# Theoretical Analysis of Nonenzymatic RNA Replication within the Virtual Circular Genome Scenario

Ludwig Burger and Ulrich Gerland*
*Physics of Complex Biosystems, Department of Bioscience, School of Natural Sciences, Technical University of Munich, 85748 Garching, Germany*
(Dated: July 17, 2024)

The transition from prebiotic chemistry to living systems requires the emergence of a reliable enzyme-free replication mechanism. Despite recent advances in template-directed polymerization, challenges like the short length of copied products and high error rates persist. In this work, we analyze a recently proposed prebiotic replication scenario, the so-called Virtual Circular Genome (VCG) [Zhou et al., RNA 27, 1-11 (2021)]: Replication takes place in a pool of oligomers, where each oligomer contains a subsequence of a circular genome, such that the oligomers encode the full genome collectively. While the sequence of the circular genome may be reconstructed based on long oligomers, short oligomers merely act as replication feedstock. We observe a competition between the predominantly error-free ligation of a short oligomer to a long oligomer and the predominantly erroneous ligation of two long oligomers. Increasing the length of long oligomers and reducing their concentration decreases the fraction of erroneous ligations, enabling high-fidelity replication in the VCG. Alternatively, the problem of erroneous products can be mitigated if only monomers are activated, such that each ligation involves at least one monomer. Surprisingly, in such systems, shorter oligomers are extended by monomers more quickly than long oligomers, a phenomenon which has already been observed experimentally [Ding et al., JACS 145, 7504-7515 (2023)]. Our work provides a theoretical explanation for this behavior, and predicts its dependence on system parameters such as the concentration of long oligomers.

## INTRODUCTION

In order to delineate possible pathways towards the emergence of life, it is necessary to understand how a chemical reaction network capable of storing and replicating genetic information might arise from prebiotic chemistry. RNA is commonly assumed to play a central role in the emergence of self-replicating reaction networks, because it can participate both as reactant and as catalyst [1–3]. While ribozymes capable of replicating strands of their own length have been demonstrated in the laboratory [4], it remains elusive how enzyme-free self-replication might have worked before the emergence of the first ribozyme.

One possible mechanism is template-directed primer extension [5–11]: In this process, a primer hybridizes to a template and gets extended by short oligonucleotides, thereby forming a (complementary) copy of the template strand. Considerable progress has been made in optimizing the protocol of template-directed primer extension, but challenges remain: (i) The produced copies are likely to be incomplete. So far, even under optimized reaction conditions, at most 12 nt have been successfully added to an existing primer [8, 10]. Moreover, as the pool of primer strands needs to emerge via random polymerization, the primer is likely to hybridize to the template at various positions, and not only to its 3'-end, leaving part of the 3'-end of the template uncopied [12]. (ii) Errors in enzyme-free copying are frequent (roughly 10%) due to the limited thermodynamic differentiation between correct Watson-Crick pairing and mismatches [13–16]. The error probability constrains the length of the genome that can be reliably replicated.

The latter problem can in principle be mitigated by making use of the kinetics of mismatch incorporation. Template-directed primer extension is known to slow down after the incorporation of a mismatch [15]. By introducing a competition between the reduced polymerization rate and a characteristic timescale of the non-equilibrium environment, it is possible to filter correct sequences from incorrect ones [16].

To address the first problem, Zhou *et al.* propose an new scenario of replication, the so-called Virtual Circular Genome (VCG) [17]. In this scenario, genetic information is stored in a pool of oligomers that are shorter than the circular genome they collectively encode: Each oligomer bears a subsequence of the circular genome, such that the collection of all oligomers encodes the full circular genome virtually. Within the pool, each oligomer can act as template or primer [17]. The oligomers hybridize to each other via Watson-Crick base pairing and form complexes that allow for templated ligation of two oligomers or extension of an oligomer by a monomer. Because the sequences of the reactants and the template are part of the genome sequence, most of the products should also retain the sequence of the genome. That way, long oligomers encoding the circular genome are produced at the expense of short oligomers [17]. Combined with continuous influx of short oligomers, the VCG pool should allow for continuous

* gerland@tum.de

replication of the virtually encoded circular genome. Importantly, replication in the VCG is expected to avoid the issue of incomplete copies. Since the genome is circular, it does not matter which part of the genome an oligomer encodes, as long as the sequence is compatible with the genome sequence. Moreover, replication should be achieveable without the need of adding many nucleotides to a primer. Provided the concentration of oligomers decreases exponentially with their length, the concentration of each oligomer in the pool can be doubled by extending each oligomer only by a few nucleotides [17]. It has already been demonstrated experimentally that extending an oligomer by a few nucleotides is possible in a VCG with pre-activated nucleotides [18]. A computational study performed by Chamanian *et al.* points out that the VCG is prone to loss of genetic information via "sequence scrambling" [19]: If the genome contains identical sequence motifs at multiple different loci, replication in the VCG will mix the sequences of these loci, thus destroying the initially defined genome.

Here, we characterize fidelity and yield of replication in VCG pools. To this end, we study a broad spectrum of prebiotically plausible and experimentally accessible oligomer pools, from pools containing only monomers and long oligomers of a single length to pools including a range of long oligomers with uniform or exponential concentration profile. The length of the included oligomers as well as their concentration are free parameters of our model.

Our findings indicate that, regardless of the details of the pool, there is a competition between three types of reactions: (i) Ligations of two short oligomers, *e.g.* monomers, form oligomers that are too short to specify a unique locus on the genome sequence. (ii) Ligations of a short oligomer to a long oligomer typically form long oligomers with sequences compatible to the genome, whereas (iii) ligations of two long oligomers are error-prone. Therefore, the fidelity is primarily determined by the competition between the error-free extension of a long oligomer by a monomer and the ligation of two long oligomers. The likelihood of the latter reaction can be minimized by lowering the concentration of long oligomers, even though this increases the fraction of short oligomers that are used up for ligations of the first type. Consequently, the fidelity can be improved at the expense of reducing the yield. The efficiency, meaning the yield attainable at a fixed high fidelity, depends on the length of oligomers in the pool: The longer the oligomers, the higher the maximal efficiency.

Alternatively, the issue of erroneous ligations is solved if only monomers are activated, such that each ligation incorporates only one monomer at a time, as in the experimental study by Ding *et al.* [18]. In this case, the VCG concentration can be chosen arbitrarily large without compromising fidelity. Interestingly, our model predicts an unexpected feature: In the limit of high VCG concentrations, the rate of primer extension is higher for shorter oligomers than for long oligomers, even though, intuitively, complexes containing longer oligomers are expected to be more stable and thus more productive. The same behavior was indeed observed experimentally [18]. We provide an explanation for this feature, and discuss its dependence on system parameter such as the length and the concentration of long oligomers in the pool.

## RESULTS

### Model

The VCG is a replication scenario where a circular genome is stored in a pool of oligomers. The oligomers are shorter than the genome that they encode. Each oligomer bears a subsequence of the circular genome, such that, collectively, the oligomers encode the full genome (Fig. 1). Modelling replication in a VCG pool makes it necessary to decide (i) which circular genome is encoded, (ii) how the genome is mapped to its pool of oligomers, and (iii) which chemical reactions govern the kinetics in the VCG.

#### *Circular Genomes*

For a given genome length $L_G$, a total of $4^{L_G}/(2L_G)$ distinct circular genomes are possible, where the factor $1/(2L_G)$ accounts for the freedom to choose the Watson or the Crick strand as reference sequence and to select its starting position. The genomes differ in terms of the subsequences that are included. A subsequence of length $L$ that appears exactly once in one genome, might appear multiple times or not at all in another genome. While missing subsequences are unproblematic, including the same subsequence in the genome multiple times can pose issues. As the VCG needs to re-construct the global sequence from the local information stored in the oligomers, it is prone to assembling a scrambled global sequence if the same local subsequence appears multiple times within the circular genome [19]. For this reason, we restrict ourselves to genomes, in which subsequences that appear multiple times are as short as possible. The corresponding minimal length of subsequences that are included at most once in the genome is denoted as $L_S$. Moreover, we assume that each subsequence shorter than $L_S$ appears at least once in the genome.

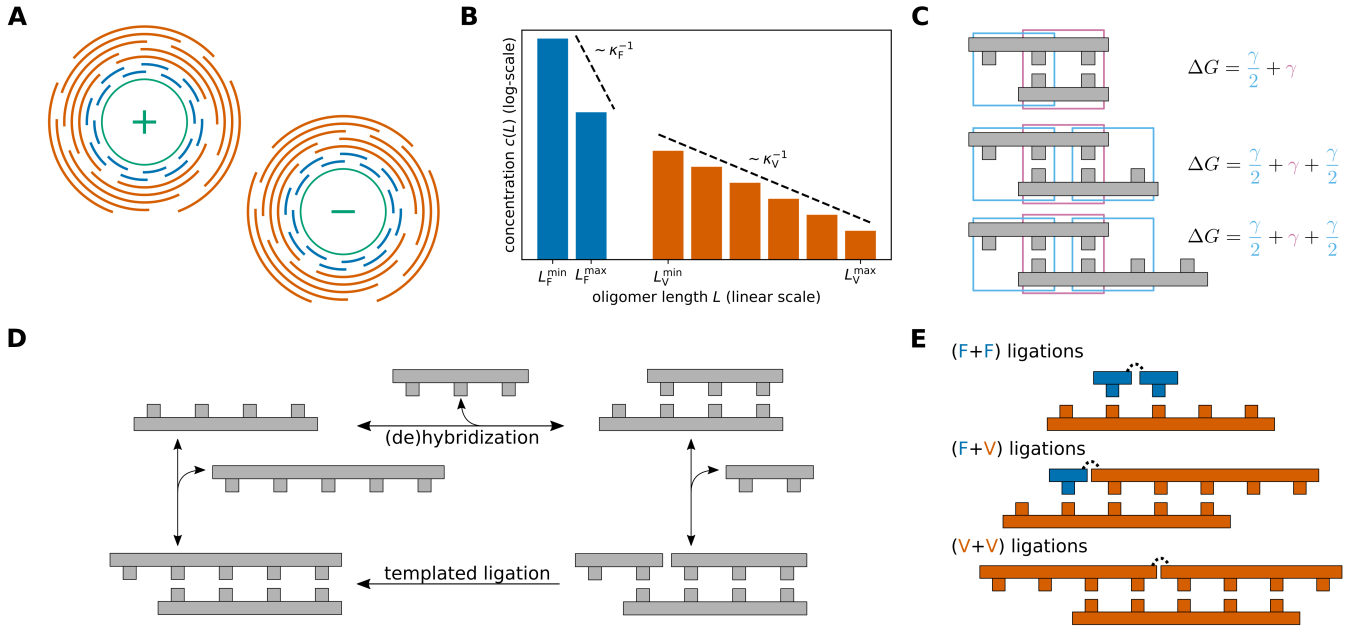



FIG. 1. Model. **A** In the Virtual Circular Genome (VCG) model, a circular genome (depicted in green) as well as its sequence complement are encoded in a pool of oligomers (depicted in blue and orange). Collectively, the pool of oligomers encodes the whole sequence of the circular genome. Depending on their length, two types of oligomers can be distinguished: Feedstock oligomers (depicted in blue) are too short to specify a unique locus on the genome, while long VCG oligomers (depicted in orange) do. **B** The length-distribution of oligomers included in the VCG pool is assumed to be exponential. The concentration of feedstock oligomers and VCG oligomers as well as their respective length scales of exponential decay $\kappa_F^{-1}$ and $\kappa_V^{-1}$ can be varied independently. The set of included oligomer lengths can be restricted via $L_F^{min}$, $L_F^{max}$ and $L_V^{min}$, $L_V^{max}$. **C** The hybridization energy of complexes is computed using a simplified nearest-neighbor model: Each full block comprised of two base pairs (depicted in pink) contributes $\gamma$, while dangling end blocks (depicted in blue) contribute $\gamma/2$. **D** Oligomers form complexes via hybridization reactions, or dehybridize from an existing complex. The ratio of hybridization and dehybridization rate is governed by the hybridization energy. If two oligomers are adjacent to each other in a complex, they can undergo templated ligation. **E** Based on the length of the reacting oligomers, we distinguish three types of templated ligation: Ligation of two feedstock oligomers (F+F), ligation of feedstock oligomer to VCG oligomer (F+V) and ligation of two VCG oligomers (V+V).

