Self-Assembly of Informational Polymers by Templated Ligation

Joachim H. Rosenberger,[1,*] Tobias Göppel,[1,*] Patrick W. Kudella,[2] Dieter Braun,[2,‖] Ulrich Gerland,[1,†] and Bernhard Altaner[1,‡,§]

[1]*Physics of Complex Biosystems, Technical University of Munich, 85748 Garching, Germany*
[2]*Systems Biophysics, Ludwig Maximilian University Munich, 80799 Munich, Germany*

(Received 25 November 2020; revised 21 June 2021; accepted 25 June 2021; published 13 September 2021)

The emergence of evermore complex entities from prebiotic building blocks is a key aspect of origins of life research. The RNA-world hypothesis posits that RNA oligomers known as ribozymes acted as the first self-replicating entities. However, the mechanisms governing the self-assembly of complex informational polymers from the shortest prebiotic building blocks were unclear. One open issue concerns the relation between concentration and oligonucleotide length, usually assumed to be exponentially decreasing. Here, we show that a competition of timescales in the self-assembly of informational polymers by templated ligation generically leads to nonmonotonic strand-length distributions with two distinct length scales. The first length scale characterizes the onset of a strongly nonequilibrium regime and is visible as a local minimum. Dynamically, this regime is governed by extension cascades, where the elongation of a "primer" with a short building block is more likely than its dehybridization. The second length scale appears as a local concentration maximum and reflects a balance between degradation and dehybridization of completely hybridized double strands in a heterocatalytic extension-reassembly process. Analytical arguments and extensive numerical simulations within a sequence-independent model allowed us to predict and control these emergent length scales. Nonmonotonic strand-length distributions confirming our theory were obtained in thermocycler experiments using random DNA sequences from a binary alphabet. Our work emphasizes the role of structure-forming processes already for the earliest stages of prebiotic evolution. The accumulation of strands with a typical length reveals a possible starting point for higher-order self-organization events that ultimately lead to a self-replicating, evolving system.

DOI: 10.1103/PhysRevX.11.031055

Subject Areas: Biological Physics, Chemical Physics
Statistical Physics

I. INTRODUCTION

A key question in research on the origins of life is how structure and biochemical complexity could emerge from unstructured conditions on early Earth. One of the most well-known hypotheses in this context is that of an "RNA world" [1–6]. In this scenario, RNA oligomers acted as both the carrier of information and "ribozymes," i.e., catalytic molecules allowing for the replication of this information and other metabolic functions. Yet, the RNA-world hypothesis does not address the question of how RNA strands that are complex enough to act as functional ribozymes came into being [7–11]. In the light of evolutionary principles, a multistep scenario of self-organization seems plausible, cf. Fig. 1. However, the intermediate steps on the way toward functional polynucleotides are still not well understood.

The smallest ribozymes known today are 50–100 nucleotides (nt) long [12–14]. A common view is that the reliable self-assembly of replicating RNA molecules required specific sequences of 20–30 nt in length [15–18]. It has been shown that selective (Watson-Crick) base pairing can lead to a vast reduction of complexity in sequence space, a phenomenon called cooperative ligation [19,20]. Moreover, a recent hypothesis suggests that a replicating catalytic network would emerge as a "virtual circular genome," which self-assembles from an initial distribution of short oligonucleotides [16].

The efficiency and viability of such catalytic networks strongly depend on the relative concentrations of oligonucleotides of different lengths. Commonly exponential length distributions are assumed [16,21]. While a decay in length is natural, the exact shape of length distributions emerging from short building blocks is not known [16]. Models and experiments have reported different observations [11,19,22].

*These authors contributed equally to this work.
†gerland@tum.de
‡bernhard.altaner@tum.de
§Spokesperson.
‖dieter.braun@lmu.de

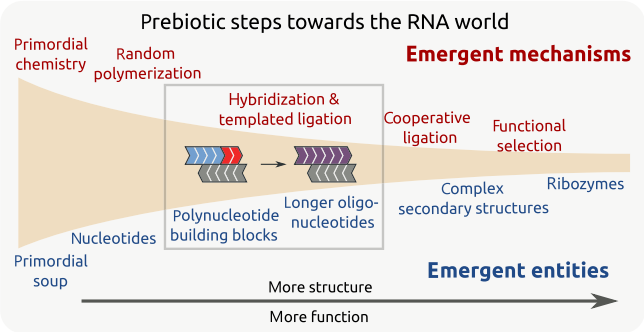

FIG. 1. Prebiotic evolution is a multistep process that creates new entities exhibiting emergent mechanisms of interaction. Our work outlines the emergence of structured oligonucleotides from the smallest building blocks.

Importantly, for low concentrations, the concatenation of oligonucleotides is dominated by a process known as *templated ligation* [11,23–25]. Unlike *random ligation*, where two oligonucleotides directly combine into a longer strand [26,27], templated ligation involves a third strand, cf. Fig. 1. The third strand, also called the "template," enables the covalent bonding of two other strands adjacently hybridized on the template. Thus, the self-assembly of oligonucleotides cannot be captured by standard polymerization theory, where exponential length distributions are well understood [28,29].

Probing the length distribution arising from templated ligation is challenging [16]. When forming covalent bonds between oligonucleotides, an energy gap needs to be overcome. In enzyme-free situations, this energy can be provided by an activation chemistry often involving imidazole or 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide [9,10,30–32]. The yields are usually small, experiments require a long time, and results can be obscured by side products [7,15,33,34]. As a consequence, experimental research has focused on so-called "primer extension," which regards the extension of a short ($\sim$10 nt) oligonucleotide "primer" on a longer template, rather than self-assembly of oligonucleotides from small building blocks [30,35–41].

An alternative for providing the energy for bond formation is using a ligase, which can be either an RNA-based ribozyme or a modern protein. The latter is not prebiotically plausible. However, these enzymes drastically increase yields and reaction speeds. While ligases require oligomers of lengths between six and ten nucleotides, enzymatic systems can still serve as conceptual models to explore the principles of self-assembly and ligation-based early replication [8,19].

In order to study the self-assembly from smallest building blocks, we employed a computational and analytical approach based on a minimal "bottom-up" model. A transition between two dynamical regimes featuring differently decaying distributions has been reported recently [22]. These results were obtained within a coarse-grained, deterministic model, which does not capture the entire complexity associated with templated ligation.

A general yet simple theory identifying the generic properties of the self-assembly from shortest oligomer building blocks has been missing. The goal of this work is to close this gap. To this end, we investigated a model that captures the elementary mechanism of self-assembly: the hybridization of strands to form arbitrary complexes on which templated ligation can occur. To focus on the assembly process alone, the dependence on oligonucleotide sequences was neglected: The binding energy of a hybridization site is proportional to its length and characterized by a single parameter γ, reflecting a typical binding energy *per nucleotide*, which emerges naturally in mean-field descriptions like the "random sequence approximation" [22].

Our main result is that the competition of timescales between (length-dependent) dehybridization, extension, and a degradation or observation timescale generically leads to a *nonmonotonic strand-length* distribution. We show that different dynamic processes govern different regions in the space of strand lengths. The boundaries between these regions are given by a local minimum at a length $L_{\min}$ and a local maximum at $L_{\max} > L_{\min}$, which can be approximated by two analytical length scales $L^* \sim L_{\min}$ and $L^\dagger \sim L_{\max}$. This accumulation of strands at the typical length scale $L_{\max}$ constitutes a novel structure-forming process. Many of the microscopic details only enter the theory via a single parameter that characterizes the effective rate of extension. This allowed us to apply our theory to experiments, where a nonmonotonic length distribution emerges from the enzymatic ligation of a random pool of DNA sequences in a thermocycler.

II. MODEL AND SIMULATION METHOD

A. State of the art

Previous theoretical work largely studied templated ligation by effective models. The description of the state space had been reduced to strand lengths, without taking into account the hybridization complexes explicitly [20,22,24,42–49]. In such a coarse-grained picture, (de) hybridization and templated ligation are not elementary reactions but are combined into an effective extension reaction. To specify the corresponding rate, the intricacies of the assembly process are neglected and *a priori* assumptions regarding the relevant configurations are made [20,24,42–45]. Many models neglect the dependence of the binding energy on the number of paired nucleotides [20,24,42–45,47,48]. Others consider a length-dependent dehybridization rate only up to some cutoff length such that the timescale of ligation is always much larger than the timescale of the dehybridization kinetics [22]. A study addressing the full complexity of the assembly was limited by small system sizes [50].

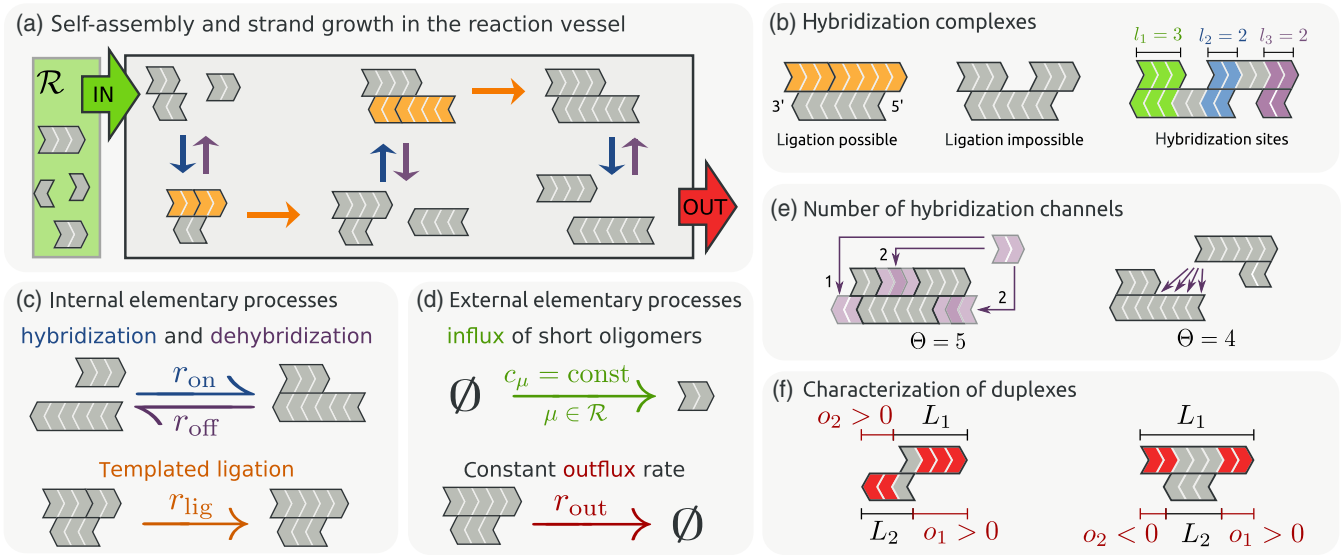

FIG. 2. (a) Short strands entering the reaction vessel from the reservoir are the initial building blocks of the system. Inside the vessel, strands form various complexes via hybridization and dehybridization. Subsequent ligation leads to longer strands eventually leaving the system. (b) Examples of higher-order complexes with multiple hybridiziation sites. (c) The internal elementary processes are hybridization, dehybridization, and templated ligation with corresponding rates r_{on}, r_{off}, and r_{lig}. (d) The external elementary processes couple the system to its environment. Short strands of length $L = \mu$ for $\mu \in \mathcal{R}$ are chemostated via the coupling to an external reservoir of initial building blocks at fixed concentrations c_μ. All complexes leave the system at a constant rate r_{out}. (e) When two complexes collide, they can form Θ different hybridization complexes. (f) Duplexes $D \coloneqq (L_1, L_2, o_1)$ are uniquely characterized by the strand lengths $L_1, L_2 \in \mathbb{N}$ and one of the overhangs $o_i \in \mathbb{Z}$ of strand S_i, $i \in \{1, 2\}$, at its $3'$ end. Overhangs o_i can be negative, as for the case of o_2 in the right-hand example.

B. Model

Figure 2(a) sketches the model dynamics. Short oligomers enter a reaction volume V, where they hybridize to form (partially) double-stranded complexes. If oligomers aggregate in suitable configurations, they may undergo templated ligation. Eventually, all complexes leave the reaction vessel at a constant rate, mimicking a flow reactor.

