

Research



Cite this article: Roy S, Sengupta S. 2021

Evolution towards increasing complexity through functional diversification in a protocell model of the RNA world. *Proc. R. Soc. B* **288**: 20212098.

<https://doi.org/10.1098/rspb.2021.2098>

Received: 22 September 2021

Accepted: 21 October 2021

Subject Category:

Evolution

Subject Areas:

theoretical biology, evolution

Keywords:

RNA world, ribozyme, protocell, template

Author for correspondence:

Supratim Sengupta

e-mail: supratim.sen@iiserkol.ac.in

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.5683800>.

Evolution towards increasing complexity through functional diversification in a protocell model of the RNA world

Suvam Roy and Supratim Sengupta

Department of Physical Sciences, Indian Institute of Science Education and Research Kolkata, Mohanpur-741246, India

SS, 0000-0001-5294-0561

The encapsulation of genetic material inside compartments together with the creation and sustenance of functionally diverse internal components are likely to have been key steps in the formation of 'live', replicating protocells in an RNA world. Several experiments have shown that RNA encapsulated inside lipid vesicles can lead to vesicular growth and division through physical processes alone. Replication of RNA inside such vesicles can produce a large number of RNA strands. Yet, the impact of such replication processes on the emergence of the first ribozymes inside such protocells and on the subsequent evolution of the protocell population remains an open question. In this paper, we present a model for the evolution of protocells with functionally diverse ribozymes. Distinct ribozymes can be created with small probabilities during the error-prone RNA replication process via the rolling circle mechanism. We identify the conditions that can synergistically enhance the number of different ribozymes inside a protocell and allow functionally diverse protocells containing multiple ribozymes to dominate the population. Our work demonstrates the existence of an effective pathway towards increasing complexity of protocells that might have eventually led to the origin of life in an RNA world.

1. Introduction

The RNA world hypothesis posits a central role for RNA in information storage, catalysis and regulation and argues that life based on RNA only existed prior to a DNA-protein world. The discovery of ribozymes [1–3], their synthesis in the laboratory using *in vitro* evolution [4–6] and abiotic synthesis of ribonucleotides [7–9] have provided indirect evidence for the plausibility of an RNA world.

RNA, instead of being exclusive players in a primordial world, coexisted with small peptides [10] and lipids [11] and such coexistence opens up the possibility of compartmentalization of RNA sequences in lipid vesicles. Such vesicles, that formed through self-aggregation of hydrophobic lipid molecules [12,13], can provide several advantages [14–16] in an RNA world if they are stable against degradation [17,18]. The increased osmotic stress experienced by vesicles containing RNA strands leads to the transfer of fatty acid micelles from empty vesicles promoting growth of the former at the expense of the latter [19]. The growth of protocells and their subsequent division owing to physical forces without significant leakage of their internal components [20] suggest a possible route towards proliferation and competition between protocells. Non-enzymatic, template directed polymerization of RNA strands inside protocells using activated nucleotides [18] indicate how such replication processes can increase the number and diversity of encapsulated sequences. Short catalytic peptides, encapsulated in vesicles, can bind to the vesicle membrane and facilitate growth in their volume by attracting lipids from non-lipophilic membranes [21,22]. The type of vesicular membrane that facilitates non-enzymatic replication of RNA strands [23] and catalytic action of encapsulated Hammerhead ribozymes [24] has also been identified. Such experiments

are gradually revealing the conditions under which vesicles can start acting as chemical factories that eventually triggered the transition from chemistry to biology and led to the emergence of the earliest protocellular life-forms.

Computer simulations [25–27] have further extended the insights obtained from experiments. RNA replicators have been shown to have enhanced likelihood of survival inside vesicles when compared to surface-based, spatially open systems [25,27] because compartments provide more effective protection against hydrolysis and take-over by parasitic counterparts. Protocells also have the advantage of enhancing the maximum error-threshold [27] that can be tolerated during replication of ribozymes, thereby making such functional sequences more robust against mutational degradation.

Even though *in vitro* experiments have provided significant insights into prebiotic processes, the real test of the efficacy of such processes can only come from situating them in plausible primordial conditions. Hence, the question of the appropriate environmental conditions for the origin of life is a challenging yet pertinent one. Those conditions need to support a multi-stage, hierarchical process of emergence of complexity that starts with prebiotic synthesis of the basic building blocks and ends with self-sustained replication of cells with heritable traits. The mineral-rich environments found near terrestrial geothermal pools provide the most compelling environment [28,29], especially in the context of our model. The mineral-rich region can act as a catalytic substrate speeding up template-directed polymerization reactions especially in the presence of alternate wet-dry cycles [28–32]. Such cycles have been shown to be effective for formation of long RNA polymers [33,34] with complex secondary structures [34]. We envisage a scenario where clay particles encapsulated in vesicles [20] can act as substrates for RNA sequences. The template-directed polymerization process on such substrates can then be facilitated by activated nucleotides [34–36].

Despite the manifold advantages of protocells, there are several questions that need to be answered before we can fully understand how RNA-based life encapsulated within protocells emerged. The emergence of the first ribozyme inside a protocell must have been a chance event. Even if the template-directed replication process produces a large number of replicates, how likely is it that one of them becomes a functional molecule like a ribozyme? Although a precise answer to this question remains elusive, recent work [37–39] suggests that this number may not be too low. Random sampling of sequences reveal that only a small set of secondary structures are obtained upon folding the sampled sequences with many sequences yielding the same secondary structure. Certain secondary structures appear much more frequently and such strong phenotype bias is evident from the large (several orders of magnitude) differences in the frequencies of the most and least common secondary structures. Remarkably, the secondary structures found in nature happen to be the ones that appear with the highest frequency [39]. Moreover, the number of random sequences that need to be sampled to generate these secondary structures is quite low ($\approx 10^5$ for $L = 126$) compared to the total number of possible sequences (4^L) which can be astronomically large for large L [39]. These results suggest that functional structures may not be very difficult to produce and can be generated by sampling a relatively small number of RNA sequences. Nevertheless, we need a better understanding of the processes of RNA replication that favoured the appearance and

subsequent proliferation of the first ribozyme. In other words, what were the processes by which the first ribozyme came into existence inside protocells? What was its functional nature and what impact did it have on protocellular evolution? If subsequently, other ribozymes appeared during the error-prone replication process, were protocells with multiple ribozymes having diverse functionality able to proliferate at the expense of those with fewer ones?

In this work, we use the rolling-circle mechanism of RNA sequence replication within protocells to address these questions. This mechanism, first observed in viruses [40–42], is more effective [43] than the non-enzymatic template-directed primer extension process in creating a large number of replicated RNA sequences. The chance creation of a replicase ribozyme with a small probability during the error-prone RNA sequence replication process inside protocells can further enhance the RNA replication rate. If other functionally distinct ribozymes also appear subsequently, under certain conditions, they can act synergistically to enhance each other's production. We speculate on the nature and hierarchy of the emergent ribozymes and show that they can not only synergistically aid each other's formation but also favour evolution of the protocell population towards increasing complexity through preferential selection of protocells with a larger number of functionally diverse ribozymes. Our work shows how the rolling circle mechanism of replication together with the chance creation of a few ribozymes can work in conjunction and point to a plausible pathway for the emergence of the earliest protocellular life-form.

