

Open questions in understanding life's origins

Christopher J. Butch^{1,2,3}, Markus Meringer⁴, Jean-Sebastien Gagnon^{5,6}  & H. James Cleaves II^{2,3,7,8}  

The chemical space of prebiotic chemistry is extremely large, while extant biochemistry uses only a few thousand interconnected molecules. Here we discuss how the connection between these two regimes can be investigated, and explore major outstanding questions in the origin of life.

As we search for habitable and inhabited planets beyond Earth, defining life and understanding how it originates is critical to designing life detection missions¹. Though scientists from many fields have tried to understand the origins of life, and many hypotheses exist, a precise definition of life remains elusive², and we do not presently know how life began.

From interstellar observations and carbonaceous meteorites, it is known that complex organic chemistry occurs widely in primitive solar system environments (e.g., ref. ³). Conversely, we have the single data point of the chemistry produced by our biosphere. The space between these data points is sparsely filled by experiment, model, and hypothesis. Experimentally addressing the chemical origins of life is complicated by the size of organic chemical space⁴, and the tandem sparsity and complexity of reactions which could give rise to autocatalytic, replicative and ultimately living chemistry. A large amount of chemistry remains to be explored, and it is likely the field will benefit from a combination of experimental, observational and computational studies. For example, computational chemists can algorithmically explore chemical space using graph “grammars”⁵ much more rapidly than “wet” chemists can experimentally, though such computations are still hampered by accuracy and computational capacity⁶.

Origins of life models, regardless of biases along heterotrophic/autotrophic axes⁷, all depend on the origin of chemical reaction *networks*. But life is more than a collection of reactions and compounds, it is a *systemic* phenomenon characterized by feedbacks that modulate kinetics. Within reaction networks, slight differences in reactivity can cause large systemic effects. *Network closure*, in which the edges (in this case reactions) and nodes (here, chemical compounds) of a network form a single connected component⁸, is a unifying concept defining hierarchically functional and selectable biological units (e.g., metabolic pathways, genes, organelles, cells,

¹ Department of Biomedical Engineering, College of Engineering and Applied Sciences, Nanjing University, Nanjing, China. ² Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, Japan. ³ Blue Marble Space Institute for Science, Seattle, WA, USA. ⁴ German Aerospace Center (DLR), Earth Observation Center (EOC), Oberpfaffenhofen-Wessling, Germany. ⁵ Physics Department, Norwich University, Northfield, VT, USA. ⁶ Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA. ⁷ Institute for Advanced Study, Princeton, NJ, USA. ⁸ Center for Chemical Evolution, Georgia Institute of Technology, Atlanta, GA, USA. ✉email: henderson.cleaves@gmail.com

species, ecosystems, etc.^{9,10}). Network development may also lead to the emergence of novel phenomena⁹: new graph rewriting rules may be created by networks, for example by creating new phases which alter some reactions' kinetics. How networks achieve closure and simultaneously bring about “internal causation,” in which species are created by catalytically closed reaction networks, is a complex but addressable question. Previous models have demonstrated how these types of emergent systemic properties may have contributed to the origins of life^{11–13}, but models with more precise chemical predictivity are needed.

There are several fundamental problems in understanding the chemical origins of life which require study in the context of networks and their closure (summarized in Fig. 1).

1. *Understanding how individual chemical reactions concatenate to expand reaction networks.* To understand the transition from prebiotic chemistry to biochemistry, it is important to first understand the generation of complex chemical networks from some kind of “primordial” feedstock. “Wet” chemists approach this problem by doing experiments, and defining rules to predict outcomes of reactions based on these experiments. This approach is of course the most realistic, but studying the whole chemical possibility space could be time consuming (if at all possible). A more time efficient approach is to formalize chemical rules using “graph grammars”^{5,6}, and use those rules to predict “real-world” chemistry in silico. However,

this type of approach leaves many problems unresolved, for example important reaction pathways may be excluded because they are not intuited using human or machine-learned screening. The expansion of networks using formalized grammars creates complex reaction networks, but kinetics affect the abundances of products, which influence downstream network dynamics. Reaction networks primarily grow and self-limit when they run out of feedstocks, not because they are limited by “possibility space”.

2. *Exploring the relationship between networks and chemically realistic catalysis.* As reaction networks grow, they may create new compounds capable of influencing network development by acting as catalysts, which enable new reactions or enhance one or more reactions relative to the network. The creation of network-influencing catalysts by reaction networks has been addressed by Kauffman's binary polymer model, which examines how catalysts endowed with randomly-assigned kinetic enhancement properties affect network closure¹². Diversity-generating reactions might not be open-ended because implementation of all possible reaction rules on all generated substrates leads to no new implementable reactions or products. Some reaction networks may never grow or change their explorable properties because they generate no new phenomena (e.g. catalysts or phases) or reactions, while others become truly open-ended.