Strands and complexes.—The basic element of our dynamics is a directed oligomer called a strand, which consists of covalently linked nucleotides. All linear conformations that can arise from a set of strands are allowed; see Fig. 2(b). Only self-folding and branched hybridization structures are excluded. While these effects might become important for longer strands, they can be neglected when dealing with the self-assembly from short strands. The overlapping region between two strands is referred to as a hybridization site. Single strands are called m-mers. We explicitly refer to monomers, dimers, trimers, and tetramers for $m = 1, 2, 3, 4$.

Elementary processes and parameters.—The internal elementary reactions are hybridization, dehybridization, and templated ligation; see Fig. 2(c). Hybridization and dehybridization are assumed to be elementary and reversible reactions with rates r_{on} and r_{off}. Thermodynamic consistency [51,52] connects r_{on} and r_{off} to the free energy ΔG_b° of a hybridization site:

$$\frac{r_{\mathrm{off}}}{r_{\mathrm{on}}} = (V N_A c^\circ) e^{\beta \Delta G_b^\circ}, \tag{1}$$

where $\beta = (k_B T)^{-1}$, k_B is Boltzmann's constant and T denotes the absolute temperature, N_A is the Avogadro constant and $c^\circ = 1$ mol/l is the standard concentration.

When two strands of length L_1 and L_2 are hybridized adjacently on a third strand, they may ligate to a new strand of length $L_1 + L_2$. The ligation rate r_{lig} is assumed to be independent of L_1, L_2, the directionality of the strands, and microscopic details. The uniform ligation rate can be interpreted as an effective average. A more detailed model could reflect that short oligomers predominately ligate to $3'$ ends [16,53] due to the underlying chemistry [54,55] or include stalling effects [39,56,57]. Since template-free ligation is a much slower process than templated ligation [23], it is neglected. Moreover, two external reactions connect the system to its environment, cf. Fig. 2(d). (i) A coupling to a reservoir fixes the concentrations c_m of m-mers with $m \in \mathcal{R}$. (ii) Each complex exits the system at a constant rate r_{out}.

Thermodynamics and kinetics of hybridization.—The binding energy ΔG_b° of a hybridization is assumed to be directly proportional to the length of the binding site l, see Fig. 2(b),

$$\beta \Delta G_b^\circ(l) = \gamma l, \tag{2}$$

where $\gamma < 0$ is a parameter that gives the (negative) binding energy per unit length in units of $k_B T$.

Equation (1) thermodynamically constrains the ratio of r_{on} and r_{off}. An additional kinetic parameter is needed for a full parametrization of the rates. Here, we use a constant rate of collision between two complexes $r_{coll} = (VN_A c^\circ t_0)^{-1}$, where $t_0 = (r_0)^{-1}$ is a microscopic, intensive, collision timescale; see Sec. S1 of the Supplemental Material (SM) [58]. All times are measured in units of t_0. In general, two colliding complexes can form multiple configurations via Θ distinct hybridization channels; see Fig. 2(e). The probability of choosing each of these channels is assumed to be equal:

$$p_{hyb} = 1/\Theta, \qquad \Theta > 0. \tag{3}$$

Hence, the hybridization rate via a given channel is

$$r_{on} = r_{coll} p_{hyb}. \tag{4}$$

For the dehybridization rate we obtain from Eq. (1)

$$r_{off} = \frac{1}{\Theta} e^{\gamma l}. \tag{5}$$

In reality, the collision rate depends on the properties of the colliding complexes, the solvent, and temperature. A parametrization where the binding energy γl is attributed to the dehybridization rate r_{off} is a common kinetic assumption, and has been confirmed experimentally [59–61]. The kinetic assumptions Eqs. (4) and (5) reduce the computational complexity, while still maintaining the sampling of all configurations thermodynamically consistent; see Sec. S1 in the SM [58].

In addition to this standard model, we also consider a "bounded" variant, where the dehybridization rate cannot become smaller than a minimal rate r_{cut}, such that $r_{off} = r_{cut}$ if $e^{\gamma l}/\Theta < r_{cut}$. The bounded model can be thought of as an effective implementation of a system subjected to an external mechanism causing dehybridization of *all* complexes with a timescale of $\tau \sim (r_{cut})^{-1}$. Such a situation can be realized by the thermal cycling in a "thermal trap" situated in a hydrothermal vent or be the consequence of other naturally occurring cycles [22,62–66].

Standard parameters.—Our primary focus is a scenario where the building blocks entering from the reservoir are dimers only. If not indicated otherwise, $c_2 = 2$ mM. The volume V is chosen such that 10^4 single-stranded dimers are present. This dimer-only scenario is the simplest model allowing for templated ligation and makes analytical considerations easier. If not otherwise stated, $\gamma = -0.5$, $r_{lig} = e^{-6}$, and $r_{out} = 5 \times 10^{-9}$. In the bounded model r_{cut} is a further parameter.

Implementation.—The model dynamics were implemented in C++ using the Gillespie algorithm [67–69] (see Sec. S1 in the SM for details [58]).

III. SIMULATION RESULTS AND ANALYSIS

The main observable in this work is the length distribution of strands $\rho(L)$. It expresses the concentration of strands of length L, irrespective of the complexes they belong to.

A. Self-enhancing catalysis leads to long-tailed distributions

Self-assembly via templated ligation is a self-enhancing mode of growth, where long strands facilitate their own formation. This process competes with degradation. For large outflux rates, strands remain inside the reaction volume only for short times and participate in few or even no templated ligations. The resulting stationary length distribution is therefore expected to be short tailed.

In contrast, for a small outflux rate, strands spend more time inside the system and thus have a higher chance to serve as a template or to get ligated, leading to the formation of longer strands. These longer strands again allow for larger hybridization sites and are better templates. Consequently, we expect the existence of a crossover value for the outflux rate $r_{out} = r_{out}^c$, where the formation of longer strands is dominantly self-enhancing. In the Appendix A, we derive the value of the crossover rate,

$$r_{out}^c = 2(c_2)^2(e^{-4\gamma} + 2e^{-3\gamma}) r_{lig}, \tag{6}$$

under the assumption that (i) short-tailed distributions are dominated by the smallest building blocks and (ii) timescales of the dehybridization of these building blocks are small compared to the timescale of ligation.

We probed the stationary distribution for various values of the outflux rate r_{out}. Simulation results for the standard model are shown in Fig. 3(a). Figure 3(b) gives the analogous results for the bounded model.

Since the derivation of Eq. (6) does not rely on the dynamics of long strands affected by the cutoff, we expect the same transition from short- to long-tailed distributions in both scenarios. For sufficiently large outflux rates the resulting short-tailed length distributions look quantitatively similar. The curves for the crossover outflux rate $r_{out}^c = 3.24 \times 10^{-7}$ obtained from Eq. (6) are indicated as dashed (orange) lines. The long-tailed distributions for small outflux rates differ significantly: In the standard model, Fig. 3(a), a local minimum and maximum emerge. In contrast, the long-tailed distributions in the bounded model, Fig. 3(b), decay monotonically.

This behavior is rationalized in the right-hand column of Fig. 3, where we sketch the dependence of the (effective) rates of the processes affecting the strand length. The crucial effective growth process is the extension reaction, i.e., hybridization of a third strand followed by ligation. The effective rate is denoted by r_{ext}. In the unbounded model, the dehybridization rate r_{off} intersects the horizontal

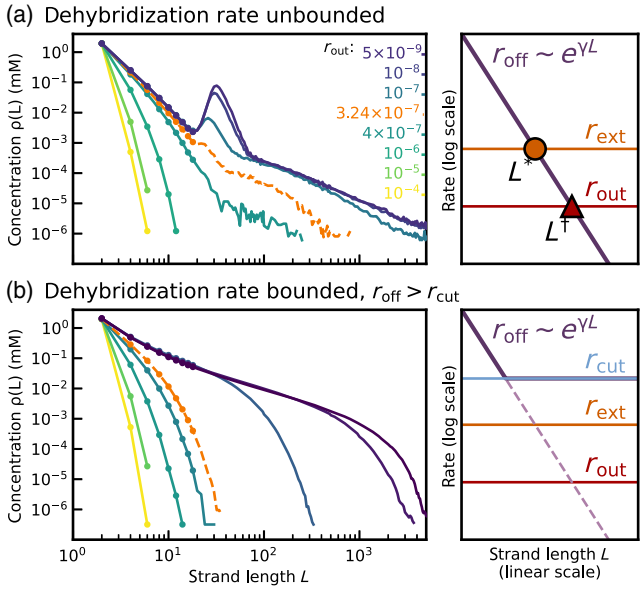

FIG. 3. Stationary length distributions (left) and competition of timescales (right) for the standard model (a) and its bounded variant (b) for different values of the outflux rate r_{out}. In the bounded model, dehybridization cannot become smaller than $r_{cut} = 0.05$. Dehybridization is thus faster than ligation ($r_{lig} = 2.5 \times 10^{-3}$) for all lengths. In both models, the length distributions develop long tails when decreasing the outflux rate r_{out}. The orange (dashed) curves correspond to a system where the outflux rate takes the crossover value $r_{out} = 3.24 \times 10^{-7}$, cf. Eq. (6). For outflux rates below the transition value, the unbounded model exhibits a nonmonotonic length distribution with a local minimum and maximum at L_{min} and L_{max}.

lines corresponding to constant extension and outflux rates at two distinct length scales L^* and $L^\dagger$. This already hints at the two emergent length scales L_{min} and L_{max} in the length distribution. This intersection does not occur for the bounded model.

An analogous argument to Eq. (6) for the transition from long to short tails was made in Ref. [22]. There, the authors studied assembly in a model, where strands break by cleavage. The crucial difference between their work and our unbound model is that in their model ligation is always the slowest process.

B. Competition of timescales enables extension cascades and persisting complexes

We now focus on the standard model without effective thermal cycling and with a sufficiently low outflux rate. It already became clear that the nonmonotonic behavior stems from complexes for which dehybridization is not necessarily the fastest process. If the binding energy of a duplex is close to zero, it dehybridizes quickly. In contrast, if the binding energy has a large absolute value, the duplex is stable and the extension with a third strand becomes probable. The extended complex is even more stable and another extension becomes even more probable. We call this phenomenon an *extension cascade*.

Disregarding dehybridization and outflux for now, an extension cascade only stops when no further extension is possible. In our model this is only the case for a fully hybridized duplex consisting of two maximally overlapping strands of the same length. These duplexes persist for long times. The fate of such a long-lived complex is determined by either dehybridization or outflux.

Structure of complexes.—We partition complexes into different classes by distinguishing between single strands, duplexes, and higher-order complexes, cf. Fig. 4(a). We further subdivide duplexes according to "parity": Fully hybridized duplexes have zero parity. In contrast, duplexes with even or odd overhangs have even or odd parity. Note that in the dimer-only model, mixed parities are excluded, because all strand lengths are even.

Extension cascades only reach a terminal fully hybridized duplex when they start from even duplexes. Duplexes with odd parity will undergo quasi-infinite extension cascades. Figures 4(a) and 4(b) show the partitioned length distribution. Short strands are mostly single stranded. In contrast, the concentration peak is dominated by fully hybridized duplexes. The effect of quasi-infinite extension cascades is visible in the tail. Higher-order complexes are

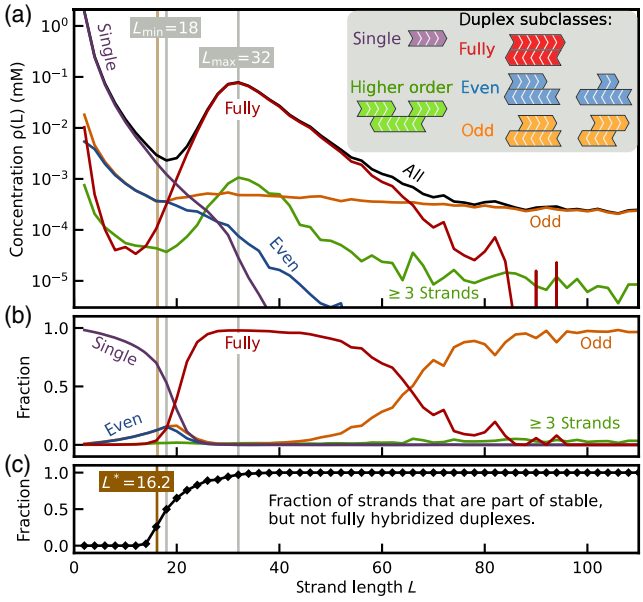

FIG. 4. (a) Partitioning the contributions of the different subgroups to the strand-length distribution reveals the dominant configurations: Short strands are mostly single stranded. Strands with lengths around the peaks are in the persistent fully hybridized zero-parity configuration. In the dimer-only model, odd duplexes never reach a fully hybridized state and cause the long tail of the distribution. (b) The probability of different complex types conditioned on strand length. (c) The probability that a duplex with nonzero parity is stable conditioned on strand length. Around $L = L^*$ [cf. Eq. (13)] this probability increases rapidly.

less abundant and do not contribute significantly to the shape of the distribution.

