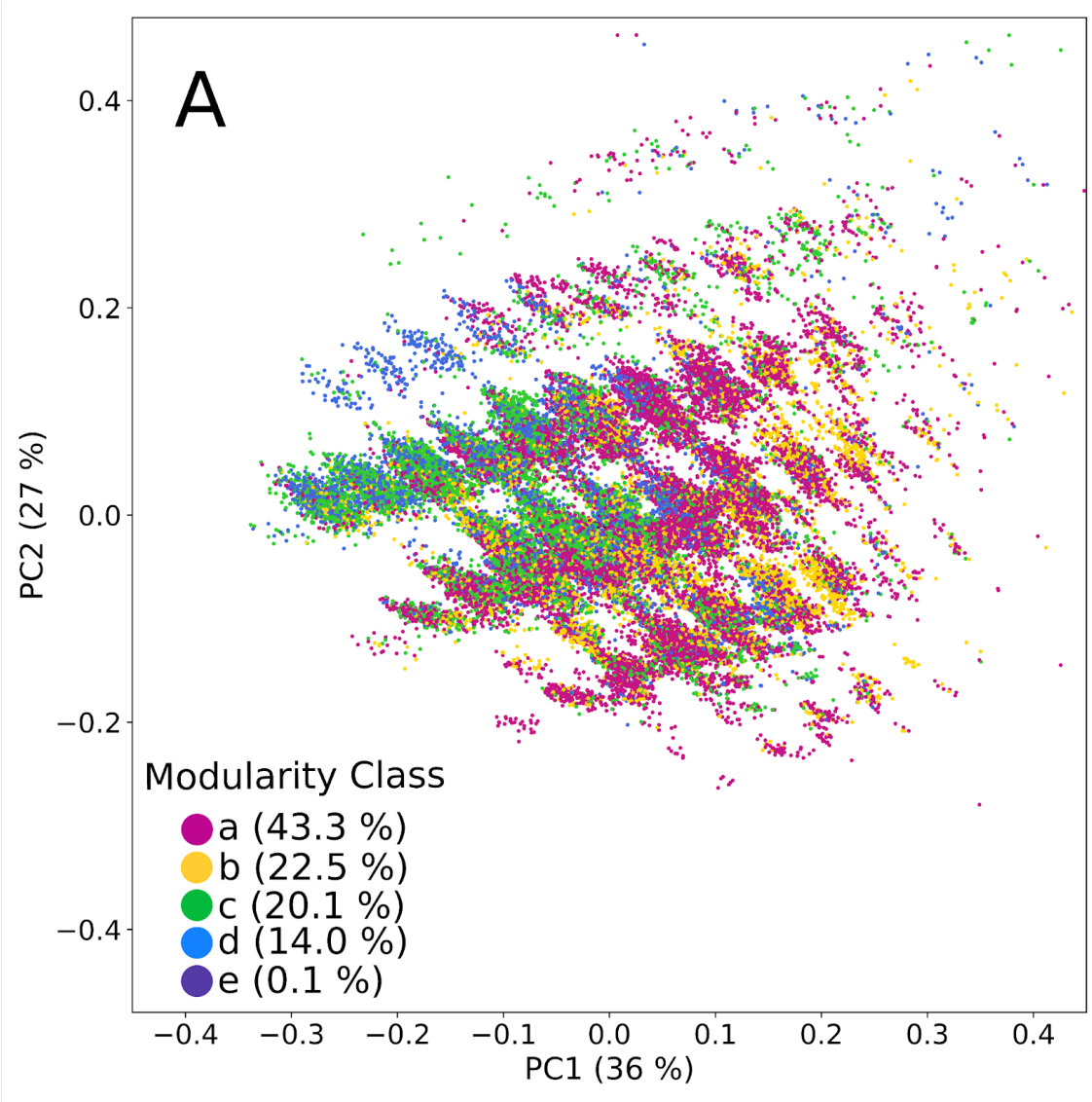

Fig. SI18: Principal component analysis (PCA) of the cheminformatic descriptor variables colored by modularity class computed using the methods described in (Blondel et al. 2008) and visualized using Gephi (Bastian et al. 2009).

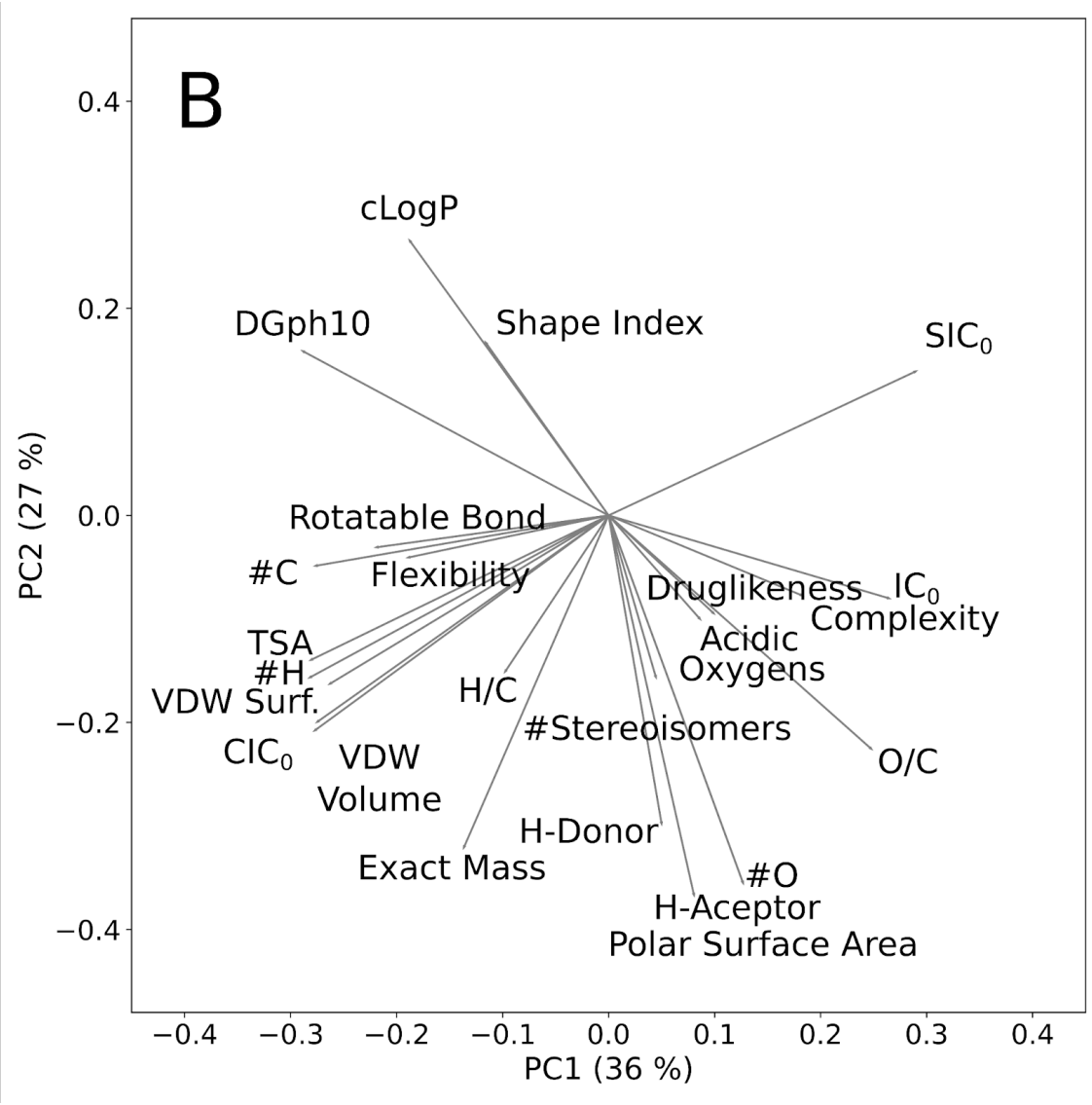

Fig. SI19: PCA loading plot showing how strongly each descriptor influences component identification. IC0, SIC0 and CIC0 are information theoretic indices with neighborhood order equal to zero. IC represents the Information Content of a molecule graph, SIC is related to the Structural Information Content and CIC is the complementary information content that is defined as the difference between the maximum possible complexity of a molecule graph and its Information Content (IC). A more detailed definition of these descriptors can be found in (Basak et al. 2000). There is a positive correlation among the modularity class and cLogP/ΔG(pH = 10) and ΔG(pH = 10)/Shape Index. There is also a positive correlation among the Number of Stereoisomers and the Number of Oxygen Atoms, Polar Surface Area and Number of H-Acceptors or H-Donors. In the vertical direction the observation apparently clusters in relation to Molecular Polarity and Exact Mass (PC2 Component). The PC1 component is more related to the Molecular Complexity descriptor and leads to the formation of sub clusters related principally to the number of carbon atoms and the number of rotatable bonds or the flexibility of the molecule. There is also a positive correlation among Molecular Complexity, IC0 and O/C Ratio.

