Version October 26, 2024
Preprint typeset using LaTeX style openjournal v. 09/06/15

RULE-IT: AN ONLINE PLATFORM FOR REACTION NETWORK EXPLORATIONS FOR CHEMICAL EVOLUTION

Bruno Cuevas-Zuviria
Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Consejo Superior de Investigaciones Científicas (CSIC-INIA), Pozuelo de Alarcón, Madrid, Spain 28223

Tymofii Sokolskyi
Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA 53715
Version October 26, 2024

Abstract

Exploring the chemical space of prebiotic reactions is crucial for understanding the emergence of life. However, the combinatorial explosion of compounds and the lack of user-friendly computational tools pose significant challenges. To address these issues, we developed Rule-it, a web platform that enables prebiotic chemists to build, expand, prune, and visualize chemical reaction networks. The platform integrates experimental data and provides a comprehensive framework for studying chemical evolution. By simplifying the generation and analysis of reaction networks, Rule-it aims to advance research in prebiotic chemistry.

Subject headings: Origins of Life, Chemical Reaction Networks, Chemoinformatics, Prebiotic Chemistry

1. INTRODUCTION

Finding the principles that allowed the chemistry of ancient Earth to give place to life might be one of the most ambitious research goals of modern science. The classical view that started this field considered the creation of biomolecules from prebiotic components as a milestone that could precede the emergence of life (Miller, 1953). After 70 years of research, numerous prebiotic pathways leading to biomolecules have been discovered (Ferris et al., 1972; Kitadai & Maruyama, 2018; Ruiz-Mirazo et al., 2014). Various organic compounds can be produced under hydrothermal conditions (McCollom & Seewald., 2007), delivered by carbonaceous chondrite meteorites(Engel & Nagy, 1982), iron meteorites (Elsila et al., 2021), comets (Dworkin et al., 2001) or produced by mineral catalysis (Graaf et al., 2023). Due to such a diversity of options for prebiotically-available molecules, scientists can now shift their focus to the principles of self-organization and evolution in chemical systems (Baum & Vetsigian, 2017).

A major problem of organic chemistry is the combinatorial explosion of compounds under certain conditions (Warr 1997). This phenomenon can be observed, for example, in the classic Milley-Urey experiment (Wollrab 2016) and the formose reaction - a relatively well-studied autocatalytic cycle (Colón-Santos et al., 2019). However, recent research has shown that there is a certain level of control among the alleged combinatorial explosions depending on reaction conditions (Baltussen et al., 2024; Robinson et al., 2022), e.g., the presence of minerals can constrain the diversity of formose reaction products (Colón-Santos et al., 2019). In parallel to experimental approaches, computational techniques can also help explore the chemical space spanned by prebiotic chemistry by generating reaction networks (Arya et al., 2022; Roszak et al., 2024; Sharma et al., 2021; Wołos et al., 2020). Reaction networks can be grown from rules guiding how different functional groups of the seed compounds can react with each other. In the resulting networks, which can be extensive, we should find at least a partial agreement with reality by incorporating experimental information. However, a critical gap around applying such methodologies is a general lack of integration of computational approaches into prebiotic chemistry, especially if chemists are assumed to know little or nothing about coding. Therefore, we have created Rule-it, a web platform for prebiotic chemists to build, explore, constrain, and compare chemical reaction networks with their experimental evidence.

bruno.czuviria@upm.es
sokolskyi@wisc.edu

2. RESULTS

2.1. *Rule-It: a user-friendly chemical network expansion tool*

Applying rule-based generation of chemical-reaction networks is not easy enough for most users outside the computational chemistry community. To diminish the technical barrier preventing the wide adoption of this methodology, we have implemented our whole network expansion pipeline (**Figure 1**) into a user-friendly graphical interface platform named *Rule-it*. The platform allows easy and modular reaction rule crafting, reaction network expansion (**Figure 1b**), pruning (**Figure 1c**), annotation, and converting chemical names to rules,

