life

MDPI

Article

Thermodynamic and Kinetic Sequence Selection in Enzyme-Free Polymer Self-Assembly inside a Non-equilibrium RNA Reactor

Tobias Göppel, Joachim H. Rosenberger, Bernhard Altaner and Ulrich Gerland *

Physics of Complex Biosystems, Technical University of Munich, 85748 Garching, Germany; tobias.goeppel@tum.de (T.G.); joachim.h.rosenberger@tum.de (J.H.R.); bernhard.altaner@tum.de (B.A.)
* Correspondence: gerland@tum.de

Abstract: The RNA world is one of the principal hypotheses to explain the emergence of living systems on the prebiotic Earth. It posits that RNA oligonucleotides acted as both carriers of information as well as catalytic molecules, promoting their own replication. However, it does not explain the origin of the catalytic RNA molecules. How could the transition from a pre-RNA to an RNA world occur? A starting point to answer this question is to analyze the dynamics in sequence space on the lowest level, where mononucleotide and short oligonucleotides come together and collectively evolve into larger molecules. To this end, we study the sequence-dependent self-assembly of polymers from a random initial pool of short building blocks via templated ligation. Templated ligation requires two strands that are hybridized adjacently on a third strand. The thermodynamic stability of such a configuration crucially depends on the sequence context and, therefore, significantly influences the ligation probability. However, the sequence context also has a kinetic effect, since non-complementary nucleotide pairs in the vicinity of the ligation site stall the ligation reaction. These sequence-dependent thermodynamic and kinetic effects are explicitly included in our stochastic model. Using this model, we investigate the system-level dynamics inside a non-equilibrium 'RNA reactor' enabling a fast chemical activation of the termini of interacting oligomers. Moreover, the RNA reactor subjects the oligomer pool to periodic temperature changes inducing the reshuffling of the system. The binding stability of strands typically grows with the number of complementary nucleotides forming the hybridization site. While shorter strands unbind spontaneously during the cold phase, larger complexes only disassemble during the temperature peaks. Inside the RNA reactor, strand growth is balanced by cleavage via hydrolysis, such that the oligomer pool eventually reaches a non-equilibrium stationary state characterized by its length and sequence distribution. How do motif-dependent energy and stalling parameters affect the sequence composition of the pool of long strands? As a critical factor for self-enhancing sequence selection, we identify kinetic stalling due to non-complementary base pairs at the ligation site. Kinetic stalling enables cascades of self-amplification that result in a strong reduction of occupied states in sequence space. Moreover, we discuss the significance of the symmetry breaking for the transition from a pre-RNA to an RNA world.

Keywords: emergence of life; templated ligation; enzyme-free self-assembly; informational polymers; prebiotic evolution; enzyme-free replication; RNA reactor; autocatalytic set

check for updates

Citation: Göppel, T.; Rosenberger, J.H.; Altaner, B.; Gerland, U. Thermodynamic and Kinetic Sequence Selection in Enzyme-Free Polymer Self-Assembly inside a Non-equilibrium RNA Reactor. *Life* **2022**, *12*, 567. https://doi.org/10.3390/life12040567

Academic Editors: Oliver Trapp, Alexander Ruf and Louis B. Le Sergeant D'Hendecourt

Received: 15 March 2022
Accepted: 6 April 2022
Published: 10 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Extant biological systems use different molecules for the storage of genetic information than for the catalysis of biomolecular reactions. Inevitably, the question arises of which came first, the informational polymer carrying the instructions for the enzymes, or the enzymes assembling the polymers? As RNA cannot only store genetic information but also fold into catalytically active structures [1–5], it is central to one of the most prominent hypotheses for the emergence of living systems [6–11]. However, this *RNA* *world hypothesis* does not explain the origin of the catalytic RNA molecules, called ribozymes [12]. While recent experimental work revealed potential prebiotic pathways to synthesize nucleotides [13–15], the mechanisms assembling these building blocks into functional molecules are only beginning to be explored [16–21]. The smallest ribozymes known today are 30 to 100 nucleotides long [22–24]. More complex ribozymes that could, e.g., assist replication are likely to have a minimum length of more than 150 nucleotides [25–27]. For polymers of a length between 30 to 150, a total of 10^{18} to 10^{90} distinct sequences are possible. However, the subset of catalytically active sequences is generally believed to be tiny. Hence, the spontaneous emergence (and maintenance) of a functional ribozyme in a random pool of oligonucleotides seems highly unlikely. Therefore, it remains unclear how the transition from a pre-RNA world lacking ribozymes to an RNA world where essential processes are driven by ribozymes could occur.

To understand this transition, one must study the dynamics in sequence space, which emerges when mononucleotides and short oligonucleotides inside a reaction volume collectively self-assemble into longer strands [28,29]. The self-assembly is governed by the process of *templated ligation* [30–34]. In this process, two strands that are hybridized adjacently on a third template strand are covalently linked. In contrast to random ligation concatenating two arbitrary strands, the process of templated ligation is sequence-selective for two reasons, a thermodynamic and a kinetic one. First, complementary nucleotide pairs at the hybridization sites increase the stability of the complex of strands, and thereby also increase the probability for a new covalent bond to be formed. However, complementary sequences of the same length comprising different sequence motifs can have different binding stabilities. The stability of a complex is determined by the hydrogen bonding between complementary base pairs and the stacking interactions between neighboring base pairs [35]. Changing the order of the base pairs or flipping one of the pairs generally alters the complex's stability. Hence, certain sequence motifs can be favored over others thermodynamically [35–37]. Second, the kinetics of the ligation step are motif-selective: Non-complementary nucleotide pairs in the vicinity of the ligation site stall the formation of a new covalent bond. As a result, the formation of new strands from shorter fragments that do not match the template strand is also suppressed kinetically [32,38–40].

Experimentally probing the enzyme-free self-assembly of long strands from a random pool of mononucleotides and short fragments is challenging. Typically, the experiments require long times, while the reaction yields remain low and undesired side products obscure the results. Moreover, tracking the evolution of the whole sequence pool simultaneously remains an unsolved technical challenge [20,21,41,42]. Due to these constraints, non-enzymatic self-assembly experiments either employed initial oligonucleotides with precisely designed sequences limiting the product space [43–47] or focused on *primer extension* scenarios. In the latter scenario, a defined primer that is statically bound to a defined longer template strand gets extended by mononucleotides and short oligomers [38,48–54]. Two explorative experimental studies investigated the emergence of progressively longer strands from DNA-oligomers [18,55], both using DNA ligases to accelerate the assembly dynamics and to obtain better yields, and employing temperature cycling for strand separation. In Ref. [55], all possible 12-mers that can be formed from a binary alphabet of A and T are present initially. The assembly dynamics give rise to structured sequence pools characterized by a reduced sequence entropy compared to a random pool. The emerging longer strands are either characterized by a large A or T content since mixed strands are more prone to self-inhibition due to hairpin formation. In Ref. [18], three pairs of carefully designed complementary sequences composed of 20 nucleotides were used as basic building blocks. The authors demonstrated that certain subsets of sequence motifs composed of two basic building blocks form cooperative networks. Since the initial building blocks are already quite long in both studies, the binding energies of bound strands are large, such that small differences in the stacking energies associated with adjacent nucleotide pairs become irrelevant. However, subtle differences in the stacking energies might trigger sequence selection already on the level of the shortest oligomers, i.e., dimers and trimers, for which dissociation occurs spontaneously, rather than being induced externally. Since a sequence bias emerging early on might feedback onto itself, it could have a substantial impact on the pool of longer strands at later stages. In summary, an experimental study exploring growth dynamics into longer polymers starting from a pool of small building blocks is still missing.

Investigating the collective growth from small building blocks theoretically or by means of computer simulations in a model including the essential features of self-assembly, i.e., sequence-dependent (de)hybridization and ligation dynamics, is also challenging: First, the number of possible complex configurations grows exponentially fast as strands become longer. Second, there is an intrinsic separation of time scales between the fast dissociation of short and the slow dissociation of long hybridization sites and the slow ligation step. To date, no theoretical study on self-assembly via templated ligation accounted for the motif-dependent thermodynamic and kinetic aspects of hybridization and bond formation (see Section 4.2). Therefore, the following questions remained open: (1) What are the emerging dynamics in sequence space as strands grow longer? (2) Which are critical factors that enable self-enhancing sequence selection? (3) How do motif-dependent thermodynamic and kinetic parameters affect the selection process?

In Ref. [56], we developed a simulation method that partially handled the complexity of the self-assembly process. In this first study, we treated the sequence dependence of the (de)hybridization dynamics in a mean-field picture, in which the dissociation rate only depends on the length of the hybridization site. Our study identified several growth regimes arising from the competition of timescales for dissociation and extension. Moreover, we showed that, depending on external control parameters, the strand length distribution in the stationary state can exhibit a non-monotonous shape characterized by a distinct strand length. For the present study, we extended the simulation method to explicitly treat sequences, including sequence-dependent thermodynamics and kinetics. The 'RNA reactor' simulations that we report here assume a closed reaction volume, initialized with mononucleotides and a few dinucleotides, with an unbiased nucleotide distribution (symmetric initial condition in sequence space). Within the RNA reactor, oligomers grow via templated ligation and degrade via hydrolysis. Eventually, the sequence pool converges to a non-equilibrium stationary state characterized by its length and sequence distribution.

To address the above questions, we consider different model variants. We start with a simple reference scenario, where kinetic stalling is absent and the stacking energies for all complementary neighboring nucleotide pairs are identical. This scenario distinguishes solely between complementary and non-complementary pairings. We then introduce thermodynamic and kinetic sequence selection, both separately and in combination, and compare the resulting four different scenarios. Our main finding is that, under the conditions assumed here, thermodynamic discrimination within hybridized strands is not sufficient by itself to promote self-enhanced sequence selection that drives the sequence pool significantly away from the random state. However, distinct patterns in sequence space arise if non-complementary strand termini at the ligation site slow down the ligation step significantly (kinetic stalling). In this case, a small thermodynamic bias for certain sequence-motifs triggers a self-enhancing dynamics, such that the thermodynamically favored sequence motif dominates the stationary state.

2. Models and Methods

2.1. Strands and Complexes

We consider a binary alphabet composed of two complementary nucleotides, denoted as X and Y for generality. A molecule containing L nucleotides linked covalently is called a *strand* of length L (see Figure 1a). A single nucleotide is a strand of $L = 1$. Strands are directed and point from the $5'$ to the $3'$ end, which we also refer to as the $-$ and the $+$ ends. We allow strands to hybridize to each other, but do not account for the possibility of self-folding. An entity formed by several hybridized strands is referred to as a *complex*. All staggered conformations that can arise from a set of single strands are allowed inside the

RNA reactor, regardless of the number of strands and *mismatches*, i.e., non-complementary nucleotide pairs (see Figure 1b,c and Figure A1 in Appendix B). However, branched hybridization structures and other nonlinear complexes involving loops are excluded.

We call a complex that contains two or three strands a *duplex* or *triplex*, respectively. The overlapping horizontal region between two strands is referred to as a *hybridization site*. Moreover, the vertical interface between two strands hybridized adjacently on a third strand is called a *ligation site*.

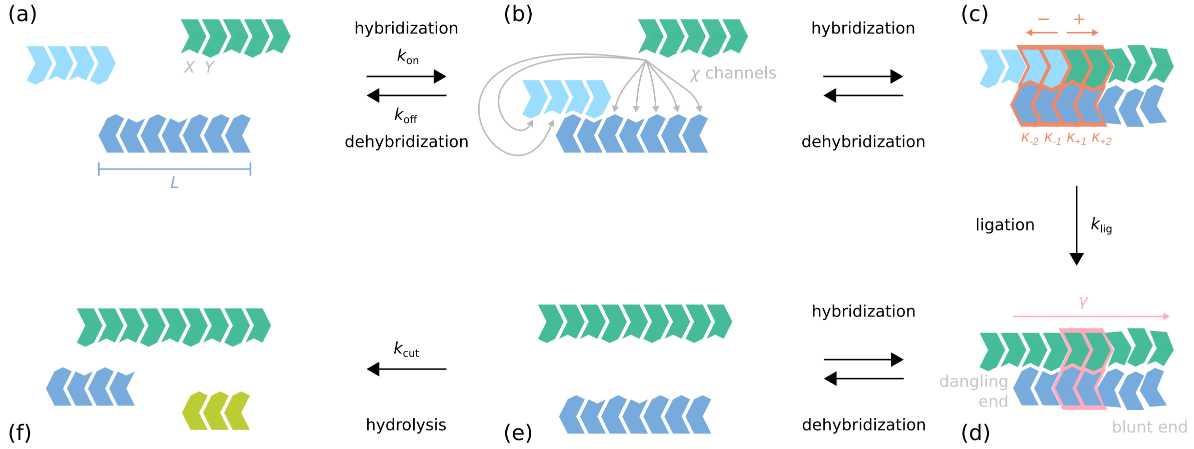

Figure 1. Schematic illustration of the dynamics inside the RNA reactor. The elementary processes are hybridization, dehybridization, ligation on the template, and hydrolysis with corresponding elementary rates k_{on}, k_{off} and k_{lig}, and k_{cut}. The elementary rates k_{on}, k_{off}, k_{lig} are functions of the sequence context: (**a**) Strands have a binary sequence and are directed. L denotes their length. (**b**) When two molecules collide, they can form χ different hybridization complexes. (**c**) Hybridization sites within complexes (horizontal interfaces) can contain mismatches. Two strands ($-$ and $+$ strand) located adjacently on another strand may get joined covalently via templated ligation. The speed of the ligation reaction depends on the complementarity κ of the nucleotide pairs at the ± 1 and ± 2 position (red box). Non-complementary pairings lead to kinetic stalling. (**d**) The stability of a hybridization site is governed by the hybridization energy ΔG_{hyb}. ΔG_{hyb} is obtained by summing over stacking energies γ associated with nearest-neighbor blocks (purple box) and considering terminal nucleotide pairs. ΔG_{hyb} and γ depend on the structural and sequential context. Mismatches weaken the binding. (**e**,**f**) Covalent bonds within single strands or single-stranded segments may get cleaved via hydrolysis at a constant rate. The resulting unactivated strand termini are assumed to be rapidly reactivated.

2.2. Elementary Reactions

Strands and complexes form new complexes via *hybridization*, *dehybridization*, *templated ligation* and *hydrolysis* (see Figure 1). All reactions are assumed to be elementary and occur with sequence- and structure-dependent rates k_{on}, k_{off}, k_{lig}, and k_{cut}. Assuming constant environmental conditions, k_{on} and k_{off} are related to the *hybridization energy* ΔG_{hyb} associated with a hybridization site via the thermodynamic consistency requirement [57]

$$\frac{k_{\text{off}}}{k_{\text{on}}} = V N_A c_\circ \, e^{\beta \Delta G_{\text{hyb}}} \,, \tag{1}$$

where $\beta = (k_{\text{B}} T)^{-1}$, k_{B} is Boltzmann's constant, and T denotes the (absolute) temperature, V and N_A are the reaction volume and Avogadro constant and $c_\circ = 1\,\text{mol/L}$ is the reference concentration. We will express all concentrations as a multiple of the reference concentration. Moreover, in the following, we use the dimensionless hybridization energy

$$\Gamma = \beta \Delta G_{\text{hyb}}. \tag{2}$$

Γ is obtained by summing over dimensionless motif-dependent stacking energies of nearest-neighbor blocks [35–37] (see Section 2.4). Γ thus reflects the number of complementary and non-complementary nucleotide pairs and their arrangement. Generally, mismatches increase the hybridization energy, therefore reducing the stability of a complex.

The formation of new covalent bonds requires energy, which needs to be provided by the environment in the form of an activation chemistry [42,58,59]. We assume that the RNA reactor is constantly fueled with the activation chemistry, enabling a fast chemical (re)activation of the termini of all strands present in the system. With that, two strands that are located next to each other on a third strand can always ligate. The fast activation step is not modeled explicitly. The rate k_{lig} at which two neighboring strands ligate depends on the paired nucleotides in the vicinity of the ligation site. Mismatches lead to *kinetic stalling*, i.e., a reduction of the ligation speed [32,38–40].

We model the kinetic stalling using the *kinetic stalling factors* $\Phi_{\pm} \leq 1$. The stalling factors $\Phi_{\pm}$ are functions of the *complementarities* $\kappa_{\pm i} \in \{1, 0\}$ of the paired nucleotides in the vicinity of the ligation site. The value 1 indicates a complementary pair, whereas the value 0 indicates a non-complementary pair of nucleotides. Φ_{-} takes the complementarities κ_{-1}, κ_{-2} of the two nucleotides in the $-$ direction of the ligation site into account, while Φ_{+} is an equivalent expression for the two nucleotides in the $+$ direction (see Figure 1c,d and Section 2.5 for more details). The two stalling factors are then multiplied with the *basal ligation rate* λ. With that, the ligation rate k_{lig} becomes

$$k_{\text{lig}} = \lambda\, \Phi_{-}(\kappa_{-1}, \kappa_{-2})\, \Phi_{+}(\kappa_{+1}, \kappa_{+2}). \tag{3}$$

Since random ligation of two strands in the absence of a template is weak compared to templated ligation [31–34], we neglect it in our model.

While covalent bonds within double-stranded parts of complexes are assumed to be stable against hydrolysis, bonds within single-stranded sections get cleaved [60–63] (see Figure 1e,f). The corresponding rate is assumed to be sequence-independent,

$$k_{\text{cut}} = \text{const.} \tag{4}$$

with that, the overall degradation rate for a single strand of length L is $(L-1)k_{\text{cut}}$, for example. In real systems, k_{cut} varies by several orders of magnitude as a function of environmental parameters and crucially depends on the polymer's backbone chemistry [60,62–65]. Note that templated ligation and cleavage are irreversible, since the respective reverse reactions (random ligation and "templated cleavage") are absent in our model.

2.3. Kinetics of Hybridization and Dehybridization

Since Equation (1) only constrains the ratio of k_{on} and k_{off}, an additional kinetic parameter is required to fix the kinetics of the model. However, the chosen parametrization has only a minor effect on the global kinetics, given that ligation and hydrolysis are rare compared to hybridization and dehybridization (see Section 2.8). Our approach uses a constant rate of collision between two complexes $k_{\text{coll}} = (V N_A c_{\circ} t_0)^{-1}$, where t_0 is the collision time scale. In the following, we express all times in units of collision time scale t_0.

In general, two colliding complexes can form multiple hybridization configurations via χ distinct hybridization channels (see Figure 1b). We assume no bias for any channel, such that the probability of choosing one particular channel is

$$p_{\text{hyb}} = 1/\chi. \tag{5}$$

Hence, the rate for a hybridization via a given channel is

$$k_{\text{on}} = k_{\text{coll}}\, p_{\text{hyb}}, \tag{6}$$

whereas the dehybridization rate becomes

$$k_{\text{off}} = \frac{1}{\chi} e^{\Gamma}. \tag{7}$$

If $\chi = 0$, no hybridization can occur. This is the case if one of the colliding complexes is a duplex without any overhang. A parametrization attributing the hybridization energy Γ to the dehybridization rate is common in theoretical approaches and is consistent with experiments [66–68]. The kinetic model resulting from the specific choice of k_{on} and k_{off}, i.e., Equations (6) and (7) was developed, described in detail, and rationalized in Ref. [56], where we studied self-assembly in a sequence-independent model with hybridization energies simply being proportional to the overlap lengths. The kinetic assumptions Equations (6) and (7) reduce the computational complexity considerably, while still sampling complexes in a thermodynamically consistent way (see Appendix C).