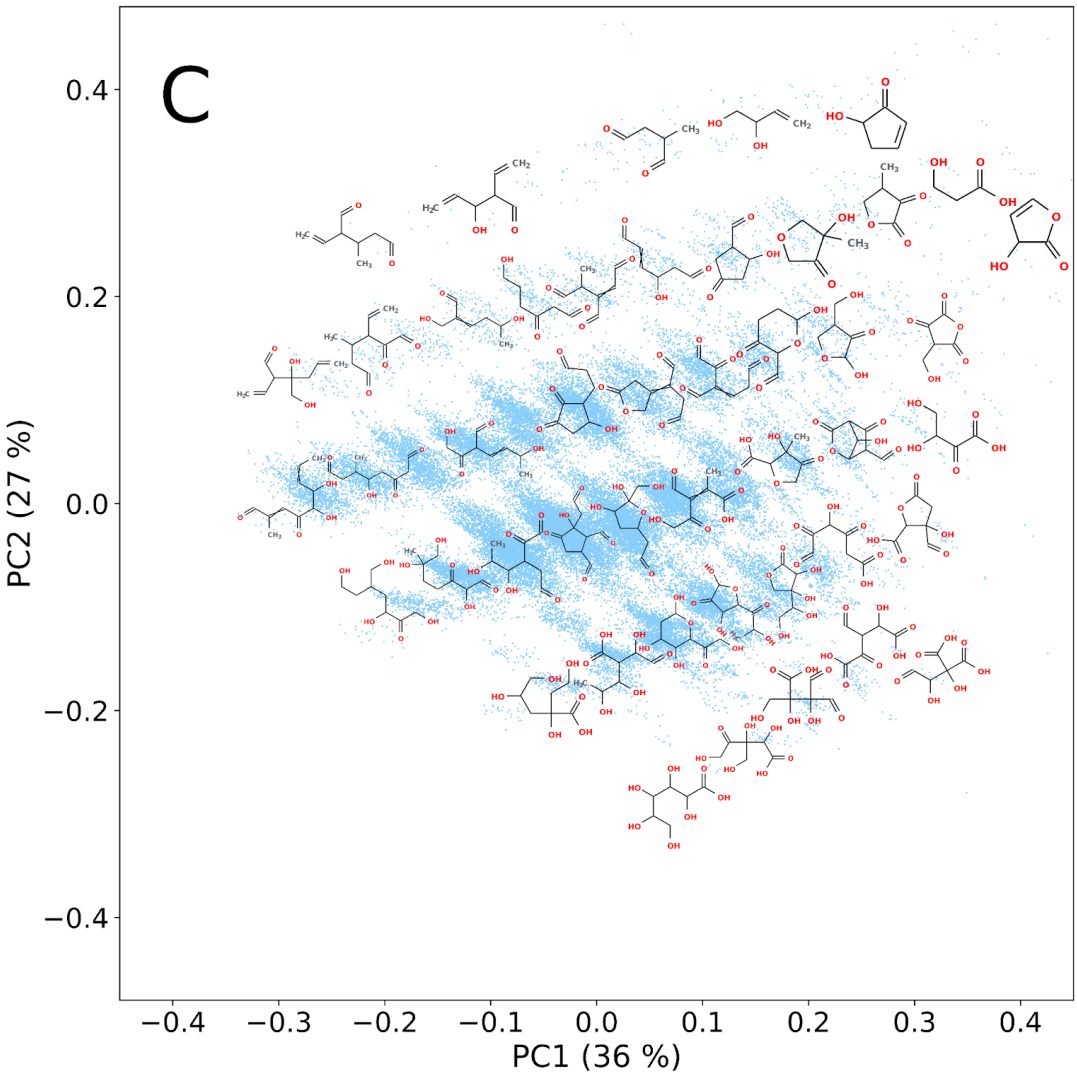

Fig. SI20: Representative structures of the molecules in each cluster. A general trend of decreasing flexibility from left to right is evident, as is increasing polarity from top to bottom.