The minimum at $L = L_{\min}$ is due to the increase of the concentration of fully hybridized duplexes at a characteristic length scale $L^* \lesssim L_{\min}$, which we derive below. Figure 4(c) shows that L^* is the typical length scale on which duplexes become stable enough for extension cascades to start.

Kinetics of duplexes.—Since the dehybridization rate depends on the length of the hybridization site, it connects timescales to length scales. As such, the characteristic scales L^* and $L^\dagger$ also divide the length distribution into different dynamical regimes. Since the length distribution is dominated by single strands and duplexes, we now consider the kinetics of duplexes in detail.

A duplex consisting of strands S_1 and S_2 with lengths L_1 and L_2 is fully characterized by the 3-tuple $D := (L_1, L_2, o_1)$. The number o_1 is the (positive or negative) overhang of strand S_1 on its $3'$ end; see Fig. 2(f). When the two strands collide, they can form $\Theta = L_1 + L_2 - 1$ different duplexes. Applying this to Eq. (5), the dehybridization rate becomes

$$r_{\text{off}}^{\text{dupl}}(D) = \frac{1}{L_1 + L_2 - 1} e^{\gamma l(D)}. \tag{7}$$

First, we formally derive the onset of extension cascades at L^*: Hybridization of a short m-mer can occur on one of the two nonzero overhangs o_i, $i \in (1, 2)$, of the duplex D and results in a triplex T_i. If the m-mer is subsequently ligated to its neighboring strand, we call the combined process an *extension*. In that case, the length of the hybridization site of the m-mer with the duplex is $z_i = \min(|o_i|, m)$. To calculate an effective rate for this process, we assume that the dynamics of the m-mer hybridization is fast compared to the ligation rate and to the dehybridization rate of the duplex D. Consequently, the concentrations of the duplex D and the triplex T_i can be assumed to be at a binding equilibrium and we obtain

$$c_{T_i} = c_D c_m e^{-\gamma z_i}. \tag{8}$$

With this, we define the effective extension rate with an m-mer as the ratio of the rate of ligations from that triplex and the duplex concentration, i.e., $r_{\text{ext},m} = r_{\text{lig}} c_T / c_D$. Using Eq. (8) and taking into account that there are generally two ligation sites ($i = 1, 2$), the extension rate reads

$$r_{\text{ext},m}(D) = r_{\text{lig}} c_m \sum_{\substack{i \in \{1,2\} \\ o_i \neq 0}} e^{-\gamma z_i}. \tag{9}$$

Hence, the extension rate with a short m-mer is

$$r_{\text{ext}}(D) = \sum_m r_{\text{ext},m}(D). \tag{10}$$

The ratio of r_{ext} and r_{off} gives the condition for the onset of extension cascades for the duplex D, $1 < r_{\text{ext}}(D)/r_{\text{off}}^{\text{dupl}}(D)$. As dimers are the most abundant species, we approximate $r_{\text{ext}}(D) \gtrsim r_{\text{ext},2}(D)$, yielding the lower bound:

$$1 < \frac{r_{\text{ext},2}(D)}{r_{\text{off}}^{\text{dupl}}(D)}. \tag{11}$$

To be more systematic, we now consider a system containing only strand lengths smaller or equal to some fixed value L_0. We then determine the minimal L_0 such that duplexes appear which can undergo extension cascades. Using Eqs. (7) and (9) we write the ratio in Eq. (11) as

$$\frac{r_{\text{ext},2}(D)}{r_{\text{off}}^{\text{dupl}}(D)} = (L_1 + L_2 - 1) c_2 r_{\text{lig}} \sum_{\substack{i \in \{1,2\} \\ o_i \neq 0}} e^{-\gamma(l + z_i)}. \tag{12}$$

This ratio is largest for symmetric duplexes with $L_1 = L_2 = L_0$, where $l(D) + z_i = L_0$. The two duplex configurations maximizing the ratio are thus the odd duplex $D_{\pm 1} = (L_0, L_0, \pm 1)$ and the even duplex $D_{\pm 2} = (L_0, L_0, \pm 2)$. The smallest L_0, for which extension cascades are possible, defined as L^*, are found by solving

$$1 = 2(2L^* - 1) c_2 r_{\text{lig}} e^{-\gamma L^*}, \tag{13}$$

which yields $L^* \approx 16.2$. As the shortest building blocks are dimers, $L_\bullet^* = \lceil L^* \rceil$ is calculated by ceiling L^* to the next even integer, i.e., $L_\bullet^* = 18$.

For strong binding, i.e., $\gamma < -1$, the subexpontial length dependence which enters via the channel number Θ can be neglected. To leading order one then has

$$L^* \approx \ln\left(c_2 \frac{r_{\text{lig}}}{r_0}\right) \gamma^{-1}, \tag{14}$$

where we made the dependence of the microscopic kinetic parameter r_0 explicit.

The distinct peak in the strand-length distribution is caused by fully hybridized duplexes $(L, L, 0)$ being end points of extension cascades. These duplexes persist until they dehybridize or leave the system. This gives rise to two different fates depending on their length. For L smaller than some critical value $L^\dagger$, $r_{\text{off}}^{\text{dupl}}(L, L, 0) > r_{\text{out}}$, duplex production in the stationary state is mostly balanced by dehybridization. For long duplexes with $L > L^\dagger$, we have $r_{\text{off}}^{\text{dupl}}(L, L, 0) < r_{\text{out}}$, and hence the stationary concentration is mostly determined by a balance of their production with the outflux. The outflux rate r_{out} is independent of L, whereas r_{off} decreases exponentially with L. We thus expect the existence of two different regimes where the stationary concentration of the fully hybridized duplexes exhibits a different dependence on L. We can find the length where the dehybridization rate becomes smaller than the outflux rate by

$$r_{\mathrm{off}}(L^{\dagger},\gamma)=\frac{e^{\gamma L^{\dagger}}}{2L^{\dagger}-1}=r_{\mathrm{out}}. \tag{15}$$

Solving this equation numerically for the standard parameters, we obtain $L^{\dagger}=30.07$. Ceiling to the next even integer yields $L^{\dagger}_{\blacktriangle}=\lceil L^{\dagger}\rceil=32$.

As above, we may ignore the logarithmic kinetic dependence on the length for strong binding and obtain

$$L^{\dagger}\approx\ln\left(\frac{r_{\mathrm{out}}}{r_0}\right)\gamma^{-1}. \tag{16}$$

From Fig. 4(a) we see that $L^{\dagger}$ coincides with the position of the maximum L_{max}, whereas L^{*} serves as a proxy for the position of the minimum L_{min}.

In Appendix B we perform an extensive screening of the parameter space demonstrating that the analytical estimates Eqs. (13) and (15) are generally valid. Moreover, the transient behavior of the length distribution in a closed system is discussed in Appendix C. There, the global transient observation time τ_{obs} limits the maximal lifetime of any complex and thus plays the same role as the global degradation timescale r_{off}^{-1} in an open system.

C. Monomer-dimer mixtures

So far, we have studied systems using dimers as initial building blocks. While this made our analytical considerations easier, only strands of even length appeared in the system. These strands enabled infinite extension cascades starting from duplexes with odd parity. As a result, the length distribution had a heavy tail; see Fig. 4(a).

Figure 5 shows the length distribution for a reservoir containing monomers and dimers at a total building block concentration of $c_{\mathrm{tot}}=c_1+c_2=2$ mM. The fraction of monomers $f_m\coloneqq c_1/c_{\mathrm{tot}}$ is set to 70%. Now, infinite extension cascades are suppressed and the long tail collapses. The partitioning of complexes into various substructures shows that fully hybridized duplexes again dominate the tail of the distribution. As above, duplexes with finite overlap are distinguished by the parity of their overhangs, with the addition of mixed parity duplexes, having different parities at the different sites.

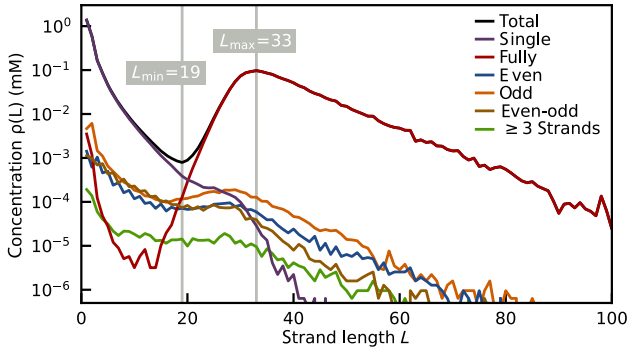

FIG. 5. Partitioned length distribution for a system coupled to a reservoir containing monomers and dimers at a monomer fraction of $f_m=0.7$. In contrast to Fig. 4, virtually all strands with $L>L^{*}$ belong to a fully hybridized duplex.

The general understanding of the characteristic features of the length distribution developed above remains valid. Repeating the calculations leading to Eq. (13) for the onset of extension cascades with the combined extension rate for both monomers and dimers leads to the same equation, with the dimer concentration c_2 replaced by the total concentration c_{tot}; see Appendix D. The position of the maximum does not depend on the building block concentration, cf. Eq. (15). Hence, the peak in Fig. 5 is roughly at the same position as in the dimer-only system. We confirm the validity of this result by probing different monomer fractions f_m in Appendix D.

D. Growth of complexes

The length scale $L^{\dagger}$ relates the dehybridization to the outflux timescale. However, its role in the dynamics is not straightforward. We now show that $L^{\dagger}$ is a typical scale where self-enhancing processes leading to the growth of strands and complexes break down.

Trajectories of stable duplexes.— In what follows, we investigate the trajectories of extension cascades starting from stable duplexes until they leave the system in a fully hybridized configuration. We sample trajectories from the steady state of the monomer-dimer system with a monomer fraction of $f_m=70\%$; see Fig. 5. Our sampling algorithm is consistent with the events that occur in a steady state and is explained in Sec. S2 of the SM [58].

An initial stable duplex consists of a long strand of size L_{long} and a short strand of size $L_{\mathrm{short}}\leq L_{\mathrm{long}}$. It has an overlap l_{initial} and a length $C_{\mathrm{initial}}=L_{\mathrm{long}}+L_{\mathrm{short}}-l_{\mathrm{initial}}$, cf. Fig. 6(a). These stable duplexes are the starting point for extension cascades and eventually become fully hybridized duplexes of length $C_{\mathrm{final}}\geq C_{\mathrm{initial}}$. If the length of the initial complex is the same as that of the final complex, $C_{\mathrm{final}}=C_{\mathrm{initial}}$, no extension occurs beyond the length of the original duplex and we speak of *pure primer extension*. In contrast, if $C_{\mathrm{final}}>C_{\mathrm{initial}}$, processes occurred that extended the length of the initial duplex and we speak of *duplex extension*. A detailed look at extension cascades involving duplex extension can be found in Appendix E.

Figure 6(b) shows the joint probability distribution $p(C_{\mathrm{initial}},l_{\mathrm{initial}})$. It is maximal for $C_{\mathrm{initial}}\sim L^{\dagger}$ and $l_{\mathrm{initial}}\sim L^{*}$. The accumulation of probability at this point characterizes a *typical initial configuration*, but does not determine the length of the individual strands.