2.4. Hybridization Energy

Detailed models for the free energy of given RNA and DNA secondary structures [35–37] build on the so-called stacking interactions of neighboring nucleotide pairs at the hybridization sites [35,69]. Every nearest-neighbor interaction, i.e., every block of two adjacent nucleotide pairs, is associated with a motif-dependent stacking energy. These stacking energies additively contribute to the total free energy. Additional contributions to the total free energy take into account nonlinearities of secondary structures such as loops, branching points, and particular end configurations.

Our coarse-grained model that excludes nonlinear complex structures conserves the essential feature of the detailed nearest-neighbor models. The central element of our energy model is the stacking interaction of two neighboring nucleotides pairs P_i and P_{i+1} with

$$P_i \text{ and } P_{i+1} \in \left\{ \begin{matrix} X \\ \cdot \\ Y \end{matrix}, \begin{matrix} Y \\ \cdot \\ X \end{matrix}, \begin{matrix} X \\ \\ X \end{matrix}, \begin{matrix} Y \\ \\ Y \end{matrix} \right\}, \tag{8}$$

where dots symbolize hydrogen bonds between complementary nucleotides. To every block of adjacent nucleotide pairs $[P_i P_{i+1}]$, we assign a dimensionless stacking energy $\gamma([P_i\, P_{i+1}])$. (Note that the last two pairs are non-complementary. Therefore, the nucleotides are not connected via a dot.) The hybridization energy is then given by the sum over all stacking energies and contributions ϵ_- and ϵ_+ accounting for the terminal nucleotide pairs at the $-$ and the $+$ end of double-stranded segment (see Figure 1d), i.e.,

$$\Gamma = \sum_{i \,\in\, \text{blocks}} \gamma_i + \epsilon_- + \epsilon_+. \tag{9}$$

The contributions $\epsilon_\mp$ for the $\mp$ end also depend on the structural and sequence context. If the $\mp$ terminal nucleotide pair forms a *dangling end*, i.e., is preceded or followed by an unpaired nucleotide, we have $\epsilon_\mp \neq 0$. If the terminal nucleotide pair is part of a ligation site, there also is a contribution $\epsilon_\mp \neq 0$. If otherwise, it corresponds to *blunt end* of a complex, and we have $\epsilon_\mp = 0$ (see Appendix A for details).

For simplicity, we assume symmetric stacking energies, i.e., $\gamma([P_i\, P_{i+1}]) = \gamma([P_{i+1}\, P_i])$. Moreover, complementary nearest-neighbor blocks are either *alternating* if

$$[P_i\, P_{i+1}] \in \left\{ \begin{bmatrix} X - Y \\ \cdot \quad \cdot \\ Y - X \end{bmatrix}, \begin{bmatrix} Y - X \\ \cdot \quad \cdot \\ X - Y \end{bmatrix} \right\}, \tag{10}$$

or *homogeneous* if

$$[P_i\, P_{i+1}] \in \left\{ \begin{bmatrix} X - X \\ \cdot \quad \cdot \\ Y - Y \end{bmatrix}, \begin{bmatrix} Y - Y \\ \cdot \quad \cdot \\ X - X \end{bmatrix} \right\}. \tag{11}$$

Here, the $-$ symbol stands for a covalent bond. We denote stacking energies assigned to alternating and homogeneous blocks by γ_{alt} and γ_{hom}. Motivated by the observation that $\gamma_{\text{alt}} \neq \gamma_{\text{hom}}$ in DNA and RNA systems (see Table 1) [36,37], we treat the energy difference

$$\Delta\gamma = \gamma_{\text{alt}} - \gamma_{\text{hom}}. \tag{12}$$

as a variable parameter. Without loss of generality, we assume $\Delta\gamma \leq 0$ for our model. Moreover, we assume constant stacking energies $\gamma_{1\text{nc}}$ and $\gamma_{2\text{nc}}$ for nearest neighbor blocks containing one or two non-complementary nucleotide pairs. Since blocks containing mismatches weaken the binding, their stacking contributions are positive. The contribution for a block with two mismatches is larger than for a block with only one mismatch. In summary, the block-wise contributions obey the hierarchy

$$\gamma_{\text{alt}} \leq \overline{\gamma_{\text{com}}} \leq \gamma_{\text{hom}} < 0 < \gamma_{1\text{nc}} < \gamma_{2\text{nc}}, \tag{13}$$

where $\overline{\gamma_{\text{com}}}$ is the average energy value of complementary blocks (see Table 2), i.e.,

$$\overline{\gamma_{\text{com}}} = (\gamma_{\text{alt}} + \gamma_{\text{hom}})/2. \tag{14}$$

Table 1. $\overline{\gamma_{\text{com}}}$: mean stacking energies for complementary nearest neighbor blocks in binary RNA and DNA systems at a reference temperature of 37 °C in units of k_BT [36,37]. $\Delta\gamma$: difference between alternating and homogeneous blocks (see Equation (12)). Note that the sign of $\Delta\gamma$ depends on whether A and U or T or G and C are considered for the binary system.

System	**RNA**		**DNA**	
Nucleotides	**A, U**	**C, G**	**A, T**	**C, G**
$\overline{\gamma_{\text{com}}}$	−1.74	−5.00	−1.40	−3.26
$\Delta\gamma$	−0.46	0.60	0.42	−0.65

2.5. Kinetic Stalling

Our kinetic stalling model describes the experimentally observed sequence dependence [32,38,39] in a simplified way, using only two parameters, σ_1, σ_2. Mismatches directly at the ligation site affect the ligation speed more substantially than distant ones. If the nucleotide pair at the ±1 position is non-complementary ($\kappa_{\pm 1} = 0$), a mismatch at the ±2 position ($\kappa_{\pm 2} = 0$) amplifies the stalling effect. Otherwise, a mismatch at the ±2 position has no effect, i.e.,

$$\Phi_{\pm}(\kappa_{\pm 1}, \kappa_{\pm 2}) = \begin{cases} 1 & \text{for } \kappa_{\pm 1} = 1 \wedge \kappa_{\pm 2} \in \{0,1\} \\ \sigma_1 & \text{for } \kappa_{\pm 1} = 0 \wedge \kappa_{\pm 2} = 1 \\ \sigma_1\sigma_2 & \text{for } \kappa_{\pm 1} = 0 \wedge \kappa_{\pm 2} = 0 \end{cases}, \tag{15}$$

where $\sigma_1 \leq \sigma_2$. If the hybridization site in the + or − direction contains only one nucleotide pair (see Figure 1c), we use Equation (15) with $\kappa_{\pm 2} = 1$.

The strength of the stalling effect depends on the underlying activation chemistry as well as the type of nucleotides being used [32,38,39]; therefore, we treat σ_1 and σ_2 as variable parameters (see Table 2).

2.6. Effective Cyclic Environment

According to the energy model defined in Equation (9), hybridization energies for long, primarily complementary hybridization sites become arbitrarily negative. Hence, the corresponding dehybridization rates converge to zero exponentially. As a result, strands can be bound in duplexes without single-stranded overhangs over long times. This effect is called template inhibition and leads to freezing of the dynamics [41,61,70]. To overcome this problem, we assume cyclic variations of the physico-chemical conditions (temperature, pH , or salt concentrations) inside the RNA reactor such that all hybridized strands separate within the period time τ [19,71]. Aforesaid oscillatory conditions arise for example due to convection flows induced by temperature gradients or micro scale water cycles at a heated gas–liquid interface. Both scenarios arise naturally in rock fissure in the vicinity of hydrothermal vents [72–76]. They are modeled effectively by introducing a lower bound for the dehybridization rate [56,77], i.e., modifying Equation (7) such that

$$k_{\text{off}} = \max\left\{\frac{1}{\chi}e^{\Gamma}, k_{\text{low}}\right\}, \tag{16}$$

where $\tau = k_{\text{low}}^{-1}$. With that, the (dis)assembly dynamics of long complementary complexes do not obey the thermodynamic consistency requirement Equation (1) anymore. Nonetheless, the kinetics are still plausible: The constant collision rate is a reasonable approximation for a collision process with a diffusion coefficient decaying with length compensated by a cross section growing with length [56].

Table 2. Summary of parameters used in Section 3.

Process	Parameter	Value
hybridization	k_{coll} c_{tot}	1 $0.01\, c_{\circ}$
dehybridization	$\overline{\gamma_{\text{com}}}, \gamma_{1\text{nc}}, \gamma_{2\text{nc}}, \Delta\gamma$ l_{low}	$-1.25, 0.375, 0.75, [-0.3, 0]$ 7
ligation	l_{lig} σ_1, σ_2	10 $[0,1]$, $[0.1,1]$
hydrolysis	l_{cut}	18.5

2.7. Validity of Our Model and Application to Primer Extension

In the Results section, we focus on self-assembly scenarios where all strands (apart from monomers) are equally important because there are no distinct template, primer, and substrate strands as in typical primer-extension situations. However, in Appendix F, we show that our modeling of the kinetic stalling and the (de)hybridization kinetics in combination leads to copying dynamics in primer-extension situations consistent with the experimental literature.

2.8. Parametrization of Rates

We can parametrize every rate constant k_* introduced so far by a dimensionless length l_* such that

$$k_* = e^{\overline{\gamma_{\text{com}}}\, l_*}. \tag{17}$$

This presentation will prove convenient in the later analysis of the results as it connects time scales to length scales. For example, $l_{\text{low}} = 7$ tells us that entirely complementary hybridization sites composed of more than seven nucleotides dissociate as quickly as altogether complementary hybridization sites comprising exactly seven nucleotides. Parameters used in the following are summarized in Table 2. Moreover, $l_{\text{lig}} = 10$ signifies that the timescale of a dehybridization for a hybridization site counting more than ten nucleotides would be slower than the bare ligation timescale if the lower bound with l_{low} would not have been introduced.

2.9. Implementation

To simulate the model dynamics in C++, we use an extension of the framework developed in [56], based on an optimized Gillespie algorithm [78–80]. The simulation only keeps those species in memory that have a non-zero copy number. If a species appears (vanishes), the corresponding species object is created (deleted) dynamically.

3. Results

3.1. Boundary Conditions and Observables

We aim to investigate the model dynamics starting on the lowest level, i.e., where mononucleotide and a few short oligonucleotides collectively evolve into larger entities. Will the sequences of longer strands be random, or will they show patterns? We chose the arguably simplest setting for our study, which is a closed reaction volume that does not exchange complexes with the environment. In such a setting, we expect the dynamics to settle to a stationary state eventually. We initialize the reaction volume symmetrically with 5000 nucleotides distributed over 4920 mononucleotides and 40 dimers (see Figure 2a). Moreover, we adjust the reaction volume such that the total nucleotide concentration is given by $c_{\text{tot}} = 0.01\, c_{\circ}$. The ratio of the initial monomer to dimer concentration is $c_1^{\text{init}} : c_2^{\text{init}} = 123 : 1$.

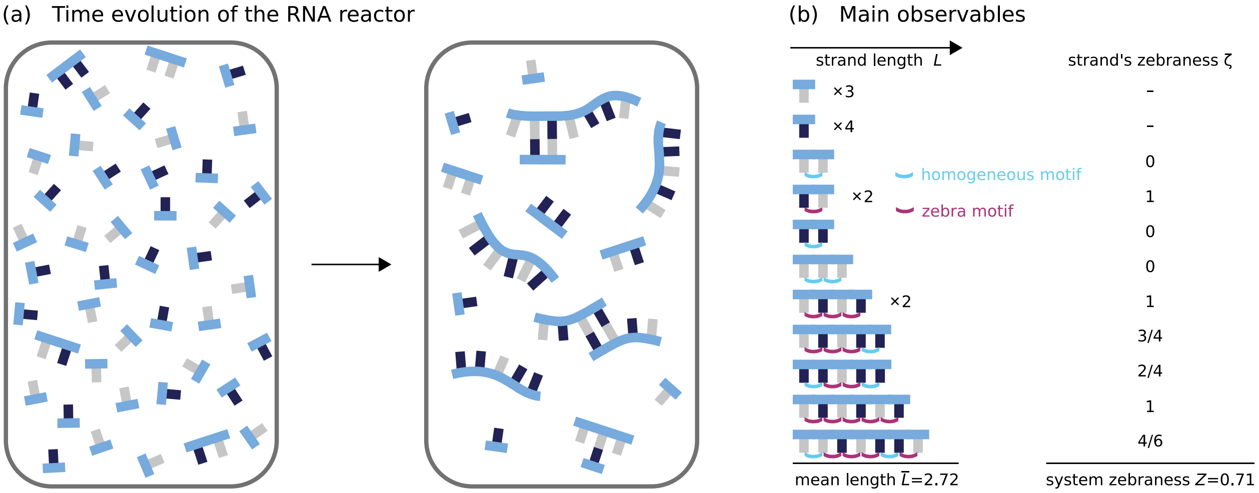
(a) Time evolution of the RNA reactor
(b) Main observables
strand length L
strand's zebraness ζ
×3
×4
×2
×2
homogeneous motif
zebra motif
-
-
0
1
0
0
1
3/4
2/4
1
4/6
mean length $\overline{L}$=2.72
system zebraness Z=0.71

Figure 2. (**a**) Schematic illustration of the time evolution of the non-equilibrium RNA reactor. The reactor is initialized symmetrically with mononucleotides and a few dimers such that the amounts of X and Y nucleotides are equal and that all four dimer sequences have the same concentrations (see Section 3.1). Within the RNA reactor, oligomers grow via templated ligation and degrade via hydrolysis. Eventually, the sequence pool converges to a non-equilibrium stationary state characterized by its length and sequence distribution (see Figure 1); (**b**) To characterize the dynamics in sequence space, we introduce the zebraness ζ on the level of single strands and the system-level zebraness Z. The zebraness ζ of a single strand is the fraction of zebra motifs, i.e., alternating binary motifs contained in the strand. In contrast, the system-level zebraness Z measures how zebra-like, i.e., alternating or homogeneous, the pool is as a whole. Z corresponds to the total number of zebra motifs spread over all strands normalized with respect to the overall number of binary motifs within all strands present in the reactor.

Our focus is on the evolution of the length distribution and the dynamics in sequence space. The length distribution c_L expresses the concentration of strands of length L, irrespective of whether they are part of a complex or not. We denote the mean length by $\overline{L}$. To describe the dynamics in sequence space, we aim for a simple observable with an intuitive and straightforward interpretation. Therefore, we introduce the *zebraness* as a characterization of a strand's sequence. The zebraness $\zeta(S)$ of a strand S of length L_S is the number of alternating "zebra" submotifs $X - Y$ or $Y - X$ within its sequence divided by the number of binary motifs $L_S - 1$ (see Figure 2b). With that, a random sequence S_r is expected to have $\zeta(S_r) = 0.5$ on average. Moreover, the system-level zebraness Z characterizes how zebra-like the ensemble of strands is. It is given by

$$Z = \frac{\sum_S \zeta(S)\,(L_S - 1)}{\sum_S (L_S - 1)}, \tag{18}$$

where the summation is performed over all individual strands with $L > 1$. A system containing homogeneous strands only would have $Z = 0$, whereas, for a system exclusively composed of strands with alternating sequences, we would have $Z = 1$. All plots show ensemble averages which are taken over 20 independent realizations of the dynamics.

3.2. Overview of Key Findings

Before we present the detailed analysis of the four different scenarios outlined in the introduction, we briefly summarize our key findings. First, we study the simplest variant of our model where both kinetic stalling and energetic bias are absent, i.e., $\Delta\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$. This scenario only distinguishes between nearest-neighbor blocks containing zero, one, or two mismatches. Alternating and homogenous blocks have identical energetic properties. While non-complementary pairings decrease the complex's stability, erroneous pairings at the ligation site do not reduce the bare ligation rate. The first model variant does not give rise to motif selection; the composition of the sequence pool remains entirely random, i.e., $Z = 0.5$.

In the second scenario, we introduce an energetic bias $\Delta\gamma < 0$ for alternating blocks while still assuming a non-discriminative ligation. The energetic bias favors the hybridization of strands with zebra-like sequences. This time, a weak zebra pattern with $Z > 0.5$ is induced transiently during the initial growth phase. However, the pattern vanishes almost completely as the system approaches the steady-state (see Figure 3).

The dynamics in sequence space change drastically if kinetic stalling with $\sigma_1, \sigma_2 < 1$ is applied. If non-complementary nucleotide pairs at the ligation site slow down the formation of a covalent bond, distinct patterns in sequence space can emerge. In the third scenario, we investigate the correlation between the strength of the kinetic stalling effect and the reduction of possible states in sequence space assuming identical energetic properties for alternating and homogeneous blocks, i.e., $\Delta\gamma = 0$. Within this setting, we observe a spontaneous symmetry breaking in sequence space. Independent realizations of the dynamics evolve to stationary states, dominated by either zebra motifs with $Z < 0.5$ or homogeneous motifs with $Z > 0.5$ (see Figure 4). Moreover, we see that a dominant pattern emerges such that $Z \to 0$ or 1 if the stalling effect is strong enough.

In the fourth scenario, we show that a slight energetic bias $\Delta\gamma < 0$ can become self-amplifying if kinetic stalling is present (see Figure 5). Depending on the strength of the kinetic stalling, the system converges to either a partial or pure zebra state characterized by either $Z > 0.5$ or $Z \to 1$.

3.3. Reference Model without Energetic Bias and Kinetic Stalling

This section aims to answer whether the energetic discrimination of matches and mismatches alone is sufficient to give rise to spontaneous symmetry breaking in sequence space such that $Z \neq 0.5$. To this end, we study the simplest variant of our model with neither energetic bias nor kinetic stalling, i.e., $\Delta\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$.

Initially, the growth dynamics of the mean length $\overline{L}$ is slow until $t \approx 8.8 \times 10^9$ (see the dark blue curve in Figure 3a). At this point, the mean length $\overline{L}$ starts to increase rapidly. We refer to this time point as the *onset* of growth and denote it by $\hat{t}$. After the steep increase, $\overline{L}$ reaches a plateau value. The inset shows the steady-state length distribution displaying a double-exponential shape.

The ensemble average of the zebraness Z initially fluctuates and then converges to $Z = 0.5$ (see Figure 3b). Looking at single trajectories (see Figure 3c) reveals a behavior similar to the ensemble average. The initial values of $Z \lessgtr 0.5$ on the single trajectory-level are due to small numbers of strands with $L > 1$. A value of $Z = 0.5$ hints towards an entirely random sequence pool but does not exclude motif correlations on larger scales. However, analyzing distributions of longer motifs reveals that the final sequence composition is indeed random (see Appendix D).