2. Material and methods

A replication mechanism that can efficiently create a large number of RNA sequences is more likely to yield a ribozyme as the outcome of a chance replication event. Tupper & Higgs [43] have compared RNA replication by the non-enzymatic, template-directed primer extension mechanism with the rolling circle mechanism to determine their relative effectiveness. We supplement their results by incorporating the effect of temperature on replication processes. These comparison simulations (see the electronic supplementary material) reveal that exponential growth of the number of strands necessary to sustain the replication process against degradation in case of template-directed primer extension, is constrained by the requirement of very high dry-phase temperatures greater than or equal to 60°C (electronic supplementary material, figure S1). Hence, for more realistic dry phase temperatures that were prevalent in primordial earth, the template-directed primer extension process may not have been effective in sustaining high growth rate of RNA strands. By contrast, the rolling circle mechanism can lead to exponential growth of the number of both open-ended single-stranded RNA (ssRNA) and circular double-stranded RNA (dsRNA) molecules (electronic supplementary material, figure S2) and is not constrained by the requirement of a dry phase (see the electronic supplementary material for details). We therefore consider the creation of new RNA sequences inside protocells to be driven by the rolling circle mechanism.

During RNA replication by the rolling circle mechanism (figure 1) a small complementary primer (approx. 8 nt) first attaches to a circular ssRNA. The primer then extends by template-directed primer extension. Upon becoming full length it can extend further by gradually displacing its other end from the initial point. This creates a hanging tail which keeps growing in length with further primer extension. However, when the

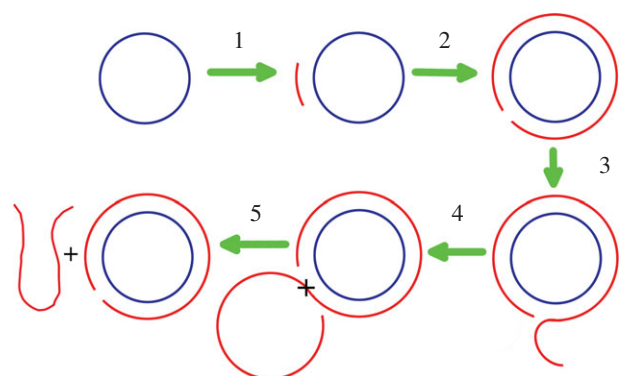


Figure 1. Schematic diagram of the rolling circle replication process. (1) A short complementary primer attaches to a circular ssRNA template. (2) It is extended by the template-directed primer extension process. (3) Upon attaining full length, the primer extends further by displacing the other end of itself from the initial point, resulting in an overhanging portion. (4) When the overhanging tail attains a length equal to that of the circular template, it breaks apart. (5) The separated tail becomes an open-ended ssRNA. (Online version in colour.)

hanging tail becomes too long, it breaks off and becomes an open-ended ssRNA.

(a) Protocell model

(i) Overview

We start with a model consisting of N lipid vesicles (protocells) each of which encapsulate one circular ssRNA molecule of length 200 nucleotides [44]. Reactions occurring inside each protocell (described below) with specified rates leads to formation of different species and quantity of RNA molecules. The evolution of the protocell population occurs following a birth–death process, when a protocell containing a number of RNA species (V_i) that is greater than a specified threshold (V_T) splits into two with its contents divided at random between the daughter cells. To keep the population size fixed, a protocell (j) is eliminated with a probability that is proportional to the size difference ($V_i - V_j$). This ensures that protocells with a larger number of RNA strands are more likely to survive across generations. However, our results do not change significantly even if the eliminated protocell is selected at random.

(ii) Details of the model

Initially when the protocells contain only circular RNA, three types of reactions are possible: conversion of circular ssRNA (s) to circular dsRNA (d) via template-directed primer extension process, production of new open-ended ssRNA (l) from circular dsRNA (d) through the rolling circle replication mechanism and degradation of all types of RNA strands. The ssRNA molecules produced through rolling circle replication will mostly fold into complex structures because of their long lengths. Owing to base-pair mismatches during replication, the replicated strands will not be exact complements of the template. We assume that a small fraction of those folded ssRNA molecules with complex structures attains enzymatic capabilities like that of a replicase (R), cyclase (C), nucleotide synthase (N) and peptidyl transferase (P). This expectation is based on the fact that error-prone replication of the same template at different times can produce sequences with distinct, complex secondary structures. The electronic supplementary material, figure S3 shows a subset of distinct secondary structures of sequences obtained by non-enzymatic replication of the same 200 mer template sequence multiple times. Such structural diversity in sequences of length 200 is not

surprising since structural diversity increases with template lengths (table S1 in electronic supplementary material). Given their complex secondary structures, as well as for reasons given in the previous section, it may not be unreasonable [37–39] to expect that some of these replicates can exhibit distinct catalytic abilities.

A replicase can catalyse the process of circular ssRNA to circular dsRNA and circular dsRNA to open-ended ssRNA production by increasing the replication rate to $K_{fast} \gg K_{rep}$, the latter being the non-enzymatic replication rate. In each protocell, this rate is chosen from the distribution $K_{rep} = K_0 e^{-0.005 \times L - 2.8 \times \text{Norm}(0.35, 0.0667)} \text{h}^{-1}$ that was obtained from sequence-level simulations of the replication process that are described in the electronic supplementary material. We use $K_{fast} \sim 0.362 \text{h}^{-1}$, which is the rate at which the fastest replicase found [6] can replicate a 200 mer. The process of replicase catalysed replication though significantly fast, is still prone to mismatches which increase sequence diversity in the protocell. Hence, we assume that new ribozymes can be created with the same probabilities even in such cases. A cyclase can circularize an open-ended ssRNA molecule to form a circular ssRNA by joining its open ends. As there are no experimental rates available for the circularization reaction, we assume this also occurs at the rate $K_{cyc} = K_{fast}$. A nucleotide synthase can create new monomers inside a protocell in situations where the finite monomer pool is inadequate for all the reactions to proceed and is therefore crucial for sustaining new RNA strand formation in monomer-poor environments. Finally, a peptidyl transferase (p) can join free amino acid molecules to form small peptide chains. Lipid molecules of the protocell membrane turn lipophilic by forming compounds with small peptide chains [21] and such lipophilic membranes attract lipid molecules from nearby protocells with non-lipophilic membranes and grow at the expense of the latter. Hence, we assume that whenever a peptidyl transferase ribozyme appears inside a protocell i , its threshold volume V_T will increase as $V_T^i = V_T^i + kp_i$ (where k is the amount of volume increase in units of RNA strand numbers brought about by a single peptidyl transferase ribozyme). In the presence of these ribozymes, there will be three extra reactions namely; replicase catalysed circular ssRNA to circular dsRNA, circular dsRNA to open-ended ssRNA and cyclase catalysed, non-catalytic open-ended ssRNA to circular ssRNA creation.