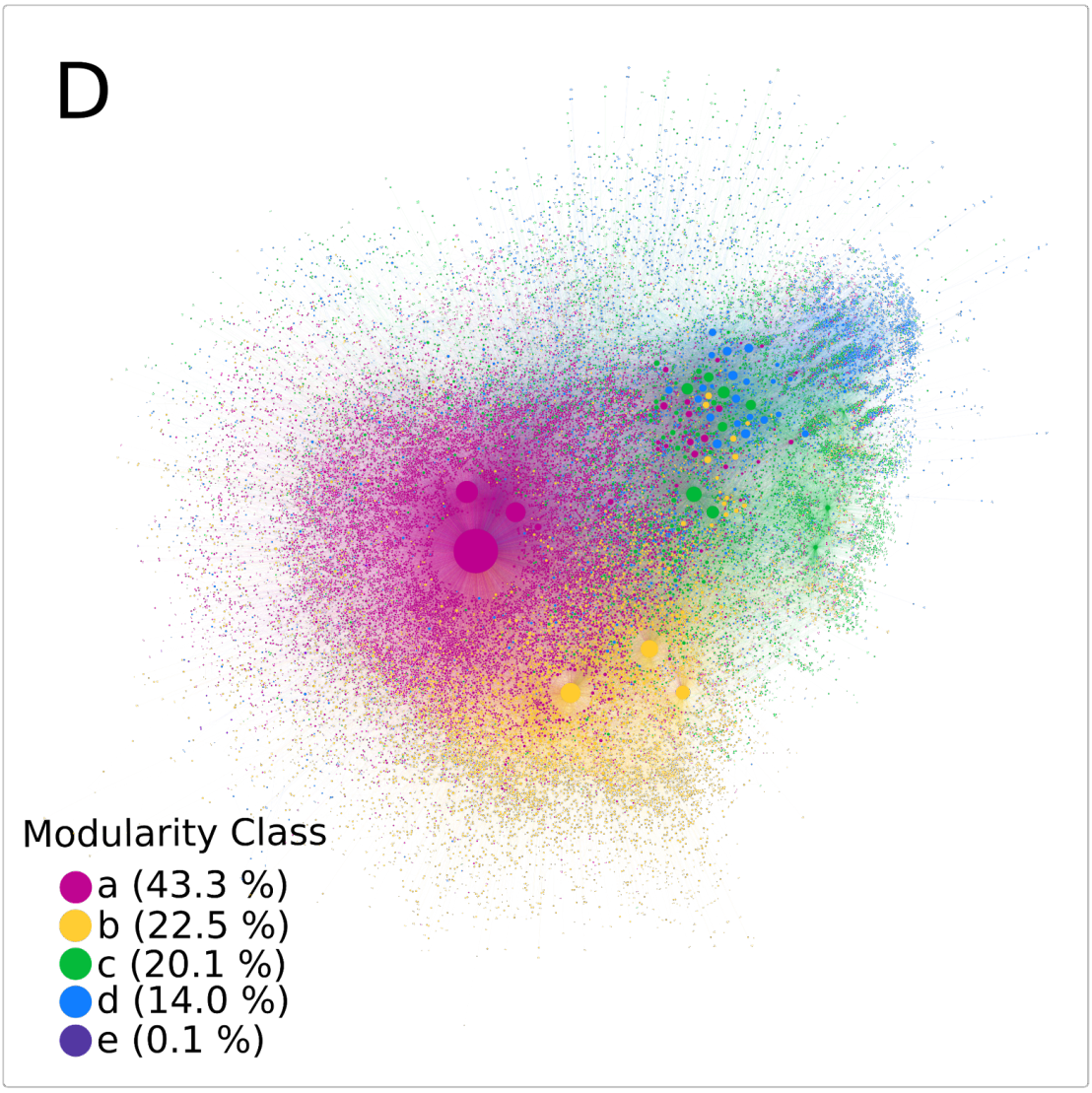

Fig. SI21: The molecules in the ADG CRNR can be grouped into five natural modularity classes using Gephi (Bastian et al. 2009) with a resolution of ~ 2.8. Modularity analysis was carried out with different resolution values and usually all gave four large groups (a-d), as well as other small subgroups. Modularity class e here contains several subgroups that do not cluster at this level of resolution (Lambiotte et al. 2008) (In this plot node size is proportional to in-degree).

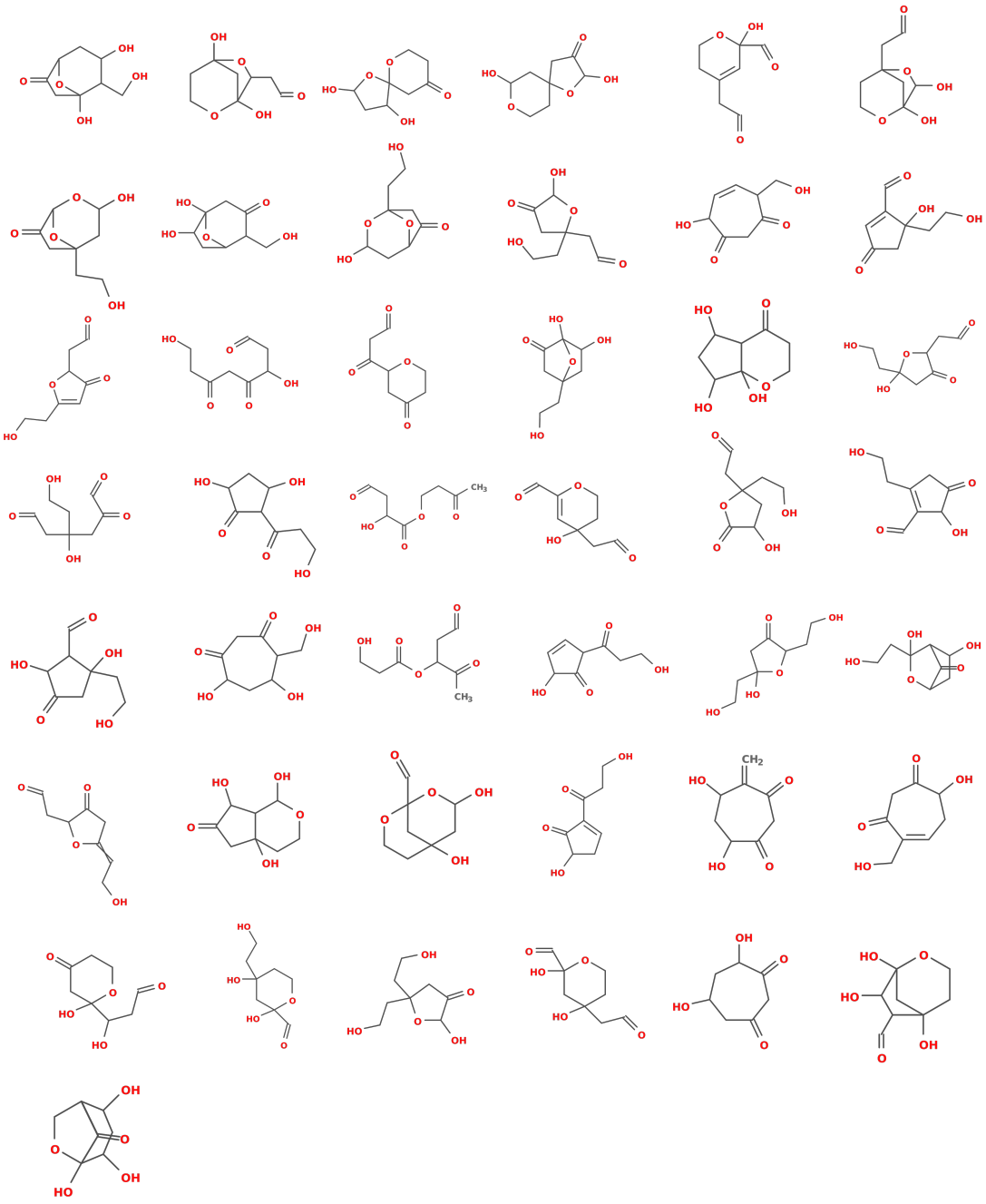

Fig. SI22: Examples of compounds present in modularity class e.

3.9 Thermochemical Calculations

We used the eQuilibrator API to calculate each compound's standard free energy of formation, $\Delta_f G'^{\circ}$, and the change in free energy for each reaction, $\Delta rG'^{\circ}$ (Noor et al. 2013), (Beber et al. 2022), estimated at standard conditions, meaning that all species are at 1 M concentration at 25° C and 1 bar (see Figure SI12). The eQuilibrator API enumerates the list of compounds from the Kyoto Encyclopedia of Genes and Genomes (KEGG, (Kanehisa and Goto 2000)). These calculations assume a physiological pH (7.4) and 0.1 M ionic strength. We used the eQuilibrator API default value for the Mg^{2+} concentration (pMg, in this case pMg = 3.0), which is required

for these calculations, but which mostly affects reactions involving ATP (Beber et al. 2022). We explored the eQuilibrator API calculations at pH 2, 7.4 and 10 (see SI File 4). The eQuilibrator software uses component contribution methods to estimate thermodynamic values. Temperature and pressure cannot be varied, because the model calibration relies on experimental measurements for $\Delta_f G$ conducted at this T and P (Beber et al. 2022). The calculation of thermodynamic properties from the group contribution method also relies on the existence of similar groups in the training data (Beber et al. 2022). It was not possible to calculate thermodynamic values for a small subset of the compounds generated in the ADG CRN produced here (i.e., 14 of 48,403 compounds ($\sim$ 0.3%), mostly acid anhydrides, involved in 305 of 100,268 (again $\sim$ 0.3%) of the network's reactions).

3.10 Detection of Autocatalytic Motifs

Autocatalysis is a phenomenon in which a reaction or sequence of reactions produces a compound catalyzed by one or more products of the reaction(s) in a system (Blokhuis et al. 2020; Gánti 1975; Peretó 2012; Preiner et al. 2019; Andersen et al. 2021). Kinetically, autocatalysis leads to super-linear growth of a species or network. However, as discussed in (Andersen et al. 2021), for a CRN, computationally identifying whether a compound produces more than one copy of itself is not straightforward. Although formalized definitions of autocatalysis that can be generally applied to CRNs remains to be defined (Andersen et al. 2021), we here searched for topologically formal autocatalytic motifs, i.e., for those in which a cyclic set of reactions produces two or more copies of one of the cycle component compounds, as inferred from the network topology. Fig. SI23 shows examples of autocatalytic motifs provided in the prior literature.

We used two main approaches to identify autocatalytic cycles. (1) an imperative programming approach, in which a graph algorithm was written to find motifs, using the Ford-Fulkerson algorithm (Ford and Fulkerson 1956); (2) a declarative programming approach, in which the target pattern is described in a query and a graph engine finds the pattern, for example using a graph database management system and its query language (a benefit of the declarative approach is that one does not need to write a graph algorithm to match a specific search pattern). There is also a third method that we will not address here but that is related to a linear algebra approach, where a current matrix is evaluated to find flows in the network (see Clarke 1975; Kolar-Anić et al. 2010; Čupić et al. 2018).