*Construction of VCG Pools*

The genome is mapped to a pool of oligomers as illustrated in Fig. 1A. In such a pool, oligomers of different lengths $L$ can appear with different total concentrations $c(L)$ (Fig. 1B). We refer to this length-dependent concentration profile as the length distribution of the pool. While the lengths and the concentrations of the oligomers included in the pool are free parameters, the sequences of the oligomers must collectively encode the full pre-defined circular genome. Therefore, for any oligomer length $L$ included in the pool, we include all possible genome subsequences of length $L$. Depending on their length, oligomers fall into two categories: (i) short feedstock oligomers and (ii) long VCG oligomers. Feedstock oligomers are oligomers that are shorter than the unique subsequence length $L_S$. Since their sequence appears multiple times on the genome, they do not encode a specific position on the genome. Thus, they serve as feedstock for the replication of the VCG rather than as information storage. Conversely, oligomers longer than the unique subsequence length $L_S$ have a unique locus on the circular genome. Collectively, these oligomers enable the reconstruction of the full genome. For this reason, these oligomers are referred to as VCG oligomers.

Due to their different role within the VCG pool, it is meaningful to distinguish between feedstock and VCG oligomers in the length distribution. The full length distribution can be expressed as a sum of the length distributions of feedstock oligomers $c_F(L)$ and VCG oligomers $c_V(L)$, $c(L) = c_F(L) + c_V(L)$. Both types of oligomers follow an exponential length distributions. The concentration of VCG oligomers can be varied independently from the concentration of feedstock oligomers, and the length scales for the exponential decay ($\kappa_F^{-1}$ vs. $\kappa_V^{-1}$) may differ between VCG oligomers and feedstock oligomers. Additionally, we can restrict the set of oligomer lengths included in the pool by setting minimal and maximal lengths for feedstock oligomers, $L_F^{min}$ and $L_F^{max}$, and for

VCG oligomers, $L_{\mathrm{V}}^{\min}$ and $L_{\mathrm{V}}^{\max}$,

$$c_{\mathrm{F}}(L) = \begin{cases} \hat{c}_{\mathrm{F}} \exp\left(-\kappa_{\mathrm{F}} L\right) & \text{if } L_{\mathrm{F}}^{\min} \leq L \leq L_{\mathrm{F}}^{\max}, \\ 0 & \text{else,} \end{cases}$$

$$c_{\mathrm{V}}(L) = \begin{cases} \hat{c}_{\mathrm{V}} \exp\left(-\kappa_{\mathrm{V}} L\right) & \text{if } L_{\mathrm{V}}^{\min} \leq L \leq L_{\mathrm{V}}^{\max}, \\ 0 & \text{else.} \end{cases}$$

This parametrization includes uniform length distributions as a special case ($\kappa_{\mathrm{F}} = 0$ and $\kappa_{\mathrm{V}} = 0$). If $L_{\mathrm{V}}^{\min}$ and $L_{\mathrm{V}}^{\max}$ are chosen similar to each other, the parametrization produces concentration profiles that are peaked. Peaked length distributions can emerge from the interplay of templated ligation, (de)hybridization and outflux in an open system [20]. For convenience, we also define the total concentration of feedstock oligomers, $c_{\mathrm{F}}^{\mathrm{tot}} = \sum_L c_{\mathrm{F}}(L)$, as well as the total concentration of VCG oligomers, $c_{\mathrm{V}}^{\mathrm{tot}} = \sum_L c_{\mathrm{V}}(L)$.

### *Kinetics of VCG Pools*

*(De)hybridization* — Oligomers can hybridize to each other to form double-stranded complexes, or dehybridize from an existing complex. For simplicity, we do not include self-folding within a strand, which is a reasonable assumption for short oligomers. The stability of a complex is determined by its hybridization energy, with lower hybridization energy indicating greater stability. We use a simplified nearest-neighbor energy model to compute the hybridization energy [20–22]: The total hybridization energy equals the sum of the energy contributions of all nearest-neighbor blocks in a given complex (Fig. 1C). The energy contribution associated with a block of two Watson-Crick base pairs (matches) is denoted $\gamma < 0$, and dangling end blocks involving one Watson-Crick pair and one free base contribute $\gamma/2$. Nearest-neighbor blocks involving mismatches increase the hybridization energy by $\gamma_{\mathrm{MM}}$ per block, thus reducing the stability of the complex. The rate constants of hybridization and dehybridization are related via a thermodynamic consistency requirement,

$$\frac{k_{\mathrm{off}}}{k_{\mathrm{on}}} = c^{\circ} \exp\left(\beta \Delta G\right), \tag{1}$$

where $c^{\circ} = 1$ M is the standard concentration, and $\Delta G$ is the free energy of hybridization. The association rate constant $k_{\mathrm{on}}$ is proportional to the encounter rate constant $k_{\mathrm{enc}} = (c^{\circ})^{-1}\ t_0^{-1}$. The associated encounter timescale $t_0$ serves as the elementary time unit of the kinetic model, with all other reaction timescales measured relative to it. The dehybridization rate $r_{\mathrm{off}}$ of an oligomer can be expressed in units of $t_0^{-1}$ via Eq. (1).

*Templated Ligation* — Two oligomers A and B that are hybridized adjacently to each other on a third oligomer C can produce a new oligomer A-B via templated ligation. Depending on the length of A and B, we distinguish three types of ligation reactions: (i) (F+F) ligations in which two feedstock oligomers ligate, (ii) (F+V) ligations where a VCG oligomer is extended by a feedstock oligomer, and (iii) (V+V) ligations involving two VCG oligomers (Fig. 1E). The formation of a covalent bond via templated ligation is not spontaneous, but requires the presence of an activation reaction. Usually, these reactions add a residual group to the 5'-end of the nucleotide which can later act as leaving group in the bond formation [10, 23]. We assume that the concentration of activating agent is sufficiently high for the activation to be far quicker than the formation of the covalent bond, and consider activation and covalent bond formation as a single effective reaction. Most of the paper discusses the behavior of pools where all oligomers are activated, but we also study pools in which only monomers are activated.

*Observables* — While (de)hybridization leaves the oligomer sequences unchanged, templated ligation produces oligomers with increased length and new sequences. Templated ligations that produce feedstock oligomers do not contribute to the replication of the circular genome encoded in the pool. We define the yield $y$ as the fraction of nucleotides that are incorporated via templated ligations that form VCG oligomers relative to the total number of nucleotides used by all templated ligations,

$$y = \frac{\#\text{ nucleotides incorporated in VCG oligomers}}{\#\text{ incorporated nucleotides}} .$$

The sequences of VCG oligomers produced via templated ligation can be compatible with the circular genome (correct), or incompatible (incorrect). Based on this distinction, fidelity is defined to be the fraction of nucleotides incorporated in correct VCG oligomers,

$$f = \frac{\#\text{ nucleotides incorporated in correct VCG oligomers}}{\#\text{ nucleotides incorporated in VCG oligomers}} .$$

Efficient replication of the VCG requires both high fidelity and high yield. Hence, we introduce the "efficiency of replication" $\eta$, an observable accounting for both quantities simultaneously,

$$\begin{aligned} \eta &= f \cdot y \\ &= \frac{\#\text{ nucleotides incorporated in correct VCG oligomers}}{\#\text{ incorporated nucleotides}} . \end{aligned}$$

In order to discern the contributions of different types of templated ligations to fidelity, yield, and efficiency, we define the ligation share $s$ of a ligation type,

$$s(\text{type}) = \frac{\#\text{ nucleotides incorporated via ligation type}}{\#\text{ incorporated nucleotides}} .$$