For a finite monomer system, we multiply each replication rate with a term $f_i = S_i/S_i^{\max}$ to account for reduction of replication rates in the absence of sufficient monomers. S_i^{\max} denotes the maximum number of monomers (in units of 200 mers) the i th protocell can hold and S_i is the number of available monomers in a cell at any instant in units of 200 mers. The initial number of free monomers in each protocell was taken to be $S_i = S_i^{\max} = aV_T^i$, where a is a constant factor. In the presence of nucleotide synthase, new monomer creation inside a cell will depend on the number of nucleotide synthase molecules inside that cell. Hence we modify the multiplication factor as $f_i = (S_i + bn_i)/S_i^{\max}$. The quantity ‘ b ’ is a measure of the number of free monomers (in units of 200 mers) a nucleotide synthase can produce. For our simulations, $b = 1$, $a = 50$ (unlimited monomer availability scenario) and $a = 0.8$ (when monomer availability is limited). Because monomers can diffuse freely across the vesicle membrane, monomers produced by nucleotide synthases inside different protocells are uniformly distributed across all the protocells in the population. Hence, the S_i values after each reaction step are updated to $S_i = (\sum_i S_i^{\text{old}} + bn_o - g)V_T^i / (\sum_i V_T^i)$, where $n_o = \sum_i n_i^{\text{old}}$ is total number of nucleotide synthase molecules in the entire population before the reaction step and g is the number of new strands created in the entire population after the reaction step.

The reactions described above lead to changes in numbers of different RNA species inside the i th protocell. The differential

equations expressing the variation in the number of each type of RNA species over time are:

$$\dot{s}_i = K_{\text{fast}} \frac{c_i l_i}{V_T^i} - K_{\text{rep}}^i s_i f_i - K_{\text{fast}} \frac{r_i s_i f_i}{V_T^i} - h s_i, \quad (2.1)$$

$$\dot{d}_i = K_{\text{rep}}^i s_i f_i + K_{\text{fast}} \frac{r_i s_i f_i}{V_T^i} - h d_i, \quad (2.2)$$

$$\begin{aligned} \dot{l}_i = & (1 - P_r - P_c - P_n - P_p) K_{\text{rep}}^i d_i f_i \\ & + (1 - P_r - P_c - P_n - P_p) K_{\text{fast}} \frac{r_i d_i f_i}{V_T^i} - K_{\text{fast}} \frac{c_i l_i}{V_T^i} - h l_i, \end{aligned} \quad (2.3)$$

$$\dot{r}_i = P_r K_{\text{rep}}^i d_i f_i + P_r K_{\text{fast}} \frac{r_i d_i f_i}{V_T^i} - h r_i, \quad (2.4)$$

$$\dot{c}_i = P_c K_{\text{rep}}^i d_i f_i + P_c K_{\text{fast}} \frac{r_i d_i f_i}{V_T^i} - h c_i, \quad (2.5)$$

$$\dot{n}_i = P_n K_{\text{rep}}^i d_i f_i + P_n K_{\text{fast}} \frac{r_i d_i f_i}{V_T^i} - h n_i, \quad (2.6)$$

$$\text{and } \dot{p}_i = P_p K_{\text{rep}}^i d_i f_i + P_p K_{\text{fast}} \frac{r_i d_i f_i}{V_T^i} - h p_i, \quad (2.7)$$

where s, d, l, r, c, n, p denote the number of circular ssRNA and dsRNA, open-ended ssRNA, replicase, cyclase, nucleotide synthase and peptidyl transferase respectively. P_r, P_c, P_n, P_p are the creation probabilities of these four types of ribozymes by the rolling circle replication process. The degradation rate is taken to be same ($h = 0.0008 \text{ h}^{-1}$) for all RNA species. The ribozyme-catalysed reactions are second-order reactions i.e. the formation rate of the corresponding RNA species will be proportional to the number of both ribozymes and substrates (circular ssRNA/dsRNA templates, open-ended RNA). However, as the left-hand-side of the rate equations denote the rate of change of the number of different types of RNA species, we divide these second-order rates with the maximum protocell volume V_T^i to match the dimensions.

(iii) Stochastic version of the model

The six types of reactions that can occur inside a protocell can be divided into two categories: (i) non-enzymatic: circular ssRNA to circular dsRNA, circular dsRNA to open-ended ssRNA and degradation of all kinds of RNA strands and (ii) enzymatic: replicase catalysed circular ssRNA to circular dsRNA, replicase catalysed circular dsRNA to open-ended ssRNA and cyclase catalysed open-ended ssRNA to circular ssRNA. In the fully stochastic version of the model, we calculate total reaction rate for each cell and then determine the time step size dt as inverse of the maximum of those rates ($dt = 1/K_{\text{tot}}^{\text{max}}$). During this time period, a cell will undergo any of the six types of reactions with probability $K_{\text{tot}}^i dt$. If a reaction occurs, the type of reaction is chosen at random based on the relative reaction propensities of these six types of reactions, which are $K_{\text{rep}}^i s_i / K_{\text{tot}}^i$, $K_{\text{rep}}^i d_i / K_{\text{tot}}^i$, $h(s_i + d_i + l_i + r_i + c_i + n_i + p_i) / K_{\text{tot}}^i$, $K_{\text{fast}}(r_i s_i f_i / V_T^i) / K_{\text{tot}}^i$, $K_{\text{fast}} \frac{r_i d_i f_i}{V_T^i} / K_{\text{tot}}^i$ and $K_{\text{fast}} \frac{c_i l_i}{V_T^i} / K_{\text{tot}}^i$, respectively. The cell division and population update steps are also implemented stochastically, as described earlier. The detailed algorithm is given in the electronic supplementary material.