Figure 6(c) shows $p(L_{\mathrm{long}},L_{\mathrm{short}})$. We see that it is restricted to the lower triangle defined by $L^{*}\lesssim L\lesssim L^{\dagger}$. The boundaries of this region reflect our analysis above: Strands shorter than L^{*} typically do not bind strongly enough to start extension cascades. In contrast, strands longer than $L^{\dagger}$ are mostly double stranded and thus not available to form the

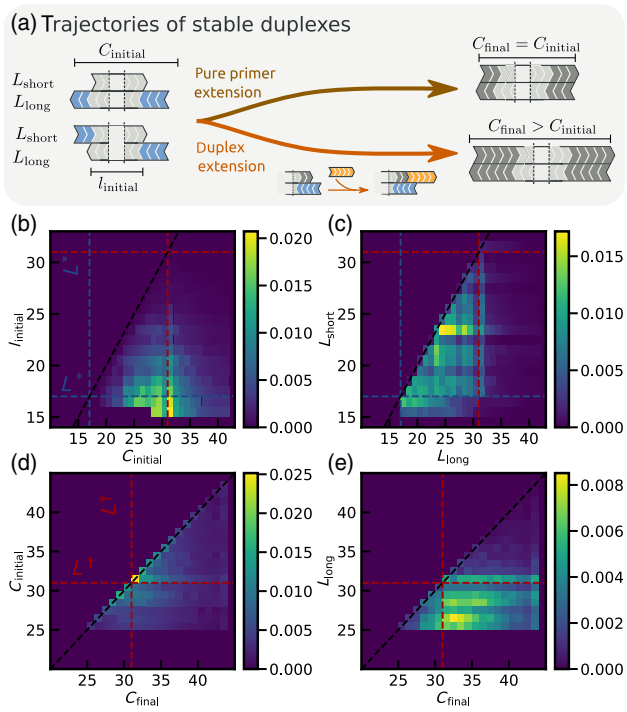

FIG. 6. (a) Trajectories start with an initial stable duplex characterized by its strand lengths L_{long} and L_{short} together with the initial overlap l_{initial} and complex length C_{initial}. Trajectories not creating new single-stranded regions beyond the initial complex are referred to as "pure primer extension." In contrast, duplex extension leads to a final complex with $C_{\text{final}} > C_{\text{initial}}$. (b)–(e) Trajectory statistics can be understood from various joint probability distributions, where $L^* \sim 17$ and $L^\dagger \sim 31$. See main text for details.

initial duplexes. The approximately uniform behavior of the distribution in that region indicates that no particular combination of strand lengths is preferred.

Next, we consider the length C_{final} of the fully hybridized duplex that marks the end of an extension cascade. Figure 6(d) shows the joint probability distribution $p(C_{\text{final}}, C_{\text{initial}})$. The diagonal line $C_{\text{initial}} = C_{\text{final}}$ indicates pure primer extension and has a total weight of ∼17%. The point $C_{\text{final}} = C_{\text{initial}} = 31 \sim L^\dagger$ has the maximal individual weight (∼2.5%).

Finally, Fig. 6(e) shows $p(C_{\text{final}}, L_{\text{long}})$. Purely autocatalytic processes, where the long strand facilitates the formation of itself, are contained on the diagonal line $L_{\text{long}} = C_{\text{final}}$. These autocatalytic trajectories constitute a fraction of about 2.5% of all extension cascades.

Autocatalytic trajectories ($L_{\text{long}} = C_{\text{final}}$) are a subset of trajectories with pure primer extension ($C_{\text{initial}} = C_{\text{final}}$). Most extension cascades, however, lead to fully hybridized duplexes that are longer than either of the two strands of the initial complex. In order to emphasize the cooperative effects, we refer to this more general process as heterocatalytic growth.

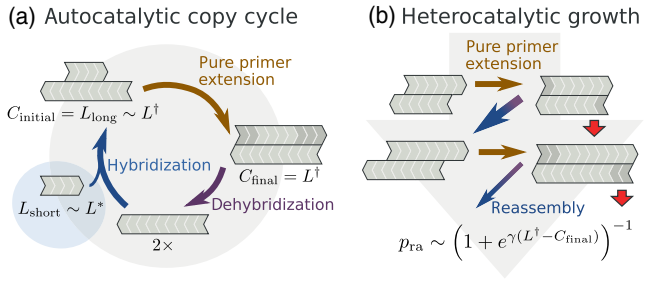

FIG. 7. Heterocatalytic (a) and autocatalytic (b) processes for the growth of strands. In the strongly nonequilibrium regime, extension cascades cover the available overhang of stable duplex and form longer fully hybridized strands. These long strands can then dehybridize and reassemble, thus creating new overhangs to be covered by extension cascades. The reassembly probability p_{ra} is determined by the balance between dehybridization and outflux and decays to zero fast for $L \gtrsim L^\dagger$.

Catalytic growth and reassembly.—Autocatalytic and heterocatalytic *cycles* are formed by combining extension cascades with a dehybridization of the final duplex and eventual reassembly. Figure 7(a) illustrates an autocatalytic cycle, while Fig. 7(b) shows heterocatalytic growth. Note that in general heterocatalytic processes involve duplex extensions.

In the following we investigate how such catalytic cycles shape the strongly nonequilibrium regime $L^* \leq L \leq L^\dagger$ of the strand-length distribution: After a fully hybridized duplex is reached at the end of an extension cascade it will dehybridize or leave the system. If it dehybridizes, it may hybridize to another single strand and create a new stable duplex with a new overhang. By this *reassembly*, long strands catalyze the formation of other long strands. The reassembly probability p_{ra} is mostly determined by the competition between outflux and dehybridization. A sigmoidal dependence on the length C_{final} follows:

$$p_{\text{ra}} \sim \frac{r_{\text{off}}}{r_{\text{off}} + r_{\text{out}}} \sim (1 + e^{\gamma(L^\dagger - C_{\text{final}})})^{-1}. \tag{17}$$

The reassembly probablity p_{ra} decays exponentially with the C_{final}. Thus, the production rate of longer strands $C_{\text{final}} \sim L^\dagger$ is drastically reduced.

In summary, the strongly nonequilibrium catalytic strand growth is constrained to strand length between L^* and $L^\dagger$. It is this self-enhancing dynamical behavior which leads to the increased production of strands with lengths $L^* < L < L^\dagger$. This effect directly yields a region in the strand length distribution, where concentration increases with length. To the right of the peak at $L \sim L^\dagger$, catalytic cycles producing longer strands are too slow in order to compete with the outflux rate r_{out}. Similarly, in the analogous transient situation, the observation time τ_{obs} is too short to allow the catalytic production of strands beyond a certain length.

IV. EXPERIMENTAL SYSTEM

To test our theory experimentally, we used DNA strands of length $L_{\mathrm{bb}} = 12$ as basic building blocks in a closed volume. The strands have random sequences drawn from a binary alphabet of A (adenine) and T (thymine). As discussed above and in Appendix C, in closed systems the global transient observation time τ_{obs} plays the same role as r_{off}^{-1} in an open system.

Enzyme-free templated ligation is slow and not compatible with experimental timescales [19]. Ligases speed up the assembly process, but require the formation of complexes involving at least three strands with $L \gtrapprox 12$ and the ligase. The probability of finding such complexes decreases with temperature. Further, the ligase activity itself is temperature dependent, resulting in a nontrivial temperature dependence of the effective extension rate. In isothermal systems, one may encounter a stalemate situation. For high temperatures, the extension rate is small since the formation of the required complexes is thermally suppressed. In contrast, for low temperatures, the dehybridization rate is small and the system is effectively frozen. This stalemate can be resolved using temperature cycles [8,19]: During the cold phase, the extension rate is initially high until virtually all possible ligations in existing complexes have occurred. Hence, the hot phase is required to create new ligatable complexes. However, temperatures in the hot phase must still be such that the binding energy γ remains negative. Only then is the binding energy still proportional to the overlap length, such that the competition of timescales gives rise to a nonmonotonic length distribution.

A. Experimental method and results

Our experiment was performed using a Taq DNA ligase from New England BioLabs and a ThermoFisher ProFlex PCR system to generate the temperature profile shown in Fig. 8(a). This setup is similar to the setup used in Ref. [8]. The analysis of the length distributions is done by running the sample in a polyacrylamide gel electrophoresis, post-staining the DNA with intercalating SYBR gold dye, and taking fluorescent images of the gel in a BioRad ChemiDoc MP. Concentration quantification is performed with a custom software extracting the lane intensity from gel CCD images (see Sec. S3 D in the SM [58]). The bands visible at lengths of 16 and 24 nt for all lanes in Fig. 8(b) are artifacts from the ligation buffer and DNA synthesis, respectively.

We analyzed the length distribution for various observation times τ_{obs} for different isothermal conditions and cycling scenarios, where temperature alternated between $T_{\mathrm{cold}} = 33\,°\mathrm{C}$ at variable temperature T_{hot}, cf. Fig. 8(a). Isothermal experiments resulted in no product formation within 60 and 116.5 h, cf. Sec. S3 of the SM [58]. For low temperatures, even short duplexes with strands of length L_{bb} cannot separate. For high temperatures, the extension is suppressed because no stable ligatable complexes are formed.

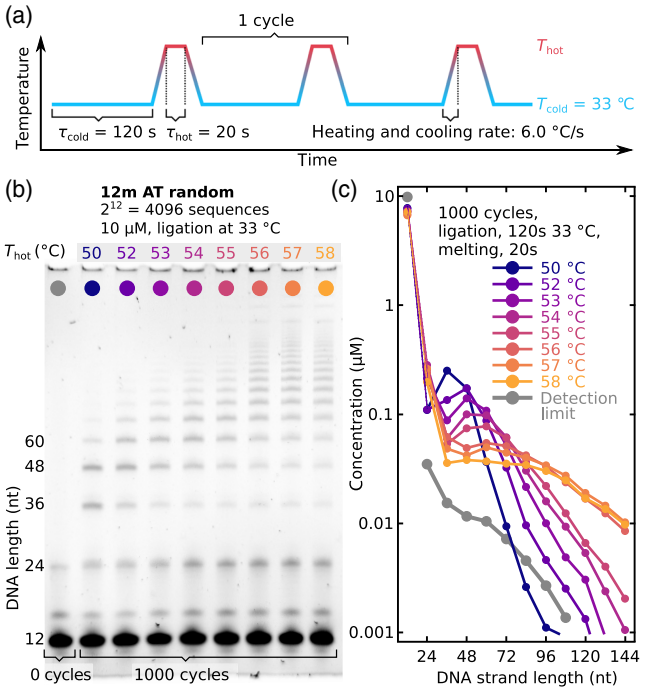

FIG. 8. Product concentration analysis for a 12 nt random sequence AT-only pool. (a) Experimental temperature profile. Ligation occurs for 120 s at 33 °C after which the sample is heated to the variable hot reassembly temperature T_{hot} for 20 s. (b) Image of a polyacrylamide gel electrophoresis. The first lane on the left shows the "baseline" sample, which is similar to the other lanes but was not subjected to temperature cycling. The other lanes have the same ligation conditions but different temperatures for dissociation. (c) Quantitative results for the length distribution. Nonmonotonic length distributions with a maximum and minimum were observed.

Cyclic conditions led to different product distributions, shown in Fig. 8(b). The length distribution decayed quickly for $T_{\mathrm{hot}} = 50\,°\mathrm{C}$, while it decayed slowly for $T_{\mathrm{hot}} = 58\,°\mathrm{C}$. All length distributions showed a nonmonotonic behavior exhibiting a local minimum L_{min} between 24 and 48 nt and a maximum L_{max} between 36 and 72 nt. For higher dissociation temperatures the peak was found to be flatter and wider. The shape of the distribution changed significantly in a limited range for T_{hot}.

B. Effective theory

In order to understand the behavior of the experimental system when varying the temperature T_{hot} in the hot phase, we consider the thermodynamics of the standard Gibbs free energy ΔG°. It enters our theory as the central temperature-dependent binding parameter $\Delta G^{\circ}/(k_B T)$. To leading order in T, the Gibbs energy can be written as $\Delta G^{\circ} = \Delta H^{\circ} - T\Delta S^{\circ}$, where the standard enthalpy ΔH° and standard entropy ΔS° are temperature-independent microscopic parameters [70,71].

The most significant effects occur when the binding energy changes sign at the critical temperature $T_c = \Delta H^{\circ}/\Delta S^{\circ}$.

Assuming that ΔH° and ΔS° scale linearly with the length of the hybridization site, T_c is independent of length. For our experiment, we estimated the expected value of T_c between 60°C and 75°C (see Sec. S3 A in the SM [58]).