The Ford-Fulkerson algorithm can be used to find a generalized structure of the autocatalytic motifs defined in Fig. SI23. This algorithm seeks to know if a given molecule can be a catalytic node within an autocatalytic cycle motif. The algorithm begins by defining the source node, and then performs a search for all reaction nodes (target nodes) that can be reached starting from the source node. For each target node found, it asks if there are at least two return routes to the source node (branches 1 and 2, see Fig. SI23). These return routes must be unique routes (that is, branches 1 and 2 only share source and target nodes). In branch 0 all possible routes are evaluated. However, in branches 1 and 2, only one of the possible branches 1 or 2 is taken. This selection is not random since it is based on the adjacency list initially constructed by reading the network of reactions in an ordered way, which associates each node with a series of successors ordered as they appear in the network exploration. Thus, branches 1 and 2 are both the first paths of all possible return paths to the catalytic node. This allows the following variables to be further extracted from an autocatalytic motif:

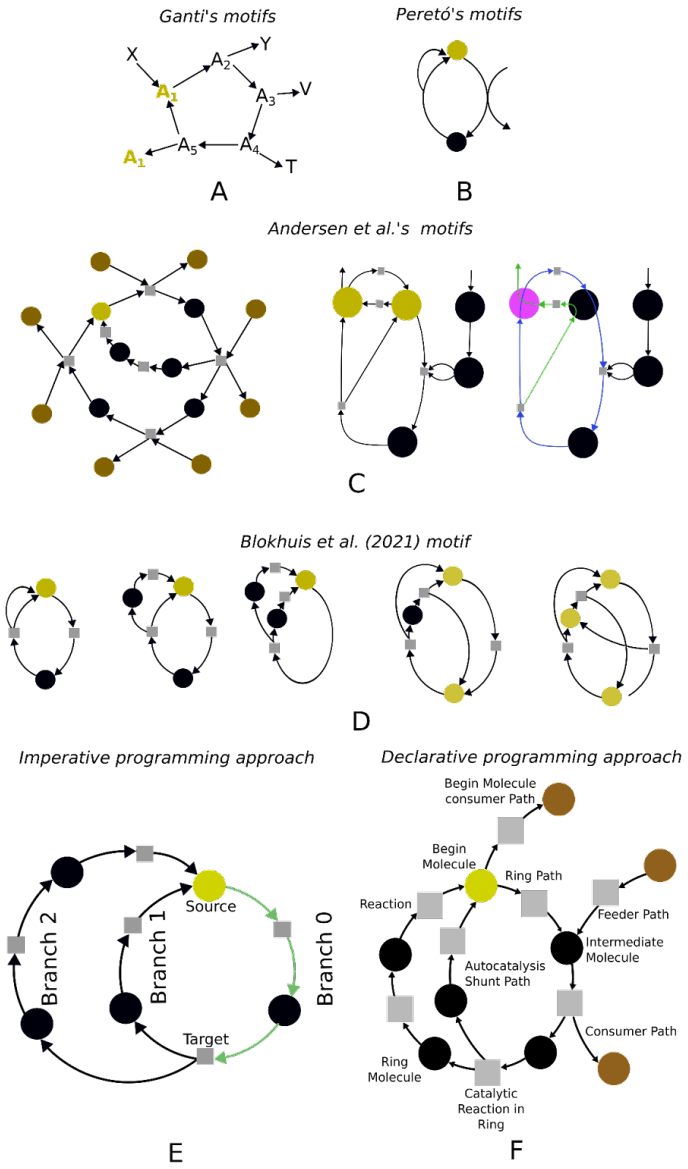

Fig. SI23: Autocatalytic cycle definitions proposed by Gánti (1975); Peretó (2012); Andersen et al. (2021) and Blokhuis et al. (2020). Circles represent molecules and squares reactions. Autocatalytically produced molecules are highlighted in yellow. Brown circles are external molecules to the cycle (e.g., feeder molecules or output molecules). Purple circles represent an "exclusive autocatalytic node" (see Andersen et al. 2021), which takes into account the edges as flows allowing for a net production of that molecule. In Fig. SI23E, the autocatalytic motif used in the imperative programming approach is illustrated. The source node represents the autocatalytic molecule and the target node represents the autocatalytic reaction. In Fig. SI23F, the autocatalytic motif used in the declarative programming approach is illustrated.

1. the thermodynamic spontaneity (as defined using the eQuilibrator API (Beber et al. 2022)) of each defined autocatalytic cycle

2. the frequency of autocatalytic cycles as a function of cycle length

3. the molecules most commonly used as cycle feedstocks, and

4. the molecules most commonly produced by identified autocatalytic cycles.

(1) is obtained using the methods described in the Thermochemical Calculations section and the cycle length;
(3) is defined as the maximum path between `Length_0` + `Length_1` and `Length_0` + `Length_2`.

For the declarative approach, the target pattern shown at the bottom of Fig. SI23 was formulated generally so that a Cypher Query Language (CQL) query could be built to find instances of this pattern in the network by executing the query in the graph database management system Neo4j. A key feature of this pattern is that the shunt pathway ensures that the cycle satisfies the definition of topologically formal autocatalysis described (Andersen et al. 2021), though without flow constraints to ensure mass balance.

This method defines potential autocatalytic reaction sequences as cyclic reaction sequences which lead back to defined starting molecules, a "ring," attached to at least one bridging reaction sequence which connects two compounds in the ring, which we call here a "shunt." The query begins at a defined "begin molecule:, where the "ring path" pattern is defined by starting and ending at "begin molecule", with a user-defined number of intermediate reactions ("ring size") to travel before arriving back at the "begin molecule". However, this rule is overly simple, as it allows for the ring to intersect itself, e.g., in Fig. SI23 producing the "pinched" pattern (with one interactional node) instead of the desired ring structure (with two interactional nodes). To avoid this, an additional constraint on the number of edges relative to vertices was added such that the number of edges must be equal to two times the number of molecules in the ring minus one, to ensure that the ring doesn't form pinch points.

The rationale behind adding the shunt path is that such shunts ensure that the target "begin molecule" is generated in a stoichiometric quantity greater than one, which is a minimal definition of self-amplification. The query for the shunt sub-structure defines that a reaction in the ring, "catalytic reaction in ring", must have a pathway to reach the "begin molecule", and must be located in the ring as a user-defined distance from the "begin molecule".

Further constraints were added to ensure that molecules in the shunt path do not appear in the ring path molecules, otherwise the shunt would not be distinct from the ring. A "feeder path" was added to the target pattern definition to allow the cycle to receive external material as feedstock, as also shown in the Peretó (2012) definition shown in Fig. SI23. A "consumer path" was added to the pattern match, similar to the definition in (Andersen et al. 2021), to allow for identification of molecules produced from the activity of possible autocatalytic cycles. Constraints were added such that the reactions and molecules in the shunt cannot be in the ring structure.

Finally, the search constraint "begin molecule consumer path" was added to the query to see how the "begin molecule" was consumed in the CRN since the "begin molecule" is presumably also produced in possibly larger stoichiometric quantities. Similar constraints were also added to ensure that the "begin molecule consumer" was distinct from the "begin molecule." By adding these constraints to the pattern match definition, autocatalytic cycle identification becomes more interesting from the perspective of network theory, and fewer and more relevant network cycles

are found by the search algorithm.