3. Results

Even though our ultimate goal is to ascertain the outcome of competition between different protocells in a population that are distinguished by the number and type of their component RNA species, it is instructive to use the above dynamical model to find out the effect of varying probabilities for creation of different types of ribozymes inside a single protocell. We start with a single circular ssRNA molecule inside a protocell with non-enzymatic rate $K_{\text{rep}} \sim 0.0096 \text{ h}^{-1}$ and solve the

equations (2.1)–(2.7) numerically to determine the time evolution of the number of different types of RNA strands inside it. For this numerical model, we ignore the process of 8 mer primer attachment to a circular ssRNA. To better understand the role of cyclase and replicase on the evolution of RNA strands inside a single protocells, we first consider the case of abundant monomer availability, use $S^{\text{max}} = 50 V_T$ and neglect the possible formation of nucleotide synthase, peptidyl transferase (i.e. $P_n = P_p = 0$). We varied the creation probabilities of replicase (P_r) and cyclase (P_c) while keeping their sum fixed to unity. A faster rate of growth of RNA strands is achieved by increasing P_c relative to P_r (see the electronic supplementary material, figure S5(A)). For a single protocell, the predominance of cyclase over a replicase is more important since the former leads to the creation of many new circular templates which enhances the likelihood for creation of dsRNA and eventually potential ribozymes (both cyclase and replicase). Because the production of ribozymes, including replicases, depend on the availability of circular templates, an insufficient number of circular templates on which the replicase can act, creates a bottleneck for production of new ssRNA strands. For a monomer deficient system ($S^{\text{max}} = 0.8 V_T$), the presence of a nucleotide synthase that catalyses monomer production can help in sustaining the growth of new RNA strands. Lack of adequate monomers can affect the growth of new strands despite the presence of cyclase and replicase and lead to saturation in number of RNA strands. Growth of RNA strands is most favoured when the formation probability of a nucleotide synthase (P_n) is similar to that of a cyclase and replicase, $P_r = P_c \sim P_n$ and $P_p = 0$ (electronic supplementary material, figure S5(B)). Finally, we also consider the creation of peptidyl transferase that catalyses formation of short peptides and in the process facilitates protocell growth via membrane transfer. We varied its creation probability (P_p) by keeping the sum of all four ribozyme formation probabilities fixed and assuming $P_r = P_c = P_n$. Increasing P_p gradually decreases RNA strand formation rate (electronic supplementary material, figure S5(C)) even though it leads to higher threshold volumes. These results suggest a hierarchy for the appearance of ribozymes with the cyclase being the most important in our framework, followed by the replicase. In monomer poor environments, the probability of formation of nucleotide synthase also needs to be larger than a threshold (see the electronic supplementary material, figure S5(B)) to ensure an adequate supply of monomers needed for sustained growth of strands. Even though the peptidyl transferase is helpful in facilitating protocell growth and diversification of its component RNA species, its impact is dependent on the presence of the other enzymes. However, these results which address the evolution of RNA species inside a single protocell, do not reveal whether competition between protocells with different RNA content selectively favour certain types of protocells. We address this issue in the subsequent sub-sections. Intriguingly, we find that the conditions for the proliferation of protocells with increasingly diverse functionality are distinct from the conditions required for the highest rate of growth of RNA strands inside a single protocell.

(a) Role of replicase and cyclase in protocell evolution

Using our fully stochastic population dynamics model, we examined the effect of replicase and cyclase first acting separately and then simultaneously, on the evolutionary dynamics of the protocell population. For these simulations, we assume

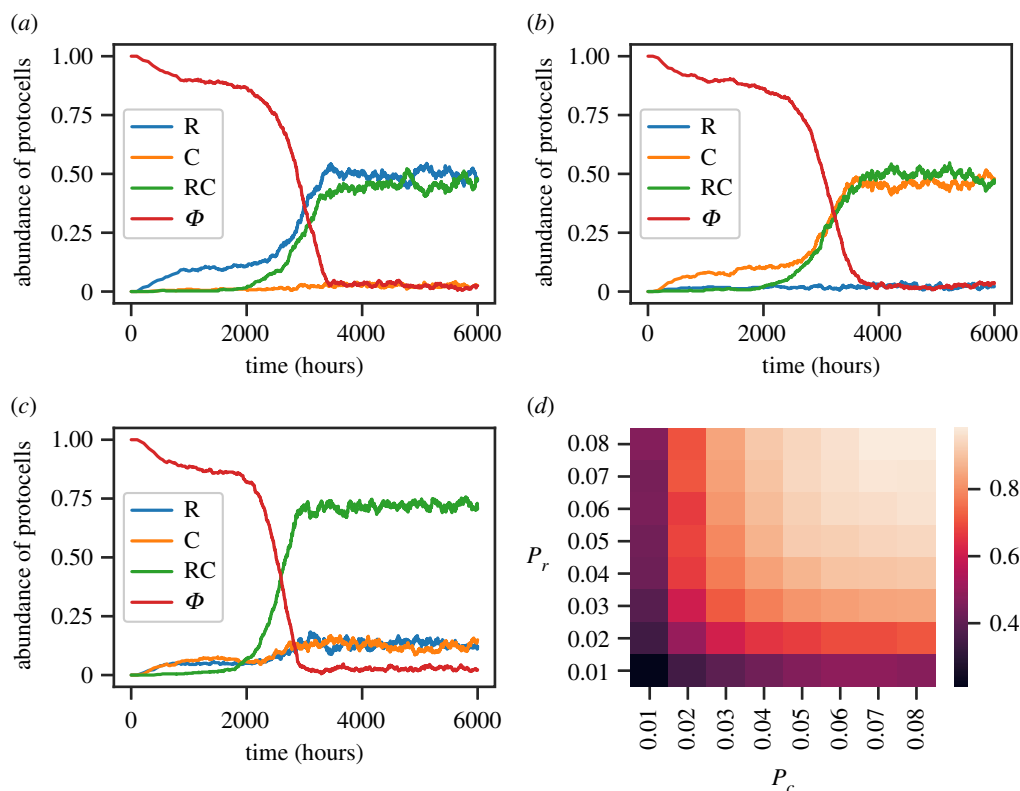


Figure 2. Fractional abundance of protocells containing replicase (R) or/and cyclase (C) or none of them (Φ) for (a) $P_r = 0.05$ and $P_c = 0.01$, (b) $P_r = 0.01$ and $P_c = 0.05$, and (c) $P_r = 0.03$ and $P_c = 0.03$ when there are sufficient monomers in the system. (d) Heatmap of the abundance of RC protocells for various value of P_r and P_c . (Online version in colour.)

effectively unlimited availability of monomers in the system by taking $S^{\max} = 50 V_T$. We carry out these simulations for $V_T = 100$ (for all protocells in the population) and for $N = 400$ protocells, with one circular ssRNA per cell initially, whose replication rate is assigned randomly from a distribution (see Material and methods section). We found that a replicase alone can never sustain ribozymes in the population for any value of the degradation rate and replicase creation probability. However, for a fixed creation probability, a cyclase alone, can sustain the population below a critical value of degradation rate (h). This happens because a replicase cannot create new circular templates which it needs to act on to be effective, it can only speed up the replication rates of existing circular strands. As a result, for any non-zero value of the degradation rate, the initial population of circular templates eventually die out without being replenished in the absence of a cyclase, leading to elimination of RNA strands from the protocell population. By contrast, a cyclase alone can create new circular templates and therefore can sustain ribozymes in the protocell population via non-enzymatic replication.