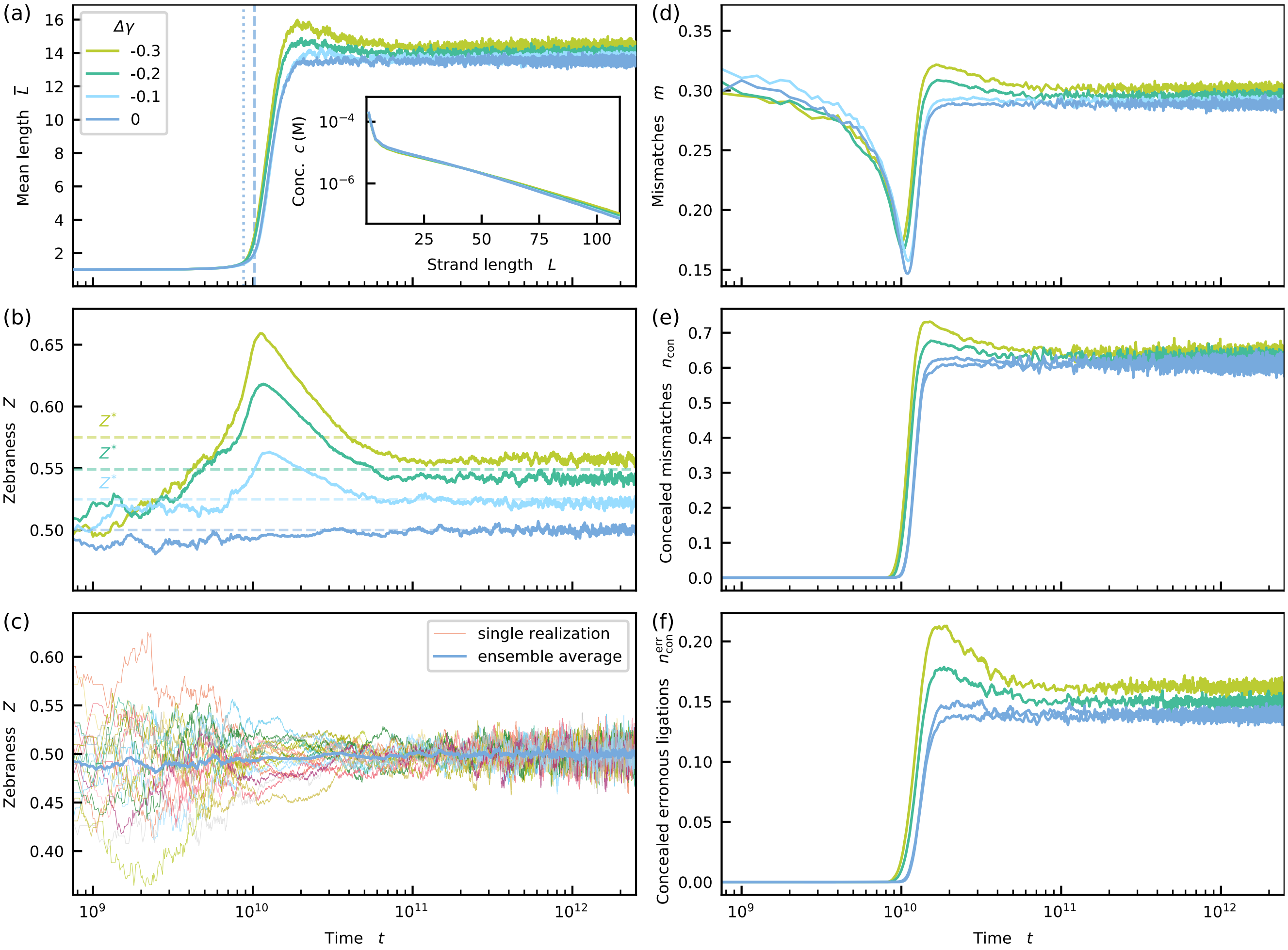

Figure 3. Mean length $\overline{L}$ and system-level zebraness Z as functions of time for $\sigma_1 = \sigma_2 = 1$ and various $\Delta\gamma$. (**a**) A sharp increase of $\overline{L}$ appears at $\hat{t}$. For $\Delta\gamma = 0$, the dashed line corresponds to $\hat{t}$ resulting from the formal definition, whereas the dotted line is the prediction obtained from Equation (19). For $\Delta\gamma < 0$, $\overline{L}$ reaches a maximum before decaying gradually to the stationary value. The inset shows the steady-state length distributions. (**b**) If there is no energetic bias, i.e., $\Delta\gamma = 0$, no distinct patterns emerge in sequence space, and hence $Z = 0.5$ (see also Appendix D). If an energetic bias $\Delta\gamma < 0$ is applied, Z grows initially and then decays when $\overline{L} \approx 7$. The final value is slightly above the random state $Z = 0.5$ and below the simple thermodynamic estimate Z^* (see Equation (22)). (**c**) Single realizations of the dynamics for $\Delta\gamma = 0$ behave similar to ensemble average. Strong fluctuations for small times stem from low numbers of strands with $L > 1$. (**d**) The fraction of mismatches m first decreases and then increases as the mean length becomes longer. (**e**) The fraction of concealed mismatches, i.e., mismatches not affected by energetic discrimination grows simultaneous with the mean length. (**f**) Over time, concealed erroneous ligations become frequent and destroy the initial sequence bias.

The evolution of the mean length shows some interesting features. After a lag phase, its increase becomes exponential at $t = \hat{t}$ (see Figure S1 in the Supplemental Material). Formally, we define the onset of growth $\hat{t}$ by intersecting the tangents to the $\overline{L}$–curve at $t = 0$ and the point where the increase is strongest (dashed line in Figure 3a; for details, see Appendix E and Figure S2). We observe that $\hat{t}$ coincides with the moment in time at which *higher-order ligations*, i.e., ligations involving at most one monomer become more abundant than ligations joining two monomers to a dimer (see Figure A3 in Appendix E).

Moreover, we can predict the onset with a relative error smaller than 15% by the following formula derived in Appendix E (dotted line in Figure 3).

$$\hat{t} \approx \log\left[\frac{c_{\text{tot}}^2 k_{\{1,1|2\}} - k_{\text{cut}}}{c_{\text{tot}} c_2^{\text{init}} k_{\{1,2|2\}}}\right] \Big/ \left(c_{\text{tot}}^2 k_{\{1,1|2\}} - k_{\text{cut}}\right) \tag{19}$$

Here, c_2^{init} is the initial dimer concentration and $k_{\{1,1|2\}}$ and $k_{\{1,2|2\}}$ are the effective rate constants at which new dimers or trimers are formed from monomers or monomers and dimers. $k_{\{1,1|2\}}$ is given by

$$k_{\{1,1|2\}} = \frac{k_{\text{lig}}}{K_{\{1,1|2\}}}, \tag{20}$$

where $K_{\{1,1|2\}}$ is the effective dissociation constant averaged over all triplexes involving two monomers and one dimer. An analogous expression exists for $k_{\{1,2|2\}}$. Repeating the computer experiment with k_{cut} and k_{lig} different from the standard values given in Table 2 confirmed the validity of Equation (19) (see Figures S3 and S4).

Moreover, altering k_{cut} and k_{lig} while keeping the other parameters fixed confirmed that the mean length $\overline{L}$ in the stationary state does not explicitly depend on these two variables but only on their ratio

$$\overline{L} = \overline{L}\left(k_{\text{lig}}/k_{\text{cut}}\right) \text{ for } t \to \infty, \, , \tag{21}$$

as expected from dimensional analysis. The dependence of the mean length on the ratio $k_{\text{lig}}/k_{\text{cut}}$ can be derived analytically for a random ligation model [81].

3.4. Energetic Bias in the Absence of Kinetic Stalling

In the previous section, we saw that the energetic discrimination between complementary and non-complementary nucleotide pairs alone is insufficient for the spontaneous emergence of patterns in sequence space. Therefore, we now ask whether an energetic bias $\Delta\gamma < 0$ favoring the binding of zebra motifs can induce zebra patterns that become self-amplifying, while kinetic stalling is still absent ($\sigma_1 = \sigma_2 = 1$).

The energetic bias, $\Delta\gamma < 0$, causes a transient overshoot in the mean length beyond the steady-state value but otherwise does not strongly affect the dynamics of the mean length (Figure 3a). The onset of growth $\hat{t}$ can still be predicted by a formula analogous to Equation (19) (see Appendix E and Figures S5–S7 for plots of single realizations). The steep increase after the lag phase is followed by a gradual descent to the steady state value, slightly below the maximum. Moreover, the steady-state length distribution remains very similar to the scenario without energetic bias.

In sequence space, a simple thermodynamic estimate Z^* for the final zebraness can be made based on a two-state system with an energy difference $\Delta\gamma$

$$Z^* = \frac{1}{1 + e^{\Delta\gamma}}. \tag{22}$$

Since this estimate neglects any correlation and feedback effects, one could naively expect that the zebranass resulting from the simulated model dynamics reaches a value larger than Z^*. Indeed, the observable initially grows beyond the estimate Z^*. However, the growth stops when the mean length reaches a value of $\overline{L} \approx 7$. At that point, the zebraness starts to decay and converges to a stationary value simultaneously with $\overline{L}$ (see Figure 3b). In the stationary state, the zebraness Z is only slightly above the result for random sequences, i.e., $Z = 0.5$, and below the simple thermodynamic estimate Z^*.

3.5. Loss of Energetic Discrimination Prevents Sequence Selection

Why are the initially emerging zebra patterns triggered by the energetic bias $\Delta\gamma < 0$ in Figure 3b neither amplified nor maintained? In the following, we analyze the growth processes in detail to give an intuitive explanation.

Strand growth requires the formation of complexes comprising at least three strands. The more negative the hybridization energies of the hybridization sites in these complexes are, the more stable the configurations become and the higher the probability for a ligation gets. Non-complementary nucleotide pairs increase the hybridization energy and, therefore, weaken the binding. To analyze the effect of these mismatches, we define the overall fraction of mismatches m as

$$m = \frac{N_{\text{non}}}{N_{\text{pairs}}}, \tag{23}$$

where N_{pairs} and N_{non} are the absolute numbers of nucleotide pairs and non-complementary nucleotide pairs in all present complexes. Initially, the fraction of mismatches m decreases since complexes mostly containing complementary nucleotides that emerge during the early growth persist longer and hence contribute more substantially to the average. However, the fraction of mismatches starts to increase when the mean length becomes larger (see Figure 3d). This increase of the mismatch fraction arises from the loss of thermodynamic discrimination induced by the cut-off k_{low} in the dehybridization rate, i.e., the effective temperature cycles (see Section 2.6). Although the hybridization energy may become arbitrarily negative for large hybridization sites, the dehybridization rate can not become smaller than the lower bound k_{low}. The length scale associated with k_{low} is $l_{\text{low}} = 7$ (see Section 2.8). This implies that an entirely complementary hybridization site comprising more than seven pairs has the same stability as a mismatch-free hybridization site composed of exactly seven pairs. Moreover, mismatches in extended hybridization sites might have no effect on the rate for unbinding because the hybridization site still contains a high number of matches. If many matches are present, the hybridization energies are strongly negative such that the lower threshold still determines the rate for dehybridization. This effect enables *concealed mismatches*. Concealed mismatches are mismatches that do not increase the dehybridization rate k_{off} of a hybridization site. Replacing a concealed mismatch with a complementary pair would not decrease k_{off} further since it is already given by the cut-off, i.e., $k_{\text{off}} = k_{\text{low}}$. The longer the strands become during the first growth phase, the more concealed mismatches emerge. With the absolute numbers of mismatches and concealed mismatches in all present complexes N_{non} and N_{con}, we now introduce the fraction of concealed mismatches n_{con} as

$$n_{\text{con}} = \frac{N_{\text{con}}}{N_{\text{non}}}, \tag{24}$$

The evolution of n_{con} shown in Figure 3e reveals that most of the occurring mismatches are concealed, once $\overline{L}$ has become approximately twice as large as l_{low}. Concealed mismatches also occur at the strand termini at ligation sites and may lead to the formation of new binary motifs which are not complementary to the templating motif at the ligation site. We call such a ligation involving at least one concealed mismatch a *concealed erroneous ligation*. Dividing the number of concealed erroneous ligations N_{err} per time by the overall number of ligations N_{lig} per time gives the fraction of concealed erroneous ligations $n_{\text{con}}^{\text{err}}$, i.e.,

$$n_{\text{con}}^{\text{err}} = \frac{N_{\text{err}}}{N_{\text{lig}}}. \tag{25}$$

Every erroneous concealed ligation mitigates the present bias in sequence space and leads to randomness. Erroneous concealed ligation is the reason why the initial sequence patterns decay almost to the random level $Z = 0.5$. However, not all hybridization sites, particularly the shorter ones, have a dehybridization rate determined by the lower bound. As the initial bias for binary zebra motifs on the system level decreases and sequences become more random, non-concealed mismatches in shorter hybridization sites become more likely. Hence, short hybridization sites not yet affected by the lower bound for the unbinding rate become less stable on average and contribute less to the growth process. This explains why the small maxima in the mean length and the other observables shown in Figure 3a,d–f disappear as the bias for binary zebra motifs fades away.

3.6. Kinetic Stalling in the Absence of Energetic Bias

We have seen above that concealed erroneous ligations suppress sequence selection during the self-assembly process. However, in the presence of kinetic stalling, concealed erroneous ligations should be reduced. In this section, we thus investigate the (more realistic) model variant, where kinetic stalling is included. We vary the strength of the kinetic stalling factor σ_1 from 0 to 0.1, while fixing $\sigma_2 = 0.1$.

Initially, the dynamics of the mean length $\overline{L}$ are qualitatively similar to the systems without kinetic stalling studied before (see Figures 3a and 4a). However, the onset of growth $\widehat{t}$ appears later. The time point of the onset $\widehat{t}$ can be predicted by a formula analogous to Equation (19), which considers the kinetic stalling effect, with an error <15%. For $\sigma_1 = 0.05$, the values for $\widehat{t}$ from the prediction and the formal definition are highlighted by the dotted and dashed lines in Figure 4a (for details, see Appendix E and Figures S8–S12). On larger timescales, the model including kinetic stalling deviates from the earlier model. After the steep increase, $\overline{L}$ does not directly settle to a steady-state. Instead, it grows gradually and converges to a constant value eventually. Visualizations with a linear x- and a logarithmic or linear y-axis reveal that the initial increase after the lag phase is approximately exponential, while the increase during the second growth phase is approximately linear (see Figures S13–S15). For $\sigma_1 = 0, 0.05$ or 0.067, similar stationary mean lengths are reached. However, the relaxation time increases with σ_1, such that it takes more than ten times longer for a system with $\sigma_1 = 0.067$ (see inset of Figure 4a) to converge to the stationary state than for a system with infinite stalling. For $\sigma_1 \leq 0.067$, the steady-state value of the mean length is more than twice as large as for the $\sigma_1 = \sigma_2 = 1$ scenario. Moreover, the length distributions in the stationary state look qualitatively similar to the ones seen earlier (see Figure S14). For $\sigma_1 = 0.1$, the increase of the mean length during the second growth phase is small during the time window of observation. From Figure 4a, we can not deduce whether the mean length already approached a stationary value, or whether it will keep on growing. If a stationary value was reached, it would be significantly smaller than for $\sigma_1 = 0.67, 0.05, 0$. For $\sigma_1 = 0.1$ (as well as for $\sigma_1 = 1$), the simulation times are large and prevented us from analyzing at longer time scales. However, plotting the curve for $\sigma_1 = 0.1$ in a coordinate system with a linear x-axis might suggest that the system has indeed already converged to a stationary state (see Figure S13). We will discuss the behavior for $\sigma_1 = 0.1$ in more detail in Section 3.9 and provide further evidence why the behavior, in this case, might be qualitatively different from the behavior for $\sigma_1 = 0.67, 0.05, 0$.

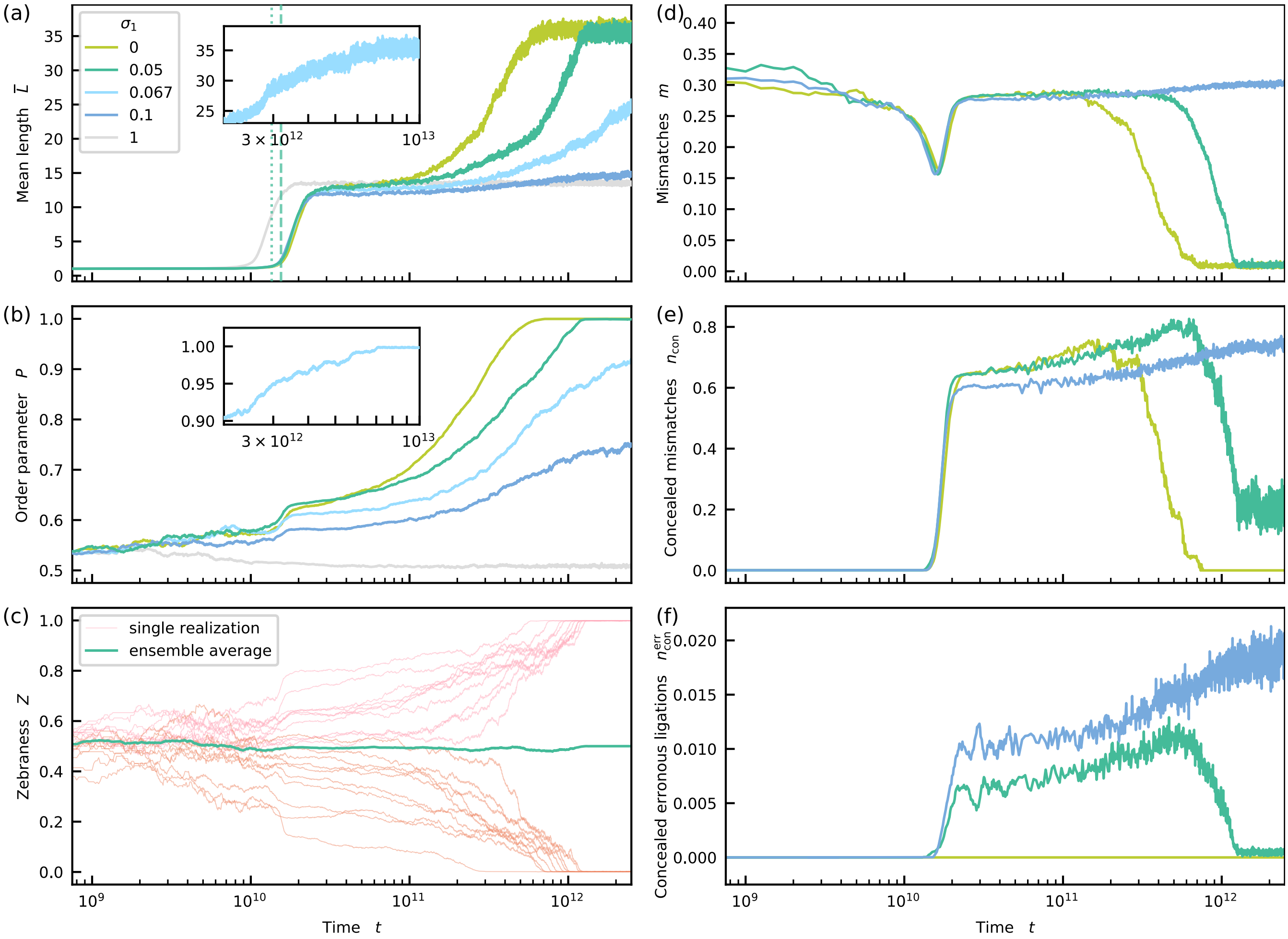

Figure 4. Evolution of mean length $\overline{L}$, sequence order parameter P, and system-level zebraness Z in a kinetic stalling scenario without energetic bias, i.e., $\Delta\gamma = 0$. For reference, we show the $\sigma_1 = \sigma_2 = 1$ curves (gray). (**a**) For $\sigma_1 = \sigma_2 < 1$, $\overline{L}$ shows two distinct growth phases. While the first one is rapid, the second one is slow. Relaxation to the steady-state for $\sigma_1 = 0.067$ appears much later (see inset). For $\Delta\gamma = 0.05$, the dashed line corresponds to $\hat{t}$ resulting from the formal definition, whereas the dotted line is the prediction. (**b**) The bias for alternating or homogeneous patterns established in the first growth phase becomes amplified during the second growth phase. For strong stalling ($\sigma_1 \leq 0.067$), the final pool comprises either pure zebra or fully homogeneous sequences. (**c**) The symmetry of the initial state is broken spontaneously. For $\sigma_1 = 0.05$, equal fractions of realizations evolve to the zebra or homogeneous state; (**c**–**f**) Dynamics of mismatches, concealed mismatches, and concealed erroneous ligations for $\sigma_1 = 0, 0.05, 0.1$. For details, see the main text.