A data loader routine was written in Python to import the MØD-generated chemical network data into Neo4j so that pattern match queries could be built. The loader routine ensured that no distinct molecule or reaction node was repeated, which gave a high-level representation on the molecule and reaction node (bipartite, or two-label graph per Neo4j) graph. Ensuring node distinctness allowed the network model to be lighter to improve computational performance since the pathways were being examined, and not the raw quantity of molecules in the network. The network model could then be thought of as a model of every distinct pathway extracted from the simulated chemical network from MØD (whose network model is physically closer to the actual chemical network representation, since this model is a level of abstraction on the MØD network). As the database populates, additional information, such as when the molecule was first generated, is stored in the .txt files. If a molecule is found again in a subsequent generation, that molecule's node is not re-recorded. Similarly, if a reaction of a certain rule already exists between molecules, that edge is not repeated. Based on the rules of the loader script, the Neo4j graph model is a species-only graph, rather than the bipartite molecule and reaction graph. The bipartite graph is useful for describing complex patterns that require constraints on the relationships between reactions and molecules, while the species-only graph is useful for network centrality measures.

The pattern match query is user-defined, and thus modifiable. The query template can further be improved to allow modification of the queried length of the shunt pathway. When querying for patterns, it was noticed that the same patterns, albeit starting with different molecules in the same cycle, were often returned each time the query was executed, which we ascribe to Neo4j's indexing system. To circumvent this, and to sample more patterns across the database, a random shuffle on the MØD data exports was implemented before importing to Neo4j to shuffle the resulting index of the molecule or reaction nodes and edges so that a new sample of the network could be taken each time. Then, the loader routine (consisting of clearing the Neo4j database, reading the MØD export data, shuffling the data and importing to the database) was repeated 10 times (this can be configured by global variables by the user) to gather cycle data from various starting points in the network.

The pattern matches were exported to a tabulated CSV file where several columns of information, e.g., the contents of each substructure path, the SMILES string of each molecule in the pattern, reaction information on each node, etc., were included for better visualization.

Neo4j can also be used to compute various network statistics, including centrality measures for example: eigenvector and betweenness centrality, random-walk betweenness and node-degree metrics example: count of incoming or outgoing edges and node degree rank by generation, and the change in node degree by generation which may help assess when certain molecules become most influential in affecting network development.

3.11 Network Data Visualization

The `Gephi` software package[8] (Bastian et al. 2009) was used to visualize the generated reaction networks. There are two principal network representations widely used to investigate chemical

[8] https://gephi.org

reaction networks: Species-species and species-reaction representations (Fig. SI24; Sandefur et al. 2013). Species-reaction representations are information-dense because they contain information about the stoichiometry and identity of specific reactants. Species-species representations lose information about stoichiometry and the kinds of reactions that connect products and reactants.

Gephi provides information in the species-reaction representation using thick edges to take molecularity into account. For example, if a reaction produces or consumes two molecules of the same compound, a thick edge connects that compound to the reaction node. If a molecule is both a reactant and a product in a specific reaction, Gephi represents that reaction as an edge that returns from the reaction node to that molecule's node (see Fig. SI24).

In Gephi plots, node size can be made proportional to any graph property for emphasis, for example the in-degree or out-degree of the nodes. Here, node sizes were scaled using a spline interpolation (Sharma and Meir 1966; de Boor 1968) function in Gephi. The spline was set as shown in Fig. SI24, which gives good size definition to large values and allows intermediate values to achieve some representative degree of importance.

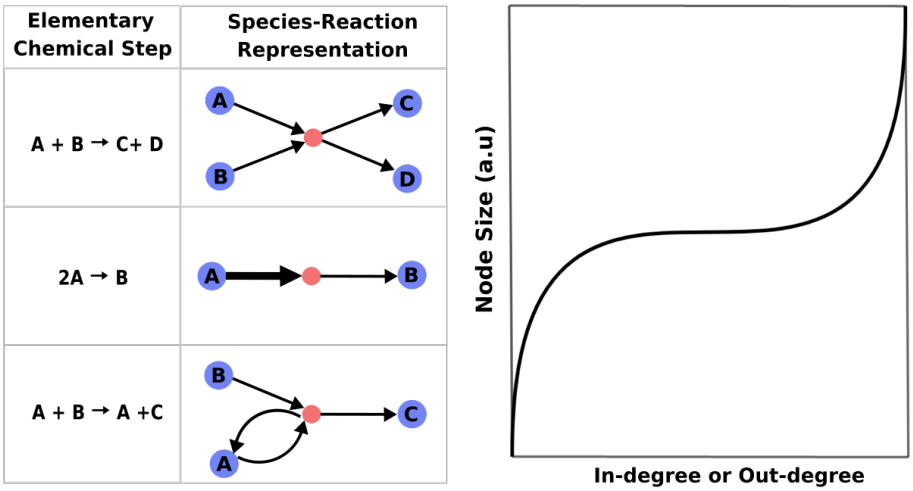

Fig. SI24: **Left:** Examples of Gephi's Species-Reaction representations, which preserve information regarding the molecularity of the reaction. **Right:** Gephi's spline interpolation function used to set the size of the nodes as a function of in-degree or out-degree.

Plot layouts were rendered using Gephi's ForceAtlas2 function by setting the gravity parameter to 0.05 and scaling parameter to 4.0, and running this for 10 minutes without avoiding overlapping and then 20 minutes avoiding overlapping. ForceAtlas2 simulates a physical system in which nodes repulse each other like charged particles, while edges attract nodes like springs, to represent a network (Jacomy et al. 2014).

4 SI Tables

Table SI1: The number of products and reactions produced per generation as a function of computation time. *The 'number of edges' presented here is equal to the number that should appear in the species representation of the network.

Generation	Cumulative number of species	Cumulative number of reaction	Differential number of edges*	Computation time
0	2	—	—	—
1	18	186	28	0.153 seconds
2	128	2192	396	1.851 seconds
3	1054	17,205	3302	59.441 seconds
4	7891	28,483	27,845	1 hour, 5 minutes
5	48,403	73,349	180,962	2 days, 20 hours

Table SI2: A list of all reaction rules encoded in our workflow.

Aldol Condensation	Beta-gamma Unsaturated Acid Decarboxylation
Benzilic Acid Rearrangement	Mannich Addition
Benzilic Acid Rearrangement (inverse)	Mannich Addition (reverse)
Retro Aldol	Hemiacetal Formation for 5 membered rings
Keto-enol migration twice	Hemiacetal Formation for 6 membered rings
Amadori/Heyns Rearrangement	Hemiacetal Formation for 7 membered rings
Ester Hydrolysis	Hemiacetal Formation for 5 membered rings, inverse
Cannizarro 1	Hemiacetal Formation for 6 membered rings, inverse
Cannizarro 2, glucose (oxidation)	Hemiacetal Formation for 7 membered rings, inverse
Cannizarro 2, glucose (reduction)	Imine to Carbonyl
Alkyne Addition, N, HH	Amidine to Amide Hydrolysis
Alkyne Addition, N, HC	Imine to Enamine
Alkyne Addition, N, CC	Nitrile Hydrolysis
Alkyne Addition, S, H	Transamination
Alkyne Addition, S, C	Transamination (inverse)
Alkyne Addition, C	Alpha-Keto Acid Decarboxylation
Knoevenagel C	Beta Decarboxylation
Knoevenagel H	Nitrile Ring Closure 5
Knoevenagel C (inverse)	Nitrile Ring Closure 6
Knoevenagel H (inv)	Ring Closure 5 membered O, O
DAMN Scission	Ring Closure 5 membered O, N
Cyanogen Ammonolysis	Ring Closure 5 membered O, S
Cyanamide Hydration	Ring Closure 5 membered N, O
Carbamylation, SC	Ring Closure 5 membered N, N
Carbamylation, SH	Ring Closure 5 membered N, S
Carbamylation, N, HH	Ring Closure 5 membered S, O
Carbamylation, N, HC	Ring Closure 5 membered S, N
Carbamylation, N, CC	Ring Closure 5 membered S, S
Cyanate from cyanogen	Ring Closure 6 membered O, O
Ketone-Catalyzed Decarboxylation	Ring Closure 6 membered O, N
Amadori Rearrangement	Ring Closure 6 membered O, S
Thial ->Thiene	Ring Closure 6 membered N, O
Sulfide to Nitrile Addition	Ring Closure 6 membered N, N
Thioamidine to Amide Hydrolysis	Ring Closure 6 membered N, S
Ketone/Aldehyde to Thioketone/Thial, HH	Ring Closure 6 membered S, O
Ketone/Aldehyde to Thioketone/Thial, HC	Ring Closure 6 membered S, N
Ketone/Aldehyde to Thioketone/Thial, CC	Ring Closure 6 membered S, S
HCN to Ketone	Ring Closure 7 membered O, O
Strecker Degradation Dicarbonyl, H, H, H, H	Ring Closure 7 membered O, N
Strecker Degradation Dicarbonyl, H, H, H, C	Ring Closure 7 membered O, S
Strecker Degradation Dicarbonyl, H, H, C, H	Ring Closure 7 membered N, O