We next consider the case where replicase and cyclase can be simultaneously present in a protocell. At each time step, we measure the fraction of protocells containing either only replicase or cyclase or both or none. Figure 2 shows the time plots of those fractions for different values of the creation probabilities of replicase (P_r) and cyclase (P_c). As evident from figure 2a when $P_r > P_c$ the abundance of protocells with only replicase (R) and with both replicase and cyclase (RC) are maximum and in equal proportions. The reverse is true for $P_c > P_r$ (figure 2b). However, for $P_r \sim P_c$, protocells with both replicase and cyclase (RC) dominates and their abundance is higher than the maximum

abundances of protocells with either of these ribozymes in the previous two cases (figure 2c). This indicates the formation of a positive feedback loop between replicase and cyclase. The cyclase saves the system from degradation by creating new circular strands on which the replicase can act to produce new ssRNA while the replicase boosts replication rates, thereby increasing the rate of creation of new replicases and cyclases. Highest abundance of RC protocells along the diagonal of figure 2d (i.e. $P_r = P_c$) further hints at the existence of such a synergistic network between the replicase and cyclase.

(b) Role of nucleotide synthase

If there is a scarcity of monomers in a protocell, the presence of a nucleotide synthase that catalyses monomer production can be helpful in sustaining growth of RNA strands within a protocell. We model this scenario by taking $S^{\max} = 0.8 V_T$ and varying the formation probability of a nucleotide synthase while keeping the formation probabilities for replicase and cyclase fixed. We found that below a threshold value ($P_n^{\text{th}} = 0.003$), the system dies out since the number of monomers are insufficient to keep the replication process going (figure 3a,b). Above this threshold probability nucleotide synthase ribozymes created inside protocells can provide the extra monomers required at each time step to sustain the replication process. A comparison between figure 3c,d makes it evident that when nucleotide synthase is created with comparable probabilities as the replicase and cyclase, protocells with all three types of ribozymes dominate the population. When the probability of formation of a nucleotide synthase is much larger compared to that of a cyclase or a replicase, protocells containing either nucleotide synthase alone or together with cyclase or replicase or with both cyclase and replicase,

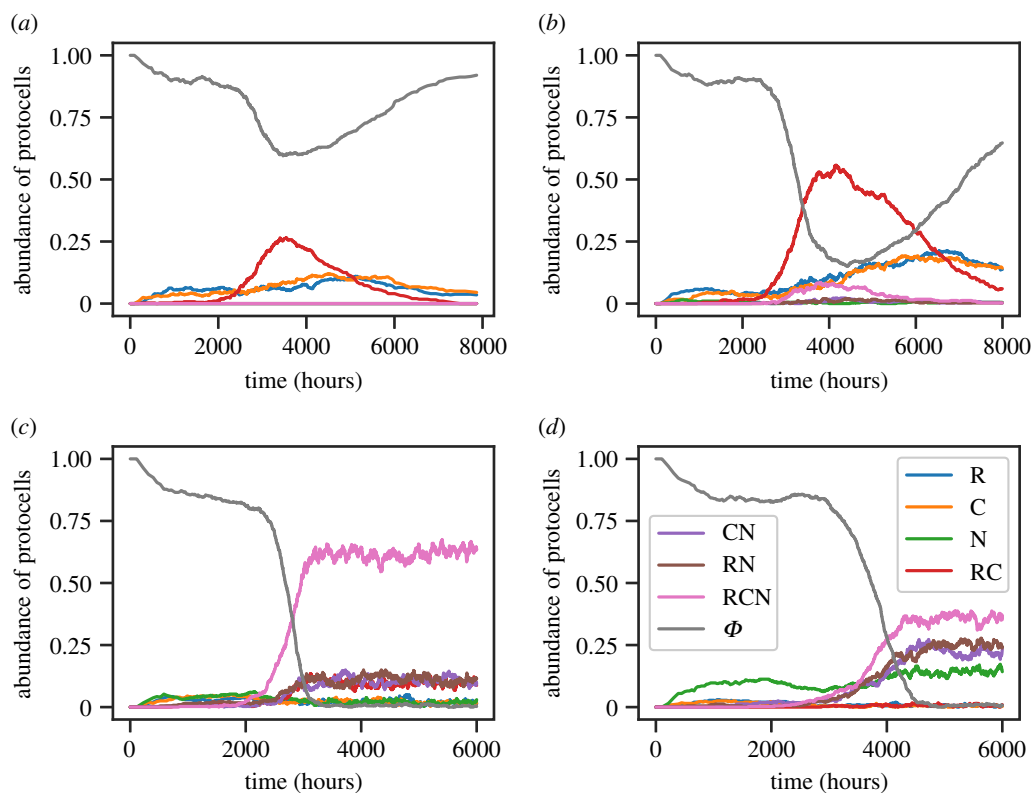


Figure 3. Fractional abundance of protocells containing different types of ribozymes (replicase: R, cyclase: C, nucleotide synthase: N, none: Φ) for (a) $P_r = P_c = 0.03$ and $P_n = 0.0$, (b) $P_r = P_c = 0.03$ and $P_n = 0.002$, (c) $P_r = P_c = 0.03$ and $P_n = 0.03$, and (d) $P_r = P_c = 0.015$ and $P_n = 0.06$, when there is a scarcity of monomers ($S^{\max} = 0.8$, $V_T = 100$, $b = 1$, $V_T = 100$). (Online version in colour.)

dominate the population; since only protocells containing nucleotide synthase can produce enough monomers necessary for new RNA strand formation, some of which may give rise to replicase(s) and/or cyclase(s).

(c) Effect of increasing the threshold volume of protocells

The threshold volume of a protocell was so far arbitrarily fixed ($V_T = 100$) in simulations described in previous sections. Increasing V_T has the advantage of ensuring proliferation of protocells with multiple distinct ribozymes for lower ribozyme formation probabilities. We used the fully stochastic model without peptidyl transferase ($P_n = 0$) to demonstrate the effect of changing threshold volume and ribozyme formation probability. We obtain the same equilibrium abundances of protocells containing ribozymes R,C,N by decreasing the probabilities as $P \propto 1/V_T$ (see the electronic supplementary material, figure S6(A-B)). Nevertheless, the advantage of increasing V_T is offset by the decreased likelihood of ribozyme production owing to the $1/V_T$ dependence of the ribozyme catalysed rates. For a particular value of degradation rate, there will be a limit up to which the threshold volume can be increased (and simultaneously the formation probabilities can be decreased), progressively making it more difficult for ribozymes to form via enzymatic processes before existing ribozymes are degraded (see the electronic supplementary material, figure S6(C)).