At positive binding energy, i.e., for $T_{\mathrm{hot}} > T_c$, strands of all lengths dissociate quickly in the hot phase. We are then in a situation akin to the bounded model discussed in Sec. III A. In order to observe a nonmonotonic distribution, the dehybridization (and thus the reassembly) rate in the hot phase must decay exponentially with strand length, which requires $T_{\mathrm{hot}} < T_c$. For our effective theory we employ a linear expansion of the binding parameter γ below the critical temperature T_c:

$$\gamma(T) = \frac{\Delta G_1^\circ}{k_B T} = \frac{T - T_c}{\xi}. \tag{18}$$

In this formula, ΔG_1° is a typical binding energy per nucleotide. The parameter ξ has units of temperature and characterizes the (inverse) slope of $\gamma(T)$ around T_c. From $\Delta G_1^\circ = \Delta H_1^\circ - T\Delta S_1^\circ$ it follows that $\xi = -k_B T_c^2/\Delta H_1^\circ \approx 30$ K for typical enthalpies and entropies (see Sec. S3 A in the SM [58]).

Unlike T_c, the parameter ξ is inversely proportional to ΔH_1° and thus depends on strand length. Our simple model is based on effective binding energies of (self-complementary) nucleotides. It is therefore questionable whether the value of $\xi \approx 30$ K obtained from standard libraries for matching nucleotides is appropriate here. Since hybridization sites also contain mismatches, the correct value of ξ describing the experimental behavior is likely smaller. Under the assumption that a nucleotide has a probability of $1/2$ to encounter its complement, an adjusted value of $\xi \approx 15$ K seems reasonable. Finally, for a full parametrization of the dehybridization rate $r_{\mathrm{off}}(L; T_{\mathrm{hot}}) \sim r_0 \exp[\gamma(T_{\mathrm{hot}})L]$, we need to specify the collision rate r_0. While the exact value of this rate depends on microscopic details, experimental evidence suggest that a value of $r_0 = 10^6$ s^{-1} is reasonable [72–74].

Figure 9 shows the length dependence of the dehybridization rate r_{off} for various values of T_{hot} below $T_c = 62$ °C and $\xi = 13$ K. One should recall that the extension rate r_{ext} is the effective rate at which a duplex binds to a third strand and subsequently ligates. As explained initially, extensions are likely to happen only in the cold phase. Because of frustrated dehybridization, we expect a single extension per duplex per cycle. Consequently, the extension rate determining L^* is given by $r_{\mathrm{ext}} \sim \tau_{\mathrm{cyc}}^{-1}$. In transient systems without outflux, the inverse observation time $\tau_{\mathrm{obs}}^{-1} = N_{\mathrm{cyc}}\tau_{\mathrm{cyc}}$ replaces r_{out} in determining $L^\dagger$.

For $\tau_{\mathrm{cyc}} = 180$ s and $N_{\mathrm{cyc}} = 1000$, we obtain the two horizontal lines in Fig. 9. The intersections with the length-dependent dehybridization rate determines the scales L^* and $L^\dagger$ as a function of T_{hot}. The big dots and triangles denote the values $L_\bullet^*$ and $L_\blacktriangle^L$ obtained by ceiling to the next integer multiple of L_{bb}.

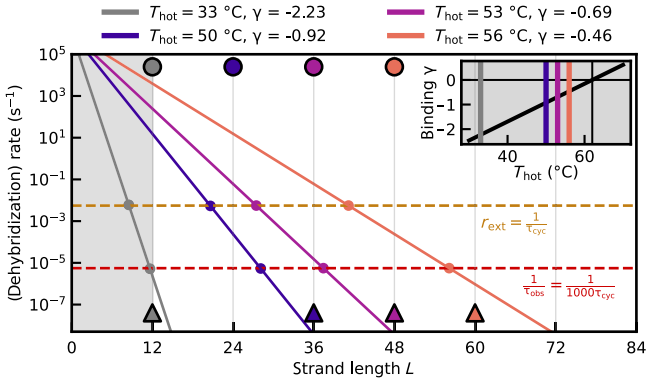

FIG. 9. Temperature dependence of the emerging length scales in the effective theory. Solid lines: dehybridization rate $r_{\mathrm{off}} = r_0 e^{\gamma(T_{\mathrm{hot}})L}$ for various values of T_{hot}. The inset shows the linear function $\gamma(T)$, Eq. (18), with $T_c = 62$ °C and $\xi = 13$ K. Horizontal dashed lines denote the effective extension rate τ_{cyc}^{-1} and the inverse observation time τ_{obs}^{-1}. As in Fig. 3(a), the competition of timescales determines the scales L^* and $L^\dagger$ as the intersection between the solid and dashed lines. They are mapped to the observable length scales L_{min} (circles) and L_{max} (triangles) by ceiling the intersection point to the next multiple of $L_{\mathrm{bb}} = 12$. By approaching T_c, the binding energy and thus the slope become smaller in magnitude, and the intersection points move to larger lengths.

We observe that the scales L^* and $L^\dagger$ shown in Fig. 9 agree well with the experimental observations for L_{min} and L_{max} shown in Fig. 8. The exact values of L^* and $L^\dagger$ depend on the exact values for the parameters r_0, ξ, and T_c, for which we used reasonable estimates. As such, they should not be confused with rigorous predictions. Nonetheless, both the order of magnitude and the qualitative dependence of experimental results on T_{hot} are fully captured by our effective theory.

However, we expect the effective theory to break down close to the critical temperature T_c. For small γ, the contributions to the dehybridization rate due to microscopic details play a more prominent role. Further, when approaching T_c, the characteristic features of the distribution shift to larger lengths. Then, both the experimental timescales and the overall oligonucleotide mass become limiting and the depletion of building blocks starts to play a role. Moreover, the quantitative evaluation of the gel plots is more difficult for long strands, cf. Fig. 8(a). For this case, other effects, like self-folding of strands, may be an important mechanism that is absent from the theory, cf. Ref. [8].

V. SUMMARY AND DISCUSSION

Since major transitions in evolution appear to have occurred when smaller entities were coming together to

form larger ones [75], a multistep scenario toward increased complexity also seems natural in a prebiotic context. While the importance of templated ligation in this scenario is clear [76], the assembly dynamics emerging from this interaction of short building blocks were not fully understood previously.

Using a minimal bottom-up model, we showed that a nonmonotonic length distribution arises from the competition of three timescales or, equivalently, the corresponding rates.

(1) The dehybridization rate r_{off} which decreases exponentially upon increasing strand length L, with the decay determined by the binding parameter per nucleotide γ.
(2) An effective extension rate r_{ext} of a duplex, which is determined by the ligation rate r_{lig}, γ, and the concentrations of building blocks.
(3) A global timescale determined by either the outflux rate r_{out} or an observational time τ_{obs}.

The competition between r_{ext} and r_{out} determines whether we see long-tailed distributions at all: If r_{out} is larger than r_{ext}, ligations are unlikely. The competition between r_{off} and r_{ext} leads to the emergence of extension cascades at a typical length scale L^{*}: As soon as strands in a hybridization complex have a length such that $r_{\mathrm{ext}} > r_{\mathrm{off}}$, they undergo extension cascades resulting in persistent configurations, which cannot extend further. The fate of such a configuration is determined by the competition between r_{off} and r_{out}: Fully hybridized duplexes shorter than $L^{\dagger}$ dehybridize before leaving the system. The single strands released in this way subsequently act as templates in newly formed primer-template complexes and thus catalyze further strand growth.

The combination of extension cascades and reassembly represents (auto or hetero)catalytic cycles producing longer strands from shorter building blocks. In the strand-length distribution, this strongly nonequilibrium regime is visible as an increase in concentration with length. Extension cascades are fast. Therefore, the dehybridization time of the fully hybridized duplex at the end of the cascade determines the completion-time scale of these cycles. For strand lengths where this timescale becomes comparable to transient or global degradation times, these catalytic cycles have no time to complete, and the length distribution decays.

The validity of this scenario was revealed using a state-of-the-art simulation. To our knowledge, no comparable simulation is available to this date. As our experiments demonstrate, the emerging length scales can be tuned by changing environmental parameters such as the melting temperature T_{hot} without changing the chemistry. On early Earth, strands of a characteristic scale $L^{\dagger}$ emerging from the self-assembly could act as building blocks of a higher level of organization. Moreover, length-dependent accumulation of such strands might trigger novel effects like phase transitions [64,65,77,78].

Advantages and shortcomings of our model.—Our minimal model allowed us to reveal universal features of the self-assembly process and to derive analytical expressions for the emerging length scales. Yet, there are several aspects that our model does not capture. It does not allow for secondary structures like hairpins and other "nonproductive" configurations [8,16]. While these (potentially functional) structures will likely be important in later stages of evolution, in the current scenario they probably have the same role as fully hybridized duplexes.

The strongest simplification in this work is the negligence of any explicit sequence dependence for the binding energy, i.e., the use of self-complementary nucleotides. However, an effective self-complementary description of hybridization arises naturally in a mean-field random sequence approximation [22]. It assumes that differences in binding energies between the (complementary and noncomplementary) nucleotide pairs are small with respect to the average binding energy γ per nucleotide.

While this scenario constitutes the extreme of vanishing sequence selectivity, a comparable situation arises in the other limit of perfect selectivity. There, only fully complementary strands bind, and a similar description is achieved using a combinatorial factor for each nucleotide, which can be incorporated by a reduction of γ or using effective concentrations (see also Sec. S1 H of the SM [58]). During extension cascades, the incorporation of mismatches is suppressed, since building blocks matching the template bind stronger and thus lead to higher extension rates. Additional stalling effects for nonmatching short building blocks likely enhance this effect [39,56,57]. Moreover, since ligation reactions are irreversible in our model, it corresponds to the high dissipation limit of sequence copying, which generally increases fidelity, cf. Ref. [79]. Consequently, we expect that in a fully sequence-dependent model, where mismatches are penalized, the maximum in the length distribution is still present and caused by fully hybridized duplexes with comparatively few errors.

Outlook.—Our work provides a first step toward understanding the emergence of structure in a kinetically and thermodynamically consistent bottom-up approach. Extending our algorithm to include sequence-dependent parameters can be a starting point for future studies.

In any explicitly sequence-dependent model, the timescales involved in the extension-reassembly process would include the sequence of the strand in addition to its length [19]. In particular, such a model extension would allow for a more direct study of evolutionary processes in sequence space. As discussed above, sequence selectivity and thus replication may arise during extension cascades. The combined heterocatalytic and autocatalytic nature of the assembly process emphasizes the importance of cooperation, cf. Sec. III D. Therefore, model extensions could

also provide a testing ground for abstract frameworks involving catalytic networks [20,80–82].

ACKNOWLEDGMENTS

U. G. thanks the KITP in Santa Barbara for hospitality on an extended visit, during which part of this work was completed. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the CRC 235 Emergence of Life (Project No. 364653263, P06) and under Germany's Excellence Strategy—EXC-2094-390783311. This research was also supported in part by NSF Grant No. PHY-1748958, NIH Grant No. R25GM067110, and the Gordon and Betty Moore Foundation Grant No. 2919.02. D. B. and P. W. K. acknowledge funding by the Simons Foundation (327125 to D. B.), the Advanced Grant (EvoTrap No. 787356) PE3, ERC-2017-ADG from the European Research Council, and the Center for NanoScience (CeNS).

APPENDIX A: ESTIMATION OF THE OUTFLUX RATE AT THE TRANSITION FROM SHORT- TO LONG-TAILED DISTRIBUTIONS

The transition from short-tailed to long-tailed distributions occurs when the direct production of long strands from reservoir strands is balanced by the production involving long strands as templates, cf. Fig. 10. In the following, the corresponding crossover value of the outflux rate r_{out} in the dimer-only model, Eq. (6), is derived.

Consider the total concentration $\rho_>$ of strands with a length larger than two, i.e., strands not provided by the reservoir. In a steady state we have

$$0 = \partial_t \rho_> = \phi - \rho_> r_{\mathrm{out}}, \tag{A1}$$

where ϕ is the concentration flux indicating processes by which $\rho_>$ grows, namely the formation of tetramers from dimers. Notice that the formation of strands with $L \geq 4$ does not change $\rho_>$. In general, this templated ligation can happen in all triplex configurations with two dimers that are adjacently hybridized. Ignoring higher-order complexes, we assume that the dominant contribution to the production of longer strands arises from a ligation reaction happening at triplexes consisting of two dimers and a templating strand of length $L \geq 2$; see Fig. 10.