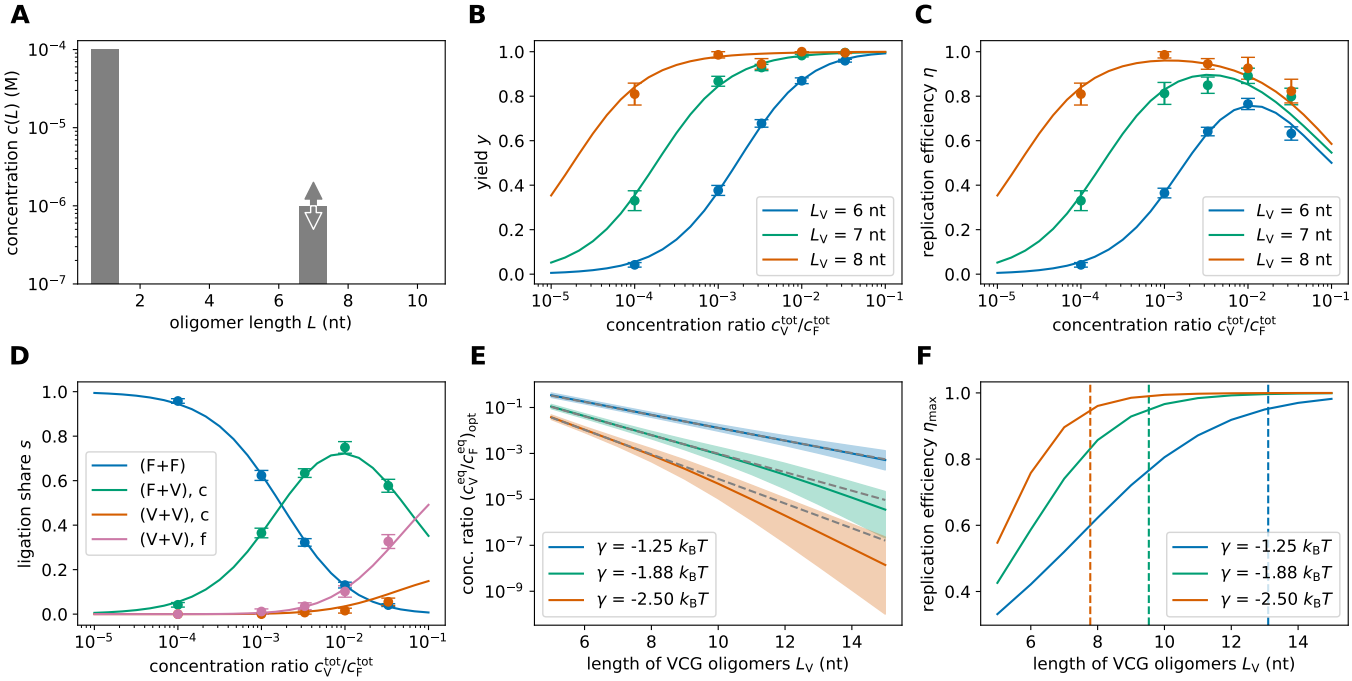



FIG. 2. Replication performance of VCG pools containing VCG oligomers of a single length (single-length VCG pools). **A** The pool contains a fixed concentration of monomers, $c_F^{tot} = 0.1$ mM as well as VCG oligomers of a single length $L_V$ at variable concentration $c_V^{tot}$. **B** The yield increases as a function of $c_V^{tot}/c_F^{tot}$, because dimerizations become increasingly unlikely for high VCG concentrations. **C** Replication efficiency is limited by the small yield for small $c_V^{tot}/c_F^{tot}$. In the limit of high $c_V^{tot}/c_F^{tot}$, replication efficiency decreases due to the growing number of error-prone (V+V) ligations. Maximal replication efficiency is reached at intermediate VCG concentration. **D** The ligation share of different ligation types depends on the total VCG concentration: In the low concentration limit, dimerization (F+F) dominates; for intermediate concentrations, (F+V) ligations reach their maximum, while, for high concentrations, a substantial fraction of reactions are (V+V) ligations. The panel depicts the behavior for $L_V = 6$ nt. **E** The optimal equilibrium concentration ratio of free VCG strands to free feedstock strands, which maximizes replication efficiency, decays as a function of length (continuous line). The analytical scaling law (Eq. (2), dashed line) captures this behavior. The window of close-to-optimal replication, within which efficiency deviates no more than 1% from its optimum (shaded areas), increases with $L_V$, facilitating reliable replication without fine-tuning to match the optimal concentration ratio. **F** Maximal replication efficiency, which is attained at the optimal VCG concentration depicted in panel E, increases as a function of $L_V$ and approaches a plateau of 100%. The oligomer length at which replication efficiency equals 95% is determined using Eq. (3) (vertical dashed lines).

### Replication efficiency reaches a maximum at intermediate concentrations of VCG oligomers

We start the characterization of replication performance with the simplest oligomer pool possible: As depicted in Fig. 2A, it contains only monomers as well as long VCG oligomers of a single length $L_V$. The total monomer concentration is fixed to $c_F^{tot} = 0.1$ mM, whereas the concentration of VCG oligomers $c_V^{tot}$ can be varied. We encode a circular genome of length $L_G = 16$ nt including both, the genome sequence as well as its complement,

```
5'-AAAGAGGACACGGCAU-3'
3'-UUUCUCCUGUGCCGUA-5'.
```

In this genome, the minimal unique subsequence length is $L_S = 3$ nt. Consequently, monomers are part of the feedstock and VCG oligomers are at least 3 nt long.

The time evolution of the VCG pool is simulated using a stochastic simulation based on the Gillespie algorithm [20–22]. The simulation is initialized with a VCG pool that follows the aforementioned length distribution (Fig. 2A). For any included oligomer length, the pool contains all possible genome subsequences of that length in equal abundance. Since the Gillespie algorithm operates on the level of counts of molecules instead of concentrations, we must assign a volume to each system. We choose volumes in the range $1\,\mu m^3$ to $1000\,\mu m^3$, with the former corresponding to the volume of a micron-sized compartment (Section SI in Supplementary Material). In addition to the volume, the reaction rate constants need to be chosen appropriately: (i) Experimentally determined association rate constants are typically around $10^6$ $M^{-1}\,s^{-1}$ [24–26]. Comparing this to the association rate constant in the simulation $k_{on} \approx (c^\circ)^{-1} t_0^{-1}$, we find $t_0 \approx 1$ µs. (ii) The timescale of dehybridization is computed via Eq. (1) using the

energy contribution $\gamma = -2.5\ k_{\mathrm{B}}T$ for a matching nearest neighbor block and $\gamma_{\mathrm{MM}} = 25.0\ k_{\mathrm{B}}T$ in case of mismatches. The value of $\gamma$ matches the average energy of all nearest neighbor blocks involving Watson-Crick base pairing or GU wobbles given by the Turner energy model of RNA hybridization [27]. The high energy penalty of nearest-neighbor blocks involving mismatches $\gamma_{\mathrm{MM}}$ suppresses the formation of mismatches. (iii) For templated ligation, we select a reaction rate constant of $k_{\mathrm{lig}} = 10^{-12}\ t_0^{-1}$. This choice of $k_{\mathrm{lig}}$ is consistent with ligation rates observed in enzyme-free template-direct primer extension experiments. Experimentally measured ligation rate constants range from around $10^{-6}$ s$^{-1}$ ($10^{-12}\ t_0^{-1}$) [9, 10] to roughly $10^{-4}$ s$^{-1}$ ($10^{-10}\ t_0^{-1}$) [23], depending on the underlying activation chemistry.

Based on the ligation events observed in the time evolution, we can compute observables that characterize the yield and fidelity of templated ligations in the oligomer pool. However, due to the small ligation rate constant, the time evolution can only be simulated for a few ligation time units. Consequently, ligation events are scarce, which leads to high variances in the computed observables. We mitigate this issue by calculating the observables based on the concentration of complexes that are in a productive configuration, even if they do not undergo templated ligation (Section SI in Supplementary Material).

Fig. 2B depicts the dependence of the yield on the concentration of VCG oligomers; the dots represent to the data obtained from the simulation. We observe that the yield approaches zero in the limit of low VCG concentration, while it reaches 100% for high $c_{\mathrm{V}}^{\mathrm{tot}}$. The concentration at which the pool transitions from low to high yield depends on the oligomer length $L_{\mathrm{V}}$: Shorter VCG oligomers require higher concentrations for high yield. The concentration dependence of the yield can be rationalized by the types of templated ligations that are involved. For low VCG concentration, most templated ligation reactions are dimerizations (1+1) with the VCG oligomers merely acting as templates (Fig. 2D). As dimers are shorter than the unique subsequence length $L_S$, their formation does not contribute to the yield, which explains the low yield in the limit of small $c_{\mathrm{V}}^{\mathrm{tot}}$. Conversely, for high VCG concentration, most of the templated ligations are (V+V) ligations, which produce oligomer that are long enough to be VCG oligomers, implying high yield.

In the limit of small VCG concentration, replication efficiency, which corresponds to the product of yield and fidelity, is low due to the low yield. In spite of the high yield for high $c_{\mathrm{V}}^{\mathrm{tot}}$, replication efficiency decreases for high VCG concentration, and reaches a maximum at intermediate VCG concentration, as shown in Fig. 2C. The reduction in efficiency for high VCG concentration can be attributed to the high fraction of (V+V) ligations

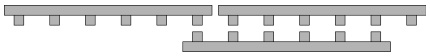
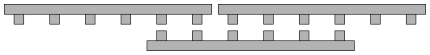
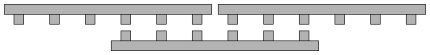
| complex | probability of correct product |
|---|---|
|  | 0.125 |
|  | 0.5 |
|  | 1.0 |

TABLE I. (V+V) ligations are prone to the formation of incorrect product due to the short overlap between educt strand and template. In general, the probability of correct product formation depends on the choice of circular genome and as well as its mapping to the VCG pool. The probabilities listed in this table refer to a VCG pool with $L_G = 16$ nt and $L_S = 3$ nt.

that produce erroneous products (Fig. 2D), leading to small fidelity. Using hexamers as an example, Tab. I illustrates why (V+V) ligations are prone to forming incorrect oligomers. In order to ensure that two oligomers only ligate if their sequences are adjacent to each other in the true circular genome, both oligomers need to share an overlap of at least $L_S$ nucleotides with the template oligomer. Otherwise, the product oligomers might form sequences incompatible with the true genome. The probability of forming incorrect products is a consequence of the combinatorics of possible subsequences in the VCG: For a genome with $L_G = 16$ nt, there are 32 different hexamers. Provided the left educt hexamer only has one nucleotide of overlap with the template, there are $1/4 \cdot 32 = 8$ possible educt oligomers. However, only one out of those 8 hexamers is the correct partner for the right educt hexamer, implying an error probability of $1 - 1/8 = 7/8$ (first row in Tab. I).

Characterizing replication efficiency using simulation data is computationally expensive. Depending on the parameters, up to a hundred simulations, each taking a few days, are needed to obtain a single data point in Fig. 2B-D to sufficient accuracy. To address this problem, we compute the observables of interest via a simplified approach which makes use of the separation of time-scales between (de)hybridization and templated ligation. Due to the slow rate of ligation, the (de)hybridization equilibrium is typically reached long before the first ligation event happens. To determine the hybridization equilibrium, we employ a coarse-grained representation of the VCG pool (Section SII in Supplementary Material): Oligomers of the same length are assumed to have identical concentration, regardless of their sequence. The concentration of complexes comprised of multiple oligomers is calculated using the law of mass action: It equals the product of the individual oligomer concentrations divided by the respective dissociation constant $K_d$, which in turn depends on the hybridization energy $\Delta G$. Depending on the length of the hybridization region, one or multiple

hybridization partners may be possible based on the oligomer sequences present in the pool. This is accounted for by weighting the complex concentrations by combinatorial multiplicities. For each oligomer length, we impose a mass conservation constraint ensuring that the equilibrium concentration of free oligomers and the concentration of oligomers bound in complexes adds up to the total concentration in the pool. Solving the law of mass action together with the mass conservation equations numerically yields the equilibrium concentrations of free VCG and feedstock strands $c_{\mathrm{V}}^{\mathrm{eq}}$ and $c_{\mathrm{F}}^{\mathrm{eq}}$ for any given total concentration $c_{\mathrm{V}}^{\mathrm{tot}}$ and $c_{\mathrm{F}}^{\mathrm{tot}}$. Knowing the equilibrium concentration of free strands allows us to compute concentrations of productive complexes and thus observables characterizing templated ligation. We refer to this method as the adiabatic approach, as it relies on the time-scale separation of (de)hybridization and ligation.

The results of the adiabatic approach agree well with the data obtained from the simulations (Fig. 2B-D). For instance, the adiabatic approach predicts the same non-monotoneous concentration dependence of replication efficiency as the simulations: VCG pools containing longer oligomers achieve higher maxima in replication efficiency at lower VCG concentration. While the available simulation data only allows for a qualitative description of this trend, the adiabatic approach enables a qualitative analysis. Fig. 2E (solid line) depicts the equilibrium concentration ratio at which replication efficiency is maximal. As expected from the qualitative trend observed in Fig. 2C, pools containing longer oligomers reach their maximum for lower concentration of VCG oligomers. The shaded area indicates the range of VCG concentrations within which the pool's efficiency deviates by no more than one percent from its optimum. We observe that the range of close-to-optimal VCG concentrations increases with $L_{\mathrm{V}}$. Thus, pools containing longer oligomers require less fine-tuning of the VCG concentration for replication with high efficiency.