The process of membrane transfer from non-lipophilic to lipophilic protocells containing RNA strands, facilitated by a dipeptide that binds to the membrane [21], can lead to increase in V_T via physical processes only. A peptidyl transferase enzyme can ligate free amino acids to form the peptides needed to initiate this process, thereby making it possible for

protocells with peptidyl transferase to increase their threshold volume. This will lead to heterogeneity in the threshold volume of protocells in the population. For a fixed set of values of the formation probabilities, a protocell with peptidyl transferase is more likely to acquire a selective advantage over those that lack such an enzyme, because of its higher threshold volume. The number of RNA molecules inside such protocells will keep growing beyond the threshold allowed for protocells lacking that ribozyme. This increases the chances of creation of more ribozymes of all types inside them, including peptidyl transferase which will increase the threshold volume even further. This explains the rise in average threshold volume of protocells with time (figure 4a) and the higher value of the relative abundance of cells with all four ribozymes compared to the cases when amino acids and peptidyl-transferase were absent. A comparison between figure 4b,c and for the case when $P_r = P_c = P_n > P_p$ (figure not shown) makes it clear that protocells with all four types of ribozymes are most abundant in the population when all ribozymes are created with comparable probabilities, thereby hinting at the existence of a synergistic network between them. The electronic supplementary material, figure S7 shows the peak of the distribution of total number of ribozymes and non-catalytic open-ended strands in dividing cells is more than twice as high as those in cells which are getting eliminated by competition. The electronic supplementary material, figure S8 shows the distribution of the relative abundance of each type of ribozyme in dividing cells.

4. Discussion and conclusion

The plausibility of life based on RNA requires RNA encapsulated within lipid vesicles to not just replicate but also acquire a variety of functions that will ensure that the protocell

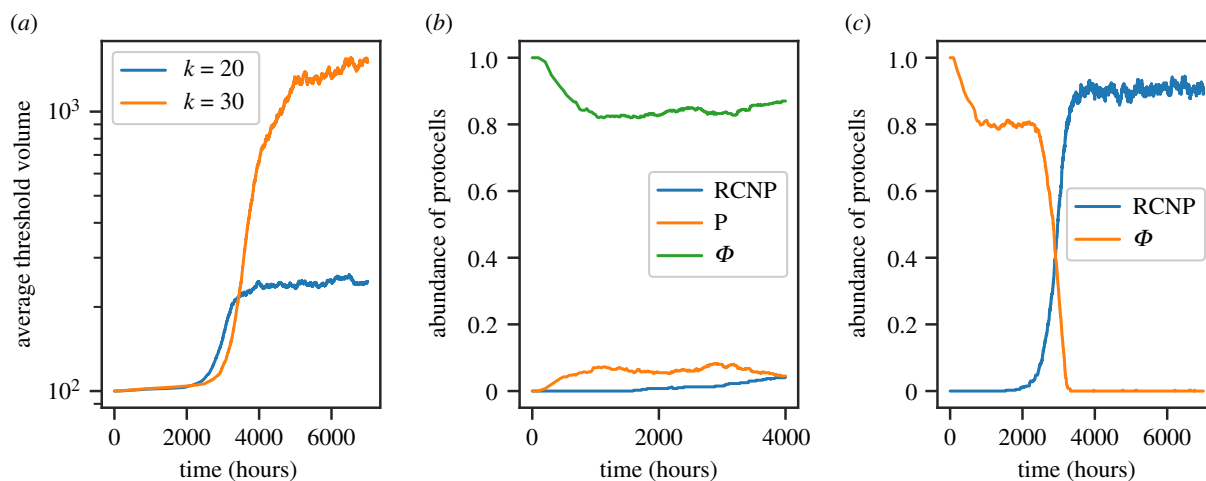


Figure 4. The presence of amino acids and creation of peptidyl-transferase. (a) Average threshold volume of protocells versus time for $V_T^i = V_T^j + 20p_i$ and $V_T^i = V_T^j + 30p_i$ (y-axis is in logscale). The fractional abundance of protocells containing ribozymes (RCNP: replicase + cyclase + nucleotide synthase + peptidyl transferase, Φ : none) versus time for (b) $P_r = P_c = P_n = 0.02$ and $P_p = 0.06$, and (c) $P_r = P_c = P_n = P_p = 0.03$; when $V_T^i = V_T^j + 20p_i$; $S_i^{\max} = 0.8V_T^i$. (Online version in colour.)

turns into a self-sustaining replicator with heritable traits. A population of protocells can then undergo Darwinian evolution in a manner that allows for the emergence and eventual proliferation of more improved variants. Emergence of functionally diverse ribozymes within a protocell requires rapid replication and we show here how the rolling circle mechanism can be harnessed to ensure exponential growth in the number of long RNA strands inside a protocell. A large number of such strands that can fold into complex secondary structures enhance the likelihood of chance emergence of ribozymes. We suggest a hierarchy for the emergence of different ribozymes that would be most beneficial for preferential selection of protocells with more functionally diverse ribozymes during the course of evolution. By varying the formation probability of different ribozymes, we identified the conditions under which evolution of the protocell population leads to proliferation of those protocells that contain larger numbers of functionally distinct ribozymes. We do not claim that the pathway to evolving complexity that we discuss is the only one possible. For example, the ribozymes that catalyse phospholipid synthesis [45] or synthesis of activated nucleotides near neutral pH [46] could also have been some of the earliest ribozymes. Nevertheless, we hope that our proposed pathway provides a framework for thinking about the hierarchical emergence of functional diversity that confers a stepwise increase in selective advantage.

The nature of evolution discussed here is not strictly evolution in the Darwinian sense where sequence-encoded, heritable, phenotypic traits are selected. In the primordial epoch under study, evolution is driven primarily by biophysical processes that make protocells with a larger number of RNA sequences grow at a faster rate. Initially, selection does not distinguish between the nature of the RNA sequence fragments inside the protocells. Eventually, however, protocells with ribozymes are selected, not because selection acts on such catalytic phenotypes, but because protocells containing such ribozymes produce more RNA sequences and therefore grow at a much faster rate than protocells lacking those ribozymes. Nevertheless, such ribozymes should not be considered as a heritable trait until they are genetically encoded. Hence, for functional diversity of ribozymes in protocells to be maintained across generations, ribozymes will

have to be produced with similar probabilities through the error-prone replication process every generation until the emergence of template specificity. To underscore this point, we carried out simulations where the ribozyme creation probabilities are reduced 10-fold after the fraction of protocells with all four ribozymes increases to 0.5. Even in such a scenario, we find that subsequent evolution can no longer sustain protocells containing these ribozymes (see the electronic supplementary material, figure S9(A)). We argue that the low fidelity of the replication process was initially advantageous in creating a diversity of complex, secondary structures from a single (or a few) template(s) that increased the chances of emergence of distinct ribozymes prior to their encoding in the template sequence. An early appearance of such genetic encoding might occur through the emergence of template specificity that allows higher-fidelity, template-directed, catalysed replication of specific ribozymes. Such templates could be thought of as quasi-genes and a collection of distinct quasi-genes, each producing a specific type of ribozyme, might have been a precursor of RNA genomes where different segments encode different functions. We modelled the emergence of template specificity and higher replication fidelity by biasing the replicase catalysed replication towards creating more replicases. Other ribozymes could still be created during the enzymatic replication process, albeit with a much lower probability. As expected, this leads to significant reduction in the fraction of protocells containing all four ribozymes but an increase in the fraction of protocells containing the replicase (see the electronic supplementary material, figure S9(B)). Eventually, the transition to Darwinian evolution will occur only through genetic encoding of heritable traits. This would require not just improved replication accuracy, brought about by the emergence of the replicase ribozyme, but also functional differentiation of replicases to ensure replication of not just catalytic RNA but also of the templates encoding such traits. Takeuchi *et al.* have shown how the division of labour between the template and catalyst [47,48] could have evolved through conflicting multi-level evolution [48] which induce the breaking of symmetry between the strands. The strand that loses its catalytic function and reduce its copy number inside a protocell eventually becomes the genome encoding for functional molecules.