As the hybridization dynamics of dimers are fast, we assume a biding equilibrium. This means that the ratio of the concentration of a triplex and its constituents is determined by its binding energy.

With the elementary rates for hybridization and dehybridization defined in Sec. II B, the binding energy of a complex C is given by

$$\beta \Delta G^{\circ}_{\mathrm{tot}}(C) = \gamma \sum_{i \in C} l_i + \sigma \ln(2), \tag{A2}$$

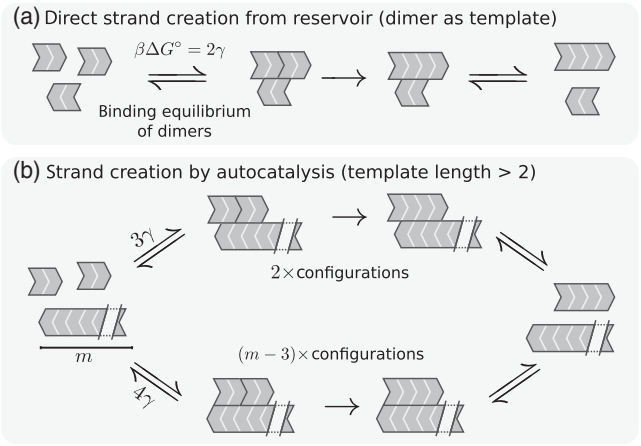

FIG. 10. (a) Formation of a tetramer from the dimer background. A total overlap of two leads to a total binding energy of $\beta\Delta G^{\circ} = 2\gamma$. (b) Templated ligation of dimers on an m-mer. There are two overhanging configurations with $\beta\Delta G^{\circ} = 3\gamma$ and $m-3$ configurations with $\beta\Delta G^{\circ} = 4\gamma$.

where we sum over all hybridization sites. The term $\sigma \ln(2)$ is a "symmetry penalty" that occurs if the complex is rotationally symmetric ($\sigma = 1$) and is zero ($\sigma = 0$) otherwise (see Sec. S1 G in the SM for details [58]).

Using Eq. (A2), the ligation flux for triplexes consisting of dimers only is $\phi_2 = (c_2)^3 e^{-2\gamma} r_{\mathrm{lig}}$; see Fig. 10(a). In contrast, the ligation flux corresponding to templates of length $m > 2$ is

$$\phi_m = (c_2)^2 [2e^{-3\gamma} + (m-3)e^{-4\gamma}] c_m r_{\mathrm{lig}}, \tag{A3}$$

where we took into account the different configurations of the relevant triplexes; see Fig. 10(b).

We separate the ligation flux into two components, $\phi = \phi_2 + \phi_>$. The first term, ϕ_2, only involves the building blocks provided by the reservoir. In contrast, the second term, $\phi_> \coloneqq \sum_{m>2} \phi_m$, involves longer strands. The transition occurs when the latter dominates the former.

Assuming that the length distribution is dominated by single strands, we approximate $\rho_> \approx \sum_{m>2} c_m$, to obtain an expression (lower bound) for $\phi_>$ as

$$\phi_>(\rho_>) \approx (c_2)^2 (2e^{-3\gamma} + e^{-4\gamma}) r_{\mathrm{lig}} \rho_>.$$

In the stationary situation the balance equation (A1) is

$$0 \approx \phi_2 + \phi^c_>(\rho_>) - \rho_> r_{\mathrm{out}}, \tag{A4}$$

which is solved by the crossover value $\rho_> = \rho^c_>$. In this approximation, autocatalysis starts to dominate the production of longer strands from the background when

$\phi^c_> = \phi(\rho^c_>) > \phi_2$. In terms of the outflux rate this means that autocatalysis dominates if

$$r_{\text{out}} < r^c_{\text{out}} = 2(c_2)^2(e^{-4\gamma} + 2e^{-3\gamma})r_{\text{lig}}. \qquad \text{(A5)}$$

APPENDIX B: EXPLORATION OF PARAMETER SPACE

To verify that our results of Sec. III B are indeed generic, we performed an extensive parameter sweep. In each row of Fig. 11, a single parameter is varied while all the other parameters are fixed at their standard values. The left-hand column in Fig. 11 shows simulated stationary length distributions. The right-hand column presents the analytical expressions for the (ceiled) values of L^* and $L^\dagger$ with the characteristic lengths L_{min} and L_{max} from the simulation result. A colored curve in the left-hand panel corresponds to the accordingly colored marker in the right-hand panel. The tails of the distributions are smoothed using a standard running-average smoothing algorithm.

Figure 11(a) shows the result for a variation of the outflux rate r_{out}. The transition from a short- to a long-tailed length distribution was already discussed in Sec. III A. As the outflux rate should not influence the onset of extension cascades, we expect the minimum to remain constant, which the simulation confirms. Increasing the outflux rate shifts L_{max} to lower lengths in a logarithmic way in accordance with Eq. (16).

In Fig. 11(b) we vary the binding energy γ. We observe that increasing the binding energy displaces the characteristic peak toward shorter strands. The behavior of both curves is roughly inversely proportional: $L \propto -\gamma^{-1}$.

Next, we vary the bare ligation rate r_{lig}; see Fig. 11(c). The position of the maximum remains unchanged, since the transition determining the fate of a fully hybridized state is not affected by the ligation rate; see Eq. (15). In accordance with Eq. (13), decreasing r_{lig} logarithmically shifts the onset of extension cascade and the position of the minimum to larger lengths. For the smallest ligation rate plotted, we cross the transition toward short-tailed distributions described in Sec. III A, and the characteristic peak in the length distribution disappears.

Figure 11(d) shows the effect of varying the dimer concentration c_2. Since reducing c_2 logarithmically reduces the effective rate of extension with a dimer, higher concentrations enable extension cascades already for duplexes consisting of shorter strands, shifting the minimum to the left. Again, the position of the peak remains constant. For the smallest concentration shown, we cross the transition toward a short-tailed distribution.

In summary, the phenomenological positions of the minimum L_{min} and the peak L_{max} are well described by the expressions for L^* and $L^\dagger$, Eqs. (13) and (15).

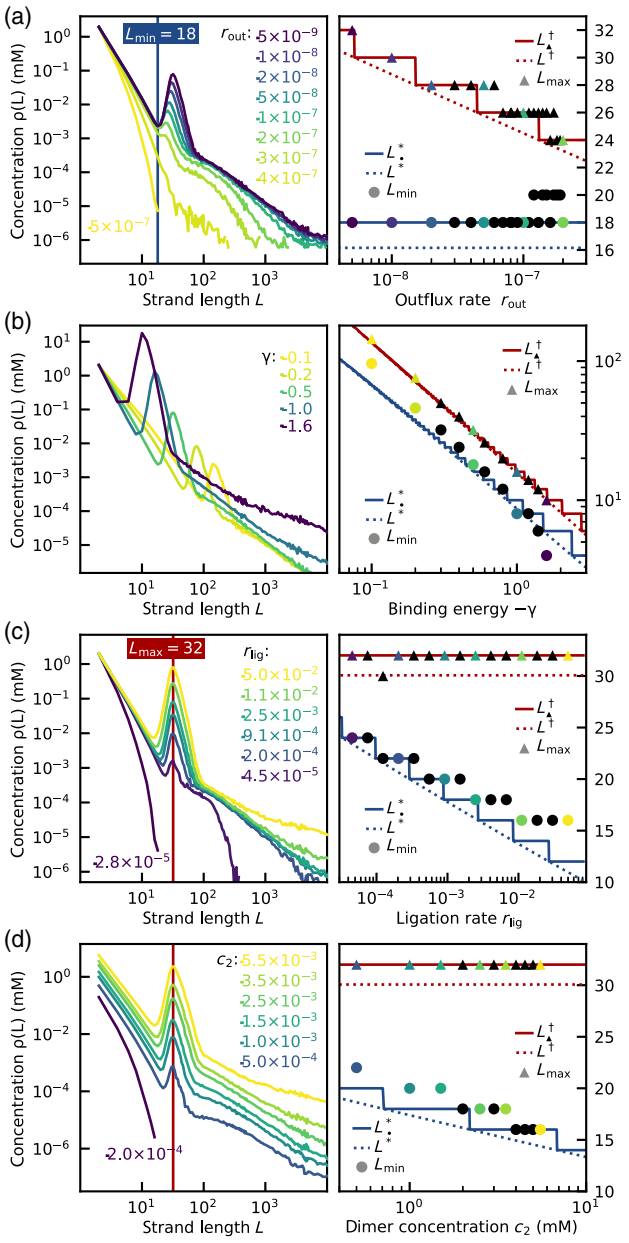

FIG. 11. Probing the parameter space of the dimer-only model. Left-hand column: stationary length distributions. Right-hand column: comparison of the observed values L_{min} and L_{max} and the predictions for L^* and $L^\dagger$ via Eqs. (13) and (15). Variable parameters are (a) the outflux rate r_{out}, (b) the dimensionless binding energy per nucleotide γ, (c) the bare ligation rate r_{lig}, and (d) the concentration of chemostated single-stranded dimers c_2.

APPENDIX C: TRANSIENT BEHAVIOR IN CLOSED SYSTEMS

Next, we investigate a closed system without influx or outflux. We prescribed the concentration of initial building blocks and let the system evolve transiently. Because of the irreversibility of the ligation reaction, closed systems are not ergodic: Short building blocks will deplete and the final

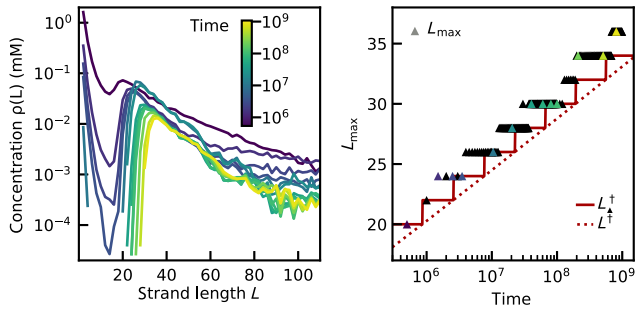

FIG. 12. Transient strand distributions. Left: temporal development of the length distribution in a closed system. Over time the concentration of short strands decreases and the minimum develops into depleted region. Right: the position of the maximum $L_{\max}$ shifts logarithmically with time toward longer lengths.

configuration contains only two very long strands. However, this stationary state will never be reached on practical timescales.

Thus we consider a transient state at intermediate times. We focus on the situation where long strands have already formed, and extension cascades are possible, with still a sufficient amount of short building blocks available. Then, the system behaves similar to the steady state in an open system with small outflux rates.

As in the stationary case, we observe a minimum and maximum in the length distribution. Figure 12(a) shows the length distribution for the standard choice of parameters for various values of the transient observation time $t = \tau_{\text{obs}}$. Figure 12(b) shows that the position of the maximum increases logarithmically with the observation time.

In order to get an intuition for this behavior, we again use an argument involving the competition of timescales. As in the open systems, strands longer than L^* will dominantly occur in fully hybridized configurations. In contrast, the second timescale is not determined by a global outflux rate and fully hybridized duplexes eventually dehybridize with a length-dependent rate $r_{\text{off}}(L)$.

Yet, dehybridization of duplexes of length L only plays a role for observation times longer than $\tau_{\text{obs}} \sim r_{\text{off}}(L)^{-1}$.

We thus expect the global transient observation time τ_{obs} to play the same role as the timescale r_{off}^{-1} in an open system. The length scale $L = L^\dagger$ that determines the peak in a closed system can then be obtained by replacing r_{off} with τ_{obs}^{-1} in Eq. (15) or (16). In that case, the position of the peak should increase logarithmically with time, consistent with the results shown in Fig. 12(b).

APPENDIX D: EXPLORING MONOMER-DIMER MIXTURES

Figure 13(a) shows the length distribution for a reservoir where the total initial building block concentration $c_{\text{tot}} = c_1 + c_2 = 2$ mM is constant. We then vary the monomer fraction $f_m := (c_1/c_{\text{tot}})$ from zero to 90%. The orange curve is the dimer-only system at standard parameters, showing the long tail caused by the infinite extension cascades. For any finite monomer concentration, infinite extension cascades are suppressed and the long tail collapses.

