life

MDPI

Article

Hitting Times of Some Critical Events in RNA Origins of Life

Caleb Deen Bastian [1,*,†] **and Hershel Rabitz** [1,2,†]

[1] Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA; hrabitz@princeton.edu
[2] Department of Chemistry, Princeton University, Princeton, NJ 08544, USA
* Correspondence: cbastian@princeton.edu
† These authors contributed equally to this work.

Abstract: Can a replicase be found in the vast sequence space by random drift? We partially answer this question through a proof-of-concept study of the times of occurrence (hitting times) of some critical events in the origins of life for low-dimensional RNA sequences using a mathematical model and stochastic simulation studies from Python software. We parameterize fitness and similarity landscapes for polymerases and study a replicating population of sequences (randomly) participating in template-directed polymerization. Under the ansatz of localization where sequence proximity correlates with spatial proximity of sequences, we find that, for a replicating population of sequences, the hitting and establishment of a high-fidelity replicator depends critically on the polymerase fitness and sequence (spatial) similarity landscapes and on sequence dimension. Probability of hitting is dominated by landscape curvature, whereas hitting time is dominated by sequence dimension. Surface chemistries, compartmentalization, and decay increase hitting times. Compartmentalization by vesicles reveals a trade-off between vesicle formation rate and replicative mass, suggesting that compartmentalization is necessary to ensure sufficient concentration of precursors. Metabolism is thought to be necessary to replication by supplying precursors of nucleobase synthesis. We suggest that the dynamics of the search for a high-fidelity replicase evolved mostly during the final period and, upon hitting, would have been followed by genomic adaptation of genes and to compartmentalization and metabolism, effecting degree-of-freedom gains of replication channel control over domain and state to ensure the fidelity and safe operations of the primordial genetic communication system of life.

Keywords: RNA world; stochastic simulation algorithm; random counting measure; measure-kernel-function; ordinary differential equation; high dimensional model representation; global sensitivity analysis; fitness and similarity sequence landscapes; hitting times; survival analysis

Citation: Bastian, C.D.; Rabitz, H. Hitting Times of Some Critical Events in RNA Origins of Life. *Life* **2021**, *11*, 1419. https://doi.org/10.3390/life11121419

Academic Editor: Bruce Damer

Received: 1 November 2021
Accepted: 11 December 2021
Published: 17 December 2021

1. Introduction

The origins of life, abiogenesis, is a matter of high importance, for it gives insight into the distribution of life in the universe. We focus on the RNA world hypothesis, where life began with self-replicating RNA molecules that can evolve under Darwinian evolution, following necessary conditions of compartmentalization and metabolism, for geometry and synthesis of nucleobases from metabolic precursors, respectively. Self-replicating sets of RNA were proposed first by Tibor Ganti [1,2] and have been studied by many others [3–6]. This is an information-centric perspective on abiogenesis, representing the putative beginning of genomic Darwinian evolution. Information centrism interprets a living organism as an operating genetic communication system in some connected domain that encodes and decodes genomic state relative to a replication channel.

While genomics, epigenomics, and transcriptomics of modern-day organisms are based on DNA, RNA, and epigenetic marks such as DNA methylation, RNA origins in their purest form concern the dual-function of RNA as an informational polymer and ribozyme. This article is similar in spirit to works in the 1970s through the 1990s, including

Manfred Eigen's works on replicating sets of RNA [7,8]. Clues to the RNA world, among others, are found in the nucleotide moieties in acetyl coenzyme A and vitamin B12, the structure of the ribosome as a ribozyme [9] and moreover the centrality of RNA to the translation system, and the existence of viroids [10]. A putative canonical RNA origins sequence involves RNA dependent RNA polymerase (RdRp) ribozyme, whose first gene is perhaps the Hammerhead (HH) ribozyme, enabling rolling-circle amplification of the sequence [11]. This setup is unnecessary, as the first role of RNA may not have been template-assisted polymerization but instead based on mechanisms of RNA recombination and networking [12]. If we assume the RdRp sequence is size 200 nucleotides, then there are $4^{200} \approx 10^{120}$ sequences. Starting from some population of interacting RNA molecules, we are interested in the times of first occurrence of critical events. Evolution seems to have concluded this search for a high-fidelity replicator in a fairly short period of time, i.e., within 400 million years of the Earth having a stable hydrosphere [5]. RdRp's are known to be very ancient enzymes, are necessary to all viruses with RNA genomes, and have been proposed to have originated from junctions of proto-tRNAs relative to the context of a primitive translation system [13]. Recent work has shown that replicative RNA and DNA polymerases have a common ancestor of a RdRp [14]. Directed evolution, selecting on polymerization, is a potential way of identifying such a ribozyme. Directed compartmentalized self-replicating systems (RNA or DNA polymerases) mimicking prebiotic evolution have been demonstrated whereby polymerases are selected on their ability to replicate their own encoding gene [15]. Directed evolution has identified a RdRp that can replicate its evolutionary ancestor, an RNA ligase ribozyme; however, at increased activity, it has reduced fidelity and cannot maintain the integrity of its information [16].

A subtlety to the RNA origins argument is that template-directed polymerization by its nature requires two copies of the RdRp sequence, one for the polymerase and another for the template. This makes the RNA origins search extend until two copies are discovered. Cross reactions with other species also may influence the search time for the high-fidelity replicators. The clay mineral montmorillonite, which is common on Earth, can catalyze RNA oligomerization [17]; however, the utility of montmorillonite in these activities is not thought to be sufficient for origins, having been extensively studied [18,19]. Interestingly, it has been proposed that clay not only promotes origins, but constitutes it, which then later gave rise to RNA-based replication [20]; such mineral life as a genetic communication system has a high mutation rate and is degenerate.