Table SI2 continued from previous page

Strecker Degradation Dicarbonyl, H, H, C, C	Ring Closure 7 membered N, N
Strecker Degradation Dicarbonyl, H, C, H, H	Ring Closure 7 membered N, S
Strecker Degradation Dicarbonyl, H, C, H, C	Ring Closure 7 membered S, O
Strecker Degradation Dicarbonyl, H, C, C, H	Ring Closure 7 membered S, N
Strecker Degradation Dicarbonyl, H, C, C, C	Ring Closure 7 membered S, S
Strecker Degradation Dicarbonyl, C, H, H, H	Ring Closure 5 membered, inverse O, O
Strecker Degradation Dicarbonyl, C, H, H, C	Ring Closure 5 membered, inverse O, N
Strecker Degradation Dicarbonyl, C, H, C, H	Ring Closure 5 membered, inverse O, S
Strecker Degradation Dicarbonyl, C, H, C, C	Ring Closure 5 membered, inverse N, O
Strecker Degradation Dicarbonyl, C, C, H, H	Ring Closure 5 membered, inverse N, N
Strecker Degradation Dicarbonyl, C, C, H, C	Ring Closure 5 membered, inverse N, S
Strecker Degradation Dicarbonyl, C, C, C, H	Ring Closure 5 membered, inverse S, O
Strecker Degradation Dicarbonyl, C, C, C, C	Ring Closure 5 membered, inverse S, N
Schiff Tautomerization	Ring Closure 5 membered, inverse S, S
Alkyne Addition, N, HH, inverse	Ring Closure 6 membered, inverse O, O
Alkyne Addition, S, H, inverse	Ring Closure 6 membered, inverse O, N
Alkyne Addition, C, inverse	Ring Closure 6 membered, inverse O, S
DAMN Scission, inverse	Ring Closure 6 membered, inverse N, O
Cyanogen Ammonolysis, inverse	Ring Closure 6 membered, inverse N, N
Cyanamide Hydration, inverse	Ring Closure 6 membered, inverse N, S
Carbamylation, SC, inverse	Ring Closure 6 membered, inverse S, O
Carbamylation, SH, inverse	Ring Closure 6 membered, inverse S, N
Carbamylation, N, HH, inverse	Ring Closure 6 membered, inverse S, S
Cyanate from cyanogen, inverse	Ring Closure 7 membered, inverse O, O
Ketone-Catalyzed Decarboxylation, inverse	Ring Closure 7 membered, inverse O, N
Amadori Rearrangement, inverse	Ring Closure 7 membered, inverse O, S
Thial ->Thiene, inverse	Ring Closure 7 membered, inverse N, O
Sulfide to Nitrile Addition, inverse	Ring Closure 7 membered, inverse N, N
Thioamidine to Amide Hydrolysis, inverse	Ring Closure 7 membered, inverse N, S
Ketone/Aldehyde to Thioketone/Thial, HH, inverse	Ring Closure 7 membered, inverse S, O
HCN to Ketone, inverse	Ring Closure 7 membered, inverse S, N
Strecker Degradation Dicarbonyl, H, H, H, H, inverse	Ring Closure 7 membered, inverse S, S
5 Membered Ring Exoamine Hydrolysis	Amine Nitrile Ring Closure 5
6 Membered Ring Exoamine Hydrolysis	Amine Nitrile Ring Closure 6
Ammonolysis of Esters	Amine Imine Ring Closure 5
Isocyanate Hydrolysis	Amine Imine Ring Closure 6
Cyanamidation, SC	Amidine Nitrile Ring Closure 6
Cyanamidation, SH	Amide Nitrile Ring Closure 6
Cyanamidation, N, HH	Amine Carbonyl Ring Closure, H, 5
Cyanamidation, N, HC	Amine Carbonyl Ring Closure, C, 5
Cyanamidation, N, CC	Amine Carbonyl Ring Closure, H, 6

Table SI2 continued from previous page

Nitrile Amination	Amine Carbonyl Ring Closure, C, 6
HCN Addition to Nitriles	Michael Addition 0,0,
CN addition to iminonitriles	Michael Addition 0,1,
CN addition to aminodinitriles	Michael Addition 0,2,
Amide Hydrolysis	Michael Addition 1,0,
Amide Formation Hydrolysis, C	Michael Addition 1,1,
Alkene Addition, CN	Michael Addition 1,2,
Elimination + enol to keto	Michael Addition 2,0,
Elimination2	Michael Addition 2,1,
Dehydration of Amines	Michael Addition 2,2,
Deydration of Amines (inverse)	Michael Addition 3,0,
Hydration of C(=O)C	Michael Addition 3,1,
Hydration of C=C(O)	Michael Addition 3,2,
Enamine Hydration and Elimination	Michael Addition 0,0, (reverse)
Deamination of Vicinal Diamine	Michael Addition 0,1, (reverse)
Deamination of Vicinal Diamine, (inverse)	Michael Addition 0,2, (reverse)
Deamination 2	Michael Addition 2,0, (reverse)
Deamination 3	Michael Addition 2,1, (reverse)
Deamination 3 (inverse)	Michael Addition 2,2, (reverse)
Cyanohydrin Formation	Strecker Degradation, HH
Strecker Synthesis	Strecker Degradation, HC
Alpha-Beta Unsaturated Acid Decarboxylation	Strecker Degradation, CC

Table SI3: Counts of reaction rules implemented per generation in the generated network.

Rule	**Gen 1**	**Gen 2**	**Gen 3**	**Gen 4**	**Gen 5**
Aldol Condensation	0	6	417	3128	14874
Elimination2	0	6	74	986	9111
Knoevenagel H	0	6	347	2197	8248
Retro Aldol	1	16	97	1180	7901
Knoevenagel C	0	5	277	2103	7528
Elimination + enol to keto	8	67	224	1437	7233
Michael Addition 0,2,	0	0	9	525	6834
Cannizarro 2, glucose (oxidation)	1	11	62	712	5392
Hemiacetal Formation for 5 membered rings	1	12	62	739	5333
Keto-enol migration twice	1	16	96	739	4001
Hydration of C(=O)C	0	9	84	525	3643
Knoevenagel H (inv)	1	11	58	661	3395
Hemiacetal Formation for 6 membered rings	1	7	36	418	3227
Cannizarro 2, glucose (reduction)	1	11	62	420	2816
Knoevenagel C (inverse)	0	0	8	409	2784
Hydration of C=C(O)	0	0	12	131	1822
Hemiacetal Formation for 7 membered rings	1	6	18	194	1530
Michael Addition 0,2, (reverse)	0	0	0	41	1109
Hemiacetal Formation for 5 membered rings, inverse	0	1	12	74	890
Ring Closure 5 membered O, O	0	1	11	54	694
Benzilic Acid Rearrangement (inverse)	0	1	9	55	641
Benzilic Acid Rearrangement	0	1	13	85	625
Hemiacetal Formation for 6 membered rings, inverse	0	1	7	45	441
Ring Closure 6 membered O, O	0	1	8	35	413
Cannizarro 1	0	1	10	57	382
Beta Decarboxylation	0	0	2	12	317
Hemiacetal Formation for 7 membered rings, inverse	0	1	6	30	228
Ring Closure 7 membered O, O	0	1	7	21	162
Alpha-Keto Acid Decarboxylation	0	0	1	10	103
Beta-gamma Unsaturated Acid Decarboxylation	0	0	0	2	33
Alpha-Beta Unsaturated Acid Decarboxylation	0	0	0	3	16

Bibliography

Andersen, J. L., Andersen, T., Flamm, C., Hanczyc, M. M., Merkle, D., and Stadler, P. F. (2013). Navigating the Chemical Space of HCN Polymerization and Hydrolysis: Guiding Graph Grammars by Mass Spectrometry Data. *Entropy*, 15(10):4066–4083.