The length distributions for finite monomer fractions look qualitatively similar. The larger f_m, the less nucleotide mass is added by the influx, and the lower the concentrations. The lower the monomer concentration, the more of the bias toward strands of even lengths is retained. The bias dominates for short strands, leading to the zigzag pattern visible in Fig. 5(a). For long strands, the bias vanishes. In accordance with Eq. (15), the position of the maximum is unchanged, as it does not depend on the building block concentration.

The minimum position, i.e., the typical length L^* for the onset of extension cascades, is derived analogously to the dimer-only model using the condition $1 < r_{\text{ext}}(D)/r_{\text{off}}^{\text{dupl}}(D)$, cf. Sec. III B. Instead of considering the extension with a dimer only, one needs to include the extension with a monomer. The extension rate is

$$r_{\text{ext}}(D) \approx r_{\text{ext},1} + r_{\text{ext},2} = r_{\text{lig}} \sum_{\substack{i\in\{1,2\} \\ o_i\neq 0}} (c_2 e^{-\gamma[\min(|o_i|,2)]} + c_1 e^{-\gamma[\min(|o_i|,1)]}). \quad \text{(D1)}$$

The criterion for extension cascades then reads

$$1 \leq (L_1 + L_2 - 1) r_{\text{lig}} \times \sum_{\substack{i\in\{1,2\} \\ o_i\neq 0}} (c_2 e^{-\gamma[l+\min(|o_i|,2)]} + c_1 e^{-\gamma[l+\min(|o_i|,1)]}). \quad \text{(D2)}$$

The right-hand side of the Eq. (D2) is maximal for the odd duplex configuration $D_{\pm 1} = (L_0, L_0, \pm 1)$, for which $l + \min(|o_i|, 2) = l + \min(|o_i|, 1) = L_0$, which leads to

$$1 \leq 2(2L_0 - 1) r_{\text{lig}} (c_2 + c_1) e^{-\gamma L_0}. \quad \text{(D3)}$$

Consequently, L^* for monomer-dimer mixtures obeys

$$1 = 2(2L^* - 1) r_{\text{lig}} c_{\text{tot}} e^{-\gamma L^*}, \quad \text{(D4)}$$

where $c_{\text{tot}} = c_1 + c_2$ is the total concentration of building blocks.

Equation (D4) is the same formula as for the dimer-only system, except that the dimer concentration c_2 is substituted by the total concentration of building blocks c_{tot}. In accordance with formula Eq. (D4), we observe that the position of the minimum is constant $L_{\min} = 19$ under variation of the monomer fraction while keeping the total concentration fixed at $c_{\text{tot}} = 2$ mM; see Fig. 13(b).

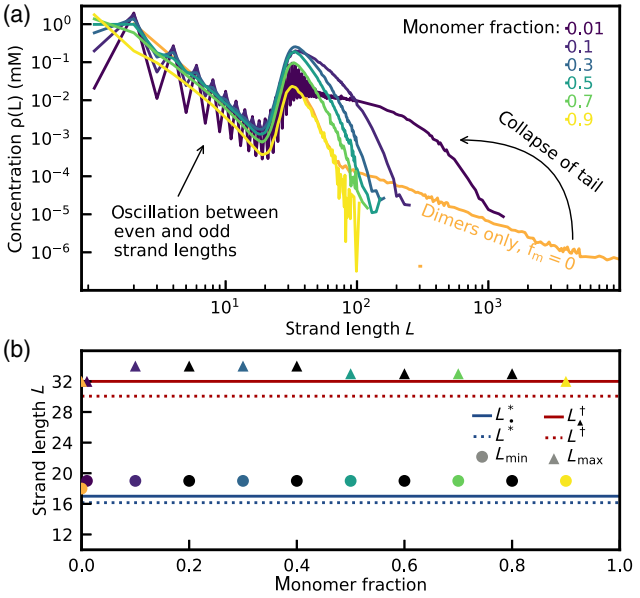

FIG. 13. (a) Length distributions for monomer-dimer mixtures. The monomer fraction f_m is varied between zero and 90% at a total concentration $c_{\text{tot}} = 2$ mM. For low f_m the concentration between even and odd strands oscillates heavily for short strands. The long tail that is present for $f_m = 0$ (orange curve, only even strand lengths shown) collapses even for very small f_m. (b) L^* in the monomer-dimer system is calculated via Eq. (D4), which is the same as the formula for the dimer-only system upon substituting the dimer concentration with the total concentration c_{tot}.

APPENDIX E: BEYOND PURE PRIMER EXTENSION

Figure 6(a) already depicted an example of an extension step leading to the growth of a complex beyond its initial length. In the following, we take a more detailed look at this phenomenon.

Growth happens essentially independently at each end of a duplex. It thus makes sense to take the perspective of a single end, since it allows us to distinguish the two strands by their roles: We call the strand whose end is overhanging the template, whereas the other strand is called the primer. Moreover, we refer to the length of the overhang at the start of a trajectory as its initial copy site length l_{cs}.

The obvious mechanism that leads to duplex extension is depicted in Fig. 14(b). It occurs when the original primer is extended with a strand that is longer than the (remaining) length of the copy site. After this extension, the roles of primer and template are reversed and a new copy site is created. We thus denote this process as *primer-template switching*.

A complex undergoing an extension cascade is not always a simple duplex. Ligation reactions can also occur away from the stable hybridization site. We say that *template extension* occurs, if another strand facilitates the extension of the template strand; see Fig. 14(c). From the perspective of the stable hybridization site, the length of its associated copy site l_{cs} has increased.

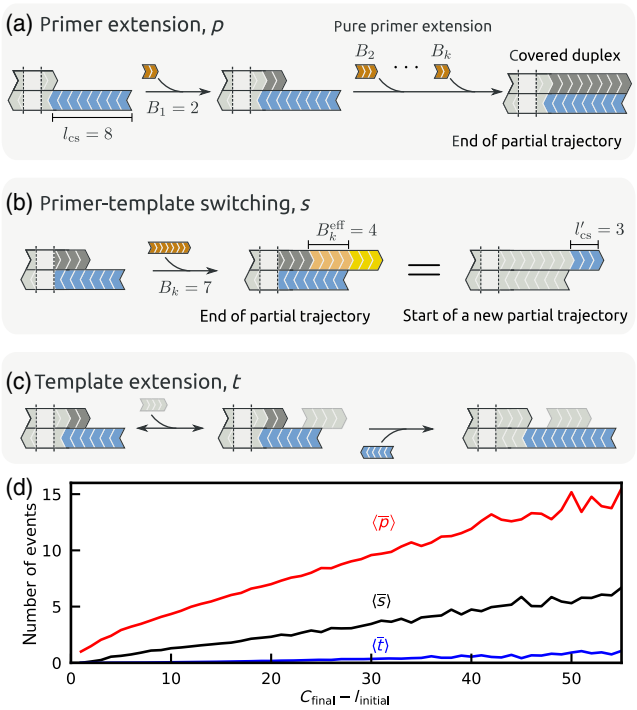

FIG. 14. Pure primer extension (a) versus duplex extension (b), (c). The overhang at the beginning of a (partial) trajectory is called a copy site (blue) with length l_{cs}. (b) In primer-template switching events a building block extends the primer beyond the original copy site. The original copy site is fully covered and a new copy site is formed. The roles of primer and template are exchanged. (c) Copy sites can grow independently of the original primer by template extension with the help of a helper strand. (d) The number of extension events occurring during the covering of the total copy site $C_{\text{final}} - l_{\text{initial}}$ split according to different types of extensions.

Figure 14(d) shows the number of extension events along a trajectory as a function of the total single-stranded length that is covered during the trajectory, $C_{\text{final}} - l_{\text{initial}}$. The standard primer-extension steps (p, red curve) are most common. In contrast, template extension (t, blue curve) is rare. For large $C_{\text{final}} - l_{\text{initial}}$, the number of events behaves strictly linear and primer-template switching (s, black curve) is approximately 3 times less likely than primer extension. For small values of $C_{\text{final}} - l_{\text{initial}}$, the relative fraction of primer-template switching increases, since a short available overhang increases the chance of primer-template switching.

[1] *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*, 3rd ed., Cold Spring Harbor Monograph Series No. 43, edited by R. F. Gesteland, T. Cech, and J. F. Atkins (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY 2006).

[2] L. E. Orgel, *Prebiotic Chemistry and the Origin of the RNA World*, Crit. Rev. Biochem. Mol. Biol. **39**, 99 (2004).

[3] G. F. Joyce, *RNA Evolution and the Origins of Life*, Nature (London) **338**, 217 (1989).

[4] F. Crick, *The Origin of the Genetic Code*, J. Mol. Biol. **38**, 367 (1968).

[5] L. E. Orgel, *Evolution of the Genetic Apparatus*, J. Mol. Biol. **38**, 381 (1968).

[6] W. Gilbert, *Origin of Life: The RNA World*, Nature (London) **319**, 618 (1986).

[7] J. W. Szostak, *The Eightfold Path to Non-Enzymatic RNA Replication*, J. Syst. Chem. **3**, 2 (2012).

[8] P. W. Kudella, A. V. Tkachenko, A. Salditt, S. Maslov, and D. Braun, *Structured Sequences Emerge from Random Pool When Replicated by Templated Ligation*, Proc. Natl. Acad. Sci. U.S.A. **118**, e2018830118 (2021).

[9] L. Zhou, D. K. O'Flaherty, and J. W. Szostak, *Assembly of a Ribozyme Ligase from Short Oligomers by Non-enzymatic Ligation*, J. Am. Chem. Soc. **142**, 15961 (2020).

[10] L. Zhou, D. K. O'Flaherty, and J. W. Szostak, *Template-Directed Copying of RNA by Non-Enzymatic Ligation*, Angew. Chem. Int. Ed. **132**, 15812 (2020).

[11] J. Derr, M. L. Manapat, S. Rajamani, K. Leu, R. Xulvi-Brunet, I. Joseph, M. A. Nowak, and I. A. Chen, *Prebiotically Plausible Mechanisms Increase Compositional Diversity of Nucleic Acid Sequences*, Nucleic Acids Res. **40**, 4711 (2012).

[12] A. R. Ferre-D'Amare and W. G. Scott, *Small Self-Cleaving Ribozymes*, Cold Spring Harbor Perspect. Biol. **2**, a003574 (2010).

[13] K. R. Birikh, P. A. Heaton, and F. Eckstein, *The Structure, Function and Application of the Hammerhead Ribozyme*, Eur. J. Biochem. **245**, 1 (1997).

[14] E. Ekland, J. Szostak, and D. Bartel, *Structurally Complex and Highly Active RNA Ligases Derived from Random RNA Sequences*, Science **269**, 364 (1995).

[15] F. Wachowius and P. Holliger, *Non-Enzymatic Assembly of a Minimized RNA Polymerase Ribozyme*, ChemSystemsChem **1**, 12 (2019).

[16] L. Zhou, D. Ding, and J. W. Szostak, *The Virtual Circular Genome Model for Primordial RNA Replication*, RNA **27**, 1 (2021).

[17] H. Mutschler, A. Wochner, and P. Holliger, *Freeze–Thaw Cycles as Drivers of Complex Ribozyme Assembly*, Nat. Chem. **7**, 502 (2015).

[18] J. Attwater, A. Raguram, A. S. Morgunov, E. Gianni, and P. Holliger, *Ribozyme-Catalysed RNA Synthesis Using Triplet Building Blocks*, eLife **7**, e35255 (2018).

[19] S. Toyabe and D. Braun, *Cooperative Ligation Breaks Sequence Symmetry and Stabilizes Early Molecular Replication*, Phys. Rev. X **9**, 011056 (2019).

[20] A. V. Tkachenko and S. Maslov, *Onset of Natural Selection in Populations of Autocatalytic Heteropolymers*, J. Chem. Phys. **149**, 134901 (2018).

[21] E. Guseva, R. N. Zuckermann, and K. A. Dill, *Foldamer Hypothesis for the Growth and Sequence Differentiation of Prebiotic Polymers*, Proc. Natl. Acad. Sci. U.S.A. **114**, E7460 (2017).

[22] A. V. Tkachenko and S. Maslov, *Spontaneous Emergence of Autocatalytic Information-Coding Polymers*, J. Chem. Phys. **143**, 045102 (2015).