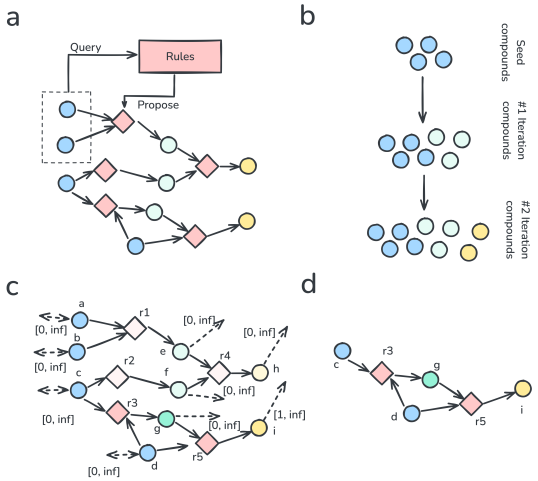

FIG. 1.— **a)** Example of reaction network under expansion. Each of the reactions is derived from the set of reaction rules after finding reactions that can be active with the existing collection of compounds; **b)** the iterative expansion network shown in panel a allows the progressive discovery of different chemicals across iterations; **c)** the pruning algorithm considers each reaction to be a flux (v) converting some inputs into products with stoichiometry S, and a set of input/output fluxes in each molecule. All molecules can be exported, but only inputs can be imported. The detected molecules (intense color) need to be exported with a flux higher than 0. The goal is to minimize the reaction flux that fulfills the steady conditions of this network (e.g. no change over time) and the constraints on input/output; **d)** result of punning the network in panel c.

among other functionalities. The methodology also visualizes chemical reaction networks as graphs. By allowing each step to generate a whole output that can be used in the next step, we avoid unnecessary technical burdens on the server side. We also give the user privacy, as data is not stored, and the ability to modify the intermediate data if necessary. The platform is available to the public at https://ruleit.streamlit.app/. The server can be self-hosted using the code at: https://github.com/brunocuevas/ruleit

2.2. *A one-stop framework to study chemical evolution in prebiotic chemistry*

To illustrate the application of Rule-it, we searched for chemical pathways to generate biologically-relevant compounds from prebiotic chemistry. Specifically, we focused on a dataset of 670 compounds extracted from an *E. coli* Flux-Balance-Analysis metabolism model (Feist 2007) and 154 compounds detected in the composition of the Murchison meteorite (Vincent 2021) as the starting conditions. This process had five steps: i) generating the chemical rules that give place to the chemical reaction network; ii) expanding from those rules using the set of initial seeds used in the experiment; iii) pruning the resulting network to generate the minimal reaction network that could produce the experimental evidence; iv) visualize the network.

The generation of reaction rules is enabled through the specification of the SMARTS format. However, generating SMARTS can be "*mind-bending*", especially if the users are unfamiliar with chemoinformatics. A common issue with crafted reaction rules is the emergence of unexpected results. Therefore, we provide a tool to help

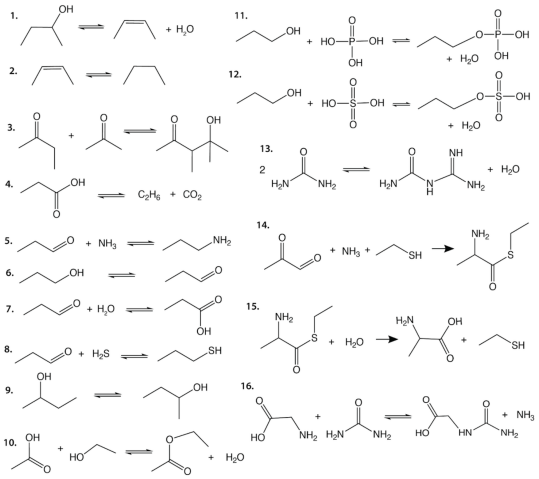

FIG. 2.— Sample reactions illustrating 16 main types of reaction rules used in this study.