In addition to the numerical results, we utilize the adiabatic approach to study the optimal equilibrium VCG concentration analytically (Section SIII in Supplementary Material). We find that, for any choice of $L_{\mathrm{V}}$, replication efficiency reaches its maximum when the fractions of dimerization (1+1) reactions and erroneous (V+V) ligations are equal (see Fig. 2D for $L_{\mathrm{V}} = 6$ nt). This criterion can used to derive a scaling law for the optimal equilibrium concentration ratio $c_{\mathrm{V}}^{\mathrm{eq}}/c_{\mathrm{F}}^{\mathrm{eq}}$ as a function of the oligomer length $L_{\mathrm{V}}$,

$$\left.\frac{c_{\mathrm{V}}^{\mathrm{eq}}}{c_{\mathrm{F}}^{\mathrm{eq}}}\right|_{\mathrm{opt}} \sim \sqrt{\frac{1}{\Lambda_{\mathrm{F+F}}} - \frac{1}{L_{\mathrm{V}}}} \exp\left(-\frac{|\gamma| L_{\mathrm{V}}}{2}\right), \qquad (2)$$

which is depicted as a dashed line in Fig. 2E. The optimal equilibrium concentration ratio decreases exponentially with $L_{\mathrm{V}}$, while the hybridization energy $\gamma/2$ sets the inverse length-scale of the exponential decay. Analytical estimate and numerical solution agree well, as long as the hybridization is weak and oligomers are sufficiently short. For strong binding and long oligomers, complexes involving more than three strands play a non-negligible role, but such complexes are neglected in the analytical approximation.

Fig. 2F illustrates the maxima of replication efficiency as a function of oligomer length. These values are obtained by computing the replication efficiency for the optimal VCG concentrations shown in Fig. 2E. Consistent with the qualitative trend observed in Fig. 2C, longer oligomers enable higher maxima in replication efficiency. Regardless of the choice of $\gamma$, replication efficiency reaches 100% if $L_{\mathrm{V}}$ is sufficiently high. Using Eq. (2), the following approximation for the maximal replication efficiency can be formulated,

$$\eta_{\max} = 1 - \eta^{\circ} \sqrt{\frac{1}{\Lambda_{\mathrm{F+F}}} - \frac{1}{L_{\mathrm{V}}}} L_{\mathrm{V}} \exp\left(-\frac{|\gamma| L_{\mathrm{V}}}{2}\right), \qquad (3)$$

where $\eta^{\circ}$ is a genome-dependent constant (Section SIV in Supplementary Material). This equation provides guidance for the construction of VCG pools with high replication efficiency: Given a target efficiency, the necessary oligomer length and hybridization energy, *i.e.* temperature, can be calculated. In Fig. 2F, we determine the oligomer length $L_{\mathrm{V}}$ necessary to achieve $\eta_{\max} = 95\%$ for varying hybridization energies $\gamma$. At higher temperature (weaker binding), VCG pools require longer oligomers for replication with high efficiency.

### Pools containing VCG oligomers of different lengths are dominated by the long oligomers

In the previous section, we characterized the behavior of pools containing VCG oligomers of a single length. We found that (F+V) ligations enable replication with high efficiency. In contrast, (V+V) ligations are error-prone due to insufficient overlap between the educt strands and the template. Introducing multiple oligomer lengths into the pool might overcome this limitation: A long oligomer could serve as a template for the ligation of two shorter VCG oligomers. If the shorter oligomers hybridize to the template with a sufficiently long hybridization region, the error-ratio of templated ligation will be significantly reduced compared to the ligation of two long strands. Thus, replication efficiency in systems with mixed oligomer lengths might surpass the performance of the single-length pools.

We analyze this question quantitatively by studying a VCG pool that contains monomers, tetramers, and octamers. The concentration of monomers is set to $c(1) = 0.1$ mM, while the concentrations of the VCG

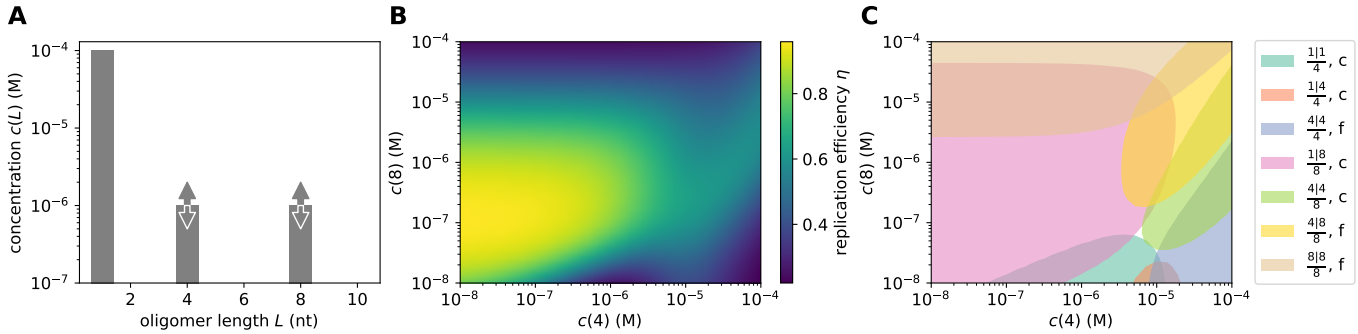



FIG. 3. Replication performance of VCG pools containing VCG oligomers of multiple lengths (multi-length VCG pools). **A** The pool contains a fixed concentration of monomers, $c_{\mathrm{F}}^{\mathrm{tot}} = 0.1$ mM, as well as tetramers and octamers at variable concentration. The hybridization energy per nearest neighbor block is $\gamma = -2.5\ k_{\mathrm{B}}T$. **B** Replication efficiency reaches its maximum for $c(8) \approx 1.3 \cdot 10^{-7}$ M and significantly lower tetramer concentration, $c(4) \approx 7.4 \cdot 10^{-12}$ M. Efficiency remains close-to-maximal on a plateau around the maximum spanning almost two orders of magnitude in tetramer and octamer concentration. In addition, efficiency exhibits a ridge of increased efficiency for high tetramer concentration and intermediate octamer concentration. **C** The concentrations of complexes that facilitate templated ligation are grouped by the length of the template and the educts, $\frac{L_{\mathrm{educt},1}|L_{\mathrm{educt},2}}{L_{\mathrm{template}}}$. We distinguish complexes producing correct (labeled "c") and false products (labeled "f"). For each relevant type of complex, we highlight the region of concentration where it contributes most significantly, *i.e.* at least 20% of the total ligation flux. The plateau of high efficiency is dominated by the ligation of monomers to octamers, whereas the ridge of increased efficiency is due to the correct ligation of two tetramers templated by an octamer.

oligomers are varied independently (Fig. 3A). Replication efficiency reaches its maximum at $c(8) \approx 1.3 \cdot 10^{-7}$ M and very low tetramer concentration, $c(4) \approx 7.4 \cdot 10^{-12}$ M, effectively resembling a single-length VCG pool containing only octamers. As shown in Fig. 3B, the maximal efficiency is surrounded by a plateau of close-to-optimal efficiency. The octamer concentration can be varied by more than one order of magnitude without significant change in efficiency. Similarly, adding tetramers does not affect efficiency as long as the tetramer concentration does not exceed the octamer concentration. Fig. 3C illustrates that the plateau of close-to-optimal efficiency coincides with the concentration regime where the ligation of a monomer to an octamer with another octamer acting as template, $\frac{1|8}{8}$, is the dominant ligation reaction. For high tetramer concentration and intermediate octamer concentration, templated ligation of tetramers on octamer templates, $\frac{4|4}{8}$, surpasses the contribution of $\frac{1|8}{8}$ ligations (green shaded area in Fig. 3C). The $\frac{4|4}{8}$ reactions give rise to a ridge of increased efficiency, which represents a local maximum, but is small compared to the plateau of close-to-optimal efficiency (Fig. 3B). Even though reactions of the type $\frac{4|4}{8}$ produce correct products in most cases, they compete with error-prone ligations like $\frac{4|8}{8}$ or $\frac{4|4}{4}$, which reduce fidelity. Increasing the binding affinity, *i.e.* lowering $\gamma$, enhances the contribution of $\frac{4|4}{8}$ ligations at the expense of $\frac{4|8}{8}$ ligations, resulting in an increased local maximum in replication efficiency (Fig. S2 in Supplementary Material). However, very strong hybridization energy, $\gamma = -5.0 k_{\mathrm{B}}T$, is necessary for $\frac{4|4}{8}$ ligations to give rise to similar efficiency as $\frac{1|8}{8}$ ligations. Thus, high efficiency replication facilitated by the ligation of short VCG oligomers on long VCG oligomers is in theory possible, but requires unrealistically strong binding affinity. Moreover, this mechanism is only effective if educts and template differ significantly with respect to their length. In case short and long oligomers are similar in size, templated ligation of two short oligomers tends to be error-prone, irrespective of the choice of binding affinity (Fig. S3 and S4 in Supplementary Material). In that case, (F+V) ligations remain the most reliable replication mechanism.

The same argument applies to VCG pools containing a range of oligomer lengths in between $L_{\mathrm{V}}^{\mathrm{min}}$ and $L_{\mathrm{V}}^{\mathrm{max}}$ (Fig. 4A). Setting $\gamma = -2.5\ k_{\mathrm{B}}T$ corresponds to the previously discussed weak binding limit, where ligations of short VCG oligomers are not expected to enhance replication efficiency. We maximize replication efficiency with respect to the concentration of VCG oligomers under the assumption that all VCG oligomers of any length have the same concentration (uniform profile). As expected, efficiency reaches its maximum when (F+V) ligations of long VCG oligomers are the predominant type of templated ligation (Fig. 4D). Fig. 4D depicts the relative contribution of different ligation types in VCG pools with fixed $L_{\mathrm{V}}^{\mathrm{max}} = 10$ nt and varying $L_{\mathrm{V}}^{\mathrm{min}} \leq 10$ nt. Adding short VCG oligomers to the pool leaves the relative contributions of the dominant ligation types almost unchanged and increases the fraction of unproductive (F+F) ligations only marginally. Replication efficiency is governed by the behavior of the long oligomers included in the pool. Decreasing $L_{\mathrm{V}}^{\mathrm{min}}$ for fixed $L_{\mathrm{V}}^{\mathrm{max}}$ reduces the achievable maximal replication efficiency only slightly. Conversely, lowering the maximal oligomer length $L_{\mathrm{V}}^{\mathrm{max}}$ while keeping $L_{\mathrm{V}}^{\mathrm{min}}$ fixed causes the efficiency to decrease significantly (Fig. 4B). This is plausible because, for any