Is the novel behavior of the mean length shown in Figure 4a related to a novel motif-selective dynamics in sequence space? We now investigate the evolution of the strands' sequences. Since the initial pool of sequences is symmetric and since neither zebra nor homogeneous binary motifs are preferred energetically, we do not expect a preference for a single realization to go to either a zebra ($Z > 0.5$) or a non-zebra ($Z < 0.5$) state. Hence, the system-level zebraness Z is not appropriate to describe an ensemble of realizations. As a meaningful observable to quantify the sequences bias on the ensemble level, we, therefore, introduce the *sequence order parameter* P as

$$P = \max\{Z, 1 - Z\}. \tag{26}$$

During the first growth phase of $\overline{L}$, a bias $P > 0.5$ is established for all values of σ_1. The dominance of the bias correlates with the strength of the kinetic stalling. During the slow second growth phase, P gradually increases and reaches a stationary value simultaneously with $\overline{L}$ (see Figures 4b and S11). For $\sigma_1 = 0, 0.05$, or 0.067, we observe a value of $P \approx 1$ in the stationary state. Hence, on the realization level, the final pool contains either (almost) entirely alternating or homogeneous sequences. We classify such states in sequence space as *pure*. For $\sigma_1 = 0.1$, the final sequence composition within the observation time window is also is non-random. However, the patterns are not pure, i.e., $0.5 < P < 1$. We refer to these states as *partially mixed* states. Whether the system has already converged to a stationary state with $P < 1$ or will further evolve to a pure state as for $\sigma_1 = 0.67, 0.05, 0$ remains unclear at this point (see above). Figure 4c displays the evolution of the zebraness of all realizations forming the steady-state for $\sigma_1 = 0.05$. On average, one half of the realizations evolves towards the $Z = 1$ state, while the other half evolves towards the $Z = 0$ state. Hence, the symmetry of the initial state is broken spontaneously: either zebra or homogeneous motifs are selected. (See Figures S8–S12 for equivalent plots of single realizations for other σ_1 values.)

3.7. Hydrolysis and Stalling Boost Sequence Selection

The previous section revealed a coupling between sequence selection and two distinct growth phases in the kinetic stalling scenario. We now interpret and explain this coupling. Here, we consider the cases $\sigma_1 = 0$ and $\sigma_1 = 0.05$, where all trajectories eventually converge to a pure state. The case where $\sigma_1 = 0.1$ is discussed in Section 3.9.

On the level of individual trajectories, fluctuations lead to a small bias in the motif composition even before the mean length starts to grow rapidly (see Figure 4c). This early asymmetry in the distribution of alternating and homogeneous motifs governs the fate of the realization as seen from Figure 4c.

During the first growth phase (see Figure 4a), monomers and short strands self-assemble into longer strands. The first growth phase ends when most of the initial monomers are consumed. At the end of this initial growth phase, a significant bias towards alternating or homogeneous motifs is present. However, a considerable fraction of binary motifs does not yet reflect system-level bias, i.e., differs from the dominant binary motifs (which are either $X - Y$ and $Y - X$ or $X - X$ and $Y - Y$). Therefore, mismatches in bound strands are still frequent (see Figure 4d). Moreover, since the average strand length is already significantly above l_{low}, most mismatches are concealed (see Figure 4e).

When monomers and short strands do not dominate the pool anymore, hydrolysis becomes important. Every time a strand breaks, an existing binary motif vanishes. During the second growth phase, fragments of broken strands are reassembled to longer strands via ligation on a template strand. (For an analysis of sequence patterns of strands of specific lengths, see Figure S17). If the kinetic stalling is strong, the ligation of two strands is (almost) impossible if a mismatch occurs at the ligation site and the fraction of concealed erroneous ligations is (close to) zero (see Figure 4f). Hence, every ligation forms a new binary motif that (almost) always complements the templating motif at the ligation site. If the templating motif is zebra-like (homogeneous), the new motif is zebra-like (homogeneous) too. Over time, all binary motifs created during the initial growth phase, particularly those that do not reflect the system bias, get destroyed at a uniform rate. At the same time, binary motifs emerging during the second growth phase likely reflect the system bias. Consequently, the bias becomes self-enhancing (see Figure 4b). Newly created binary motifs enhance the system bias even more, while all motifs that do not reflect the asymmetry in motif space become extinct eventually. As a result, the motif composition becomes more and more ordered and mismatches become rarer (see Figure 4d). Remaining mismatches are now even more unlikely to affect the dehybridization rate since the rate is determined by the lower bound k_{low} for long and primarily complementary hybridization sites. Hence, the fraction of concealed mismatches increases slightly (see Figure 4e). Moreover, if kinetic stalling is finite ($\sigma_1 > 0$), the fraction of concealed erroneous ligations also slightly increases (see Figure 4f).

Since the energetic properties are symmetric, every realization randomly approaches either the zebra-like or homogeneous state. As the pool becomes more complementary in its motif composition, triplex configurations achieve higher stability on average. This increase of stability enhances the probability of templated ligations. Therefore, the symmetry breaking in sequence space is concomitant with enhanced growth, which becomes apparent in the further increase of the mean length. For $\sigma_1 = 0$ and $\sigma_1 = 0.05$, the second growth phase ends when the system has reached an (almost) entirely alternating or homogeneous state (see Figure 4a,b). At that point, the fraction of mismatches is close to zero. The few mismatches that still occur are mostly due to strayed mononucleotides and short oligomers sitting on longer strands. Since these mononucleotides and short oligomers are far from being affected by the lower bound and unbind quickly, the fraction of concealed mismatches takes small values again.

3.8. Energetic Bias in the Presence of Kinetic Stalling

The previous section revealed that spontaneous motif selection occurs as a result of kinetic stalling. Without energetic bias, the sequence pool converges to a stationary state which is either dominated by homogenous or alternating sequences. In addition, the energetic symmetry can be broken explicitly if a small energetic bias favoring zebra motifs is applied. To understand the emergent phenomena in this setting, we study two systems, one with strong and one with weak kinetic stalling for various energetic biases $\Delta\gamma < 0$.

First, we consider the case where $\sigma_1 = 0.05$. Most parts of the description in the previous section ($\Delta\gamma = 0$) also apply here (see Figure 5a,b). Again, we can predict the onset of growth $\hat{t}$ (for details, see Appendix E and Figures S18–S20). The steady-state value of $\overline{L}$ depends weakly on the energetic bias. However, the final state is reached earlier if the bias is stronger. In sequence space, all trajectories end in a pure zebra state $Z = 1$. Hence, symmetry breaking in sequence space is now induced energetically as expected because of the explicit symmetry breaking in the energy landscape.

Second, we investigate a scenario with $\sigma_1 = \sigma_2 = 0.1$. The mean length grows strongly in the beginning as before (see Figure 5c and, for more details, see Appendix E and Figures S21–S23). The fast growth phase is followed by either a marginal increase ($\Delta\gamma = -0.1, -0.2$) or decrease ($\Delta\gamma = -0.3$) of $\overline{L}$ to a stationary value correlating with the strength of the energetic bias. A strong zebra pattern $Z > 0.5$ is induced in sequence space during the initial increase of the mean length (see Figure 5d). While $\overline{L}$ grows (decays) during the second phase, Z also grows (decays). Eventually, the sequence pool converges to a partially mixed stationary state with a significant majority of zebra motifs such that $0.5 < Z < 1$. The excess of zebra motif again correlates with the strength of the energetic bias. Moreover, from Figures S21–S23, it becomes clear that all single trajectories behave similarly to the ensemble mean, i.e., show steady-state values of the zebraness above 0.5.

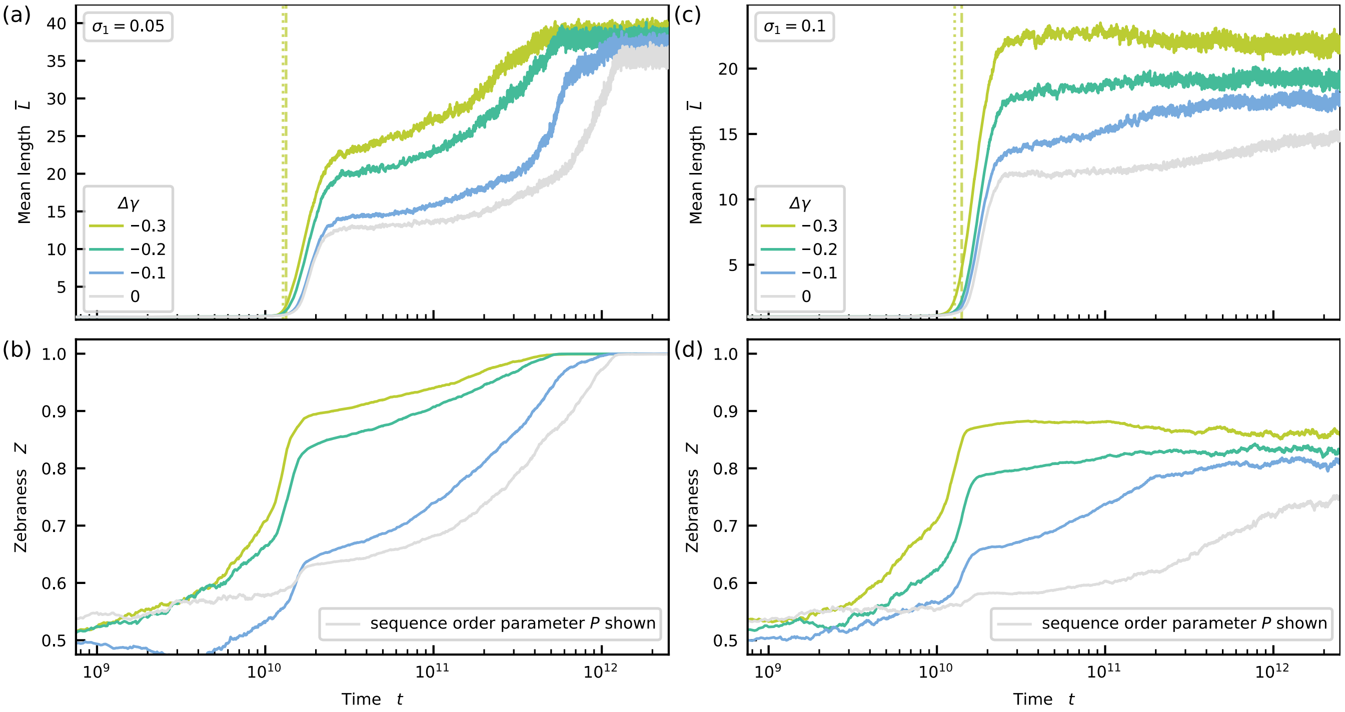

Figure 5. Left column: scenario with $\sigma_1 = 0.05$, right column: scenario with $\sigma_1 = 0.1$ (**a**) The mean length $\overline{L}$ grows again in two steps. The stronger the bias, the earlier the relaxation to the stationary value. For $\Delta\gamma = -0.3$, the dashed line corresponds to $\hat{t}$ resulting from the formal definition, whereas the dotted line is the prediction (same for (**c**)). (**b**) Pronounced zebra patterns emerge during the first growth phase. The patterns become pure during the second growth phase. (**c**) A gradual increase or decay follows the rapid growth phase. The steady-state value of $\overline{L}$ correlates with the strength of the energetic bias. (**d**) While $\overline{L}$ slightly increases (decreases), Z also increases (decreases). The sequence pool in the stationary-state shows mixed patterns dominated by zebra motifs. The fraction of zebra motifs depends on the energetic bias. (**a**–**d**) For reference, we also show the sequence order parameter P for the $\Delta\gamma = 0$ curves (gray).

3.9. Weak versus Strong Kinetic Stalling

In Section 3.6, we speculated whether the system with $\sigma_1 = 0.1$ and without energetic bias (blue curve in Figure 4) reaches a stationary state characterized by $P < 1$ in contrast to the scenarios with $\sigma_1 = 0.067, 0.5, 0$, where $P = 1$ and referred to the corresponding plot with a linear x-axis (see Figure S13). However, this plot did not allow for a clear conclusion either. It could be that the alleged partially mixed stationary state is only transient and that a pure state is reached on much larger time scales. Though the findings from Section 3.8 suggest that the stationary state of the system with $\sigma_1 = 0.1$ and without energetic bias is indeed qualitatively different from the scenarios with $\sigma_1 \leq 0.067$ and without energetic and characterized by $P < 1$. The curves for $\sigma_1 = 0.1$ and various energetic biases clearly converge to stationary states with $P < 1$ (see colored curves in Figure 5c,d). If the stationary state is partially mixed in the presence of an energetic bias, it is not too far-fetched to assume that it is also partially mixed if the energetic bias is absent.

Naturally, the question arises whether a critical value for σ_1 exists above which the system always reaches a pure state with $P = 1$ for a given value of the energetic bias $\Delta\gamma$. This first question directly leads to a second question, namely, what would be the nature of the corresponding non-equilibrium phase transition? We leave the answer to this question open for future research. However, finding an answer might be challenging since the relaxation time to the stationary state will probably diverge. At this point, we content ourselves with hypothesizing that two different regimes might exist without drawing an exact border: For *strong* kinetic stalling, the system converges to a *pure* state, while, for *weak* kinetic stalling, it converges to a *partially mixed* state.

In the light of the above hypothesis, we now analyze the dynamics of mismatches and concealed erroneous ligations in the energetically unbiased system case for $\sigma_1 = 0.1$ (see the blue curve in Figure 4d–f as for $\sigma_1 = 0$ and $\sigma_1 = 0.05$. The self-amplification of the dominant binary motif comes to a halt during the second growth phase (see Figure 4a Consequently, the fraction of concealed mismatches does not decrease again as for $\sigma_1 = 0$ and $\sigma_1 = 0.05$. Since most of the strands are long enough, most of the mismatches are concealed. Moreover, concealed erroneous ligations are not fully suppressed and still occur in the stationary state. Hence, the weak kinetic stalling scenario includes features of both dynamic regimes described in Sections 3.5 and 3.7.

4. Discussion

4.1. Summary

Our study investigated the self-assembly of prebiotic polymers via templated ligation inside a non-equilibrium RNA reactor. We identified kinetic stalling as a critical factor for self-enhanced sequence selection. The final sequence space shows no sequence patterns if the underlying stacking energies are uniform without kinetic stalling. In contrast, spontaneous symmetry breaking occurs for strong kinetic stalling. The final pool contains either entirely homogenous or zebra-like sequences. In scenarios without kinetic stalling, any energetically induced sequence bias vanishes almost completely as strands grow. In contrast, in the presence of kinetic stalling, subtle differences in the stacking energies trigger cascades of self-amplification, leading to highly ordered sequence pools. Our results hint towards the existence of two different stalling regimes. We hypothesize that, for strong kinetic stalling, the system converges to a pure state with $P = 1$, while, for weak kinetic stalling, it converges to a partially mixed state characterized by $P < 1$.

Initially, the mean length shows burst-like growth dynamics after a short lag phase. The onset of the rapid growth coincides with the time point where higher-order ligations become more abundant than ligations joining two monomers and can be predicted analytically.

4.2. Prior Work, Our Model, and Future Extensions

In an earlier model for prebiotic self-assembly, strands only grow via random ligation [82]. There, self-folding and complex formation introduced a protection mechanism against hydrolysis for double-stranded segments. Moreover, the ensemble of strands was assumed to reach a binding equilibrium immediately after a random ligation occurred. In this model, protection against hydrolysis could extend the system's sequence memory. However, the effect was only transient, and all selected patterns vanished eventually. A more recent study combines random ligation and protection against hydrolysis with growth via templated polymerization by mononucleotides [83]. The authors demonstrate that polynucleotides exceeding lengths of 100 can emerge under plausible conditions. Moreover, the authors show that a considerable fraction of the emerging strands forms ribozyme- and tRNA-like secondary structures using folding software. However, the study does not investigate correlations in sequence space on the system level.

Previous theoretical studies considering growth via templated ligation generally explored effective models that reduce the state space to (sub-)sequences without considering complex formation explicitly [33,77,84–92]. Such approaches do not treat (de-)hybridization and ligation as elementary steps. Instead, the reactions are coarse-grained into one extension process. The specification of the corresponding rate neglects the intricacies of the assembly mechanisms and requires a priori assumptions regarding the relevant configurations [33,84–87,90]. Moreover, many models ignore that the hybridization energy is a function of the number and nature of the paired nucleotides and use constant (de-)hybridization rates [33,84–87,89–91]. Other studies treat the sequence dependence of (de-)hybridization employing mean-field approximations where sequence correlations are dismissed [77]. Such simplifications result in systems effectively containing only one type of self-complementary nucleotide [56] and any form of sequence selection is necessarily absent. In contrast, our stochastic approach explicitly takes the sequence-dependent thermodynamic and kinetic aspects of templated ligation into account.

Although being already quite complex, our model also made simplifying assumptions. Future studies need to relax some of our assumptions. In particular, one has to consider nonlinear complexes containing loops and multiple branches. Such configurations can give rise to self-folding, self-templating, template inhibition, and gelation [55,82,93]. All these features potentially influence the sequence dynamics. However, we expect that these effects only become important in the long time limit once the strands have reached sufficient size for the formation of secondary structures. Consequently, the discussion of the emergence of structured sequences on shorter timescales in the limit of strong kinetic stalling is expected to hold, even if secondary structures are taken into account. In addition, one has to extend the alphabet size from two to four. The question here is whether pure states containing only a minimal number of sub motifs exist. In the future, the model could also include additional reactions such as non-templated polymerization and ligation, and recombination [82,89,94–98] or length selective environments [99]. The first two reactions probably play a role in the formation of the first short oligomers, whereas a flow-through system preferentially accumulating long strands can escalate polymerization [100]. Moreover, our study assumed a well-mixed system. Introducing a spatial component together with size-dependent diffusion constants as in Ref. [101], one could study under which conditions local clusters of sequences with specific patterns can emerge and coexist.

4.3. Plausibility of a Binary Alphabet

Our study assumed a binary alphabet following previous theoretical work [31,86,87,91,92,102–104]. While this assumption simplified the analysis, there is also evidence for a two-letter alphabet preceding the four-letter alphabet [9,10,49,105–107]. The plausibility is also underlined by the fact that functional sequences composed of only two types exist [108,109]. For the sake of generality, we referred to the two types of nucleotides appearing in our model as X and Y. This terminology was motivated by the idea that a pre-RNA, sometimes called *prebioitic XNA*, or alternative RNA nucleotides, may have existed before the modern RNA came into being [110–116]. Various backbone chemistries [117–122], non-canonical nucleotides [105,123–127], and chemical modifications [128–130] are eligible, some of which, e.g., PNA and TNA are more plausible to emerge [131–133] under the conditions on the early Earth than RNA.

4.4. Significance for the Emergence of Life

What is the origin of the first ribozymes heralding the transition from the pre-RNA to the RNA world? In Darwinian evolution, the assembly of low-level building blocks into higher-level entities triggered significant developments [134]. In the light of this evolutionary principle, a multi-step process towards greater complexity, eventually resulting in functionality, also seems natural in prebiotic evolution. Here, we studied one of the first steps following the emergence of early nucleotides. This step forms oligonucleotides displaying distinct sequence patterns that could serve as building blocks for the next higher level of self-organization towards functional ribozymes.