The protocell-division process implemented here is inspired by *in vitro* results [20,49] that show the coupling between protocell growth and division. An interesting alternative [28] takes into account the effect of phase transitions brought about by environmental cycling near terrestrial geothermal pools. In the dry phase, lipid vesicles form multilamellar structures near the edge of the pool. The confinement of monomers within the layers of such structures promote polymerization of RNA strands. During the subsequent wet-phase, such strands can be captured within lipid vesicles budding off from the multilamellar matrix. It has been speculated [28] that repeated wet-dry cycling can gradually select for protocells with a diverse set of enzymes. Validating this intriguing hypothesis through computer simulations would further underline the importance of such an ecological niche on the emergence of the earliest primordial living organisms.

Any pathway that leads to the origin of life needs to not only explain the appearance and proliferation of the earliest ribozymes but also explain the eventual emergence of a genome encoding these functional molecules. Even though our model does not address the latter phenomenon, which resulted in the encoding of heritable traits, it provides a plausible scenario for the origin and proliferation of protocells containing a

functionally diverse set of ribozymes. It also suggests a testable blueprint for the development of synthetic protocells. We hope that our work will motivate new experiments and development of realistic computational models that are informed by experimental findings in the laboratory as well as in plausible environmental settings. Such cross-talk between simulations and experiments will lead to more directed exploration and eventually better understanding of the pathway that led to the origin of life.

Data accessibility. All codes and data used to generate the results in the manuscript and in the electronic supplementary material are available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.866t1g1qs> [50].

Authors' contributions. S.R.: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing—original draft, writing—review and editing; S.S.: conceptualization, formal analysis, methodology, resources, supervision, writing—original draft, writing—review and editing.

Competing interests. We declare we have no competing interests.

Funding. S.R. is supported by an INSPIRE graduate fellowship and S.S. is partially supported by a MATRICS grant no. (MTR/2020/000446), both given by SERB, India.

Acknowledgements. We thank Julien Derr and Paul G. Higgs for interesting discussions and valuable feedback.

References

- Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982 Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell* **31**, 147–157. (doi:10.1016/0092-8674(82)90414-7)
- Stark BC, Kole R, Bowman EJ, Altman S. 1978 Ribonuclease P: an enzyme with an essential RNA component. *Proc. Natl Acad. Sci. USA* **75**, 3717–3721. (doi:10.1073/pnas.75.8.3717)
- Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. 1983 The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857. (doi:10.1016/0092-8674(83)90117-4)
- Robertson DL, Joyce GF. 1990 Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344**, 467–468. (doi:10.1038/344467a0)
- Jaeger L, Wright MC, Joyce GF. 1999 A complex ligase ribozyme evolved *in vitro* from a group I ribozyme domain. *Proc. Natl Acad. Sci. USA* **96**, 14712–14717. (doi:10.1073/pnas.96.26.14712)
- Horning DP, Joyce GF. 2016 Amplification of RNA by an RNA polymerase ribozyme. *Proc. Natl Acad. Sci. USA* **113**, 9786–9791. (doi:10.1073/pnas.1610103113)
- Powner MW, Gerland B, Sutherland JD. 2009 Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **459**, 239–242. (doi:10.1038/nature08013)
- Cafferty BJ, Fialho DM, Khanam J, Krishnamurthy R, Hud NV. 2016 Spontaneous formation and base pairing of plausible prebiotic nucleotides in water. *Nat. Commun.* **7**, 11328. (doi:10.1038/ncomms11328)
- Becker S *et al.* 2019 Unified prebiotically plausible synthesis of pyrimidine and purine RNA ribonucleotides. *Science* **366**, 76–82. (doi:10.1126/science.aax2747)
- Miller SL, Urey HC. 1959 Organic compound synthesis on the primitive Earth: several questions about the origin of life have been answered, but much remains to be studied. *Science* **130**, 245–251. (doi:10.1126/science.130.3370.245)
- Hargreaves WR, Mulvihill SJ, Deamer DW. 1977 Synthesis of phospholipids and membranes in prebiotic conditions. *Nature* **266**, 78–80. (doi:10.1038/266078a0)
- Bangham AD, Horne RW. 1964 Negative staining of phospholipids and their structural modification by surface-active agents as observed in the electron microscope. *J. Mol. Biol.* **8**, 660–668. (doi:10.1016/S0022-2836(64)80115-7)
- Gebicki JM, Hicks M. 1973 Ufasomes are stable particles surrounded by unsaturated fatty acid membranes. *Nature* **243**, 232–234. (doi:10.1038/243232a0)
- Bianconi G, Zhao K, Chen IA, Nowak MA. 2013 Selection for replicases in protocells. *PLoS Comput. Biol.* **9**, e1003051. (doi:10.1371/journal.pcbi.1003051)
- Joyce GF, Szostak JW. 2018 Protocells and RNA self-replication. *Cold Spring Harbor Perspect. Biol.* **10**, a034801. (doi:10.1101/cshperspect.a034801)
- Kamimura A, Matsubara YJ, Kaneko K, Takeuchi N. 2019 Horizontal transfer between loose compartments stabilizes replication of fragmented ribozymes. *PLoS Comput. Biol.* **15**, e1007094. (doi:10.1371/journal.pcbi.1007094)
- Black RA, Blosser MC, Stottrup BL, Tavakley R, Deamer DW, Keller SL. 2013 Nucleobases bind to and stabilize aggregates of a prebiotic amphiphile, providing a viable mechanism for the emergence of protocells. *Proc. Natl Acad. Sci. USA* **110**, 13272–13276. (doi:10.1073/pnas.1300963110)
- Adamala K, Szostak JW. 2013 Nonenzymatic template-directed RNA synthesis inside model protocells. *Science* **342**, 1098–1100. (doi:10.1126/science.1241888)
- Chen IA, Roberts RW, Szostak JW. 2004 The emergence of competition between model protocells. *Science* **305**, 1474–1476. (doi:10.1126/science.1100757)
- Hanczyc MM, Fujikawa SM, Szostak JW. 2003 Experimental models of primitive cellular compartments: encapsulation, growth, and division. *Science* **302**, 618–622. (doi:10.1126/science.1089904)
- Adamala K, Szostak JW. 2013 Competition between model protocells driven by an encapsulated catalyst. *Nat. Chem.* **5**, 495–501. (doi:10.1038/nchem.1650)
- Lai Y-C, Chen IA. 2020 Protocells. *Curr. Biol.* **30**, R482–R485. (doi:10.1016/j.cub.2020.03.038)
- O'Flaherty DK, Kamat NP, Mirza FN, Li L, Prywes N, Szostak JW. 2018 Copying of mixed-sequence RNA templates inside model protocells. *J. Am. Chem. Soc.* **140**, 5171–5178. (doi:10.1021/jacs.8b00639)
- Chen IA, Salehi-Ashtiani K, Szostak JW. 2005 RNA catalysis in model protocell vesicles. *J. Am. Chem. Soc.* **127**, 13 213–13 219. (doi:10.1021/ja051784p)
- Takeuchi N, Hogeweg P. 2009 Multilevel selection in models of prebiotic evolution II: a direct comparison of compartmentalization and spatial self-organization. *PLoS Comput. Biol.* **5**, e1000542. (doi:10.1371/journal.pcbi.1000542)