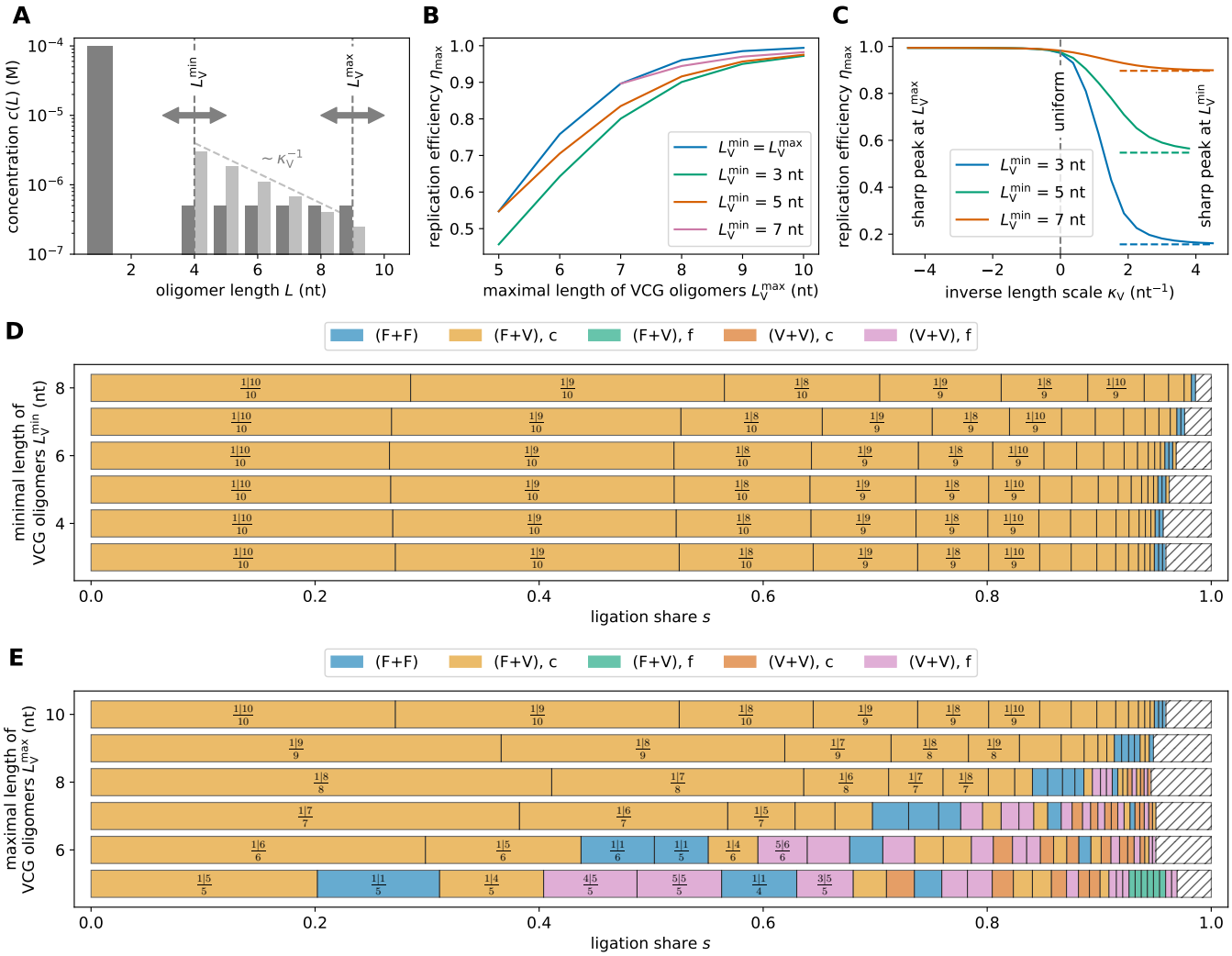



FIG. 4. Replication performance of multi-length VCG pools. **A** The pool contains a fixed concentration of monomers $c_{\mathrm{F}}^{\mathrm{tot}} = 0.1$ mM as well as long oligomers in the range $L_{\mathrm{V}}^{\mathrm{min}} \leq L_{\mathrm{V}} \leq L_{\mathrm{V}}^{\mathrm{max}}$ at variable concentration $c_{\mathrm{V}}^{\mathrm{tot}}$. The length dependence of the concentration profile is assumed to be uniform (for panels B, D, and E) or exponential (for panel C); its steepness is set by the parameter $\kappa_{\mathrm{V}}$. **B** If the length distribution is uniform, reducing $L_{\mathrm{V}}^{\mathrm{min}}$ decreases the maximal efficiency, whereas increasing $L_{\mathrm{V}}^{\mathrm{max}}$ increases it. Pools containing a range of oligomer lengths are always outperformed by single-length VCGs (blue line). **C** Assuming an exponential length distribution of VCG oligomers allows to tune from a poorly-performing regime (dominated by oligomers of length $L_{\mathrm{V}}^{\mathrm{min}}$) to a well-performing regime (dominated by oligomers of length $L_{\mathrm{V}}^{\mathrm{max}}$). In the limit $\kappa_{\mathrm{V}} \to \infty$, $\eta_{\mathrm{max}}$ approaches the replication efficiency of single-length pools containing only oligomers of length $L_{\mathrm{V}}^{\mathrm{min}}$ (dashed lines). **D** Especially for high $L_{\mathrm{V}}^{\mathrm{max}}$, replication is dominated by primer extension of the long oligomers in the VCG, here $L_{\mathrm{V}}^{\mathrm{max}} = 10$. In this limit, addition of shorter oligomers leaves the dominant (F+V) ligations almost unchanged. **E** Reducing $L_{\mathrm{V}}^{\mathrm{max}}$ for fixed $L_{\mathrm{V}}^{\mathrm{min}} = 3$ nt gives rise to increased significance of erroneous ligation reactions.

$L_{\mathrm{V}}^{\mathrm{max}}$, replication efficiency is bounded by the replication efficiency of the longest oligomer in the pool, and we know from the analysis of single-length VCG pools that short VCG oligomers lead to lower replication efficiency (see blue curve in Fig. 4B and recall Eq. (3)). In addition, for small $L_{\mathrm{V}}^{\mathrm{max}}$, short and long VCG oligomers are increasingly similar in length, which gives rise to a broad spectrum of erroneous (V+V) ligations that compete with the error-free (F+V) ligations, as depicted in Fig. 4E.

In a realistic prebiotic setup, the concentration profile of the VCG pool is typically not uniform. Depending on the mechanism that produces the pool and its coupling to the non-equilibrium environment, it might have a concentration profile that decreases or increases exponentially with length. We use the parameter $\kappa_{\mathrm{V}}$ to control this exponential length dependence (Fig. 4A): For negative $\kappa_{\mathrm{V}}$, the concentration increases as a function of length, while exponentially decaying length distributions have positive $\kappa_{\mathrm{V}}$. The previously discussed case of uniform profiles corresponds to the edge-case $\kappa_{\mathrm{V}} = 0$. We

find that replication efficiency is high if the concentration of long VCG oligomers exceeds or at least matches the concentration of short VCG oligomers (see $\kappa_{\mathrm{V}} \leq 0$ in Fig. 4C). In that case, replication efficiency is governed by the long oligomers in the pool, since these form the most stable complexes. As the concentration of long oligomers is decreased further ($\kappa_{\mathrm{V}} > 0$), the higher stability of complexes formed by longer oligomers is eventually insufficient to compensate for the reduced concentration of long oligomers. Replication efficiency is then governed by short VCG oligomers, which exhibit lower replication efficiency. In the limit $\kappa_{\mathrm{V}} \to \infty$, replication efficiency approaches the replication efficiency of a single-length VCG pool containing only oligomers of length $L_{\mathrm{V}}^{\min}$ (Fig. 4C and Fig. 2F).

### Adding dinucleotides to the feedstock decreases the efficiency of replication

So far, we focused on ensembles that contain solely monomers and no dimers as feedstock. However, examining the influence of dimers on replication in VCG pools is of interest, since dinucleotides have proven to be interesting candidates for enzyme-free RNA copying [9–11, 23]. For this reason, we study oligomer pools like those illustrated in Fig. 5A: The ensemble contains monomers, dimers and VCG oligomers of a single length $L_{\mathrm{V}}$. As our default scenario, the dimer concentration is set to 10% of the monomer concentration, but this ratio can be modified by changing $\kappa_{\mathrm{V}}$.

Fig. 5B illustrates the replication efficiency of a pool with ($L_{\mathrm{F}}^{\max} = 2$) or without dimers ($L_{\mathrm{F}}^{\max} = 1$). Pools that are rich in VCG oligomers, *i.e.* high $c_{\mathrm{V}}^{\mathrm{tot}}/c_{\mathrm{F}}^{\mathrm{tot}}$, exhibit low efficiency of replication. As erroneous (V+V) ligations are the most dominant type of reaction in this limit, the same efficiency is obtained regardless of the presence of dimers. In contrast, when the pool is rich in feedstock, pools with dimers behave differently than without: If only monomers are included, the efficiency approaches zero, as dimerizing monomers do not contribute to the yield, and thus not to the efficiency. However, including dimers enables the ligation of monomers and dimers with dimers serving as templates. Given the low VCG concentration, dimers are more abundant than VCG oligomers and thus more likely to act as templates. These ligations produce trimers and tetramers, leading to a non-zero yield. As dimers serve as templates, educt oligomers can only hybridize to the template with a single nucleotide-long hybridization region. Consequently, many products are incompatible with the circular genome, causing low efficiency. Replication efficiency reaches its maximum at intermediate VCG concentrations, where replication is dominated by (F+V) ligations. Notably, the maximal attainable efficiency is significantly lower for ensembles with dimers than without, as dimers increase the number and the stability of possible complex configurations that can form incorrect products (Tab. II). For $L_{\mathrm{F}}^{\max} = 1$, ligation products are erroneous only if the overlap between the VCG educt oligomer and template is shorter than the unique subsequence length $L_S$. For $L_{\mathrm{F}}^{\max} = 2$, however, dangling end dimers can cause incorrect products even in case of long overlap of educt oligomer and template (see right column in Tab. II). The stability of the latter complexes depends on the length of the VCG oligomers $L_{\mathrm{V}}$, whereas complexes facilitating incorrect monomer addition have constant stability, independent of oligomer length. This difference is reflected in the scaling of the effective association constant with $L_{\mathrm{V}}$ (Fig. S5 in Supplementary Material). While the effective association constant of complexes enabling incorrect (1+V) ligations $\mathcal{K}^{a}_{1+\mathrm{V},f}$ is length-independent, $\mathcal{K}^{a}_{2+\mathrm{V},f}$ scales exponentially with $L_{\mathrm{V}}$,

$$\mathcal{K}^{a}_{2+\mathrm{V},f} = \mathcal{K}^{a,\circ}_{2+\mathrm{V},f} \exp\left(|\gamma| L_{\mathrm{V}}\right).$$

Similarly, the effective association constants for correct (1+V) and (2+V) ligations also scale exponentially with the oligomer length.

The competition between correct and incorrect (F+V) ligations sets an upper bound on the efficiency of replication (Section SVI in Supplementary Material),

$$\eta_{\max} \leq \frac{\mathcal{K}^{a,\circ}_{1+\mathrm{V},c} + 0.5\ \mathcal{K}^{a,\circ}_{2+\mathrm{V},c}\ e^{-\kappa_{\mathrm{F}}}}{\mathcal{K}^{a,\circ}_{1+\mathrm{V},c} + 0.5\ \left(\mathcal{K}^{a,\circ}_{2+\mathrm{V},c} + \mathcal{K}^{a,\circ}_{2+\mathrm{V},f}\right)\ e^{-\kappa_{\mathrm{F}}}}. \quad (4)$$

In systems without dimers, *i.e.* $\kappa_{\mathrm{F}} \to \infty$, $\eta_{\max}$ approaches 100%, indicating that perfect replication is achievable. Conversely, in systems containing dimers, efficiency asymptotically approaches a value below 100%, which depends on the concentration of dimers in the feedstock. Fig. 5C shows the dependence of maximal replication efficiency on the length of VCG oligomers in pools containing monomers, dimers and VCG oligomers. As $L_{\mathrm{V}}$ increases, $\eta_{\max}$ converges towards the upper bound defined in Eq. (4) (see dashed line in Fig. 5C).