In our study, we considered model variants with and without kinetic stalling. Since kinetic stalling is probably inevitable in non-enzymatic templated ligation [32,38–40], the no-stalling variant may appear unwarranted. However, this model variant is essential to separate the effects and identify kinetic stalling as a crucial mechanism enabling self-enhancing sequence selection (see Section 3.5). Moreover, the strength of the stalling effect depends on the underlying activation and nucleotide chemistry [32,38,58] and both weak and strong kinetic stalling scenarios, potentially leading to qualitatively different outcomes, are plausible (see Sections 3.6, 3.8 and 3.9). Furthermore, in a pre-RNA world scenario, a primitive ribozyme catalyzing ligations might have a poor ability to discriminate mismatching ends kinetically. In this case, the ribozyme would operate in a regime where thermodynamics mainly control the discrimination between complementary and non-complementary nucleotides. This regime would be close to the no-stalling model variant.

Moreover, our study revealed that minor differences in the motif-dependent stacking energies significantly affect the dynamics in sequence space. The experiments of Refs. [18,55] probably did not capture this effect, due to the design of the respective experimental systems. These studies used DNA 12- or 20-mers as basic building blocks and a ligase to catalyze bond formation. The ligase requires significant overlaps of both strands with the template to work efficiently. Moreover, the experiments were performed under temperature cycling. The applied temperature cycling was too fast for the long strands to reach a binding equilibrium. Therefore, the hybridization timescale of mostly complementary hybridization sites is always set by the duration of the cold phases. Hence, subtle variations in the stacking energies are not visible. In contrast, the effective cycling in our model is slow enough for thermodynamics to govern the (de-)hybridization of short strands leading to an amplification of the stacking bias (see Section 3.8). The cut-off of the dehybridization rate only affects longer strands emerging from the pool that is already biased.

In non-enzymatic scenarios involving kinetic stalling, the strands formed from the initial pool are already the result of a primary selection process. Selection is not imposed externally but stems from a self-organizing replication network [135]. We showed that the ability to form self-organizing and self-amplifying replication networks is inherent in template-directed growth and does not require higher-level mechanisms such as sequence-specific template inhibition as a result of self-folding, reported previously [55]. In the 1980s, Kauffmann promoted the concept of *autocatalytic sets*—sets of molecules that mutually catalyze their formations [136]—as a chemical intermediate on the way to biological life [137]. Since then, autocatalytic sets have been the subject of many theoretical and experimental studies [29,30,138–143]. Once our system has reached a stationary state in the strong kinetic stalling regime, it shows the key features of such an autocatalytic set: strands with a specific pattern promote the formation of new strands of arbitrary length, showing the same pattern. Importantly, this concrete realization of an autocatalytic set emerges naturally from an unstructured initial pool without requiring any external (pre-)engineering. This insight could bridge the gap between strand formation and self-sustaining sequence replication.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/life12040567/s1.

Author Contributions: Conceptualization, T.G., B.A. and U.G.; software, T.G. and J.H.R.; formal analysis, T.G.; investigation, T.G.; writing—original draft preparation, T.G. and U.G.; writing—review and editing, T.G. and U.G.; visualization, T.G.; supervision, B.A. and U.G.; project administration, T.G. and U.G.; funding acquisition U.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the German Research Foundation (DFG) via the TRR 235 Emergence of Life (Project-ID 364653263) and the excellence cluster ORIGINS.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets that support the findings of this study, as well as the computer code used to generate these data, are available from the corresponding author upon reasonable request.

Acknowledgments: We thank Dieter Braun, Annalena Salditt, and the members of the Gerland group for valuable feedback on this manuscript.

Conflicts of Interest: The authors declare no competing interest.

Appendix A. Terminal Base Pair Energies

The energies ϵ_j (with the index j denoting to either $-$ or $+$ end) assigned to terminal nucleotide pairs are functions of the complex structure and the sequence context as well. If a terminal nucleotide pair coincides with the end of a complex, a so-called *blunt end*, there is no additional contribution, i.e., $\epsilon_j = 0$. In contrast, a terminal nucleotide pair followed or preceded by an unpaired nucleotide, a so-called *dangling end*, adds a non-zero contribution to hybridization energy, i.e., $\epsilon_j \neq 0$ [36,37].

Again, a mismatch results in an energetic penalty $\epsilon_j > 0$, weakening the binding, while a complementary nucleotide pair leads to a reward $\epsilon_j < 0$ increasing stability. As for the stacking energies $\gamma([P_i\, P_{i+1}])$, we distinguish between complementary *alternating* terminal configurations such as

$$\begin{bmatrix} & Y \\ & \cdot \\ Y- & X \end{bmatrix} \text{ or } \begin{bmatrix} Y- & X \\ \cdot & \\ X & \end{bmatrix}, \tag{A1}$$

and complementary *homogeneous* terminal configurations as, for example,

$$\begin{bmatrix} & X \\ & \cdot \\ Y- & Y \end{bmatrix} \text{ or } \begin{bmatrix} Y & \\ \cdot & \\ X- & X \end{bmatrix}. \tag{A2}$$

We denote the associated energies by ϵ_{alt} and ϵ_{hom}. As before, we treat the energy difference

$$\delta_\epsilon = \epsilon_{\text{alt}} - \epsilon_{\text{hom}}. \tag{A3}$$

as a variable parameter. Moreover, for a dangling end configuration involving one mismatch, we assume a constant contribution ϵ_{1nc}. The terminal nucleotide pair contributions obey the inequality

$$\epsilon_{\text{alt}} \leq \overline{\epsilon_{\text{com}}} \leq \epsilon_{\text{hom}} < \epsilon_{\text{1nc}}, \tag{A4}$$

with $\overline{\epsilon_{\text{com}}} = (\epsilon_{\text{alt}} + \epsilon_{\text{hom}})/2$. Parameter values used in Section 3 are summarized in Table A1. For simplicity, we set the energy difference between alternating and homogeneous dangling end contributions to half of the value of the difference between full alternating and homogeneous blocks, i.e.,

$$\delta_\epsilon = \frac{1}{2}\Delta\gamma. \tag{A5}$$

Next, we consider the contribution to the hybridization energy resulting from a terminal nucleotide pair, part of a ligation site. While, in principle, one can choose the contributions due to dangling and blunt ends freely (within a reasonable range), contributions coming from ligation sites are constrained. The reason is that one ligation site is involved in two hybridization sites.

Table A1. Summary of energy parameters for dangling ends used in Section 3.

Process	Parameter	Value
dehybridization	$\overline{\epsilon_{\text{com}}}, \epsilon_{\text{1nc}}, \delta_\epsilon$	$-0.625, 0.375, [-0.15, 0]$

To formulate this constraint, we introduce the total hybridization energy $\beta\Delta G_{\text{tot}}(C)$ of a complex C. $\beta\Delta G_{\text{tot}}(C)$ is obtained by first summing over the stacking energies of all *continuous* nearest neighbor blocks within all hybridization sites of the complex, i.e., all blocks of the form

$$\begin{bmatrix} X-Y \\ \cdot \quad \cdot \\ Y-X \end{bmatrix} \text{ or } \begin{bmatrix} X-X \\ \cdot \quad \cdot \\ Y-X \end{bmatrix} \text{ or } \dots, \tag{A6}$$

where the nucleotide pairs are linked by two covalent bonds (represented by the $-$ symbols). Next, we add the sum over all dangling end contributions. In the last step, we add an energy contribution for every ligation site, i.e., for every *noncontinuous* nearest neighbor block of the form

$$\begin{bmatrix} X \quad Y \\ \cdot \quad \cdot \\ Y-X \end{bmatrix} \text{ or } \begin{bmatrix} X-X \\ \cdot \quad \cdot \\ Y \quad X \end{bmatrix} \text{ or } \dots, \tag{A7}$$

where one covalent bond is missing. Since the covalent bond does not affect the stacking interaction, the energy contribution to the total hybridization energy $\beta\Delta G_{\text{tot}}(C)$ of noncontinuous blocks is the same as for the corresponding blocks [37]. For γ as a function of the (non)continuous nearest neighbor blocks, we have, for example,

$$\gamma\left(\begin{bmatrix} X-Y \\ \cdot \quad \cdot \\ Y-X \end{bmatrix}\right) = \gamma\left(\begin{bmatrix} X \quad Y \\ \cdot \quad \cdot \\ Y-X \end{bmatrix}\right). \tag{A8}$$

with that, $\beta\Delta G_{\text{tot}}(C)$ reads

$$\beta\Delta G_{\text{tot}}(C) = \sum_{\substack{i \in \text{continuous} \\ \text{blocks}}} \gamma_i + \sum_{\substack{d \in \text{dangling} \\ \text{ends}}} \epsilon_d + \sum_{\substack{l \in \text{ligation} \\ \text{sites}}} \gamma_l \tag{A9}$$

Note that we would obtain an identical total hybridization energy if we would replace all non-continuous nearest neighbor blocks in the complex C with continuous blocks, i.e., if we would join the $+$ and $-$ strands at all ligation sites.

We now consider two complexes C_1 and C_2 reacting to a new complex C_3 containing one or two new ligation sites. The energy difference between the states before and after the reaction has to correspond to the hybridization energy Γ_{new} associated with the newly formed hybridization site, i.e.,

$$\Gamma_{\text{new}} \overset{!}{=} \beta\Delta G_{\text{tot}}(C_3) - [\beta\Delta G_{\text{tot}}(C_1) + \beta\Delta G_{\text{tot}}(C_2)] \tag{A10}$$

Γ_{new} could also be interpreted as the energy that is needed to reverse the reaction. The constraint formulated via Equation (A10) defines the energetic contribution of the newly formed ligation site(s) to Γ_{new} associated with the newly formed hybridization site.

In practice, the constraint Equation (A10) reduces to the difference between the stacking contribution(s) $\gamma_l^{(C_3)}$ associated with the newly formed non-continuous nearest neighbor block(s) in the new complex C_3 and the corresponding dangling end contribution(s) $\epsilon_d^{(C_1)}$ and $\epsilon_d^{(C_2)}$ of the old complexes C_1 and C_2.

Newly formed ligation sites also affect the hybridization energies of hybridization sites that were already existing in C_1 or C_2 before the reaction occurred. Hence, these hybridization energies have to be recalculated comparing the total hybridization energies of the new complex C_3 and the compounds that result from virtually dissolving the hybridization site. This procedure is explained in more detail in Appendix B, where we explicitly derive hybridization energies in several exemplary complex configurations.

Note that, in general,

$$\beta\Delta G_{\mathrm{tot}}(C) \neq \sum_{h \in \substack{\text{hyb.} \\ \text{sites}}} \Gamma_h. \tag{A11}$$

This inequation becomes an equation only if the complex C does not contain any ligation sites. Moreover, we can interpret $\beta\Delta G_{\mathrm{tot}}(C)$ as the total energy stored within the complex. It is equal to the energy difference between the fully disassembled state, where we only have single strands.

Appendix B. Examples for Hybridization Energies

To illustrate the calculation of the hybridization energy, we consider four complexes, sketched in Figure A1.

In example Figure A1a, the duplex C_1, and the single strand C_2 react to the triplex C_3. Before the reaction, the total hybridization energies of the complexes are

$$\beta\Delta G_{\mathrm{tot}}(C_1) = \epsilon_{\mathrm{hom}} + \gamma_{\mathrm{alt}} + \epsilon_{\mathrm{alt}}, \tag{A12}$$

$$\beta\Delta G_{\mathrm{tot}}(C_2) = 0. \tag{A13}$$

After the reaction, the total hybridization energy is

$$\begin{aligned} \beta\Delta G_{\mathrm{tot}}(C_3) = \epsilon_{\mathrm{hom}} + \gamma_{\mathrm{alt}} + \epsilon_{\mathrm{alt}} \\ + \epsilon_{\mathrm{alt}} + \gamma_{\mathrm{alt}} + 2\gamma_{\mathrm{hom}} \end{aligned} \tag{A14}$$

As stated above, the hybridization energy Γ_{new} associated with the new hybridization site is the difference between the total hybridization energies before and after the reaction, i.e.,

$$\begin{aligned} \Gamma_{\mathrm{new}} &= \beta\Delta G_{\mathrm{tot}}(C_3) - [\beta\Delta G_{\mathrm{tot}}(C_1) + \beta\Delta G_{\mathrm{tot}}(C_2)] \\ &= \epsilon_{\mathrm{alt}} + \gamma_{\mathrm{alt}} + 2\gamma_{\mathrm{hom}}. \end{aligned} \tag{A15}$$

Since the reaction did not lead to a new ligation site, Γ_{new} corresponds to the sum over all nearest-neighbor blocks plus the danging end contributions.

In example Figure A1b, two single strands C_1 and C_2 form a new duplex C_3. The total hybridization energy before the reaction is zero. Hence, the hybridization energy Γ_{new} of the new hybridization site equals the total hybridization energy $\beta\Delta G_{\mathrm{tot}}(C_3)$ after the reaction, i.e.,

$$\begin{aligned} \Gamma_{\mathrm{new}} &= \beta\Delta G_{\mathrm{tot}}(C_3) - 0 \\ &= \epsilon_{\mathrm{alt}} + \gamma_{\mathrm{alt}} + 2\gamma_{\mathrm{1nc}} + \gamma_{\mathrm{hom}} + \epsilon_{\mathrm{hom}} \end{aligned} \tag{A16}$$

Again, since no ligation sites are involved, Γ_{new} is equivalent to the sum over stacking and end contributions.

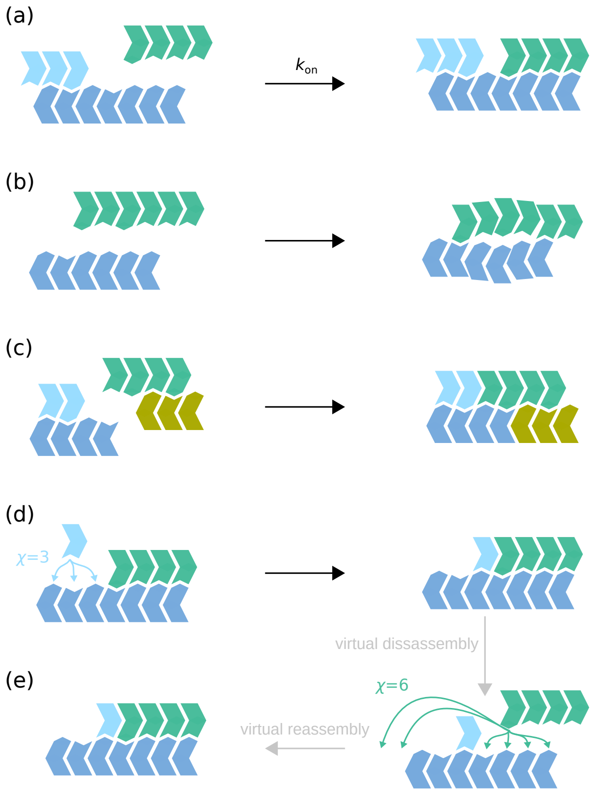

Figure A1. (**a**) A duplex C_1 with two dangling ends and a single strand C_2 form a new triplex C_3 with a blunt end. C_3 has no ligation site. (**b**) Two single strands C_1 and C_2 react to a duplex C_3 displaying a mismatch and two dangling ends. (**c**) Two duplexes C_1 and C_2 hybridize to new complex C_3 featuring two new ligation sites. (**d**) A mononucleotide C_1 hybridizes onto a duplex C_2. The new triplex has a new ligation site. (**e**) We need to update the channel factor χ to renew the dehybridization rate k_{off} associated with the hybridization site that already existed before the binding of the monomer. To this end, we virtually dissolve this hybridization site and directly reassemble the complex and recount the possible reaction channels and obtain a new (integer) value for the channel factor χ (see main text).

In example Figure A1c, the emerging complex C_3 features two new ligation sites. The total hybridization energies of the initial complexes C_1 and C_2 are

$$\beta \Delta G_{\text{tot}}(C_1) = \gamma_{\text{alt}} + \epsilon_{\text{alt}}, \tag{A17}$$

$$\beta \Delta G_{\text{tot}}(C_2) = \gamma_{\text{alt}} + 2\epsilon_{\text{alt}}. \tag{A18}$$

The total energy of complex C_3 is

$$\beta \Delta G_{\text{tot}}(C_3) = 5\gamma_{\text{alt}} + \epsilon_{\text{alt}}. \tag{A19}$$

As before, the difference between the total hybridization energies before and after the reaction determines the hybridization site energy Γ_{new} of the new hybridization site,

$$\begin{aligned} \Gamma_{\text{new}} &= \beta \Delta G_{\text{tot}}(C_3) - [\beta \Delta G_{\text{tot}}(C_1) + \beta \Delta G_{\text{tot}}(C_2)] \\ &= 3\gamma_{\text{alt}} - 2\epsilon_{\text{alt}}. \end{aligned} \tag{A20}$$

In the last example, Figure A1d, a mononucleotide C_1 hybridizes to the duplex C_2 resulting in the triplex C_3. The hybridization energies before and after the reaction are

$$\beta \Delta G_{\text{tot}}(C_1) = 0, \tag{A21}$$

$$\beta \Delta G_{\text{tot}}(C_2) = \epsilon_{\text{alt}} + \gamma_{\text{alt}} + 2\epsilon_{\text{hom}}, \tag{A22}$$

$$\beta \Delta G_{\text{tot}}(C_3) = \epsilon_{\text{alt}} + 2\gamma_{\text{alt}} + 2\epsilon_{\text{hom}}. \tag{A23}$$

Consequently, the hybridization energy Γ_{new} assigned to the mononucleotide is

$$\begin{aligned} \Gamma_{\text{new}} &= \beta\Delta G_{\text{tot}}(C_3) - [\beta\Delta G_{\text{tot}}(C_1) + \beta\Delta G_{\text{tot}}(C_2)] \\ &= \gamma_{\text{alt}}. \end{aligned} \tag{A24}$$

In the examples Figure A1c,d, we also have to recalculate the hybridization energy of the hybridization sites that were present before the reactions since they now also involve an energetic contribution from a ligation site. To this end, we virtually dissolve the hybridization site(s) that were already existing and then virtually reassemble these hybridization sites again. Reassembling the hybridization site(s), we apply the same procedure of calculating the hybridization energy as described before. The updated hybridization energy (energies) obtained that way now include(s) the correct contribution of the newly formed ligation site(s). Moreover, in the examples Figure A1a,c,d, we have to reconsider the channel factors associated with already existing hybridization sites. To update these channel factors, we again virtually disassemble the complex into its compounds (see Figure A1e). We then virtually reassemble the parts and thereby count the number of different configurations that would be possible. With the updated channel factors and hybridization energies, we recompute the dehybridization rates according to Equation (7).

Appendix C. Thermodynamics of Hybridization

With the elementary rates defined in Equations (6) and (7), the total free energy $\Delta\mathcal{G}_{\text{tot}}(C)$ of a complex C is found to be

$$\beta\Delta\mathcal{G}_{\text{tot}}(C) = \beta\Delta G_{\text{tot}}(C) + \rho\ln(2), \tag{A25}$$

for constant environmental conditions [56]. The first term on the right-hand side is the total hybridization energy defined in Equation (A9), and the second term is a *symmetry penalty* that occurs if the complex is rotationally symmetric ($\rho = 1$) and is zero ($\rho = 0$), otherwise. The free energy $\beta\Delta\mathcal{G}_{\text{tot}}(C)$ is linked to the dissociation constant K_D occurring in a mass-action approach where all concentrations are expressed in units of the standard concentration $c_\circ$ via [57]

$$\beta\Delta\mathcal{G}_{\text{tot}}(C) = \ln(K_D). \tag{A26}$$

Thermodynamically, the symmetry penalty can be understood as a decrease in the (standard internal) entropy by a factor of $\ln(2)$ due to the rotational symmetry.