Andersen, J. L., Flamm, C., Merkle, D., and Stadler, P. F. (2012). Maximizing output and recognizing autocatalysis in chemical reaction networks is NP-complete. *J. Syst. Chem.*, 3(1):1–9.

Andersen, J. L., Flamm, C., Merkle, D., and Stadler, P. F. (2016). A Software Package for Chemically Inspired Graph Transformation. In *Graph Transformation*, pages 73–88. Springer, Cham, Switzerland.

Andersen, J. L., Flamm, C., Merkle, D., and Stadler, P. F. (2021). Defining Autocatalysis in Chemical Reaction Networks. *arXiv*.

Andersen, J. L. and Merkle, D. (2020). A Generic Framework for Engineering Graph Canonization Algorithms. *ACM J. Exp. Algorithmics*, 25:1–26.

Basak, S. C., Balaban, A. T., Grunwald, G. D., and Gute, B. D. (2000). Topological Indices: Their Nature and Mutual Relatedness. *J. Chem. Inf. Comput. Sci.*, 40(4):891–898.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Beber, M. E., Gollub, M. G., Mozaffari, D., Shebek, K. M., Flamholz, A. I., Milo, R., and Noor, E. (2022). eQuilibrator 3.0: a database solution for thermodynamic constant estimation. *Nucleic Acids Res.*, 50(D1):D603–D609.

Beu, S. C., Blakney, G. T., Quinn, J. P., Hendrickson, C. L., and Marshall, A. G. (2004). Broadband Phase Correction of FT-ICR Mass Spectra via Simultaneous Excitation and Detection. *Anal. Chem.*, 76(19):5756–5761.

Blakney, G. T., Hendrickson, C. L., and Marshall, A. G. (2011). Predator data station: A fast data acquisition system for advanced FT-ICR MS experiments. *Int. J. Mass Spectrom.*, 306(2):246–252.

Blokhuis, A., Lacoste, D., and Nghe, P. (2020). Universal motifs and the diversity of autocatalytic systems. *Proc. Natl. Acad. Sci. U.S.A.*, 117(41):25230–25236.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.*, 2008(10):P10008.

Boogerd, F. C., Bruggeman, F. J., Richardson, R. C., Stephan, A., and Westerhoff, H. V. (2005). Emergence and Its Place in Nature: A Case Study of Biochemical Networks. *Synthese*, 145(1):131–164.

Bredt, J. (1924). Über sterische Hinderung in Brückenringen (Bredtsche Regel) und über die meso-trans-Stellung in kondensierten Ringsystemen des Hexamethylens. *Justus Liebigs Ann. Chem.*, 437(1):1–13.

Čupić, Ž., Maćešić, S., Novakovic, K., Anić, S., and Kolar-Anić, L. (2018). Stoichiometric network analysis of a reaction system with conservation constraints. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(8):083114.

Clarke, B. L. (1975). Stability of topologically similar chemical networks. *J. Chem. Phys.*, 62(9):3726–3738.

Cordella, L. P., Foggia, P., Sansone, C., and Vento, M. (2004). A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(10):1367–1372.

de Boor, C. (1968). On the convergence of odd-degree spline interpolation. *Journal of approximation theory*, 1(4):452–463.

Delaney, J. S. (2004). ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.*, 44(3):1000–1005.

Dhaked, D. K., Ihlenfeldt, W.-D., Patel, H., Delannée, V., and Nicklaus, M. C. (2020). Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2. *J. Chem. Inf. Model.*, 60(3):1253–1275.

Dworkin, J. P. and Miller, S. L. (2000). A kinetic estimate of the free aldehyde content of aldoses. *Carbohydr. Res.*, 329(2):359–365.

Ehrig, H., Ehrig, K., Prange, U., and Taentzer, G. (2006). *Fundamentals of Algebraic Graph Transformation (Monographs in Theoretical Computer Science. An EATCS Series)*. Springer-Verlag, Berlin, Germany.

Emmett, M. R., White, F. M., Hendrickson, C. L., Shi, S. D.-H., and Marshall, A. G. (1998). Application of micro-electrospray liquid chromatography techniques to FT-ICR MS to enable high-sensitivity biological analysis. *J. Am. Soc. Mass Spectrom.*, 9(4):333–340.

Ertl, P., Rohde, B., and Selzer, P. (2000). Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.*, 43(20):3714–3717.

Ford, L. R. and Fulkerson, D. R. (1956). Maximal Flow Through a Network. *Can. J. Math.*, 8:399–404.

Gánti, T. (1975). Organization of chemical reactions into dividing and metabolizing units: The chemotons. *Biosystems*, 7(1):15–21.

Geissman, T. (1944). The cannizzaro reaction. *Organic reactions*, 2:94–113.

Grosshans, P. B., Shields, P. J., and Marshall, A. G. (1991). Comprehensive theory of the Fourier transform ion cyclotron resonance signal for all ion trap geometries. *J. Chem. Phys.*, 94(8):5341–5352.

Gugisch, R., Kerber, A., Kohnert, A., Laue, R., Meringer, M., Rücker, C., and Wassermann, A. (2015). Chapter 6 - molgen 5.0, a molecular structure generator. In Basak, S. C., Restrepo, G., and Villaveces, J. L., editors, *Advances in Mathematical Chemistry and Applications*, pages 113–138. Bentham Science Publishers.

Hannis, J. C. and Muddiman, D. C. (2000). A dual electrospray ionization source combined with hexapole accumulation to achieve high mass accuracy of biopolymers in Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.*, 11(10):876–883.

Himsolt, M. (1997). Gml: A portable graph file format. Technical report, Technical report, Universitat Passau.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS One*, 9(6):e98679.

Jain, S. and Krishna, S. (2001). A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proc. Natl. Acad. Sci. U.S.A.*, 98(2):543–547.

Kaiser, N. K., Quinn, J. P., Blakney, G. T., Hendrickson, C. L., and Marshall, A. G. (2011a). A Novel 9.4 Tesla FTICR Mass Spectrometer with Improved Sensitivity, Mass Resolution, and Mass Range. *J. Am. Soc. Mass Spectrom.*, 22(8):1343–1351.

Kaiser, N. K., Savory, J. J., and Hendrickson, C. L. (2014). Controlled Ion Ejection from an External Trap for Extended m/z Range in FT-ICR Mass Spectrometry. *J. Am. Soc. Mass Spectrom.*, 25(6):943–949.

Kaiser, N. K., Savory, J. J., McKenna, A. M., Quinn, J. P., Hendrickson, C. L., and Marshall, A. G. (2011b). Electrically Compensated Fourier Transform Ion Cyclotron Resonance Cell for Complex Mixture Mass Analysis. *Anal. Chem.*, 83(17):6907–6910.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28(1):27–30.

Kebukawa, Y., Kilcoyne, A. L. D., and Cody, G. D. (2013). Exploring the Potential Formation of Organic Solids in Chondrites and Comets through Polymerization of Interstellar Formaldehyde. *Astrophys. J.*, 771(1):19.

Kerber, A., Laue, R., Meringer, M., and Rücker, C. (2005). Molecules in silico: potential versus known organic compounds. *MATCH Commun. Math. Comput. Chem*, 54(2):301–312.

Kier, L. B. (1989). An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.-Act. Relat.*, 8(3):221–224.

Kochev, N. T., Paskaleva, V. H., and Jeliazkova, N. (2013). Ambit-Tautomer: An Open Source Tool for Tautomer Generation. *Mol. Inf.*, 32(5-6):481–504.