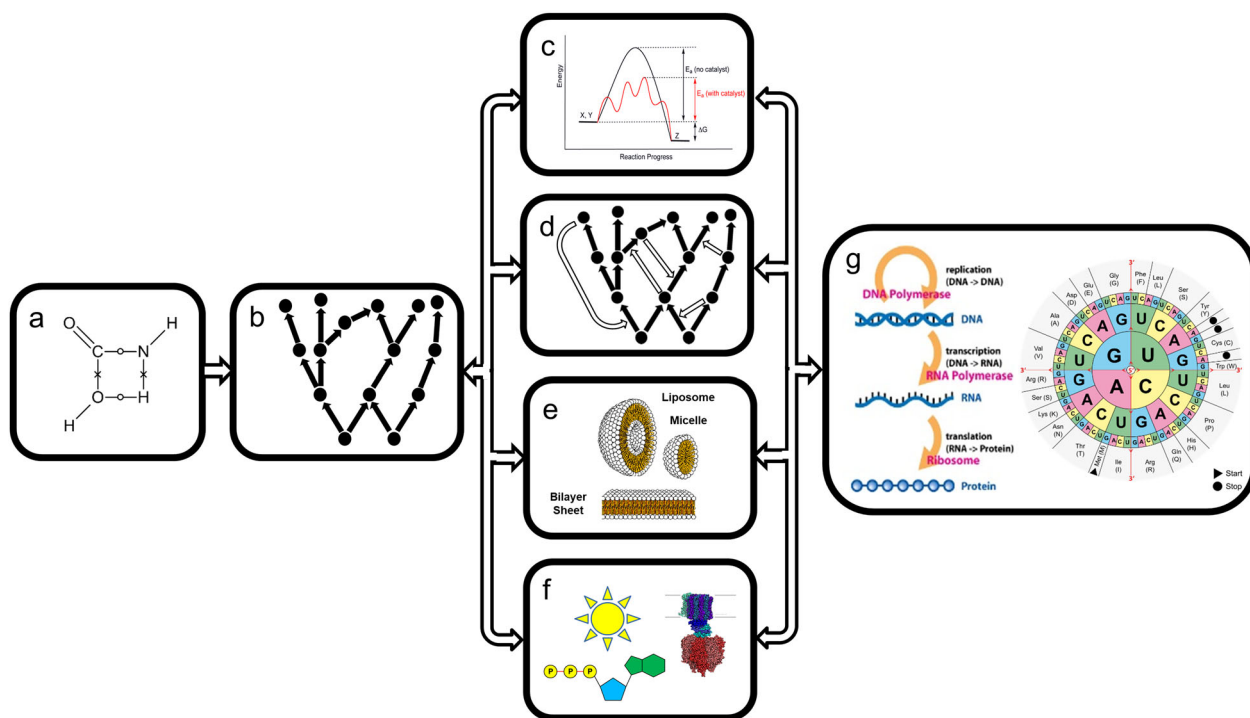


Fig. 1 Workflow for studies of the chemistry of life's origins. Prebiotic reactions proceed according to mechanisms defined by physical organic chemistry (a) to make complex product suites (b), as determined by kinetics and thermodynamics. Products of these simple networks may change the rules and restrictions of their generation causing the network to become more self-directed and life-like. Products may act as catalysts which lower energy barriers and enable the discovery of new chemistry (c) or steer the flux of compounds through certain preferred pathways which can become mutually reinforcing, leading to network closure (d). Products may also contribute to phase separation and another form of closure, encapsulation (e). Emergence of still more complex phenomena, such as activating chemistries and cellular capacitors, likely depend on multiple layers of network feedbacks (f). Finally the emergence of language-like processes may occur in a network with sufficient control to decouple structure and function (g). Feedbacks giving rise to such processes are presently hard to predict. Images are sourced as follows: Panel c: public domain licensing (<https://commons.wikimedia.org/wiki/File:CatalysisScheme.png>); Panel e: public domain licensing (https://commons.wikimedia.org/wiki/File:Phospholipids_aqueous_solution_structures.svg); Panel f: redrawn from Creative Commons Attribution-Share Alike 3.0 Unported licensing (https://commons.wikimedia.org/wiki/File:Atp_synthese.PNG); Panel g: Creative Commons Attribution-Share Alike 3.0 Unported licensing (https://commons.wikimedia.org/wiki/File:Aminoacids_table.svg).

Unlike in the Kauffman model, however, real catalysis is an inherently three dimensional problem wherein the catalytic molecule interacts with a reactant or transition state to alter the energetics of the reaction. Because of this three dimensional nature, it is then reasonable to assume that catalysis may be transferable to other reactions with similar shape and electrotopological character. It is presently unknown how this type of three dimensional biasing effects the behavior of reaction networks. Rule-based reaction network expansion does not have an inbuilt mechanism for the discovery of this type of catalysis, or the estimation of its kinetic effects. Reaction-expansion methods presently poorly predict these kinds of reaction feedbacks, including ones that may steer stereochemistry.