26. Derr J, Manapat ML, Rajamani S, Leu K, Xulvi-Brunet R, Joseph I, Nowak MA, Chen IA. 2012 Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences. *Nucleic Acids Res.* **40**, 4711–4722. (doi:10.1093/nar/gks065)
27. Shah V, de Bouter J, Quinn P, Tupper A, Higgs PG. 2019 Survival of RNA replicators is much easier in protocells than in surface-based, spatial systems. *Life* **9**, 65. (doi:10.3390/life9030065)
28. Damer B, Deamer D. 2015 Coupled phases and combinatorial selection in fluctuating hydrothermal pools: a scenario to guide experimental approaches to the origin of cellular life. *Life* **5**, 872–887. (doi:10.3390/life5010872)
29. Damer B, Deamer D. 2020 The hot spring hypothesis for an origin of life. *Astrobiology* **20**, 429–452. (doi:10.1089/ast.2019.2045)
30. Rajamani S, Vlassov A, Benner S, Coombs A, Olasagasti F, Deamer D. 2007 Lipid-assisted synthesis of RNA-like polymers from mononucleotides. *Origins Life Evol. Biospheres* **38**, 57–74. (doi:10.1007/s11084-007-9113-2)
31. DeGuzman V, Vercoutere W, Shenasa H, Deamer D. 2014 Generation of oligonucleotides under hydrothermal conditions by non-enzymatic polymerization. *J. Mol. Evol.* **78**, 251–262. (doi:10.1007/s00239-014-9623-2)
32. Forsythe JG, Yu S-S, Mamajanov I, Grover MA, Krishnamurthy R, Fernández FM, Hud NV. 2015 Ester-mediated amide bond formation driven by wet-dry cycles: a possible path to polypeptides on the prebiotic Earth. *Angew. Chem. Int. Ed.* **54**, 9871–9875. (doi:10.1002/anie.201503792)
33. Higgs P. 2016 The effect of limited diffusion and wet–dry cycling on reversible polymerization reactions: implications for prebiotic synthesis of nucleic acids. *Life* **6**, 24. (doi:10.3390/life6020024)
34. Roy S, Bapat NV, Derr J, Rajamani S, Sengupta S. 2020 Emergence of ribozyme and tRNA-like structures from mineral-rich muddy pools on prebiotic Earth. *J. Theor. Biol.* **506**, 110446. (doi:10.1016/j.jtbi.2020.110446)
35. Bapat NV, Rajamani S. 2015 Effect of co-solutes on template-directed nonenzymatic replication of nucleic acids. *J. Mol. Evol.* **81**, 72–80. (doi:10.1007/s00239-015-9700-1)
36. Jin L, Engelhart AE, Zhang W, Adamala K, Szostak JW. 2018 Catalysis of template-directed nonenzymatic RNA copying by iron(II). *J. Am. Chem. Soc.* **140**, 15016–15021. (doi:10.1021/jacs.8b09617)
37. Dingle K, Schaper S, Louis AA. 2015 The structure of the genotype–phenotype map strongly constrains the evolution of non-coding RNA. *Interface Focus* **5**, 20150053. (doi:10.1098/rsfs.2015.0053)
38. Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017 Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 1–7. (doi:10.1038/s41559-017-0127)
39. Dingle K, Ghaddar F, Sulc P, Louis AA. In press. Phenotype bias determines how natural RNA structures occupy the morphospace of all possible shapes. *Mol. Biol. Evol.* (doi:10.1093/molbev/msab280)
40. Kusumoto-Matsuo R, Kanda T, Kukimoto I. 2010 Rolling circle replication of human papillomavirus type 16 DNA in epithelial cell extracts. *Genes Cells* **16**, 23–33. (doi:10.1111/j.1365-2443.2010.01458.x)
41. Daròs J-A, Elena SF, Flores R. 2006 Viroids: an Ariadne's thread into the RNA labyrinth. *EMBO Rep.* **7**, 593–598. (doi:10.1038/sj.embor.7400706)
42. Flores R, Gas M-E, Molina-Serrano D, Nohales M-Á, Carbonell A, Gago S, De la Peña M, Daròs J-A. 2009 Viroid replication: rolling-circles, enzymes and ribozymes. *Viruses* **1**, 317–334. (doi:10.3390/v1020317)
43. Tupper AS, Higgs PG. 2021 Rolling-circle and strand-displacement mechanisms for non-enzymatic RNA replication at the time of the origin of life. *J. Theor. Biol.* **527**, 110822. (doi:10.1016/j.jtbi.2021.110822)
44. Hassenkam T, Damer B, Mednick G, Deamer D. 2020 AFM images of viroid-sized rings that self-assemble from mononucleotides through wet–dry cycling: implications for the origin of life. *Life* **10**, 321. (doi:10.3390/life10120321)
45. Budin I, Szostak JW. 2011 Physical effects underlying the transition from primitive to modern cell membranes. *Proc. Natl Acad. Sci. USA* **108**, 5249–5254. (doi:10.1073/pnas.1100498108)
46. Martin L, Unrau P, Müller U. 2015 RNA synthesis by *in vitro* selected ribozymes for recreating an RNA world. *Life* **5**, 247–268. (doi:10.3390/life5010247)
47. Takeuchi N, Hogeweg P, Koonin EV. 2011 On the origin of DNA genomes: evolution of the division of labor between template and catalyst in model replicator systems. *PLoS Comput. Biol.* **7**, e1002024. (doi:10.1371/journal.pcbi.1002024)
48. Takeuchi N, Hogeweg P, Kaneko K. 2017 The origin of a primordial genome through spontaneous symmetry breaking. *Nat. Commun.* **8**, 250. (doi:10.1038/s41467-017-00243-x)
49. Zhu TF, Szostak JW. 2009 Coupled growth and division of model protocell membranes. *J. Am. Chem. Soc.* **131**, 5705–5713. (doi:10.1021/ja900919c)
50. Roy S, Sengupta S. 2021 Data from: Evolution towards increasing complexity through functional diversification in a protocell model of the RNA world. Dryad Digital Repository. (doi:10.5061/dryad.866t1g1qs)