Kinetically, it is rationalized by looking at the interaction probability of two complexes belonging to the same species versus the interaction probability of two complexes representing different species. For equal concentrations, complexes representing different species interact twice as often as complexes belonging to the same species. While the complex resulting from a collision between distinguishable complexes is never symmetric, the interaction of identical molecules always leads to a complex with a rotational symmetry [56]. Hence, the $\rho\ln(2)$-term arises naturally in any collision based kinetic model [81]. We emphasize that the symmetry penalty is due to a reduced product-formation rate rather than a decrease in stability of the complex.

Moreover, the symmetry penalty also appears in the standard databases for free energies of hybridized oligonucleotides [36,37]. These databases also add a constant *initiation penalty* to the total free energy $\beta\Delta\mathcal{G}_{\text{tot}}(C)$. The initiation penalty is a constant multiplied by $(n-1)$, where n is the number of strands forming the complex. The initiation penalty accounts for the loss of system entropy due to the fusion of separate entities into one new complex. However, this penalty term can be set to zero through a rescaling of concentrations and therefore does not occur in our approach [56].

Appendix D. Distribution of Longer Motifs

In Section 3.3, we discussed a scenario without kinetic stalling ($\sigma_1 = \sigma_2 = 1$) and without energetic bias ($\Delta\gamma = 0$). There, the zebraness converged to $Z = 0.5$ in all individual realizations, pointing to an entirely random sequence pool (see Figure 3c). However, this result does not exclude a non-random distribution of longer (sub)motifs in the steady-state. To rule out correlations on larger scales, we analyze the distributions of (sub)motifs of lengths $n > 2$ in detail. Therefore, we compare the (sub)motif distributions $\mathcal{P}_n$ from the simulated model dynamics to the (sub)motif distributions $\mathcal{R}_n$ obtained from a true random process. To this end, we count the number of (sub)motifs of size n contained in all strand with $L \geq n$ from the simulation output at every time point. (A strand of length $L \geq n$ contains $L - n + 1$ (sub)motifs of size n.) For every time point, we also generate an equal number of random motifs of size n. We then analyze the evolution of $\mathcal{P}_n$ and $\mathcal{R}_n$ by means of the relative entropy D_n (also called Kullback–Leibler divergence) with respect to the uniform distribution $\mathcal{U}_n$ [144].

For the distributions $\mathcal{P}_n$ and $\mathcal{U}_n$, the relative entropy $D_n(\mathcal{P}_n,\mathcal{U}_n)$ is given by

$$D_n(\mathcal{P}_n,\mathcal{U}_n) = \sum_{m \in M_n} \mathcal{P}_n(m) \, \log_2\left[\frac{\mathcal{P}_n(m)}{\mathcal{U}_n(m)}\right], \tag{A27}$$

where the sum runs over all possible motifs $m \in M_n$ of the given length n. Using that,

$$\mathcal{U}_n(m) = \frac{1}{2^m}, \tag{A28}$$

Equation (A27) simplifies to

$$D_n(\mathcal{P}_n,\mathcal{U}_n) = m + \sum_{m \in M_n} \mathcal{P}_n(m) \, \log_2[\mathcal{P}_n(m)]. \tag{A29}$$

The relative entropy measures how much the distribution $\mathcal{P}_n$ deviates from the distribution $\mathcal{U}_n$. The smaller $D_n(\mathcal{P}_n,\mathcal{U}_n)$, the more similar are $\mathcal{P}_n$ and $\mathcal{U}_n$. Since $\mathcal{U}_n$ is the uniform distribution, i.e., the distribution with the largest entropy, we have that $D_n(\mathcal{P}_n,\mathcal{U}_n) \geq 0$. Only a few long strands exist at early times and the occupation numbers for most motifs are zero. For this reason, we can not directly compare $\mathcal{P}_n$ and $\mathcal{R}_n$ by means of the relative entropy. The definition of relative entropy Equation (A27) requires that the relative frequency of any motif $m \in M_n$ is always finite for the second distribution. Therefore, $D_n(\mathcal{P}_n, \mathcal{R}_n)$ would be ill-defined as long as are not all motifs present.

Figure A2a–c show the evolution of $D_n(\mathcal{P}_n,\mathcal{U}_n)$ (green lines) and $D_n(\mathcal{R}_n,\mathcal{U}_n)$ (blue lines) for $n = 4, 6, 8$. The high values of the relative entropies at early times stem from small numbers of (sub)motifs. As the number of (sub) motifs grows with time, $\mathcal{P}_n$ and $\mathcal{R}_n$ become more similar to the uniform distribution, and the relative entropies decay. However, the decay of $D_n(\mathcal{P}_n,\mathcal{U}_n)$ is slower, indicating that sequence correlations exist during early strand growth. Eventually, $D_n(\mathcal{P}_n,\mathcal{U}_n)$ and $D_n(\mathcal{R}_n,\mathcal{U}_n)$ converge to the same stationary value. This result implies that the motif distribution resulting from the model dynamics is entirely random. The insets in Figure A2 show the relative frequencies f_n of the motifs sorted by abundance for the last time point. Simulation output and random process give similar results:

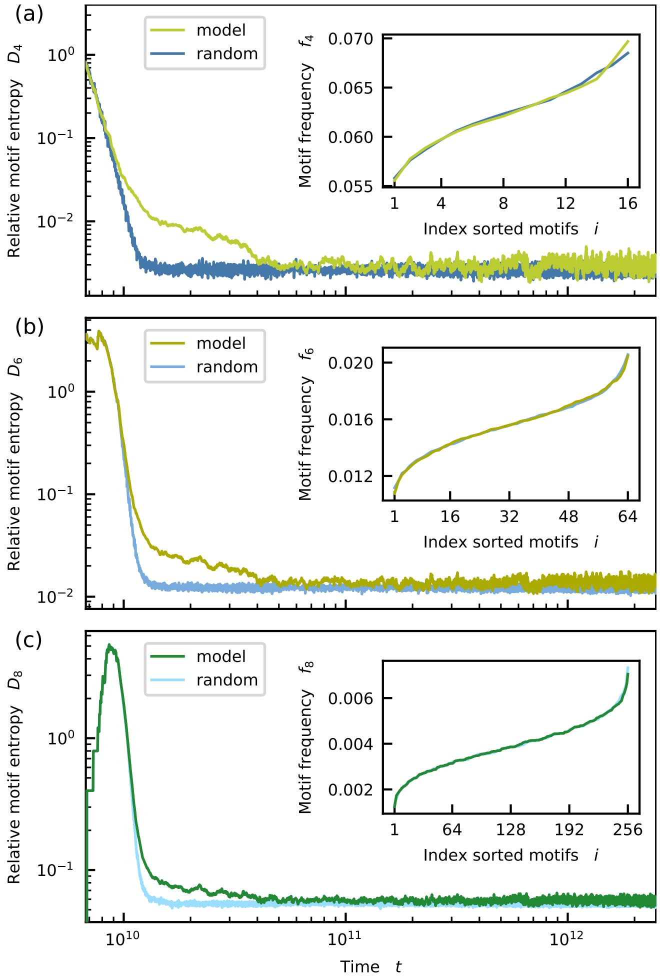

Figure A2. Relative entropies of the distributions of (sub)motifs of size four (**a**); six (**b**); and eight (**c**) as a function of time. Green curves: data obtained from simulations of the full model dynamics. Blue curves: data generated by the corresponding random process. At every time point, the number of (sub) motifs generated by the random process equals the number of (sub)motifs found in the simulation output. For small times, correlations in sequence lead to an increased relative entropy in the model dynamics. For large times, model dynamics and random processes yield similar results. The insets show the normalized frequency of (sub)motifs sorted by abundance in for the last time point.

Appendix E. Onset of Growth

In Sections 3.3–3.8, we have seen that the mean length $\overline{L}$ grows rapidly once the first new oligomers are formed from existing mononucletoides and dimers. Plotting the data with a logarithmic y-axis reveals that the growth dynamics is approximately exponential (see Figure S1). Moreover, increasing the strength of kinetic stalling (the stacking bias) shifts the onset of the rapid growth phase to later (earlier) times. The goal of this section is to derive an approximative formula for the onset of growth $\widehat{t}$.

To this end, we first focus on the scenario with $\Delta\gamma = 0$ and $\sigma_1 = \sigma_2 = 1$ discussed in Section 3.3 (see Figure 3a). Formally, we define $\widehat{t}$ as the abscissa of the intersection of the tangents to $\overline{L}(t)$ at $t = 0$ and the point where the growth is strongest (see dotted lines in Figure A3a).

To gain a microscopic picture of the growth dynamics, we look at the statistics of ligation events over time and distinguish between *first-order* and *higher-order ligations*. First-order ligations correspond to reactions where two monomers ligate to a dimer on a template of arbitrary length $L \geq 2$. In contrast, higher-order ligations involve at most one monomer and lead to new strands with $L \geq 3$. (We can further differentiate higher-order ligations into *second* and *third-order* ligations depending on whether the reaction comprises one or no monomer). Figure A3b reveals that $\widehat{t}$ coincides approximately with the time point t_{high} where higher-order ligations become more frequent than first-order ligations, i.e., for $t = t_{\mathrm{high}}$ such that we have

$$r_h(t) = r_f(t) \tag{A30}$$

where $r_h(t) = \frac{N_h(t)}{\nu \Delta t}$ and $r_f(t) = \frac{N_f(t)}{\nu \Delta t}$ are the numbers N_f and N_h of first- and higher-order ligations per time interval $\Delta t = 2.5 \times 10^8\, t_0$. Moreover, ν is a normalization constant. (In fact, Figure A3b reveals that $\widehat{t}$ corresponds precisely to the time point, where third-order reactions become more abundant than first-order ligations).

To obtain an analytic estimate t_{est} for t_{high}, we consider the dynamical mean-field rate-equation for the evolution of the dimer concentration $c_2(t)$. This equation neglects the explicit sequence dependence and coarse-grains (de)hybridization and ligation into one effective *extension*. It is given by

$$\begin{aligned} \dot{c_2} = {} & (c_1)^2 c_2 k_{\{1,1|2\}} + (c_1)^2 \sum_{L \geq 3} c_L k_{\{1,1|L\}} \\ & - c_1 (c_2)^2 k_{\{1,2|2\}} - c_2 \sum_{L_1 \geq 2} \sum_{L_2 \geq 2} c_{L_1} c_{L_2} k_{\{2,L_1|L_2\}} \\ & - c_2 k_{\mathrm{cut}} + \sum_{L \geq 3} 2 c_L k_{\mathrm{cut}}. \end{aligned} \tag{A31}$$

The first term on the right-hand side of Equation (A31) describes the creation of a new dimer via a first-order ligation with a dimer serving as the template. $k_{\{1,1|2\}}$ is the rate constant for this effective extension process. We explain it below, together with the rate constants for the other extension processes. The second term also relates to the formation of a new dimer but using a template with $L \geq 3$. The reaction is again a first-order ligation. The third term is a loss term accounting for the ligation of a monomer to a dimer, on a dimeric template. This reaction is a higher-order ligation. The fourth term is again a loss term describing a higher-order ligation of a dimer to a strand with $L_1 \geq 2$ on a template with $L_2 \geq 2$. The second to last term accounts for the loss of dimers due to cleavage with rate constant k_{cut}. The last term represents the gain of dimers due to cleavage of strands of length $L \geq 3$. There, a dimer can break apart at either side of the longer strand. Equation (A31) is valid if (1) concentrations are small enough such that the total strand concentration is approximately equivalent to the concentration of single strands, (2) complexes composed of more than three strands are negligible and, (3) the time scales for ligation and dehybridization are separated such that $k_{\mathrm{lig}} \ll k_{\mathrm{off}}$. The assumptions (1)–(3) are satisfied (see Figure S24 and Table 2).

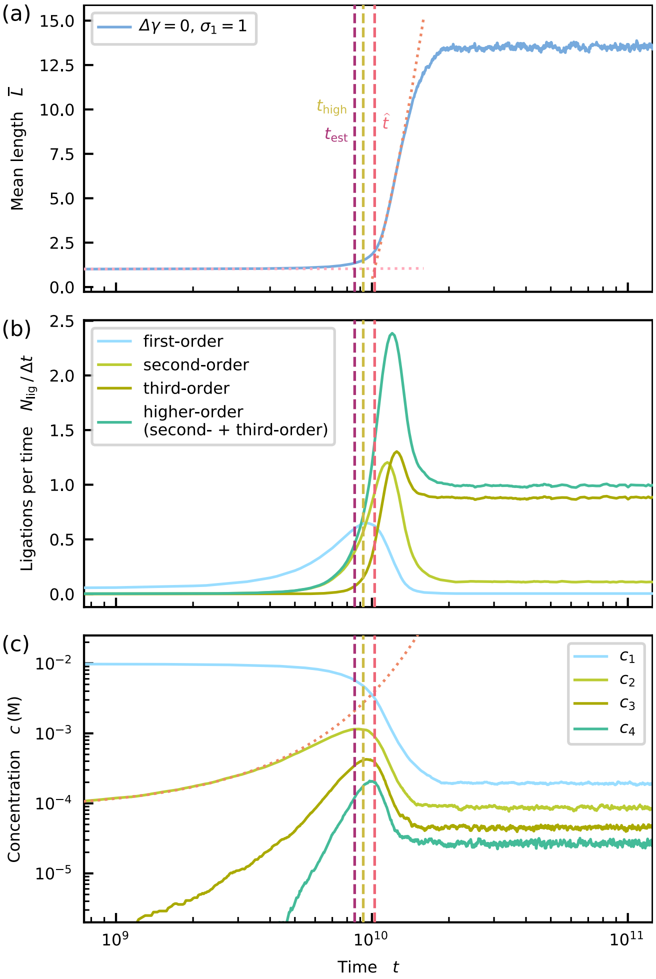

Figure A3. (**a**) We formally define the onset of growth $\hat{t}$ (dashed line) by intersecting the tangents to $\overline{L}(t)$ at $t = 0$ and the point where the increase is steepest (dotted lines). (**b**) t_{high} is defined as the time point where higher-order ligations become more frequent than first-order ligations. Our estimate t_{est} obtained from Equation (A39) matches t_{high} well (compare dashed lines). Curves are normalized such that, on average, there is one ligation per time interval in the steady-state. (**c**) The initial exponential growth of the dimer concentration is described by Equation (A37) (dotted line). The dimer concentration has a maximum at t_{high}.

We derive the effective rate constant for an extension $k_{\{1,1|2\}}$ as follows. First, we compute the average Boltzmann factor over the set of all complexes comprising exactly two monomers and one dimer $\mathcal{C}_{\{1,1|2\}}$. The Boltzmann factor associated with a specific complex $C \in \mathcal{C}_{\{1,1|2\}}$ is determined by its total binding energy $\Delta\mathcal{G}_{\text{tot}}(C)$ (see Appendix C). The sequence-averaged Boltzmann factor then defines the sequence-averaged dissociation constant $K_{\{1,1|2\}}$ via

$$\frac{1}{K_{\{1,1|2\}}} = \sum_{C \in \mathcal{C}_{\{1,1|2\}}} \frac{e^{-\beta\Delta\mathcal{G}_{\text{tot}}(C)}}{\left|\mathcal{C}_{\{1,1|2\}}\right|}. \tag{A32}$$

Recall that all concentrations are expressed as a multiple of the reference concentration $c_{\circ} = 1\,\text{mol/L}$. For that reason, the dissociation constant appears as a dimensionless quantity in Equation (A32). For complexes formed of exactly one dimer and two monomers, the total binding energy reduces to stacking energy associated with one nearest neighbor block such that

$$\begin{aligned}\frac{1}{K_{\{1,1|2\}}} = \frac{1}{16}\Bigg[&\exp\left\{-\gamma\left(\begin{bmatrix} X & X \\ & \\ X- & X \end{bmatrix}\right)\right\} + \exp\left\{-\gamma\left(\begin{bmatrix} Y & X \\ \cdot & \\ X- & X \end{bmatrix}\right)\right\} + \\ &\exp\left\{-\gamma\left(\begin{bmatrix} X & Y \\ & \cdot \\ X- & X \end{bmatrix}\right)\right\} + \exp\left\{-\gamma\left(\begin{bmatrix} Y & Y \\ \cdot & \cdot \\ X- & X \end{bmatrix}\right)\right\} + \dots\Bigg].\end{aligned} \tag{A33}$$

This equation further reduces to

$$\frac{1}{K_{\{1,1|2\}}} = \frac{1}{4}\left[e^{-\gamma_{2\mathrm{nc}}} + 2e^{-\gamma_{1\mathrm{nc}}} + e^{-\overline{\gamma_{\mathrm{com}}}}\right]. \tag{A34}$$

Second, we multiply the inverse of the average dissociation constant with the ligation rate k_{lig}. The result is

$$k_{\{1,1|2\}} = \frac{k_{\mathrm{lig}}}{K_{\{1,1|2\}}}. \tag{A35}$$

Coarse-gaining (de)hybridization and ligation into an effective extension this way is valid as long as the ligation timescale is much slower than the dehybridization time scales such that there is enough time for the (de)hybridization dynamics to equilibrate. By our choice of the ligation rate (see Section 2.6), this premise is clearly fulfilled for monomers and dimers.

The rate constant $k_{\{1,2|2\}}$ is obtained analogously. The other rate constants are less trivial. However, we will neglect them later on anyway.

For $t < t_{\mathrm{high}}$, mostly monomers and dimers populate the pool. We, therefore, assume that t_{high} roughly matches the time point where the loss terms related to higher-order ligations balance the gain terms in Equation (A31). Equation (A31) has no analytic solution. We thus have to make (crude) approximations to obtain the estimate t_{est} for t_{high}. First, we neglect all terms that involve strands of length $L > 2$ or do not include at least one monomer. The resulting simplified equation then reads:

$$\dot{c_2} = (c_1)^2 c_2 k_{\{1,1|2\}} - c_1 (c_2)^2 k_{\{1,2|2\}} - c_2 k_{\mathrm{cut}}. \tag{A36}$$

Equation (A36) is a cubic ordinary differential equation since $c_1 + 2c_2$ is a constant. Hence, the simplified equation still does not allow for a closed analytic solution for $c_2(t)$, and further (crude) approximations will be necessary.

For $t \to 0$, the dimer concentration grows exponentially (see the dotted line in Figure A3), i.e.,

$$c_2(t) \approx c_2(0)\exp\left[\left([c_1(0)]^2 k_{\{1,1|2\}} - k_{\mathrm{cut}}\right)t\right]. \tag{A37}$$

We now assume the monomer concentration to remain constant, i.e., $c_1(t) = c_1(0) \approx c_{\mathrm{tot}}$. For $t \to \infty$, the dimer concentration then approaches a stationary state concentration $\widetilde{c}_2$, which is given by

$$\widetilde{c}_2 = \frac{c_{\mathrm{tot}}^2 k_{\{1,2|2\}} - k_{\mathrm{cut}}}{c_{\mathrm{tot}} k_{\{1,1|2\}}}. \tag{A38}$$

with that, we estimate the time point for which the right-hand side of Equation (A36) vanishes by equating Equations (A37) and (A38). We obtain

$$t_{\mathrm{est}} = \frac{\log\left[c_{\mathrm{tot}}^2 k_{\{1,1|2\}} - k_{\mathrm{cut}}\right] - \log\left[c_{\mathrm{tot}} c_2(0) k_{\{1,2|2\}}\right]}{c_{\mathrm{tot}}^2 k_{\{1,1|2\}} - k_{\mathrm{cut}}}. \tag{A39}$$

t_{est} from Equation (A39) is an estimate for the time point t_{high} where higher-order ligations become more frequent than first-order ligations. Comparing this estimate to the exact value extracted from simulation data in Figure A3b, we see that t_{est} is only slightly smaller than t_{high}. Hence, Equation (A39) yields a solid estimate for the transition from the first-order to the higher-order regime. Moreover, Figure A3a shows that t_{est} matches the $\widehat{t}$ with less then 10% error. We conclude that Equation (A39) is a useful approximation for the onset of the rapid growth. In addition, t_{est} corresponds well to the time point where the dimer concentration c_2 reaches a maximum (see Figure A3c) underlining the validity of our initial assumptions. In Figure S3, we show that the prediction Equation (A39) is robust under parameter variations.