3. *Understanding the principles of spontaneous phase separation leading to cellularization.* Diversity-generating reaction networks may create products capable of producing *phase separation*, which provides new playgrounds for network growth, including covalent and non-covalent aggregation and compartmentalization, which forms the basis of phenomena like Polymerization-Induced Self Assembly¹⁴. The types of phases prebiotic chemistry is capable of generating, and the nature of their interactions, are likely more complex than generally appreciated¹⁵, and new phases may in turn create novel network dynamics⁶.
4. *Understanding the origins of activation and group transfer processes.* Biochemistry enables formally thermodynamically impossible reactions to occur by coupling activation to environmental energy sources and novel catalytic processes. Network generation assumes reactions proceed in thermodynamically favored directions, thus they must be fed. How chemical energy was supplied to primitive environments would have affected how prebiotic reaction networks developed. Before chemical networks were able to produce and take advantage of the multiple benefits created by the emergent phenomena discussed above, it is unclear how networks adapted to the changing availability of chemical energy provided by the environment. There was likely a temporal and qualitative order to the development of energy exploitation processes. Using models employing realistic thermodynamics, England¹⁶ has suggested how external energy inputs can drive systems towards increased local order, while increasing entropy generation globally.
5. *Understanding the origins of hierarchically structurally decoupled catalytic encoding.* The central dogma describes how contemporary biochemical information flows essentially unidirectionally between genotype and phenotype occurs from DNA to RNA to protein in cells¹⁷, providing a connection between genetic inheritance, mutation and natural selection. How this mapping arose is perhaps the largest open question in the emergence of life. This information flow is mediated by sophisticated covalent and non-covalent interactions and belies the possibility that there may have been alternative earlier flows during the early chemical and biochemical evolution.

Language is one of the “breakout” phenomena which define humanity¹⁸. A language is a codified system of representation in which objects are represented as symbols. “Grammars” are basic rules which allow for “syntaxes” (concatenated rules which make sense within a language’s rules) at each level of representation to function. The connection between lower-level precisely-defined rules and more abstract language rules means that low-level rules interactively and stochastically construct “meaning” in languages. Chemically, this allows for indirect coupling of structure-based

catalysis with larger reflex arcs. Viewed in this light, the phenomenon of hierarchical emergence is reminiscent of the analogical mapping between the genetic code and human language.

Language involves the mental encoding of concept, followed by expression via speech, followed by auditory or visual reception, and finally mental decoding into received “meaning.” Molecular interactions may not be directly “about” anything, molecules react and interact according to rules. The concept of “something being about something else” depends on concatenated rules and systemic context¹⁹, and is a form of *meta-catalysis*.

Outlook

Undoubtedly, understanding these interlinked and emergent phenomena will benefit from close collaboration between experimental and computational chemists. For example, discovering small molecule catalysts requires algorithms which can evaluate potential intermolecular interactions and their effects on network kinetics. Literature-based prediction is currently only poorly capable of predicting the effects of such catalysis, however medicinal chemists routinely use docking techniques to rapidly screen non-covalent interactions of drug candidates with enzymes²⁰. These methods may be adaptable to rapid screening for potential catalytic interactions among reaction network products expanded using graph transformation rules.

At the level of understanding phase separation and self-assembly, the molecular properties (e.g., K_{ow} , LogP, LogD, etc.) that enable non-covalent molecular aggregation can be estimated using cheminformatics techniques²¹. However, these methods are presently imprecise since the formation of micelles or vesicles depends on understanding higher order solvent-interaction effects, requiring the use of still other computational techniques. Such effects require their own parameterization and evaluation, but it should be possible to cull the local regions of network chemistry which can give rise to them to save computation time, and careful in vitro screening of large amphiphile libraries would improve the predictivity of computational methods.

Rapid methods for the discovery of *meta-catalysis* may provide insight into the observed “jumps” evident in evolution¹⁸. *Meta-catalysis* is *indirect*, and may help explain the origins of heritable and mutable information coding, which enable the responsiveness and adaptivity of reaction networks to external stimuli. To discover *meta-catalytic* phenomena, in silico generative networks need to be analyzed using computationally intensive tests at the level of interactional catalysis, which is not presently simple, and then re-evaluated with every other molecule in a network, to see if new “meaningful” hypergraphs are created. The models in turn need careful vetting using wet chemistry.