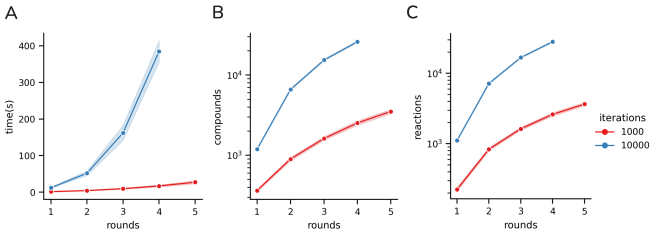

FIG. 3.— Expansion experiments using the Murchison meteorite seeds and the set of 48 reactions. Each data point corresponds to the average of 100 repeats, except for four rounds and 10000 iterations, representing 32 repeats. Each subplot represents the **A)** Time, **B)** Number of compounds discovered, **C)** Number of reactions discovered

the user debug their own reaction rules by proposing positive and negative cases together with the SMARTS expression. The user can obtain a collection of useful SMARTS by manually crafting these rules and possibly using SMARTS from previous work (e.g., Hartenfeller et al., 2011). In our study-case, we used the tool to craft 48 reaction rules representing 16 types of reversible reactions after around 10 iterations of trial and error (**Figure 2; Table S1-2**). For simplicity, we avoided reactions producing any types of rings and focused mostly on carbonyl and nitrogen-containing organics.

The second step requires using those SMARTS to generate a chemical reaction network. In our specific case, we use the compounds detected in the Murchison meteorite (**Table S3**) and the rules from **Table S1-2**. **Figure 3** summarizes the number of compounds discovered under different iterations of this expansion algorithm. Note that the differences are due to the expansion algorithm's stochastic nature; however, such variability does not seem to impact the final number of compounds discovered at every iteration.

These chemical reaction networks are too big to give insight into which reactions could occur in the system or which structures could exist within the chemical network. Therefore, to answer our question of how many compounds in *E. coli* metabolism can be derived from meteorite composition, we pruned our meteorite-derived chemical reaction network with the *E.*

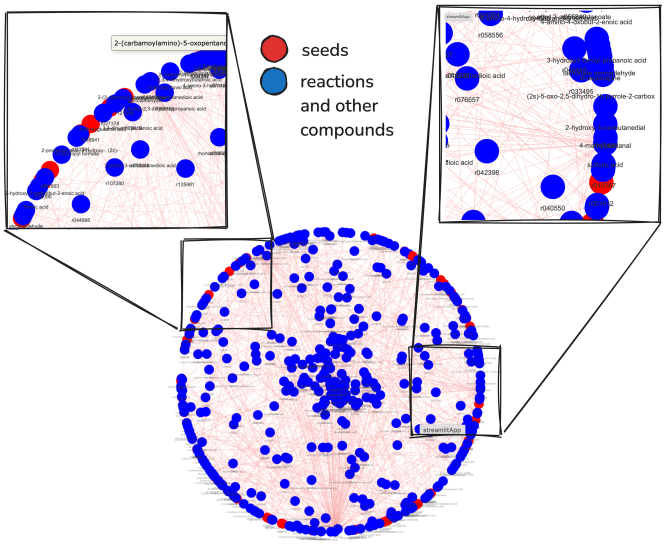

FIG. 4.— Pruned network visualization in the Ruleit server. Red nodes represent seeds; blue nodes represent each of the generated compounds and reactions

coli dataset (**Table S4**). The result was a minimal reaction network that produced as many E. coli compounds as possible, starting from Murchison meteorite soluble organic matter. Based on constrained linear programming, this method is preceded by searching SMILES from the molecular names derived from Vincent et al., 2021 for the meteorite composition and from Feist et al. 2007 for the metabolic compounds. This step can also be executed in our app, which obtains the SMILES by querying PubChem.

Overall, in the original network produced from 4 iterations of expansion on the meteorite seed set there are 132625 reactions with 118592 compounds. Following pruning, we detected 73 E. coli compounds (**Table S5**) absent in the meteorite seed set produced from 303 reactions. Producing those 73 compounds requires 401 intermediate compounds. As the total number of compounds selected for pruning is 670, this represents a success rate of 11%, which is surprisingly high given the simplicity of our ruleset.

These much smaller networks are more manageable to analyze and understand. In the fourth step, we generate a chemical reaction network representation for the resulting network (**Figure 4**) that the user can easily explore. There are also options to download the resulting information in a format that enables visualization in other software, such as Gephi (Bastian 2009).