We now turn to the more general case involving energetic bias $\Delta\gamma < 0$ and kinetic stalling and $\sigma_1 = \sigma_2 < 1$. The energetic bias is automatically accounted for by averaging over all complex configurations. To include kinetic stalling, we introduce the effective sequence-averaged dissociation constant $\widetilde{K}_{\{1,1|2\}}$ analogous to Equation (A32) as

$$\frac{1}{\widetilde{K}_{\{1,1|2\}}} = \sum_{C \in \mathcal{C}_{\{1,1|2\}}} \left[\frac{e^{-\Delta \mathcal{G}_{\text{tot}}(C)}}{\left|\mathcal{C}_{\{1,1|2\}}\right|} \mathcal{S}(C) \right], \tag{A40}$$

where we weight every term in the sum on the right-hand side with the *overall effective stalling* factor

$$\mathcal{S}(C) = \Phi_-(\kappa_{-1}(C), 1)\Phi_+(\kappa_{+1}(C), 1). \tag{A41}$$

$\mathcal{S}(C)$ considers both the + and the − monomer at the ligation site (see Equation (15)). An analogous definition holds for $\widetilde{K}_{\{1,2|2\}}$. To compute $k_{\{1,1|2\}}$ and $k_{\{1,2|2\}}$ for the rate equation Equation (A36), we now use $\widetilde{K}_{\{1,1|2\}}$ and $\widetilde{K}_{\{1,2|2\}}$ (see Equation (A35)). Figures S5–S12 and S18–S23 show that the generalized approach also yields a reasonable estimate for the onset of growth.

Appendix F. Application to Primer Extension

This section investigates a typical *primer extension* scenario, where a *primer* bound to one longer *template* becomes extended stepwise in the absence of thermal cycling. We assume that the primer-template complex is stable, i.e., does not dissociate, and that the solution surrounding this complex only contains mononucleotides. The first assumption implies that the primer is long enough, such that the dehybridization timescale is much larger than the (effective) extension timescale (see Section 2.6). Moreover, we focus on ligations involving the (partially extended) primer and neglect the formation of new dimers from mononucleotides on the template. It is well known experimentally that non-complementary nucleotides at the primer's end slow down the extension process and trigger the accumulation of misincorporation. These effects stem from the interplay of two contributions. Kinetic stalling reduces the bare ligation rate. In addition, mismatches at the primer's end weaken the monomer binding, increase its dehybridization rate, and render the next extension less probable. Moreover, non-complementarities at the primer's end reduce the thermodynamic discrimination of a hybridized monomers resulting in an increased fraction of misincorporations.

To develop a quantitative description based on our model for (de)hybridization and ligation (see Section 2), we compare the two situations sketched in Figure A4. In Figure A4a, no mismatches occur, and the extension (hybridization and subsequent ligation) proceeds in an unperturbed way. In Figure A4b, the primer terminates with a mismatch. According to our model energy model (see Section 2.4 and Appendix A), the hybridization site energy assigned to the mononucleotide in Figure A4b is less negative than in Figure A4a. The single nucleotide in Figure A4b will therefore unbind faster. Hence, primer extension becomes less likely. We call this effect *thermodynamic stalling*. In addition, the bare ligation rate is multiplied by a factor $\Phi_-(\kappa_{-2} = 1, \kappa_{-1} = 0) = \sigma_1 \leq 1$ in the situation of Figure A4b as described in Section 2.5. Therefore, primer extension becomes even more unlikely. In the main text, we referred to this contribution as *kinetic stalling*. Note that, even in the scenario without kinetic stalling ($\sigma_1 = \sigma_2 = 1$) in Section 3.4, thermodynamic stalling due to enhanced dehybridization rates is always present.

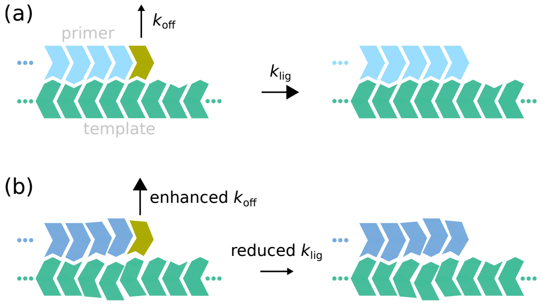

Figure A4. A monomer hybridized adjacent to matching primer terminus (**a**) has a lower dehybridization rate than a monomer bound next to a non-complementary terminus (**b**), leading to thermodynamic stalling. Moreover, kinetic stalling reduces the ligation rate.

We now quantify the thermodynamic stalling contribution. To this end, we compare the dehybridization rates $k_{\text{off}}^{(c)}$ and $k_{\text{off}}^{(n)}$ of the mononucleotides for the complementary and non-complementary primer termini. For simplicity, we assume an energetically unbiased scenario where $\Delta\gamma = \delta_\epsilon = 0$. The dehybridization rates depend exponentially on the hybridization energies $\Gamma^{(c)}$ and $\Gamma^{(n)}$. Applying the procedure to compute hybridization energies described in Appendix B, $\Gamma^{(c)}$ and $\Gamma^{(n)}$ are given by

$$\Gamma^{(c)} = \overline{\gamma_{\text{com}}} + \overline{\epsilon_{\text{com}}} - \overline{\epsilon_{\text{com}}} = \overline{\gamma_{\text{com}}}, \tag{A42}$$

$$\Gamma^{(n)} = \gamma_{\text{1nc}} + \overline{\epsilon_{\text{com}}} - \epsilon_{\text{1nc}}. \tag{A43}$$

The ratio of $k_{\text{off}}^{(c)}$ and $k_{\text{off}}^{(n)}$ now defines defines the thermodynamic stalling effect ϑ, i.e.,

$$\vartheta = \frac{k_{\text{off}}^{(c)}}{k_{\text{off}}^{(n)}} = \exp\left[\Gamma^{(c)} - \Gamma^{(n)}\right] \tag{A44}$$

$$= \exp[\overline{\gamma_{\text{com}}} + \epsilon_{\text{1nc}} - (\overline{\epsilon_{\text{com}}} + \gamma_{\text{1nc}})] \tag{A45}$$

Note that channel factors do not play a role here since they cancel out.

Multiplying the kinetic and the thermodynamic stalling factors yields the *combined stalling* factor. For the stacking parameters used in the main text, we obtain $\vartheta \approx 0.5$. For $\sigma_1 = 0.1$, the combined stalling factor has a value of 0.05. This result aligns with overall stalling factors ranging between 0.1 and 0.003 measured in non-enzymatic primer extension experiments [32,38,39].

Primer extension experiments typically also consider the error fraction ω, which is defined as the ratio of the rate for an erroneous extension and the overall extension rate. We now quantify the error fraction that results from our model. To this end, we introduce the dehybridization rates $k_{\text{off,r}}^{(c)}$ and $k_{\text{off,w}}^{(c)}$ for *right* and *wrong* mononucleotides hybridized adjacent to a complementary primer terminus. Assuming that monomer concentrations are sufficiently low such that there is no competition for the extension site and that equal amounts of both nucleotide types are present, the error fraction is

$$\omega^{(c)} = \left(1 + \frac{k_{\text{off,w}}^{(c)}}{k_{\text{off,r}}^{(c)}\,\sigma_1}\right)^{-1}. \tag{A46}$$

Relating the hybridization rates to the hybridization energies as before, Equation (A46) becomes

$$\omega^{(c)} = \left(1 + \frac{1}{\sigma_1}\exp[\gamma_{\text{1nc}} + \epsilon_{\text{1nc}} - (\overline{\gamma_{\text{com}}} + \overline{\epsilon_{\text{com}}})]\right)^{-1}. \tag{A47}$$

Using the same stacking parameters as above and $\sigma_1 = 0.1$, we obtain $\omega^{(n)} = 0.72\%$. This result agrees with the experimentally observed error fraction of 0.8% in a binary DNA system containing C and G monomers [38].

The error fraction $\omega^{(n)}$ for the extension of a primer ending with mismatch is derived analogously and reads

$$\omega^{(c)} = \left(1 + \frac{1}{\sigma_1} \exp[\gamma_{2\mathrm{nc}} + \epsilon_{1\mathrm{nc}} - (\gamma_{1\mathrm{nc}} + \overline{\epsilon_{\mathrm{com}}})]\right)^{-1}. \tag{A48}$$

For $\sigma_1 = 0.1$, we find that $\omega^{(n)} \approx 3\omega^{(c)}$. This observation is consistent with the experimental finding that non-complementarities at the primer terminus significantly increase the error fraction in the subsequent (stalled) extension process [32].

References

1. Doudna, J.A.; Szostak, J.W. RNA-catalysed synthesis of complementary-strand RNA. *Nature* **1989**, *339*, 519–522. [CrossRef] [PubMed]
2. Lorsch, J.R.; Szostak, J.W. In vitro evolution of new ribozymes with polynucleotide kinase activity. *Nature* **1994**, *371*, 31–36. [CrossRef] [PubMed]
3. Johnston, W.K. RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension. *Science* **2001**, *292*, 1319–1325. [CrossRef] [PubMed]
4. Wochner, A.; Attwater, J.; Coulson, A.; Holliger, P. Ribozyme-Catalyzed Transcription of an Active Ribozyme. *Science* **2011**, *332*, 209–212. [CrossRef]
5. Attwater, J.; Raguram, A.; Morgunov, A.S.; Gianni, E.; Holliger, P. Ribozyme-catalysed RNA synthesis using triplet building blocks. *eLife* **2018**, *7*, e35255. [CrossRef]
6. Gesteland, R.F.; Cech, T.; Atkins, J.F. (Eds.) *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*, 3rd ed.; Number 43 in Cold Spring Harbor Monograph Series; OCLC: ocm60856160 ; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2006.
7. Orgel, L.E. Prebiotic Chemistry and the Origin of the RNA World. *Crit. Rev. Biochem. Mol. Biol.* **2004**, *39*, 99–123. [CrossRef]
8. Joyce, G.F. RNA evolution and the origins of life. *Nature* **1989**, *338*, 217–224. [CrossRef]
9. Crick, F. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379. [CrossRef]
10. Orgel, L.E. Evolution of the genetic apparatus. *J. Mol. Biol.* **1968**, *38*, 381–393. [CrossRef]
11. Gilbert, W. Origin of life: The RNA world. *Nature* **1986**, *319*, 618. [CrossRef]
12. Szostak, J.W. The Narrow Road to the Deep Past: In Search of the Chemistry of the Origin of Life. *Angew. Chem. Int. Ed.* **2017**, *56*, 11037–11043. [CrossRef] [PubMed]
13. Powner, M.W.; Gerland, B.; Sutherland, J.D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **2009**, *459*, 239–242. [CrossRef] [PubMed]
14. Xu, J.; Chmela, V.; Green, N.; Russell, D.; Janicki, M.; Góra, R.; Szabla, R.; Bond, A.; Sutherland, J. Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides. *Nature* **2020**, *582*, 60–66. [CrossRef] [PubMed]
15. Becker, S.; Thoma, I.; Deutsch, A.; Gehrke, T.; Mayer, P.; Zipse, H.; Carell, T. A high-yielding, strictly regioselective prebiotic purine nucleoside formation pathway. *Science* **2016**, *352*, 833–836. [CrossRef] [PubMed]
16. Mutschler, H.; Wochner, A.; Holliger, P. Freeze–thaw cycles as drivers of complex ribozyme assembly. *Nat. Chem.* **2015**, *7*, 502–508. [CrossRef]
17. Briones, C.; Stich, M.; Manrubia, S.C. The dawn of the RNA World: Toward functional complexity through ligation of random RNA oligomers. *RNA* **2009**, *15*, 743–749. [CrossRef]
18. Toyabe, S.; Braun, D. Cooperative Ligation Breaks Sequence Symmetry and Stabilizes Early Molecular Replication. *Phys. Rev. X* **2019**, *9*, 011056. [CrossRef]
19. Salditt, A.; Keil, L.M.; Horning, D.; Mast, C.; Joyce, G.; Braun, D. Thermal Habitat for RNA Amplification and Accumulation. *Phys. Rev. Lett.* **2020**, *125*, 048104. [CrossRef]
20. Edeleva, E.; Salditt, A.; Stamp, J.; Schwintek, P.; Boekhoven, J.; Braun, D. Continuous nonenzymatic cross-replication of DNA strands with in situ activated DNA Oligonucleotides.*Chem. Sci.* **2019**, *10*, 5807–5814. [CrossRef]
21. Wachowius, F.; Holliger, P. Non-Enzymatic Assembly of a Minimized RNA Polymerase Ribozyme. *ChemSystemsChem* **2019**, *1*, 12–15. [CrossRef]
22. Ferre-D'Amare, A.R.; Scott, W.G. Small Self-cleaving Ribozymes. *Cold Spring Harb. Perspect. Biol.* **2010**, *2*, a003574. [CrossRef]
23. Birikh, K.R.; Heaton, P.A.; Eckstein, F. The Structure, Function and Application of the Hammerhead Ribozyme. *Eur. J. Biochem.* **1997**, *245*, 1–16. [CrossRef] [PubMed]
24. Scott, W.G.; Murray, J.B.; Arnold, J.R.P.; Stoddard, B.L.; Klug, A. Capturing the Structure of a Catalytic RNA Intermediate: The Hammerhead Ribozyme. *Science* **1996**, *274*, 2065–2069. [CrossRef] [PubMed]

25. Horning, D.P.; Joyce, G.F. Amplification of RNA by an RNA polymerase ribozyme. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 9786–9791. [CrossRef]
26. Doudna, J.A.; Couture, S.; Szostak, J.W. A Multisubunit Ribozyme That Is a Catalyst of and Template for Complementary Strand RNA synthesis. *Science* **1991**, *251*, 1605–1608. [CrossRef]
27. Joyce, G.F. Directed Evolution of Nucleic Acid Enzymes. *Annu. Rev. Biochem.* **2004**, *73*, 791–836. [CrossRef] [PubMed]
28. Walker, S.I. Origins of life: A problem for physics, a key issues review. *Rep. Prog. Phys.* **2017**, *80*, 092601. [CrossRef]
29. Ameta, S.; Matsubara, Y.J.; Chakraborty, N.; Krishna, S.; Thutupalli, S. Self-Reproduction and Darwinian Evolution in Autocatalytic Chemical Reaction Systems. *Life* **2021**, *11*, 308. [CrossRef]
30. Sievers, D.; von Kiedrowski, G. Self-replication of complementary nucleotide-based oligomers. *Nature* **1994**, *369*, 221–224. [CrossRef]
31. Derr, J.; Manapat, M.L.; Rajamani, S.; Leu, K.; Xulvi-Brunet, R.; Joseph, I.; Nowak, M.A.; Chen, I.A. Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences. *Nucleic Acids Res.* **2012**, *40*, 4711–4722. [CrossRef]
32. Leu, K.; Kervio, E.; Obermayer, B.; Turk-MacLeod, R.M.; Yuan, C.; Luevano, J.M.; Chen, E.; Gerland, U.; Richert, C.; Chen, I.A. Cascade of Reduced Speed and Accuracy after Errors in Enzyme-Free Copying of Nucleic Acid Sequences. *J. Am. Chem. Soc.* **2013**, *135*, 354–366. [CrossRef] [PubMed]
33. Manapat, M.L.; Chen, I.A.; Nowak, M.A. The basic reproductive ratio of life. *J. Theor. Biol.* **2010**, *263*, 317–327. [CrossRef] [PubMed]
34. Kanavarioti, A.; White, D.H. Kinetic analysis of the template effect in ribooligoguanylate elongation. *Orig. Life Evol. Biosph.* **1987**, *17*, 333–349. [CrossRef] [PubMed]
35. Yakovchuk, P. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucl. Acids Res.* **2006**, *34*, 564–574. [CrossRef] [PubMed]
36. SantaLucia, J., Jr.; Hicks, D. The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 415–440. [CrossRef]
37. Turner, D.H.; Mathews, D.H. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* **2010**, *38*, D280–D282. [CrossRef]
38. Rajamani, S.; Ichida, J.K.; Antal, T.; Treco, D.A.; Leu, K.; Nowak, M.A.; Szostak, J.W.; Chen, I.A. Effect of Stalling after Mismatches on the Error Catastrophe in Nonenzymatic Nucleic Acid Replication. *J. Am. Chem. Soc.* **2010**, *132*, 5880–5885. [CrossRef]
39. Leu, K.; Obermayer, B.; Rajamani, S.; Gerland, U.; Chen, I.A. The prebiotic evolutionary advantage of transferring genetic information from RNA to DNA. *Nucleic Acids Res.* **2011**, *39*, 8135–8147. [CrossRef]
40. Blain, J.C.; Szostak, J.W. Progress toward synthetic cells. *Annu. Rev. Biochem.* **2014**, *83*, 615–640. [CrossRef]
41. Szostak, J.W. The eightfold path to non-enzymatic RNA replication. *J. Syst. Chem.* **2012**, *3*, 2. [CrossRef]
42. Sosson, M.; Richert, C. Enzyme-free genetic copying of DNA and RNA sequences. *Beilstein J. Org. Chem.* **2018**, *14*, 603–617. [CrossRef] [PubMed]
43. Zhou, L.; O'Flaherty, D.K.; Szostak, J.W. Assembly of a Ribozyme Ligase from Short Oligomers by Nonenzymatic Ligation. *J. Am. Chem. Soc.* **2020**, *142*, 15961–15965. [CrossRef] [PubMed]
44. Zhou, L.; O'Flaherty, D.K.; Szostak, J.W. Template-Directed Copying of RNA by Non-enzymatic Ligation. *Angew. Chem. Int. Ed.* **2020**, *132*, 15812–15817. [CrossRef]
45. Zielinski, W.S.; Orgel, L.E. Autocatalytic synthesis of a tetranucleotide analogue. *Nature* **1987**, *327*, 346–347. [CrossRef]
46. Zielinski, W.S.; Orgel, L.E. Oligoaminudeoside phosphoramidates. Oligomeilzation of dimers of 3′-amino-3′-deoxy-nucleotides (GC and CG) in aqueous solution. *Nucleic Acids Res.* **1987**, *15*, 1699–1715. [CrossRef]
47. von Kiedrowski, G. A Self-Replicating Hexadeoxynucleotide. *Angew. Chem. Int. Ed.* **1986**, *25*, 932–935. [CrossRef]
48. Sosson, M.; Pfeffer, D.; Richert, C. Enzyme-free ligation of dimers and trimers to RNA primers. *Nucleic Acids Res.* **2019**, *47*, 3836–3845. [CrossRef]
49. Hänle, E.; Richert, C. Enzyme-Free Replication with Two or Four Bases. *Angew. Chem. Int. Ed.* **2018**, *57*, 8911–8915. [CrossRef]
50. Deck, C.; Jauker, M.; Richert, C. Efficient enzyme-free copying of all four nucleobases templated by immobilized RNA. *Nat. Chem.* **2011**, *3*, 603–608. [CrossRef]
51. Jauker, M.; Griesser, H.; Richert, C. Copying of RNA Sequences without Pre-Activation. *Angew. Chem. Int. Ed.* **2015**, *54*, 14559–14563. [CrossRef]
52. Kervio, E.; Hochgesand, A.; Steiner, U.E.; Richert, C. Templating efficiency of naked DNA. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 12074–12079. [CrossRef] [PubMed]
53. Prywes, N.; Blain, J.C.; Del Frate, F.; Szostak, J.W. Nonenzymatic copying of RNA templates containing all four letters is catalyzed by activated oligonucleotides. *eLife* **2016**, *5*, e17756. [CrossRef] [PubMed]
54. Li, L.; Prywes, N.; Tam, C.P.; O'Flaherty, D.K.; Lelyveld, V.S.; Izgu, E.C.; Pal, A.; Szostak, J.W. Enhanced Nonenzymatic RNA Copying with 2-Aminoimidazole Activated Nucleotides. *J. Am. Chem. Soc.* **2017**, *139*, 1810–1813. [CrossRef]
55. Kudella, P.W.; Tkachenko, A.V.; Salditt, A.; Maslov, S.; Braun, D. Structured sequences emerge from random pool when replicated by templated ligation. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2018830118. [CrossRef] [PubMed]
56. Rosenberger, J.H.; Göppel, T.; Kudella, P.W.; Braun, D.; Gerland, U.; Altaner, B. Self-Assembly of Informational Polymers by Templated Ligation. *Phys. Rev. X* **2021**, *11*, 031055. [CrossRef]