### Monomers preferentially ligate to short VCG oligomers in VCG-rich pools

Efficiency of replication is limited by the same mechanism regardless of the specifics of the underlying VCG pool: Two oligomers that have an insufficiently short region of hybridization with their template undergo templated ligation and are likely to produce an incorrect product. In the previous sections, we minimized the contribution of such ligations by fine-tuning the concentration of VCG oligomers. Alternatively, incorrect ligations can be significantly reduced if only monomers are considered to be activated, while longer oligomers remain inactive. Such scenario has already been investigated experimentally [18], and is conceivable in a prebiotic system: Activated monomers are produced by an in situ activation chemistry in the environment,

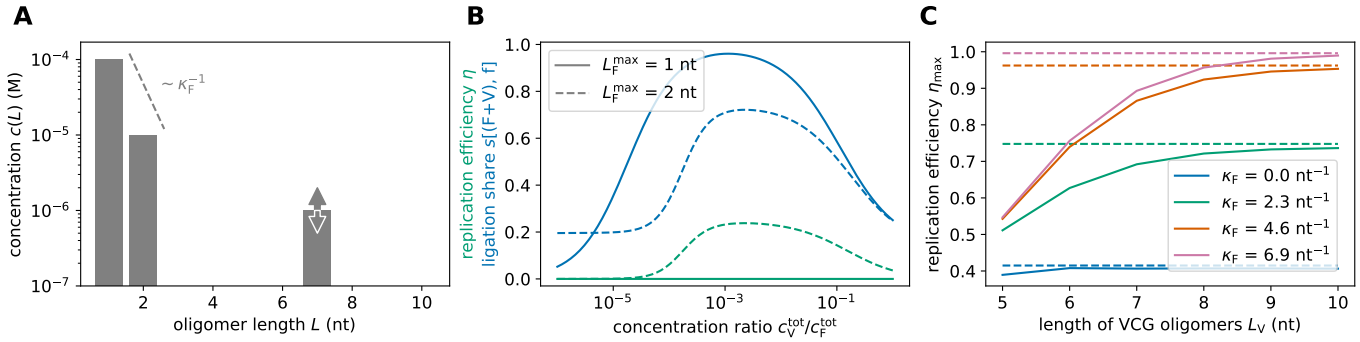



FIG. 5. Replication performance of single-length VCG pools containing monomers and dimers as feedstock. **A** The pool contains a fixed total concentration of feedstock oligomers, $c_{\mathrm{F}}^{\mathrm{tot}} = 0.1$ mM, partitioned into monomers and dimers, as well as VCG oligomers of a single length $L_{\mathrm{V}}$. The proportion of monomers and dimers can be adjusted via $\kappa_{\mathrm{F}}$, and the concentration of the VCG oligomers is a free parameter $c_{\mathrm{V}}^{\mathrm{tot}}$. **B** Replication efficiency exhibits a maximum at intermediate VCG concentration in systems with (dashed blue curve) and without dimers (solid blue curve). The presence of dimers reduces replication efficiency significantly, as they enhance the ligation share of incorrect (F+V) ligations (dashed green curve). The panel depicts the behavior for $L_{\mathrm{V}} = 7$ nt and $\kappa_{\mathrm{F}} = 2.3$ nt$^{-1}$. **C** Optimal replication efficiency increases as a function of oligomer length $L_{\mathrm{V}}$ and asymptotically approaches a plateau $\eta_{\max}^{\infty}$ (Eq. (4), dashed lines). The value of this plateau is determined by the competition between correct and false (2+V) reactions, both of which grow exponentially with $L_{\mathrm{V}}$. Thus, $\eta_{\max}^{\infty}$ depends on the relative concentration of the dimers in the pool: the more dimers are included, the lower is $\eta_{\max}^{\infty}$.

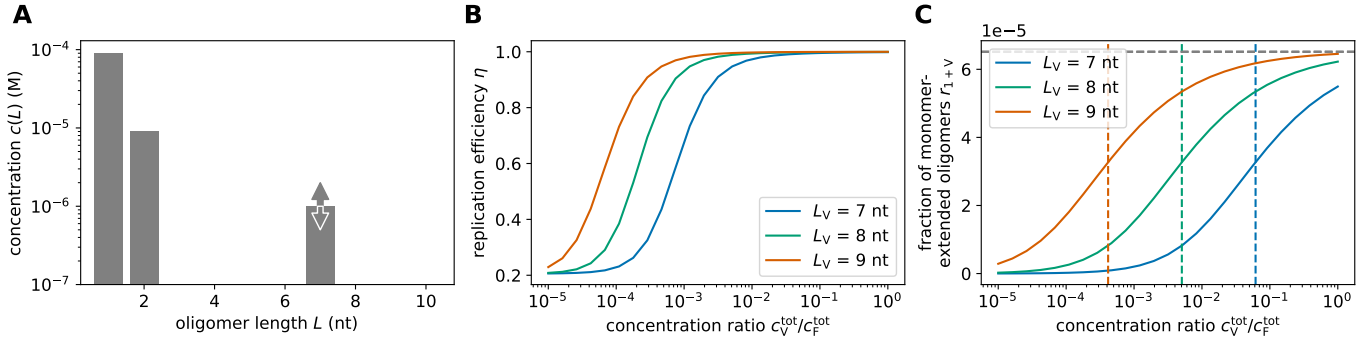



FIG. 6. Replication performance of single-length VCG pools, in which only the monomers are activated. **A** The pool contains activated monomers alongside non-activated dimers and VCG oligomers of a single length. The concentrations of monomers and dimers are fixed, $c(1) = 0.091$ mM and $c(2) = 9.1$ µM, adding up to a total feedstock concentration of $c_{\mathrm{V}}^{\mathrm{tot}} = 0.1$ mM, while the concentration of VCG oligomers $c_{\mathrm{V}}^{\mathrm{tot}}$ can be varied. **B** Unlike in pools in which all oligomers are activated, replication efficiency does not decrease at high VCG concentration if only monomers are activated, as erroneous (V+V) ligations are impossible. Instead, replication efficiency approaches an asymptotic value of 1. **C** The fraction of monomer-extended oligomers depends on the total concentration of VCG oligomers. At low VCG concentration, most oligomers are single-stranded, and extension of oligomers by monomers is scarce. At high VCG concentration, $r_{1+\mathrm{V}}$ approaches an asymptotic value $r_{1+\mathrm{V}}^{\infty}$ (Eq. (5), grey dashed line). In this limit, almost all oligomers form duplexes, which facilitate monomer addition upon hybridization of a monomer. Thus, the asymptotic fraction of oligomers that gets extended by monomers is not determined by the oligomer length, but by the binding affinity of monomers to existing duplexes. Conversely, the threshold concentration at which $r_{1+\mathrm{V}} = r_{1+\mathrm{V}}^{\infty}/2$ depends on oligomer length (colored dashed lines): Longer oligomers reach higher $r_{1+\mathrm{V}}$ at lower VCG concentration.

and subsequently transported into the compartment containing the VCG. Within the compartment, however, no activation takes place [28, 29].

We consider a pool of oligomers, consisting of activated monomers as well as non-activated dimers and VCG oligomers of a single length (Fig. 6A). The concentration of VCG oligomers can be varied, while the feedstock concentration is kept constant. Efficiency is low for low VCG concentration, as the ligation of two monomers or monomer and dimer are the most likely ligations (Fig. 6B). Conversely, high $c_{\mathrm{V}}^{\mathrm{tot}}$ facilitates the formation of complexes, in which VCG oligomers are extended by monomers, which implies high efficiency (Fig. 6B). Note that, unlike in systems where all oligomers are activated, replication efficiency does not decrease for high VCG concentration as erroneous (V+V) are impossible. Instead, perfect replication efficiency (100%) is reached for sufficiently high $L_{\mathrm{V}}$. While replication efficiency characterizes the relative amount of nucleotides used for the cor-