Finally, the user can explore processes relevant to the Origin of Life within the resulting networks. Among them, the presence of autocatalysis is an important parameter, as it enables molecular self-replication and possibly evolution (Gagrani, 2024; Gagrani et al., 2024). The app generates input required to run efficient autocatalysis detection using autocatalyticsubnetworks program (github.com/vblancoOR/autocatatalyticsubnetworks) - a matrix of stoichiometric coefficients for each reaction in the network.

3. DISCUSSION

Digesting large and messy chemical reaction networks is critical to understanding the emergence of chemical evolution. However, doing so requires a certain degree of heterodoxy. While classical organic chemistry experiments are useful for understanding the diversity of prebiotic synthesis pathways, it is important to highlight that early Earth environments were considerably "messier" than laboratory conditions (Guttenberg et al., 2017). Additionally, even trace amounts of certain compounds could have had a significant role in the origins of life — just one molecule of a member species is sufficient to start an autocatalytic cycle (Peng et al., 2022). This means that disentangling the mess of possible chemical reactions, including rare reactions often overlooked by chemists, is critical for understanding life's origins.

To tackle this problem, we provide an open-source tool that can help chemists to understand their complex data and to design new experiments. Combined with new sources of prebiotic chemistry knowledge, we expect that Rule-it can fuel a certain degree of advancement in the field.

4. METHODS

We combine an expansion algorithm with rule-based reaction generation to explore the chemical pathways that could have produced the detected compounds from the initial seed set. The expansion algorithm has been applied in many different publications (Goldford et al., 2017, 2019, 2022; Peng et al., 2022; Roszak et al., 2024; Wołos et al., 2020), and it consists of the successive addition of new chemicals discovered or reachable from previous chemicals. Rule-based chemical reaction generation enables the exploration of the chemical space by taking advantage of user-provided reaction mechanisms (oxidations, esterifications, etc). These two tools generate large reaction networks that require careful information extraction to be compared with experimental results.

4.1. *Network expansions*

In an expansion algorithm, each round uses the compounds generated by the previous round and the initial seed compounds to look for new active reactions. The output of those reactions enlarges the pool of compounds under use (**Figure 1a, 1b**). The generation can occur by considering a static set of initial reactions, similar to pruning a pre-existing chemical reaction network (CRN), or by generating a set of reactions using reaction rules or any other generative mechanism. In our case, we use the second mechanism, as it allows us to explore a more extensive chemical space (Arya et al., 2022; Roszak et al., 2024; Sharma et al., 2021; Wołos et al., 2020). However, one common problem of generative network expansions, where new reactions can be considered in each iteration of the network, is the possible combinatorial explosion of products across different expansion rounds. Therefore, heuristics to enable the efficient exploration of that space are advisable. In our case, we use a simple random exploration of the chemical space to define a set of reactions r_i and their probabilities p_i. At each iteration, we sample from the distribution of reaction probabilities and compute a reaction on a set of prefiltered reactants. After an N number of iterations, we remove all the redundant reactions. We also enable filtering reactions based on a discriminator, which can be any mechanism

removing a function because of some property. We implement discrimination based on detecting compounds with uncommon valences and molecular masses that are too high.

4.2. *Reaction rule generation*

We use SMARTS to define chemical patterns that enable chemical transformations. However, crafting these SMARTS can be tricky, as it requires a certain degree of mental gymnastics. Therefore, we first implemented a graphical program to help debug SMARTS to test positive and negative examples.

We implemented the generation of reactions using the RdKit library. Before any expansion, we pre-compute lists of valid chemical reactants for each reaction reactant template and proceed to the stochastic sampling during each iteration of the reaction network algorithm.