[23] K. Leu, E. Kervio, B. Obermayer, R. M. Turk-MacLeod, C. Yuan, J.-M. Luevano, E. Chen, U. Gerland, C. Richert, and I. A. Chen, *Cascade of Reduced Speed and Accuracy after Errors in Enzyme-Free Copying of Nucleic Acid Sequences*, J. Am. Chem. Soc. **135**, 354 (2013).

[24] M. L. Manapat, I. A. Chen, and M. A. Nowak, *The Basic Reproductive Ratio of Life*, J. Theor. Biol. **263**, 317 (2010).

[25] A. Kanavarioti and D. H. White, *Kinetic Analysis of the Template Effect in Ribooligoguanylate Elongation*, Origins Life Evol. Biosphere **17**, 333 (1987).

[26] B. Obermayer, H. Krammer, D. Braun, and U. Gerland, *Emergence of Information Transmission in a Prebiotic RNA Reactor*, Phys. Rev. Lett. **107**, 018101 (2011).

[27] C. B. Mast, S. Schink, U. Gerland, and D. Braun, *Escalation of Polymerization in a Thermal Gradient*, Proc. Natl. Acad. Sci. U.S.A. **110**, 8030 (2013).

[28] P. J. Flory, *Principles of Polymer Chemistry*, 19th ed. (Cornell University Press, Ithaca, NY, 2006).

[29] S. Lahiri, Y. Wang, M. Esposito, and D. Lacoste, *Kinetics and Thermodynamics of Reversible Polymerization in Closed Systems*, New J. Phys. **17**, 085008 (2015).

[30] M. Sosson, D. Pfeffer, and C. Richert, *Enzyme-Free Ligation of Dimers and Trimers to RNA Primers*, Nucleic Acids Res. **47**, 3836 (2019).

[31] W. S. Zielinski and L. E. Orgel, *Autocatalytic Synthesis of a Tetranucleotide Analogue*, Nature (London) **327**, 346 (1987).

[32] E. Kervio, M. Sosson, and C. Richert, *The Effect of Leaving Groups on Binding and Reactivity in Enzyme-Free Copying of DNA and RNA*, Nucleic Acids Res. **44**, 5504 (2016).

[33] E. Edeleva, A. Salditt, J. Stamp, P. Schwintek, J. Boekhoven, and D. Braun, *Continuous Nonenzymatic Cross-Replication of DNA Strands with In Situ Activated DNA Oligonucleotides*, Chem. Sci. **10**, 5807 (2019).

[34] M. Sosson and C. Richert, *Enzyme-Free Genetic Copying of DNA and RNA Sequences*, Beilstein J. Org. Chem. **14**, 603 (2018).

[35] E. Hänle and C. Richert, *Enzyme-Free Replication with Two or Four Bases*, Angew. Chem. Int. Ed. **57**, 8911 (2018).

[36] C. Deck, M. Jauker, and C. Richert, *Efficient Enzyme-Free Copying of All Four Nucleobases Templated by Immobilized RNA*, Nat. Chem. **3**, 603 (2011).

[37] M. Jauker, H. Griesser, and C. Richert, *Copying of RNA Sequences without Pre-Activation*, Angew. Chem. Int. Ed. **54**, 14559 (2015).

[38] E. Kervio, A. Hochgesand, U. E. Steiner, and C. Richert, *Templating Efficiency of Naked DNA*, Proc. Natl. Acad. Sci. U.S.A. **107**, 12074 (2010).

[39] S. Rajamani, J. K. Ichida, T. Antal, D. A. Treco, K. Leu, M. A. Nowak, J. W. Szostak, and I. A. Chen, *Effect of Stalling after Mismatches on the Error Catastrophe in Nonenzymatic Nucleic Acid Replication*, J. Am. Chem. Soc. **132**, 5880 (2010).

[40] N. Prywes, J. C. Blain, F. Del Frate, and J. W. Szostak, *Nonenzymatic Copying of RNA Templates Containing All Four Letters Is Catalyzed by Activated Oligonucleotides*, eLife **5**, e17756 (2016).

[41] L. Li, N. Prywes, C. P. Tam, D. K. O'Flaherty, V. S. Lelyveld, E. C. Izgu, A. Pal, and J. W. Szostak, *Enhanced*

Nonenzymatic RNA Copying with 2-Aminoimidazole Activated Nucleotides, J. Am. Chem. Soc. **139**, 1810 (2017).
[42] Y. J. Matsubara and K. Kaneko, *Optimal Size for Emergence of Self-Replicating Polymer System*, Phys. Rev. E **93**, 032503 (2016).
[43] L. H. G. da Silva and D. Hochberg, *Open Flow Non-Enzymatic Template Catalysis and Replication*, Phys. Chem. Chem. Phys. **20**, 14864 (2018).
[44] H. Fellermann, S. Tanaka, and S. Rasmussen, *Sequence Selection by Dynamical Symmetry Breaking in an Autocatalytic Binary Polymer Model*, Phys. Rev. E **96**, 062407 (2017).
[45] S. Tanaka, H. Fellermann, and S. Rasmussen, *Structure and Selection in an Autocatalytic Binary Polymer Model*, Europhys. Lett. **107**, 28004 (2014).
[46] Y. J. Matsubara and K. Kaneko, *Kinetic Selection of Template Polymer with Complex Sequences*, Phys. Rev. Lett. **121**, 118101 (2018).
[47] R. Mizuuchi and N. Lehman, *Limited Sequence Diversity within a Population Supports Prebiotic RNA Reproduction*, Life **9**, 20 (2019).
[48] A. Tupper, K. Shi, and P. Higgs, *The Role of Templating in the Emergence of RNA from the Prebiotic Chemical Mixture*, Life **7**, 41 (2017).
[49] P. W. Anderson, *Suggested Model for Prebiotic Evolution: The Use of Chaos.*, Proc. Natl. Acad. Sci. U.S.A. **80**, 3386 (1983).
[50] C. Fernando, G. Von Kiedrowski, and E. Szathmáry, *A Stochastic Model of Nonenzymatic Nucleic Acid Replication: Elongators Sequester Replicators*, J. Mol. Evol. **64**, 572 (2007).
[51] K. A. Dill and S. Bromberg, *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*, 2nd ed. (Garland Science, London, 2011).
[52] R. Rao and M. Esposito, *Nonequilibrium Thermodynamics of Chemical Reaction Networks: Wisdom from Stochastic Thermodynamics*, Phys. Rev. X **6**, 041064 (2016).
[53] H. Subramanian and R. A. Gatenby, *Evolutionary Advantage of Anti-Parallel Strand Orientation of Duplex DNA*, Sci. Rep. **10**, 9883 (2020).
[54] T. Walton and J. W. Szostak, *A Kinetic Model of Nonenzymatic RNA Polymerization by Cytidine-5'-Phosphoro-2-Aminoimidazolide*, Biochemistry **56**, 5739 (2017).
[55] T. Walton and J. W. Szostak, *A Highly Reactive Imidazolium-Bridged Dinucleotide Intermediate in Nonenzymatic RNA Primer Extension*, J. Am. Chem. Soc. **138**, 11996 (2016).
[56] J. Kim and M. Mrksich, *Profiling the Selectivity of DNA Ligases in an Array Format with Mass Spectrometry*, Nucleic Acids Res. **38**, e2 (2010).
[57] G. Lohman, R. J. Bauer, N. M. Nichols, L. Mazzola, J. Bybee, D. Rivizzigno, E. Cantin, and T. C. Evans, *A High-Throughput Assay for the Comprehensive Profiling of DNA Ligase Fidelity*, Nucleic Acids Res. **44**, e14 (2016).
[58] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.11.031055 for a comprehensive description of the implementation, additional parameter sweeps, a description of the analytical framework used to analyze trajectories in Sec. III D, and further experimental details.
[59] J. J. Hopfield, *Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity*, Proc. Natl. Acad. Sci. U.S.A. **71**, 4135 (1974).
[60] B. Rauzan, E. McMichael, R. Cave, L. R. Sevcik, K. Ostrosky, E. Whitman, R. Stegemann, A. L. Sinclair, M. J. Serra, and A. A. Deckert, *Kinetics and Thermodynamics of DNA RNA, and Hybrid Duplex Formation*, Biochemistry **52**, 765 (2013).
[61] T. E. Ouldridge, *The Importance of Thermodynamics for Molecular Systems and the Importance of Molecular Systems for Thermodynamics*, Nat. Comput. **17**, 3 (2018).
[62] A. Ianeselli, C. B. Mast, and D. Braun, *Periodic Melting of Oligonucleotides by Oscillating Salt Concentrations Triggered by Microscale Water Cycles inside Heated Rock Pores*, Angew. Chem. Int. Ed. **58**, 13155 (2019).
[63] A. Mariani, C. Bonfio, C. M. Johnson, and J. D. Sutherland, *p*H-*Driven RNA Strand Separation under Prebiotically Plausible Conditions*, Biochemistry **57**, 6382 (2018).
[64] C. B. Mast and D. Braun, *Thermal Trap for DNA Replication*, Phys. Rev. Lett. **104**, 188102 (2010).
[65] M. Kreysing, L. Keil, S. Lanzmich, and D. Braun, *Heat Flux across an Open Pore Enables the Continuous Replication and Selection of Oligonucleotides towards Increasing Length*, Nat. Chem. **7**, 203 (2015).
[66] L. Geyrhofer and N. Brenner, *Coexistence and Cooperation in Structured Habitats*, BMC Ecol. **20**, 14 (2020).
[67] D. T. Gillespie, *Exact Stochastic Simulation of Coupled Chemical Reactions*, J. PHYS. CHEM **81**, 2340 (1977).
[68] D. T. Gillespie, *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions*, J. Comput. Phys. **22**, 403 (1976).
[69] M. A. Gibson and J. Bruck, *Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels*, J. Phys. Chem. A **104**, 1876 (2000).
[70] J. SantaLucia and D. Hicks, *The Thermodynamics of DNA Structural Motifs*, Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).
[71] D. H. Turner and D. H. Mathews, *NNDB: The Nearest Neighbor Parameter Database for Predicting Stability of Nucleic Acid Secondary Structure*, Nucleic Acids Res. **38**, D280 (2010).
[72] I. I. Cisse, H. Kim, and T. Ha, *A Rule of Seven in Watson-Crick Base-Pairing of Mismatched Sequences*, Nat. Struct. Mol. Biol. **19**, 623 (2012).
[73] I. Schoen, H. Krammer, and D. Braun, *Hybridization Kinetics Is Different inside Cells*, Proc. Natl. Acad. Sci. U.S.A. **106**, 21649 (2009).
[74] S. Howorka, L. Movileanu, O. Braha, and H. Bayley, *Kinetics of Duplex Formation for Individual DNA Strands within a Single Protein Nanopore*, Proc. Natl. Acad. Sci. U.S.A. **98**, 12996 (2001).
[75] E. Szathmáry and J. M. Smith, *The Major Evolutionary Transitions*, Nature (London) **374**, 227 (1995).
[76] D. Sievers and G. von Kiedrowski, *Self-Replication of Complementary Nucleotide-Based Oligomers*, Nature (London) **369**, 221 (1994).
[77] M. Todisco, T. P. Fraccia, G. P. Smith, A. Corno, L. Bethge, S. Klussmann, E. M. Paraboschi, R. Asselta, D. Colombo,

G. Zanchetta, N. A. Clark, and T. Bellini, *Nonenzymatic Polymerization into Long Linear RNA Templated by Liquid Crystal Self-Assembly*, ACS Nano **12**, 9750 (2018).

[78] M. Morasch, D. Braun, and C. B. Mast, *Heat-Flow-Driven Oligonucleotide Gelation Separates Single-Base Differences*, Angew. Chem. Int. Ed. **55**, 6676 (2016).

[79] D. Andrieux and P. Gaspard, *Nonequilibrium Generation of Information in Copolymerization Processes*, Proc. Natl. Acad. Sci. U.S.A. **105**, 9516 (2008).

[80] M. Eigen and P. Schuster, *A Principle of Natural Self-Organization: Part A: Emergence of the Hypercycle*, Naturwissenschaften **64**, 541 (1977).

[81] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford University Press, New York, 1993).

[82] A. Blokhuis, D. Lacoste, and P. Nghe, *Universal Motifs and the Diversity of Autocatalytic Systems*, Proc. Natl. Acad. Sci. U.S.A. **117**, 25230 (2020).