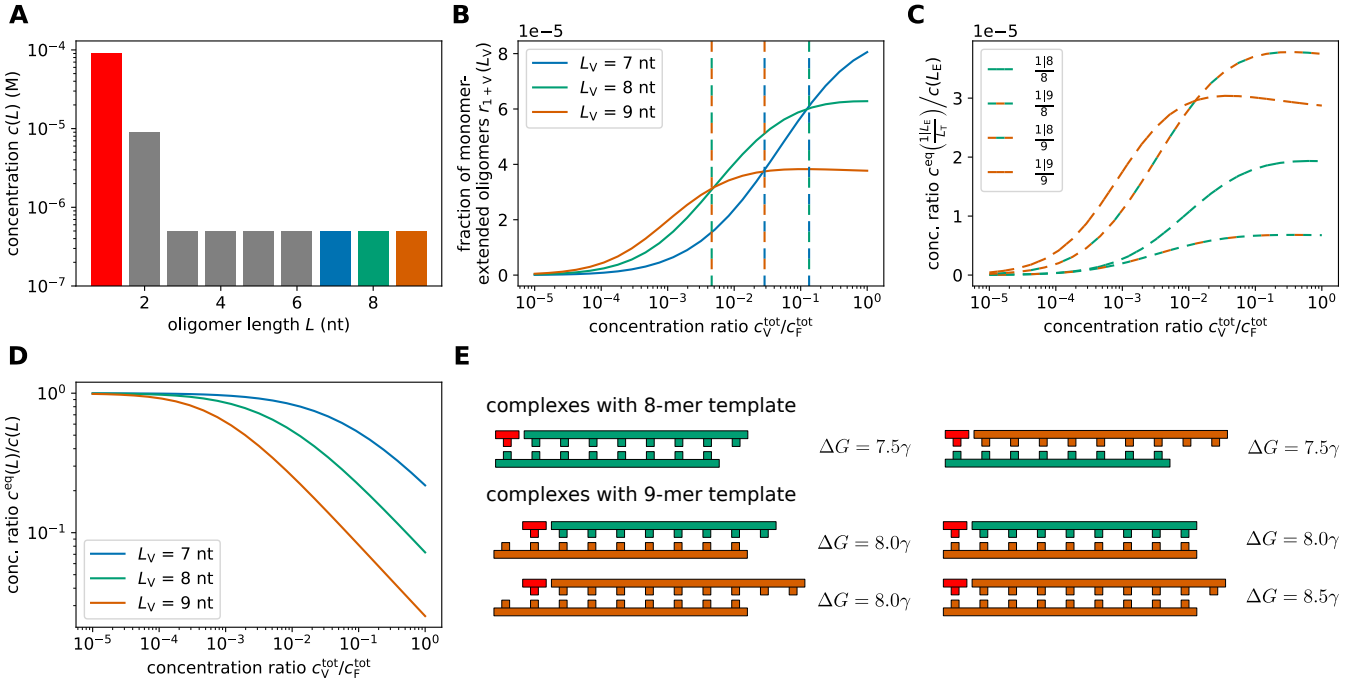



FIG. 7. Replication performance of multi-length VCG pools, in which only the monomers are activated. **A** The pool contains activated monomers as well as non-activated dimers and VCG oligomers. The concentrations of monomers and dimers are fixed, $c(1) = 0.091$ mM and $c(2) = 9.1$ µM, adding up to a total feedstock concentration of $c_V^{tot} = 0.1$ mM, while the total concentration of VCG oligomers $c_V^{tot}$ can be varied. All VCG oligomers are assumed to have the same concentration. **B** At low VCG concentration, the fraction of monomer-extended oligomers is higher for longer oligomers than for short oligomers, while, at high VCG concentration, monomers preferentially ligate to short oligomers ("inversion of productivity"). The threshold concentrations at which a short oligomer starts to outperform a longer oligomer depend on the lengths of the compared oligomers (dashed lines). **C** The mechanism underlying inversion of productivity can be understood based on the pair-wise competition of a short, *e.g.* 8-mer, and a long VCG oligomer, *e.g.* 9-mer. Over the entire range of VCG concentrations, complexes in which a 8-mer serves as a template have a lower relative equilibrium concentration than complexes with 9-mer template. Moreover, complexes with 9-mer template undergo inversion of productivity: For sufficiently high VCG concentration, the fraction of 8-mers that are extended by monomers using a 9-mer as a template exceeds the fraction of extended 9-mers. **D** The equilibrium concentration of free oligomer decreases with increasing $c_V^{tot}$. Long oligomers have a lower equilibrium concentration ratio of free oligomers, as they can form more stable complexes with longer hybridization sites. **E** Complexes, in which 8-mers serve as template, are less stable than complexes with 9-mer templates, explaining the higher relative equilibrium concentration of the latter complex type (see panel **C**). Complexes with 9-mer template have similar stability regardless of the length of the educt oligomer, *i.e.* $\frac{1|8}{9}$ and $\frac{1|9}{9}$ are similarly stable. This similar stability together with the higher concentration of free 8-mers compared to 9-mers (see panel **D**) is the reason why the fraction of monomer-extended 8-mers exceeds the one of 9-mers (see panel **C**).

rect elongation of VCG oligomers, we can also analyze which fraction of available VCG oligomers is extended by a nucleotide,

$$r_{1+V} = c_{1+V}^{eq}/c_V^{tot}$$

Here, $c_{1+V}^{eq}$ denotes the equilibrium concentration of all complexes enabling the addition of a monomer to any VCG oligomer. We find that $r_{1+V}$ depends on the VCG concentration in the same way as the efficiency: $r_{1+V}$ is small for small VCG concentration, but approaches value $r_{1+V}^{\infty}$ asymptotically for high VCG concentration (Fig. 6C). In this limit, almost all VCG oligomers are part of a duplex. Monomers can bind to these duplexes to form complexes allowing for (1+V) ligations. Since almost all VCG oligomers are already part of an duplex, increasing $c_V^{tot}$ further does not increase the fraction of VCG oligomers that can be extended by monomers. Instead, the asymptotic value is determined by the concentration of monomers and their binding affinity $K_d(1)$ to an existing duplex (Section VII in Supplementary Material),

$$r_{1+V}^{\infty} \approx \frac{1}{6}\left(\frac{1}{K_d(1)} + 1 - \frac{2}{3}K_d(1)\right)c(1). \qquad (5)$$

Noteworthy, the asymptotic value does not depend on the length of the VCG oligomers. On the contrary, the threshold VCG concentration, at which $r_{1+V} = r_{1+V}^{\infty}/2$, scales exponentially with the $L_V$ (Section VII in Supplementary Material),

$$c_V^{tot} \sim \exp(-L_V|\gamma|).$$

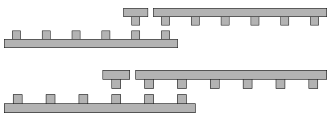
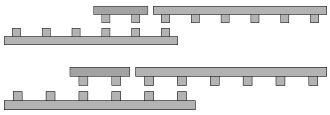
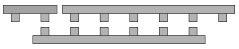
| | $L$-independent stability | $L$-dependent stability |
|---|---|---|
| $\frac{1\|L}{L}$ |  | none |
| $\frac{2\|L}{L}$ |  |  |

TABLE II. Stability of complexes facilitating incorrect (F+V) ligations. **Left column** Erroneous (1+V) ligations are possible if the educt oligomer has a short overlap region with the template. The hybridization energy for such configurations is small, and independent of the length of the VCG oligomers. (2+V) ligations may produce incorrect products via the same mechanism. **Right column** In addition, erroneous (2+V) ligations can be caused by complexes in which two VCG oligomers hybridize perfectly to each other, but the dimer has a dangling end. The stability of these complexes increases exponentially with oligomer length.

We find that longer oligomers require exponentially lower VCG concentration to achieve a given ratio $r_{1+V}$ (Fig. 6C) as their greater length allows them to form more stable complexes. This observation implies that, at a given VCG concentration, longer oligomers will always be more productive than shorter oligomers, provided there is no mutual interaction of oligomers of different lengths.

The behavior of (1+V) ligations changes qualitatively in pools containing VCG oligomers of multiple lengths due to the mutual interactions of oligomers of different length. To illustrate this, we consider an ensemble comprised of activated monomers as well as non-activated dimers and VCG oligomers of length $L_{\mathrm{V}}^{\mathrm{min}} = 3$ nt to $L_{\mathrm{V}}^{\mathrm{max}} = 9$ nt (Fig. 7A). For simplicity, we assume that the length-distribution of VCG oligomers is uniform. We study the fraction of monomer-extended oligomers grouped by oligomer length,

$$r_{1+\mathrm{V}}(L) = c_{1+V}^{\mathrm{eq}}(L)/c(L),$$

where $c_{1+V}^{\mathrm{eq}}(L)$ denotes the equilibrium concentration of all complexes (arbitrary template length) facilitating monomer-extension of VCG oligomers of length $L$. At low VCG concentration, longer oligomers are more likely to get extended by monomers than short oligomers. Increasing VCG concentration reverses this length-dependence of productivity: The fraction of short oligomers extended by monomers exceeds that of long oligomers (Fig. 7B). We elucidate the underlying mechanism using the example of 8-mers and 9-mers. Focusing on oligomers of length 8 and 9, four possible reactive triplexes can be formed (see labels in Fig. 7C). Across the entire range of VCG concentrations, triplexes using the 8-mer as a template incorporate monomers more rarely than triplexes with 9-mer templates (Fig. 7C). This is expected as complexes of the form $\frac{1|\mathrm{L}}{8}$ have weaker binding affinity than $\frac{1|\mathrm{L}}{9}$ complexes (Fig. 7E). While complexes with the shorter template (8-mer) do not exhibit any inversion of productivity, complexes with the longer template (9-mer) do: The fraction of 8-mers that can be extended by a monomer in a complex of the form $\frac{1|8}{9}$ exceeds the fraction of monomer-extendable 9-mers in complexes of the form $\frac{1|9}{9}$. Two factors contribute to this feature: (i) Triplexes of type $\frac{1|8}{9}$ and $\frac{1|9}{9}$ have similar stability (Fig. 7E), and (ii) the equilibrium concentration of free 8-mers is higher than that of 9-mers (Fig. 7D). Consequently, it is more likely for a 8-mer than for a 9-mer to hybridize to an existing duplex $\frac{1|\mathrm{empty}}{9}$, and, given the stability of the complex $\frac{1|8}{9}$, it is almost as likely for a 8-mer to stay hybridized as for a 9-mer. In summary, short oligomers sequester long oligomers as templates to enhance their monomer-extension rate, but long oligomers cannot make use of the short oligomers as templates due to the instability of complexes of the form $\frac{1|9}{8}$.

Even though the pair-wise competition of oligomers explains the mechanism of inverted productivity qualitatively, it falls short in providing quantitative predictions. To achieve this, we need to go beyond the pair-wise competition of oligomers $L$ and $M$ and account for the contributions of oligomer lengths not equal to $L$ or $M$. Despite this additional complexity, it is possible to derive a semi-analytical expression of the threshold VCG concentration, at which the fraction of monomer-extended oligomers of length $L$ equals the fraction of monomer-extended oligomers of length $M$ (Sections SVIII and SIX in Supplementary Material). The threshold concentrations predicted by the analytical approach agree well with those obtained from the numerical solution (Fig. 7B).

It is noteworthy that Ding *et al.* have already experimentally observed the aforementioned inversion of productivity [18]. In their study, they included activated monomers, activated imidazolium-bridged dinucleotides and oligomers up to a length of 12 nt, and observed that the primer extension rate for short primers is higher than the extension rate of long primers. Even though our model differs from their setup in several aspects, *e.g.* we choose a different circular genome and do not model bridged dinucleotides, evaluating our model using parameters similar to those of the experimentally studied system predicts inversion of productivity that qualitatively agrees with the experimental findings (Section SX in Supplementary Material). It is therefore to be assumed that the mechanism behind inversion of productivity described in this section also applies to the experimental observations.

## DISCUSSION AND CONCLUSION

In recent years, significant progress has been made in understanding the prebiotic formation of (ribo)nucleotides [30–34] and characterizing prebiotically relevant ribozymes [4, 35–37]. Nevertheless, constructing a convincing replication mechanism that bridges the gap between prebiotic chemistry and ribozyme-catalyzed replication still presents a considerable challenge. Because of the limitations of commonly investigated scenarios such as template-directed primer extension, Zhou *et al.* proposed a new scenario of replication called the Virtual Circular Genome (VCG) [17]. Our analysis of this scenario reveals a trade-off between fidelity and yield of replication: At low concentration of VCG oligomers, most of the ligations will produce products that are too short to specify a unique locus on the genome, resulting in a low yield of replication. At high VCG concentration, erroneous templated ligations limit the fidelity of replication. We propose two solutions to the trade-off: (i) constructing the VCG pool to optimize its replication behavior within the bounds of the fidelity-yield trade-off, and (ii) breaking the fidelity-yield trade-off by assuming that only monomers are activated.

The first strategy relies on maximizing the yield of replication for fixed fidelity. In pools that contain only monomers and VCG oligomers of a single length, replication efficiency may be maximized by increasing the length of VCG oligomers and decreasing their concentration. This reduces the likelihood of error-prone templated ligation of long oligomers. When the pool contains VCG oligomers of multiple lengths, replication efficiency is typically governed by the longest oligomer in the pool. Including dimers as feedstock for the replication turned out to increase the error fraction significantly, as dimers forming a dangling end as they hybridize to a template are prone to form incorrect product.

The second strategy involves eliminating error-prone templated ligation of two VCG oligomers by assuming that each ligation involves at least one monomer. Therefore, both fidelity and yield remain high even for high VCG concentrations, effectively breaking the fidelity-yield trade-off. For VCG pools containing only VCG oligomers of a single length along with feedstock oligomers, the maximum fraction of monomer-extended oligomers is limited by the binding affinity of the monomers to the productive complexes. Consequently, the fraction of oligomers that are extended by monomers is length-independent. However, once multiple VCG oligomer lengths compete with each other in the pool, the fraction of oligomers extended by monomers can be higher for shorter oligomers than for longer ones. We note that this surprising behavior had already been observed experimentally [18]. The feature is caused by an asymmetry in the productive interaction between short and long oligomers: While short oligomers can sequester longer oligomers as templates for their extension by a monomer, short oligomers are unlikely to serve as templates for longer oligomers.