4.3. *Reaction network pruning*

The resulting reaction networks can have thousands of reactions and molecules, from which only a few might be relevant. We wish to reduce our initial chemical reaction networks to smaller networks reproducing experimental evidence: the initial set of compounds and the detected compounds. We implemented a pruning algorithm based on three steps. First, we need to convert complex analytical techniques to produce a set of computer-readable molecules. We use the conversion of molecule names from a given experimental dataset to SMILES through the PubChem database API. Second, we look for those compounds in the generated CRN, as only a part of the network may be necessary to cover some or all detected compounds. We do this by removing factors in SMILES that allow the recognition of identical molecules (e.g., charge, stereoisomers) and by producing truncated InChiKeys. Finally, we attempt to solve the minimal number of reactions that connect the detected outputs with the seeds. For this, we state the problem as a linear programming solvable flux problem, in which seed compounds are the only fluxes in the system, reactions convert any molecules into other molecules, and all molecules can be exported. Adding the condition of a minimal positive flux in the output compounds and minimizing the overall reaction flux enabled the generation of minimal chemical reaction networks that explain the products (**Figure 1c, 1d**).

ACKNOWLEDGMENTS

The authors would like to thank Praful Gagrani, David Baum for valuable feedback on this work. TS was funded by the National Science Foundation grant number DEB 2218817. BCZ was funded by the Universidad Politécnica de Madrid (UPM) Margarita Salas Fellowship, supported by the "Unión Europea-NextGeneration EU"(code UP2021-035).

BIBLIOGRAPHY

Arya, A., Ray, J., Sharma, S., Simbron, R. C., Lozano, A., Smith, H. B., Andersen, J. L., Chen, H., Meringer, M., & Cleaves, H. J. (2022). An open source computational workflow for the discovery of autocatalytic networks in abiotic reactions. *Chemical Science*. https://doi.org/*10*.1039/d2sc00256f

Baltussen, M. G., Jong, T. J. de, Duez, Q., Robinson, W. E., & Huck, W. T. S. (2024). Chemical reservoir computation in a self-organizing reaction network. *Nature*, 1–7. https://doi.org/10.1038/s41586-024-07567-x

Baum, D. A., & Vetsigian, K. (2017). An Experimental Framework for Generating Evolvable Chemical Systems in the Laboratory. *Origins of Life and Evolution of Biospheres*, *47*(4), 481–497. https://doi.org/10.1007/s*11*084-016-9526-x

Colón-Santos, S., Cooper, G. J. T., & Cronin, L. (2019). Taming the Combinatorial Explosion of the Formose Reaction via Recursion within Mineral Environments. *ChemSystemsChem*, *1*(3). https://doi.org/10.1002/syst.201900014

Dworkin, J. P., Deamer, D. W., Sandford, S. A., & Allamandola, L. J. (2001). Self-assembling amphiphilic molecules: Synthesis in simulated interstellar/precometary ices. *Proceedings of the National Academy of Sciences*, *98*(3), 815–819. https://doi.org/10.1073/pnas.98.3.815

Elsila, J. E., Johnson, N. M., Glavin, D. P., Aponte, J. C., & Dworkin, J. P. (2021). Amino acid abundances and compositions in iron and stony-iron meteorites. *Meteoritics & Planetary Science*, *56*(3), 586–600. https://doi.org/10.1111/maps.13638

Engel, M. H., & Nagy, B. (1982). Distribution and enantiomeric composition of amino acids in the Murchison meteorite. Nature, *296*(5860), 837–840. https://doi.org/10.1038/296837a0

Ferris, J. P., Donner, D. B., & Lotz, W. (1972). Chemical evolution. IX. Mechanism of the oligomerization of hydrogen cyanide and its possible role in the origins of life. *Journal of the American Chemical Society*, *94*(20), 6968–6974. https://doi.org/10.1021/ja00775a018

Gagrani, P. (2024). Evolution of complexity and the origins of biochemical life. *ArXiv*. https://doi.org/10.48550/arxiv.2407.*11*728

Gagrani, P., Blanco, V., Smith, E., & Baum, D. (2024). Polyhedral geometry and combinatorics of an autocatalytic ecosystem. *Journal of Mathematical Chemistry*, *62*(5), 1012–1078. https://doi.org/10.1007/s10910-024-01576-x

Goldford, J. E., Hartman, H., Marsland, R., & Segrè, D. (2019). Environmental boundary conditions for the origin of life converge to an organosulfur metabolism. *Nature Ecology & Evolution*, *3*(12), 1715–1724. https://doi.org/10.1038/s41559-019-1018-8