57. Dill, K.A.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*, 2nd ed.; Garland Science: London, UK; New York, NY, USA, 2011.
58. Kervio, E.; Sosson, M.; Richert, C. The effect of leaving groups on binding and reactivity in enzyme-free copying of DNA and RNA. *Nucleic Acids Res.* **2016**, *44*, 5504–5514. [CrossRef]
59. Walton, T.; Szostak, J.W. A Kinetic Model of Nonenzymatic RNA Polymerization by Cytidine-5′-phosphoro-2-aminoimidazolide. *Biochemistry* **2017**, *56*, 5739–5747. [CrossRef]
60. Schroeder, G.K.; Lad, C.; Wyman, P.; Williams, N.H.; Wolfenden, R. The time required for water attack at the phosphorus atom of simple phosphodiesters and of DNA. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 4052–4055. [CrossRef]
61. Zhou, L.; Ding, D.; Szostak, J.W. The virtual circular genome model for primordial RNA replication. *RNA* **2021**, *27*, 1–11. [CrossRef]
62. Lindahl, T.; Andersson, A. Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. *Biochemistry* **1972**, *11*, 3618–3623. [CrossRef]
63. Li, Y.; Breaker, R.R. Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2′-Hydroxyl Group. *J. Am. Chem. Soc.* **1999**, *121*, 5364–5372. [CrossRef]
64. Komiyama, M.; Takeda, N.; Shigekawa, H. Hydrolysis of DNA and RNA by lanthanide ions: Mechanistic studies leading to new applications. *Chem. Commun.* **1999**, *16*, 1443–1451. [CrossRef]
65. Basile, L.A.; Raphael, A.L.; Barton, J.K. Metal-activated hydrolytic cleavage of DNA. *J. Am. Chem. Soc.* **1987**, *109*, 7550–7551. [CrossRef]
66. Hopfield, J.J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proc. Natl. Acad. Sci. USA* **1974**, *71*, 4135–4139. [CrossRef] [PubMed]
67. Rauzan, B.; McMichael, E.; Cave, R.; Sevcik, L.R.; Ostrosky, K.; Whitman, E.; Stegemann, R.; Sinclair, A.L.; Serra, M.J.; Deckert, A.A. Kinetics and Thermodynamics of DNA, RNA, and Hybrid Duplex Formation. *Biochemistry* **2013**, *52*, 765–772. [CrossRef]
68. Ouldridge, T.E. The importance of thermodynamics for molecular systems, and the importance of molecular systems for thermodynamics. *Nat. Comput.* **2018**, *17*, 3–29. [CrossRef]
69. Jhunjhunwala, A.; Ali, Z.; Bhattacharya, S.; Halder, A.; Mitra, A.; Sharma, P. On the Nature of Nucleobase Stacking in RNA: A Comprehensive Survey of Its Structural Variability and a Systematic Classification of Associated Interactions. *J. Chem. Inf. Model.* **2021**, *61*, 1470–1480. [CrossRef]
70. Luther, A.; Brandsch, R.; von Kiedrowski, G. Surface-promoted replication and exponential amplification of DNA analogues. *Nature* **1998**, *396*, 245–248. [CrossRef]
71. Keil, L.M.R.; Möller, F.M.; Kieß, M.; Kudella, P.W.; Mast, C.B. Proton gradients and pH oscillations emerge from heat flow at the microscale. *Nat. Commun.* **2017**, *8*, 1897. [CrossRef]
72. Mariani, A.; Bonfio, C.; Johnson, C.M.; Sutherland, J.D. pH-Driven RNA Strand Separation under Prebiotically Plausible Conditions. *Biochemistry* **2018**, *57*, 6382–6386. [CrossRef]
73. Kreysing, M.; Keil, L.; Lanzmich, S.; Braun, D. Heat flux across an open pore enables the continuous replication and selection of oligonucleotides towards increasing length. *Nat. Chem.* **2015**, *7*, 203–208. [CrossRef] [PubMed]
74. Damer, B.; Deamer, D. The Hot Spring Hypothesis for an Origin of Life. *Astrobiology* **2020**, *20*, 429–452. [CrossRef] [PubMed]
75. Mast, C.B.; Braun, D. Thermal Trap for DNA Replication. *Phys. Rev. Lett.* **2010**, *104*, 188102. [CrossRef] [PubMed]
76. Ianeselli, A.; Mast, C.B.; Braun, D. Periodic Melting of Oligonucleotides by Oscillating Salt Concentrations Triggered by Microscale Water Cycles Inside Heated Rock Pores. *Angew. Chem. Int. Ed.* **2019**, *131*, 13289–13294. [CrossRef]
77. Tkachenko, A.V.; Maslov, S. Spontaneous emergence of autocatalytic information-coding polymers. *J. Chem. Phys.* **2015**, *143*, 045102. [CrossRef]
78. Gillespie, D.T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem* **1977**, *81*, 2340–2361. [CrossRef]
79. Gillespie, D.T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **1976**, *22*, 403–434. [CrossRef]
80. Gibson, M.A.; Bruck, J. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J. Phys. Chem. A* **2000**, *104*, 1876–1889. [CrossRef]
81. Krapivsky, P.L.; Redner, S.; Ben-Naim, E. *A Kinetic View of Statistical Physics*; CBO9780511780516; Cambridge University Press: Cambridge, UK, 2010. [CrossRef]
82. Obermayer, B.; Krammer, H.; Braun, D.; Gerland, U. Emergence of Information Transmission in a Prebiotic RNA Reactor. *Phys. Rev. Lett.* **2011**, *107*, 018101. [CrossRef]
83. Roy, S.; Bapat, N.V.; Derr, J.; Rajamani, S.; Sengupta, S. Emergence of ribozyme and tRNA-like structures from mineral-rich muddy pools on prebiotic earth. *J. Theor. Biol.* **2020**, *506*, 110446. [CrossRef]
84. Matsubara, Y.J.; Kaneko, K. Optimal size for emergence of self-replicating polymer system. *Phys. Rev. E* **2016**, *93*, 032503. [CrossRef] [PubMed]
85. Gonçalves da Silva, L.H.; Hochberg, D. Open flow non-enzymatic template catalysis and replication. *Phys. Chem. Chem. Phys.* **2018**, *20*, 14864–14875. [CrossRef]
86. Fellermann, H.; Tanaka, S.; Rasmussen, S. Sequence selection by dynamical symmetry breaking in an autocatalytic binary polymer model. *Phys. Rev. E* **2017**, *96*, 062407. [CrossRef] [PubMed]

87. Tanaka, S.; Fellermann, H.; Rasmussen, S. Structure and selection in an autocatalytic binary polymer model. *EPL* **2014**, *107*, 28004. [CrossRef]
88. Matsubara, Y.J.; Kaneko, K. Kinetic Selection of Template Polymer with Complex Sequences. *Phys. Rev. Lett.* **2018**, *121*, 118101. [CrossRef]
89. Mizuuchi, R.; Lehman, N. Limited Sequence Diversity Within a Population Supports Prebiotic RNA Reproduction. *Life* **2019**, *9*, 20. [CrossRef]
90. Tkachenko, A.V.; Maslov, S. Onset of natural selection in populations of autocatalytic heteropolymers. *J. Chem. Phys.* **2018**, *149*, 134901. [CrossRef]
91. Tupper, A.; Shi, K.; Higgs, P. The Role of Templating in the Emergence of RNA from the Prebiotic Chemical Mixture. *Life* **2017**, *7*, 41. [CrossRef]
92. Anderson, P.W. Suggested model for prebiotic evolution: The use of chaos. *Proc. Natl. Acad. Sci. USA* **1983**, *80*, 3386–3390. [CrossRef]
93. Morasch, M.; Braun, D.; Mast, C.B. Heat-Flow-Driven Oligonucleotide Gelation Separates Single-Base Differences. *Angew. Chem. Int. Ed.* **2016**, *55*, 6676–6679. [CrossRef]
94. Zhou, L.; Kim, S.C.; Ho, K.H.; O'Flaherty, D.K.; Giurgiu, C.; Wright, T.H.; Szostak, J.W. Non-enzymatic primer extension with strand displacement. *eLife* **2019**, *8*, e51888. [CrossRef] [PubMed]
95. Mutschler, H.; Taylor, A.I.; Porebski, B.T.; Lightowlers, A.; Houlihan, G.; Abramov, M.; Herdewijn, P.; Holliger, P. Random-sequence genetic oligomer pools display an innate potential for ligation and recombination. *eLife* **2018**, *7*, e43022. [CrossRef] [PubMed]
96. Tupper, A.S.; Higgs, P.G. Rolling-circle and strand-displacement mechanisms for non-enzymatic RNA replication at the time of the origin of life. *J. Theor. Biol.* **2021**, *527*, 110822. [CrossRef] [PubMed]
97. Blokhuis, A.; Lacoste, D. Length and sequence relaxation of copolymers under recombination reactions. *J. Chem. Phys.* **2017**, *147*, 094905. [CrossRef]
98. Göppel, T.; Palyulin, V.V.; Gerland, U. The efficiency of driving chemical reactions by a physical non-equilibrium is kinetically controlled. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20135–20143. [CrossRef]
99. Göppel, T.; Obermayer, B.; Chen, I.A.; Gerland, U. A kinetic error filtering mechanism for enzyme-free copying of nucleic acid sequences. *Evol. Biol.* **2021**, preprint. [CrossRef]
100. Mast, C.B.; Schink, S.; Gerland, U.; Braun, D. Escalation of polymerization in a thermal gradient. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 8030–8035. [CrossRef]
101. Walker, S.I.; Grover, M.A.; Hud, N.V. Universal Sequence Replication, Reversible Polymerization and Early Functional Biopolymers: A Model for the Initiation of Prebiotic Sequence Evolution. *PLoS ONE* **2012**, *7*, e34166. [CrossRef]
102. Nowak, M.A.; Ohtsuki, H. Prevolutionary dynamics and the origin of evolution. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 14924–14927. [CrossRef]
103. Andrieux, D.; Gaspard, P. Nonequilibrium generation of information in copolymerization processes. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 9516–9521. [CrossRef]
104. Manapat, M.; Ohtsuki, H.; Bürger, R.; Nowak, M.A. Originator dynamics. *J. Theor. Biol.* **2009**, *256*, 586–595. [CrossRef] [PubMed]
105. Wachtershauser, G. An all-purine precursor of nucleic acids. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 1134–1135. [CrossRef] [PubMed]
106. Levy, M.; Miller, S.L. The stability of the RNA bases: Implications for the origin of life. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 7933–7938. [CrossRef]
107. Rogers, J.; Joyce, G.F. A ribozyme that lacks cytidine. *Nature* **1999**, *402*, 323–325. [CrossRef] [PubMed]
108. Reader, J.S.; Joyce, G.F. A ribozyme composed of only two different nucleotides. *Nature* **2002**, *420*, 841–844. [CrossRef]
109. Schlosser, K.; Li, Y. DNAzyme-mediated catalysis with only guanosine and cytidine nucleotides. *Nucleic Acids Res.* **2009**, *37*, 413–420. [CrossRef]
110. Joyce, G.F.; Schwartz, A.W.; Miller, S.L.; Orgel, L.E. The case for an ancestral genetic system involving simple analogues of the nucleotides. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 4398–4402. [CrossRef]
111. Schoning, K.U. Chemical Etiology of Nucleic Acid Structure: The alpha -Threofuranosyl-(3'rightarrow 2') Oligonucleotide System. *Science* **2000**, *290*, 1347–1351. [CrossRef]
112. Hud, N.V. Searching for lost nucleotides of the pre-RNA World with a self-refining model of early Earth. *Nat. Commun.* **2018**, *9*, 5171. [CrossRef]
113. Hud, N.; Cafferty, B.; Krishnamurthy, R.; Williams, L. The Origin of RNA and "My Grandfather's Axe". *Chem. Biol.* **2013**, *20*, 466–474. [CrossRef]
114. Joyce, G.F. The antiquity of RNA-based evolution. *Nature* **2002**, *418*, 214–221. [CrossRef] [PubMed]
115. Orgel, L.E. Did template-directed nucleation precede molecular replication? *Orig. Life Evol. Biosph.* **1986**, *17*, 27–34. [CrossRef] [PubMed]
116. Higgs, P.G.; Lehman, N. The RNA World: Molecular cooperation at the origins of life. *Nat. Rev. Genet.* **2015**, *16*, 7–17. [CrossRef] [PubMed]
117. Wachowius, F.; Attwater, J.; Holliger, P. Nucleic acids: Function and potential for abiogenesis. *Q. Rev. Biophys.* **2017**, *50*, e4. [CrossRef]

118. Georgiadis, M.M.; Singh, I.; Kellett, W.F.; Hoshika, S.; Benner, S.A.; Richards, N.G.J. Structural Basis for a Six Nucleotide Genetic Alphabet. *Proc. Natl. Acad. Sci. USA* **2015**, *137*, 6947–6955. [CrossRef]
119. Nielsen, P.E. DNA Analogues with Nonphosphodiester Backbones. *Annu. Rev. Biophys. Biomol. Struct.* **1995**, *24*, 167–183. [CrossRef]
120. Ura, Y.; Beierle, J.M.; Leman, L.J.; Orgel, L.E.; Ghadiri, M.R. Self-Assembling Sequence-Adaptive Peptide Nucleic Acids. *Science* **2009**, *325*, 73–77. [CrossRef]
121. Lescrinier, E.; Esnouf, R.; Schraml, J.; Busson, R.; Heus, H.; Hilbers, C.; Herdewijn, P. Solution structure of a HNA–RNA hybrid. *Chem. Biol.* **2000**, *7*, 719–731. [CrossRef]
122. Kim, S.C.; O'Flaherty, D.K.; Giurgiu, C.; Zhou, L.; Szostak, J.W. The Emergence of RNA from the Heterogeneous Products of Prebiotic Nucleotide Synthesis. *J. Am. Chem. Soc.* **2021**, *143*, 3267–3279. [CrossRef]
123. Cafferty, B.J.; Fialho, D.M.; Khanam, J.; Krishnamurthy, R.; Hud, N.V. Spontaneous formation and base pairing of plausible prebiotic nucleotides in water. *Nat. Commun.* **2016**, *7*, 11328. [CrossRef]
124. Kolb, V.; Dworkin, J.; Miller, S. Alternative bases in the RNA world: The prebiotic synthesis of urazole and its ribosides. *J. Mol. Evol.* **1994**, *38*, 549–557. [CrossRef] [PubMed]
125. Piccirilli, J.A.; Benner, S.A.; Krauch, T.; Moroney, S.E.; Benner, S.A. Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **1990**, *343*, 33–37. [CrossRef] [PubMed]
126. Kim, S.C.; O'Flaherty, D.K.; Zhou, L.; Lelyveld, V.S.; Szostak, J.W. Inosine, but none of the 8-oxo-purines, is a plausible component of a primordial version of RNA. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 13318–13323. [CrossRef] [PubMed]
127. Chen, J.J.; Cai, X.; Szostak, J.W. N2′ → P3′ Phosphoramidate Glycerol Nucleic Acid as a Potential Alternative Genetic System. *J. Am. Chem. Soc.* **2009**, *131*, 2119–2121. [CrossRef] [PubMed]
128. O'Flaherty, D.K.; Zhou, L.; Szostak, J.W. Nonenzymatic Template-Directed Synthesis of Mixed-Sequence 3′-NP-DNA up to 25 Nucleotides Long Inside Model Protocells. *J. Am. Chem. Soc.* **2019**, *141*, 10481–10488. [CrossRef]
129. Heuberger, B.D.; Pal, A.; Del Frate, F.; Topkar, V.V.; Szostak, J.W. Replacing Uridine with 2-Thiouridine Enhances the Rate and Fidelity of Nonenzymatic RNA Primer Extension. *J. Am. Chem. Soc.* **2015**, *137*, 2769–2775. [CrossRef]
130. Winnacker, M.; Kool, E.T. Artificial Genetic Sets Composed of Size-Expanded Base Pairs. *Angew. Chem. Int. Ed.* **2013**, *52*, 12498–12508. [CrossRef]
131. Nelson, K.E.; Levy, M.; Miller, S.L. Peptide nucleic acids rather than RNA may have been the first genetic molecule. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 3868–3871. [CrossRef]
132. Orgel, L. A Simpler Nucleic Acid. *Science* **2000**, *290*, 1306–1307. [CrossRef]
133. Colville, B.W.F.; Powner, M.W. Selective Prebiotic Synthesis of α-Threofuranosyl Cytidine by Photochemical Anomerization. *Angew. Chem. Int. Ed.* **2021**, *60*, 10526–10530. [CrossRef]
134. Szathmáry, E.; Smith, J.M. The major evolutionary transitions. *Nature* **1995**, *374*, 227–232. [CrossRef] [PubMed]
135. Blokhuis, A.; Lacoste, D.; Nghe, P. Universal motifs and the diversity of autocatalytic systems. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 25230–25236. [CrossRef] [PubMed]
136. Nghe, P.; Hordijk, W.; Kauffman, S.A.; Walker, S.I.; Schmidt, F.J.; Kemble, H.; Yeates, J.A.M.; Lehman, N. Prebiotic network evolution: Six key parameters. *Mol. Biosyst.* **2015**, *11*, 3206–3217. [CrossRef] [PubMed]
137. Kauffman, S.A. Autocatalytic sets of proteins. *J. Theor. Biol.* **1986**, *119*, 1–24. [CrossRef]
138. Lincoln, T.A.; Joyce, G.F. Self-Sustained Replication of an RNA Enzyme. *Science* **2009**, *323*, 1229–1232. [CrossRef]
139. Hordijk, W.; Hein, J.; Steel, M. Autocatalytic Sets and the Origin of Life. *Entropy* **2010**, *12*, 1733–1742. [CrossRef]
140. Hordijk, W.; Steel, M. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *J. Theor. Biol.* **2004**, *227*, 451–461. [CrossRef]
141. Hordijk, W.; Steel, M.; Kauffman, S. The Structure of Autocatalytic Sets: Evolvability, Enablement, and Emergence. *Acta Biotheor.* **2012**, *60*, 379–392. [CrossRef]
142. Vasas, V.; Fernando, C.; Santos, M.; Kauffman, S.; Szathmáry, E. Evolution before genes. *Biol. Direct* **2012**, *7*, 1. [CrossRef]

143. Vaidya, N.; Manapat, M.L.; Chen, I.A.; Xulvi-Brunet, R.; Hayden, E.J.; Lehman, N. Spontaneous network formation among cooperative RNA replicators. *Nature* **2012**, *491*, 72–77. [CrossRef]
144. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]