Kolar-Anić, L., Čupić, Ž., Schmitz, G., and Anić, S. (2010). Improvement of the stoichiometric network analysis for determination of instability conditions of complex nonlinear reaction systems. *Chem. Eng. Sci.*, 65(12):3718–3728.

Lambiotte, R., Delvenne, J.-C., and Barahona, M. (2008). Laplacian Dynamics and Multiscale Modular Structure in Networks. *arXiv*.

Landrum, G. (2013). Rdkit documentation. *Release*, 1(1-79):4.

Levy, M., Miller, S. L., Brinton, K., Bada, J. L., and Bada, J. L. (2000). Prebiotic synthesis of adenine and amino acids under Europa-like conditions. *Icarus*, 145(2):609–613.

Liebig, J. (1838). Ueber Laurent's Theorie der organischen Verbindungen. *Ann. Pharm.*, 25(1):1–31.

Liu, M., Grinberg Dana, A., Johnson, M. S., Goldman, M. J., Jocher, A., Payne, A. M., Grambow, C. A., Han, K., Yee, N. W., Mazeau, E. J., Blondal, K., West, R. H., Goldsmith, C. F., and Green, W. H. (2021). Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.*, 61(6):2686–2696.

Lobodin, V. V., Juyal, P., McKenna, A. M., Rodgers, R. P., and Marshall, A. G. (2013). Tetramethylammonium Hydroxide as a Reagent for Complex Mixture Analysis by Negative Ion Electrospray Ionization Mass Spectrometry. *Anal. Chem.*, 85(16):7803–7808.

Moriwaki, H., Tian, Y.-S., Kawashita, N., and Takagi, T. (2018). Mordred: a molecular descriptor calculator. *J. Cheminf.*, 10(1):1–14.

Morooka, S., Wakai, C., Matubayasi, N., and Nakahara, M. (2005). Hydrothermal Carbon-Carbon Bond Formation and Disproportionations of C1 Aldehydes: Formaldehyde and Formic Acid. *J. Phys. Chem. A*, 109(29):6610–6619.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.*, 103(23):8577–8582.

Noor, E., Haraldsdóttir, H. S., Milo, R., and Fleming, R. M. T. (2013). Consistent Estimation of Gibbs Energy Using Component Contributions. *PLoS Comput. Biol.*, 9(7):e1003098.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *J. Cheminf.*, 3(1):1–14.

Omran, A., Menor-Salvan, C., Springsteen, G., and Pasek, M. (2020). The Messy Alkaline Formose Reaction and Its Link to Metabolism. *Life*, 10(8):125.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peixoto, T. P. (2017). The graph-tool python library.

Peretó, J. (2012). Out of fuzzy chemistry: from prebiotic chemistry to metabolic networks. *Chem. Soc. Rev.*, 41(16):5394–5403.

Preiner, M., Xavier, J. C., do Nascimento Vieira, A., Kleinermanns, K., Allen, J. F., and Martin, W. F. (2019). Catalysts, autocatalysis and the origin of metabolism. *Interface Focus*, 9(6):20190072.

Sabio, M., Topiol, S., and Lumma, W. C. (1990). An investigation of tautomerism in adenine and guanine. *Journal of Physical Chemistry*, 94(4):1366–1372.

Sajed, T., Marcu, A., Ramirez, M., Pon, A., Guo, A. C., Knox, C., Wilson, M., Grant, J. R., Djoumbou, Y., and Wishart, D. S. (2016). ECMDB 2.0: A richer resource for understanding the biochemistry of E. coli. *Nucleic Acids Res.*, 44(D1):D495–D501.

Sandefur, C. I., Mincheva, M., and Schnell, S. (2013). Network representations and methods for the analysis of chemical and biochemical pathways. *Mol. Biosyst.*, 9(9):2189–2200.

Sander, T., Freyss, J., von Korff, M., and Rufener, C. (2015). DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.*, 55(2):460–473.

Savory, J. J., Kaiser, N. K., McKenna, A. M., Xian, F., Blakney, G. T., Rodgers, R. P., Hendrickson, C. L., and Marshall, A. G. (2011). Parts-Per-Billion Fourier Transform Ion Cyclotron Resonance Mass Measurement Accuracy with a "Walking" Calibration Equation. *Anal. Chem.*, 83(5):1732–1736.

Sayle, R. A. (2010). So you think you understand tautomerism? *J. Comput.-Aided Mol. Des.*, 24(6):485–496.

Schmitt-Kopplin, P., Gabelica, Z., Gougeon, R. D., Fekete, A., Kanawati, B., Harir, M., Gebefuegi, I., Eckel, G., and Hertkorn, N. (2010). High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. *Proc. Natl. Acad. Sci. U.S.A.*, 107(7):2763–2768.

Sharma, A. and Meir, A. (1966). Degree of Approximation of Spline Interpolation. *Journal of Mathematics and Mechanics*, 15(5):759–767.

Sharma, S., Arya, A., Cruz, R., and Cleaves Ii, H. J. (2021). Automated Exploration of Prebiotic Chemical Reaction Space: Progress and Perspectives. *Life*, 11(11):1140.

Shi, S. D.-H., Drader, J. J., Freitas, M. A., Hendrickson, C. L., and Marshall, A. G. (2000). Comparison and interconversion of the two most common frequency-to-mass calibration functions for Fourier transform ion cyclotron resonance mass spectrometry22Dedicated to Bob Squires for his many seminal contributions to mass spectrometry and ion chemistry. *Int. J. Mass Spectrom.*, 195-196:591–598.

Stanton, D. T., Dimitrov, S., Grancharov, V., and Mekenyan, O. G. (2002). Charged partial surface area (CPSA) descriptors QSAR applications. *SAR QSAR Environ. Res.*, 13(2):341–351.

Staudt, C. L., Sazonovs, A., and Meyerhenke, H. (2014). NetworKit: A Tool Suite for Large-scale Complex Network Analysis. *arXiv*.

Taylor, P. J., van der Zwan, G., and Antonov, L. (2013). Tautomerism: Introduction, History, and Recent Developments in Experimental and Theoretical Methods. In *Tautomerism*, pages 1–24. John Wiley & Sons, Ltd.

Tolmachev, A. V., Robinson, E. W., Wu, S., Smith, R. D., and Paša-Toli, L. (2011). Trapping Radial Electric Field Optimization in Compensated FTICR Cells. *J. Am. Soc. Mass Spectrom.*, 22(8):1334–1342.

von Korff, M. and Sander, T. (2019). Molecular Complexity Calculated by Fractal Dimension - Scientific Reports. *Sci. Rep.*, 9(967):1–8.

Warr, W. A. (2010). Tautomerism in chemical information management systems. *J. Comput.-Aided Mol. Des.*, 24(6):497–520.

Weber, A. L. (2004). Kinetics of Organic Transformations under Mild Aqueous Conditions: Implications for the Origin of Life and its Metabolism. *Origins Life Evol. Biosphere*, 34(5):473–495.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36.

Wiberg, K. B. (1986). The Concept of Strain in Organic Chemistry. *Angew. Chem., Int. Ed. Engl.*, 25(4):312–322.

Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R., and Steinbeck, C. (2017). The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.*, 9(1):1–19.

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C., and Scalbert, A. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, 46(D1):D608–D617.

Xian, F., Hendrickson, C. L., Blakney, G. T., Beu, S. C., and Marshall, A. G. (2010). Automated Broadband Phase Correction of Fourier Transform Ion Cyclotron Resonance Mass Spectra. *Anal. Chem.*, 82(21):8807–8812.

Yang, B. Y. and Montgomery, R. (1996). Alkaline degradation of glucose: effect of initial concentration of reactants. *Carbohydr. Res.*, 280(1):27–45.

Zeng, J., Cao, L., Chin, C.-H., Ren, H., Zhang, J. Z. H., and Zhu, T. (2020). ReacNetGenerator: an automatic reaction network generator for reactive molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, 22(2):683–691.