Goldford, J. E., Hartman, H., Smith, T. F., & Segrè, D. (2017). Remnants of an Ancient Metabolism without Phosphate. *Cell*, *168*(6), 1126-1134.e9. https://doi.org/10.1016/j.cell.2017.02.001

Goldford, J. E., Smith, H. B., Longo, L. M., Wing, B. A., & McGlynn, S. E. (2022). *Continuity between ancient geochemistry and modern metabolism enabled by non-autocatalytic purine biosynthesis*. https://doi.org/10.1101/2022.10.07.511356

Graaf, R. de, Decker, Y. D., Sojo, V., & Hudson, R. (2023). Quantifying Catalysis at the Origin of Life**. *Chemistry – A European Journal*, *29*(53), e202301447. https://doi.org/10.1002/chem.202301447

Guttenberg, N., Virgo, N., Chandru, K., Scharf, C., & Mamajanov, I. (2017). Bulk measurements of messy chemistries are needed for a theory of the origins of life. *Philosophical Transactions of the Royal Society A: Math-*

ematical, Physical and Engineering Sciences, *375*(2109), 20160347. https://doi.org/10.1098/rsta.2016.0347

Hartenfeller, M., Eberle, M., Meier, P., Nieto-Oberhuber, C., Altmann, K.-H., Schneider, G., Jacoby, E., & Renner, S. (2011). A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. *Journal of Chemical Information and Modeling*, *51*(12), 3093–3098. https://doi.org/10.1021/ci200379p

Kitadai, N., & Maruyama, S. (2018). Origins of building blocks of life: A review. *Geoscience Frontiers*, *9*(4), 1117–1153. https://doi.org/10.1016/j.gsf.2017.07.007

Miller, S. L. (1953). A Production of Amino Acids Under Possible Primitive Earth Conditions. Science, 117(3046), 528–529. https://doi.org/10.1126/science.117.3046.528

Peng, Z., Linderoth, J., & Baum, D. A. (2022). The hierarchical organization of autocatalytic reaction networks and its relevance to the origin of life. *PLoS Computational Biology*, *18*(9), e1010498. https://doi.org/10.1371/journal.pcbi.1010498

Robinson, W. E., Daines, E., Duppen, P. van, Jong, T. de, & Huck, W. T. S. (2022). Environmental conditions drive self-organization of reaction pathways in a prebiotic reaction network. *Nature Chemistry*, *14*(6), 623–631. https://doi.org/10.1038/s41557-022-00956-7

Roszak, R., Wołos, A., Benke, M., Gleń, Ł., Konka, J., Jensen, P., Burgchardt, P., Żadło-Dobrowolska, A., Janiuk, P., Szymkuć, S., and Grzybowski, B. A. (2024). Emergence of metabolic-like cycles in blockchain-orchestrated reaction networks. Chem, 10(3), 952–970. https://doi.org/10.1016/j.chempr.2023.12.009

Ruiz-Mirazo, K., Briones, C., & Escosura, A. de la. (2014). Prebiotic Systems Chemistry: New Perspectives for the Origins of Life. *Chemical Reviews*, *114*(1), 285–366. https://doi.org/10.1021/cr2004844

Sharma, S., Arya, A., Cruz, R., & II, H. C. (2021). Automated Exploration of Prebiotic Chemical Reaction Space: Progress and Perspectives. Life, 11(11), 1140. https://doi.org/10.3390/life11111140

Sokolskyi, T., Gargulak, S., Allen, E., & Baum, D. (2024). Measuring heritability in messy prebiotic chemical systems. *BioRxiv*, 2024.08.03.606486. https://doi.org/10.1101/2024.08.03.606486

Wołos, A., Roszak, R., Żadło-Dobrowolska, A., Beker, W., Mikulak-Klucznik, B., Spólnik, G., Dygas, M., Szymkuć, S., and Grzybowski, B. A. (2020). Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. Science, *369*(6511), eaaw1955. https://doi.org/10.1126/science.aaw1955

This paper was built using the Open Journal of Astrophysics LaTeX template. The OJA is a journal which provides fast and easy peer review for new papers in the `astro-ph` section of the arXiv, making the reviewing process simpler for authors and referees alike. Learn more at http://astro.theoj.org.