Replication in VCG pools aims to close the gap between prebiotic chemistry and ribozyme-catalyzed replication. For this reason, it needs to be capable of replicating (parts of) ribozymes that are believed to play a role in the emergence of life. While there are cases of small ribozymes [38] or ribozymes with small active sites, such as the Hammerhead ribozyme [39], most ribozymes obtained experimentally via *in vitro* evolution are typically around a few hundred nucleotides long [4, 35–37]. We demonstrated replication by means of an example genome of length $L_G = 16$ nt. Genomes of this length might suffice to encode a small ribozyme or an active site of an ribozyme. In principle, long ribozymes can also be encoded in a single VCG if the longest oligomers in the pool are chosen long enough to exceed the minimal length of unique subsequences $L_S$. As $L_S$ can be high for long ribozymes, it might be advisable to replicate parts of the ribozyme in different VCGs individually, followed by the assembly of the ribozyme via splinted ligation [40].

We conclude that the Virtual Circular Genome model proposed by Zhou *et al.* presents a promising scenario of enzyme-free replication. Our examination of replication in VCG pools identified replication driven by activated monomers as the most efficient and robust scenario. Its feasibility has already been demonstrated experimentally for VCGs fueled by pre-activated monomers and bridged dinucleotides [18]. Further experimental and theoretical work is necessary to identify VCG pools that are capable of continuously fueled self-sustained replication with high fidelity and yield. We anticipate that continuous replication will necessitate coupling of the VCG pool to a non-equilibrium environment that provides influx of activated monomers as well as (length-dependent) outflux of product oligomers, for example via thermophoretic traps [41–44]. The constraints formulated in this work can guide the search for self-replicating VCG pools in the vast space of possible concentration profiles and non-equilibrium environments.

## ACKNOWLEDGMENTS

We thank T. Göppel, M. Battipede, J. C. Espinoza Campos, J. Harth-Kitzerow for stimulating discussions. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the CRC/TRR 392/1 Molecular Evolution (Project-ID 521256690), and under Germany's Excellence Strategy EXC-2094-390783311, ORIGINS.

[1] G. F. Joyce, RNA evolution and the origins of life, Nature **338**, 217 (1989).

[2] M. P. Robertson and G. F. Joyce, The origins of the RNA world, Cold Spring Harb Perspect Biol **4**, a003608 (2021).

[3] P. G. Higgs and N. Lehman, The RNA world: molecular cooperation at the origins of life, Nat. Rev. Genet. **16**, 7 (2015).

[4] J. Attwater, A. Wochner, and P. Holliger, In-ice evolution of RNA polymerase ribozyme activity, Nat. Chem. **5**, 1011 (2013).

[5] E. Kervio, M. Sosson, and C. Richert, The effect of leaving groups on binding and reactivity in enzyme-free copying of DNA and RNA, Nucleic Acids Res. **44**, 5504–5514 (2016).

[6] T. Walton and J. W. Szostak, A highly reactive imidazolium-bridged dinucleotide intermediate in nonenzymatic rna primer extension, J. Am. Chem. Soc. **138**, 11996–12002 (2016).

[7] N. Prywes, J. C. Blain, F. Del Frate, and J. W. Szostak, Nonenzymatic copying of rna templates containing all four letters is catalyzed by activated oligonucleotides, eLife **5**, 10.7554/elife.17756 (2016).

[8] E. Hänle and C. Richert, Enzyme-free replication with two or four bases, Angew. Chem. Int. Ed. **57**, 8911–8915 (2018).

[9] M. Sosson, D. Pfeffer, and C. Richert, Enzyme-free ligation of dimers and trimers to RNA primers, Nucleic Acids Res. **47**, 3836 (2019).

[10] G. Leveau, D. Pfeffer, B. Altaner, E. Kervio, F. Welsch, U. Gerland, and C. Richert, Enzyme-free copying of 12 bases of RNA with dinucleotides, Angew. Chem. Int. Ed. **61**, e202203067 (2022).

[11] F. Welsch, E. Kervio, P. Tremmel, and C. Richert, Prolinyl nucleotides drive enzyme-free genetic copying of RNA, Angew. Chem. Int. Ed. **62**, e202307591 (2023).

[12] J. W. Szostak, An optimal degree of physical and chemical heterogeneity for the origin of life?, Philos. Trans. R. Soc. B **366**, 2894–2901 (2011).

[13] E. Kervio, A. Hochgesand, U. E. Steiner, and C. Richert, Templating efficiency of naked dna, Proc. Natl. Acad. Sci. USA **107**, 12074–12079 (2010).

[14] K. Leu, B. Obermayer, S. Rajamani, U. Gerland, and I. A. Chen, The prebiotic evolutionary advantage of transferring genetic information from rna to dna, Nucleic Acids Res. **39**, 8135–8147 (2011).

[15] S. Rajamani, J. K. Ichida, T. Antal, D. A. Treco, K. Leu, M. A. Nowak, J. W. Szostak, and I. A. Chen, Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication, J. Am. Chem. Soc. **132**, 5880–5885 (2010).

[16] T. Göppel, B. Obermayer, I. A. Chen, and U. Gerland, A kinetic error filtering mechanism for enzyme-free copying of nucleic acid sequences, biorXiv 10.1101/2021.08.06.455386 (2021).

[17] L. Zhou, D. Ding, and J. W. Szostak, The virtual circular genome model for primordial RNA replication, RNA **27**, 1 (2021).

[18] D. Ding, L. Zhou, S. Mittal, and J. W. Szostak, Experimental tests of the virtual circular genome model for nonenzymatic RNA replication, J. Am. Chem. Soc. **145**, 7504 (2023).

[19] P. Chamanian and P. G. Higgs, Computer simulations of template-directed RNA synthesis driven by temperature cycling in diverse sequence mixtures, PLoS Comput. Biol. **18**, e1010458 (2022).

[20] J. H. Rosenberger, T. Göppel, P. W. Kudella, D. Braun, U. Gerland, and B. Altaner, Self-assembly of informational polymers by templated ligation, Phys. Rev. X **11**, 031055 (2021).

[21] T. Göppel, J. H. Rosenberger, B. Altaner, and U. Gerland, Thermodynamic and kinetic sequence selection in enzyme-free polymer self-assembly inside a nonequilibrium RNA reactor, Life **12**, 567 (2022).

[22] G. Laurent, T. Göppel, D. Lacoste, and U. Gerland, Emergence of homochirality via template-directed ligation in an RNA reactor, PRX Life **2**, 013015 (2024).

[23] T. Walton and J. W. Szostak, A kinetic model of nonenzymatic RNA polymerization by cytidine-5'-phosphoro-2-aminoimidazolide, Biochemistry **56**, 5739 (2017).

[24] J. G. Wetmur and N. Davidson, Kinetics of renaturation of DNA, J. Mol. Biol. **31**, 349 (1968).

[25] W. H. Braunlin and V. A. Bloomfield, Proton NMR study of the base-pairing reactions of d(GGAATTCC): salt effects on the equilibria and kinetics of strand association, Biochemistry **30**, 754 (1991).

[26] M. Livshits and A. Mirzabekov, Theoretical analysis of the kinetics of DNA hybridization with gel-immobilized oligonucleotides, Biophys. J. **71**, 2795 (1996).

[27] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, Proc. Natl. Acad. Sci. USA **101**, 7287 (2004).

[28] C. Kriebisch, L. Burger, O. Zozulia, M. Stasi, A. Floroni, D. Braun, U. Gerland, and J. Boekhoven, Template-based information transfer in chemically fueled dynamic combinatorial libraries, Nat. Chem. 10.1038/s41557-024-01570-5 (2024).

[29] O. D. Toparlak, L. Sebastianelli, V. Egas Ortuno, M. Karki, Y. Xing, J. W. Szostak, R. Krishnamurthy, and S. S. Mansy, Cyclophospholipids enable a protocellular life cycle, ACS Nano **17**, 23772–23783 (2023).

[30] M. W. Powner, B. Gerland, and J. D. Sutherland, Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions, Nature **459**, 239–242 (2009).

[31] J. Xu, M. Tsanakopoulou, C. J. Magnani, R. Szabla, J. E. Šponer, J. Šponer, R. W. Góra, and J. D. Sutherland, A prebiotically plausible synthesis of pyrimidine $\beta$-ribonucleosides and their phosphate derivatives involving photoanomerization, Nat. Chem. **9**, 303–309 (2016).

[32] J. Xu, N. J. Green, C. Gibard, R. Krishnamurthy, and J. D. Sutherland, Prebiotic phosphorylation of 2-thiouridine provides either nucleotides or dna building blocks via photoreduction, Nat. Chem. **11**, 457–462 (2019).

[33] D. Whitaker and M. W. Powner, Prebiotic synthesis and triphosphorylation of 3'-amino-TNA nucleosides, Nat. Chem. **14**, 766–774 (2022).

[34] D. J. Ritson and J. D. Sutherland, Thiophosphate photochemistry enables prebiotic access to sugars and terpenoid precursors, Nat. Chem. **15**, 1470–1477 (2023).

[35] W. K. Johnston, P. J. Unrau, M. S. Lawrence, M. E. Glasner, and D. P. Bartel, RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension, Science **292**, 1319 (2001).

[36] U. F. Müller and D. P. Bartel, Improved polymerase ribozyme efficiency on hydrophobic assemblies, RNA **14**, 552 (2008).

[37] A. Wochner, J. Attwater, A. Coulson, and P. Holliger, Ribozyme-catalyzed transcription of an active ribozyme, Science **332**, 209 (2011), publisher: American Association for the Advancement of Science.

[38] A. D. Pressman, Z. Liu, E. Janzen, C. Blanco, U. F. Müller, G. F. Joyce, R. Pascal, and I. A. Chen, Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating RNA, J. Am. Chem. Soc. **141**, 6213 (2019).

[39] W. G. Scott, L. H. Horan, and M. Martick, The hammerhead ribozyme, in *Prog. Mol. Biol. Transl. Sci.*, Vol. 120 (Elsevier, 2013) pp. 1–23.

[40] L. Zhou, D. K. O'Flaherty, and J. W. Szostak, Assembly of a ribozyme ligase from short oligomers by nonenzymatic ligation, J. Am. Chem. Soc. **142**, 15961 (2020).

[41] C. B. Mast and D. Braun, Thermal trap for DNA replication, Phys. Rev. Lett. **104**, 188102 (2010).

[42] C. B. Mast, S. Schink, U. Gerland, and D. Braun, Escalation of polymerization in a thermal gradient, Proc. Natl. Acad. Sci. USA **110**, 8030–8035 (2013).

[43] M. Kreysing, L. Keil, S. Lanzmich, and D. Braun, Heat flux across an open pore enables the continuous replication and selection of oligonucleotides towards increasing length, Nat. Chem. **7**, 203–208 (2015).

[44] A. Ianeselli, A. Salditt, C. Mast, B. Ercolano, C. L. Kufner, B. Scheu, and D. Braun, Physical non-equilibria for prebiotic nucleic acid chemistry, Nat. Rev. Phys. **5**, 185–195 (2023).