The RNA World concept provides an example of what is lacking in origins models. Chemists have provided ever more “prebiotically plausible” syntheses of RNA²², and SELEX experiments have shown it is easy to isolate and amplify molecules which bind target molecules²³, but a recent computational exploration of nucleic acid space found more than a million possible alternative backbones to deoxyribose/ribose life as we know it uses²⁴. None of these approaches can presently address how reaction networks make compounds in thermodynamically and kinetically feasible ways, and predict how the resulting products interact to modify the reaction networks which produce them.

Summary

The transition from reaction mechanisms predetermined by physical chemistry to catalyzed network-pruning reactions, to self-sorting phase generating reactions and indirect catalysis form

fundamental questions on the origin of life. Both experimental and computational approaches will help understand these transitions. However, numerous problems need to be solved to be able to apply computation in meaningful, tractable ways. Borrowing and adapting techniques from other disciplines is likely the most straightforward method of making progress in this area. Refining these approaches will help focus studies in experimentally testable ways.

Received: 19 November 2020; Accepted: 29 December 2020;

Published online: 02 February 2021

References

1. Neveu, M., Hays, L. E., Voytek, M. A., New, M. H. & Schulte, M. D. The ladder of life detection. *Astrobiology* **18**, 1375–1402 (2018).
2. Cleland, C. E. *The Quest for a Universal Theory of Life: Searching for Life As We Don't Know It* (Cambridge University Press, 2019).
3. Sandford, S. A., Nuevo, M., Bera, P. P. & Lee, T. J. Prebiotic astrochemistry and the formation of molecules of astrobiological interest in interstellar clouds and protostellar disks. *Chem. Rev.* **120**, 4616–4659 (2020).
4. Meringer, M. & Cleaves, H. J. Exploring astrobiology using in silico molecular structure generation. *Philos. Trans. A Math. Phys. Eng. Sci.* **375**, 20160344 (2017).
5. Andersen, J. L. et al. Navigating the chemical space of HCN polymerization and hydrolysis: guiding graph grammars by mass spectrometry data. *Entropy* **15**, 4066–4083 (2013).
6. Wolos, A., et al. Synthetic connectivity, emergence, and self-regeneration in the network of prebiotic chemistry. *Science* **369**, eaaw1955 (2020).
7. Miller, S. L., Schopf, J. W. & Lazcano, A. Oparin's "Origin of Life": sixty years later. *J. Mol. Evol.* **44**, 351–353 (1997).
8. Newman, M. E. J. *Networks: An Introduction* (Oxford University Press, Oxford, 2010).
9. Heylighen, F. in *Encyclopedia of Library and Information Sciences*, (eds Bates, M. J. & Maack, M. N.) (2008).
10. Díaz-Muñoz, S. L., Boddy, A. M., Dantas, G., Waters, C. M. & Bronstein, J. L. Contextual organismality: beyond pattern to process in the emergence of organisms. *Evolution* **70**, 2669–2677 (2016).
11. Gánti, T. Organization of chemical reactions into dividing and metabolizing units: the chemotons. *BioSystems* **7**, 15–21 (1975).
12. Kauffman, S. A. Autocatalytic sets of proteins. *J. Theor. Biol.* **119**, 1–24 (1986).
13. Maturana, H. R. & Varela, F. J. *Autopoiesis and Cognition: The Realization of the Living*. Vol. 42. (Springer Science & Business Media, 1991).
14. Penfold, N. J., Yeow, J., Boyer, C. & Armes, S. P. Emerging trends in polymerization-induced self-assembly. *ACS Macro Lett.* **8**, 1029–1054 (2019).
15. Jia, T. Z. et al. Membraneless polyester microdroplets as primordial compartments at the origins of life. *Proc. Natl Acad. Sci. USA* **116**, 15830–15835 (2019).
16. England, J. L. Statistical physics of self-replication. *J. Chem. Phys.* **139**, 121923 (2013).
17. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
18. Maynard Smith, J. & Szathmari, E. *The Major Transitions in Evolution* (Oxford University Press, 1997).
19. McCarthy, C. The Kekulé problem. *Nautilus* **19**, 22–31 (2017).
20. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
21. Israelachvili, J. N. *Intermolecular and Surface Forces* (Elsevier Science, 2011).
22. Yadav, M., Kumar, R. & Krishnamurthy, R. Chemistry of abiotic nucleotide synthesis. *Chem. Rev.* **120**, 4766–4805 (2020).
23. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
24. Cleaves, H. J., Butch, C., Burger, P. B., Goodwin, J. & Meringer, M. One among millions: the chemical space of nucleic acid-like molecules. *J. Chem. Inf. Model.* **59**, 4266–4277 (2019).

Acknowledgements

The authors wish to thank the Earth-Life Science Institute (ELSI) and the ELSI Origins Network (EON) which supported the research which brought the authors of this work into contact. EON was supported by a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Author contributions

C.B., M.M., J.G., and H.J.C. contributed equally to the development and preparation of the manuscript and figure.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.J.C.II.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021