Theories of abiogenesis study metabolism [21], cellular compartmentalization [22–24], hydrothermal vent chemical gradient energy [25], or hot springs [26–28], and so on, to define geochemical settings suitable for origins. These settings are compatible with RNA origins. Compartmentalization leading to selection on random sequences has been explored by studying environment forcing in hot springs and their effects on sequence identification [26–28], where the hypothesis is that fluctuations in environment forcing through cycling of wet, dry, and moist phases of lipid-encapsulated sequences subject sequences to combinatorial selection and identify structural and catalytic functions from the initial system state of random sequences. These functions include metabolic activity, pore formation, and structural stabilization. We assume the prebiotic molecular inventories of RNA and its precursors are provided by meteorites [29] and/or by Miller–Urey processes [30], such as from formamide [31] or many phase synthesis [32,33]. For energy and environmental factors, we consider a variant of Darwin's "warm little pond", where the putative environment for RNA origins of life is an icy pond with geothermal activity, a hot spring, or perhaps a hydrothermal vent: ice and cold temperature facilitate complexing of single strands into double strands and polymerization [34], and heat (energy) facilitates dissociation of double strands into single strands. More information on abiotic sources of organic compounds, mechanisms of synthesis and function of macromolecules, energy sources, and environmental factors can be found in the literature [35].

RNA origins have attracted many modeling efforts and analyses [36,37]. The concept of a self-replicating set of RNA molecules was initially studied by Manfred Eigen [7],

wherein he studied the error-threshold of the critical fidelity to the main information. Two-dimensional spatial modeling has been applied in which reactions occur locally with finite diffusion, suggesting a spatially localized stochastic transition [38], simulated using Gillespie's stochastic simulation algorithm (SSA) [39]. Another model is of autocatalytic sets of collectively reproducing molecules, which has been developed in reflexively autocatalytic food-generated (RAF) theory [40]. Various physics-based analyses have been conducted, such as in light of Bayesian probability, thermodynamics, and critical phenomena [41]. Systems of quasi-species based on the principle of natural self-organization called hypercycles employ non-equilibrium auto-catalytic reactions [8]. Nonlinear kinetic models for polymerization have been used to study the emergence of self-sustaining sets of RNA molecules from monomeric nucleotides [42]. Theoretical analysis has been conducted into RNA origins. Attention has been drawn to an evolving population of dynamical systems and how dynamics affect the error threshold of early replicators and possibly towards compartmentalization conveying hypercycles [43]. String-replicator dynamics have been studied and properties suggested to be necessary to RNA origins, including the ability to operate a functional genetic communication system and ecological and evolutionary stability [44–46]. A variety of pre-RNA worlds have been suggested, with RNA being preceded or augmented by alternative informational polymers, such as other nucleic acids [47], beta amyloid [48], polycyclic aromatic hydrocarbons [49], lipids [24], peptides [50], and so on. It seems that pre-RNA worlds existed independent of the RNA world in the sense that they are not ancestral to the RNA world, and that these worlds may have had non-trivial interactions with the RNA world.

A key concept of stochastic systems is that of a hitting time: the time of the first occurrence of some event. We develop a simple mathematical model at the sequence level to represent the synthesis and function of RNA molecules in order to gain insight into the hitting times of various critical events of RNA origins of life. The idea is to study the surface of hitting times in terms of the structure of the system. The model lacks many features of realism, such as sequence size variability, finite sources of "food" (activated nucleotides in our context), limited diffusion rates, poor system mixing, and so on, in order to concentrate on the process as a search problem. The notion of fitness landscapes has been studied extensively in evolutionary biology [51]. Landscape topology has been considered in an Opti-Evo theory, which assumes sufficient environmental resources and argues that fitness landscapes do not contain "traps" and globally optimal sequences form a connected level-set [52].

We describe the model in Section 2, where we define a replicating reaction network, whose random realizations are constructed using SSA. We describe hitting times as key random variables of interest and characterize polymerization as a transition kernel. In Section 3, we conduct and discuss simulation studies based on SSA, where we analyze the structure of the polymerase measures and the input–output and survival behaviors of the hitting times given the parameters of the system. In Section 4, we end with conclusions.

2. Materials and Methods

We define $M = \{\text{Adenine}, \text{Uracil}, \text{Guanine}, \text{Cytosine}\}$, abbreviated as $M = \{A, U, G, C\}$, corresponding to the RNA bases. We study the space of sequences having length n, that is, $E = M^n$ with space of possibilities (a σ-algebra) $\mathcal{E} = 2^E$, so that the pair $(E, \mathcal{E})$ is the measurable space of all RNA sequences of length n with $|E| = 4^n$. We let ν be a probability measure (distribution) on $(E, \mathcal{E})$, giving the probability space triple $(E, \mathcal{E}, \nu)$. Appendix A gives an overview of $(E, \mathcal{E}, \nu)$. A related space, though not utilized in this article, is the space of all RNA sequences up to length n, $E^* = \cup_{i=1}^{n} H^i$ with $\mathcal{E}^* = 2^{E^*}$ and size $|E^*| = \frac{4}{3}(|E| - 1)$. We denote the collection of non-negative $\mathcal{E}$-measurable functions by $\mathcal{E}_{\geq 0}$.

We build a simplified mathematical model for the time-evolution of a population of interacting RNA molecules in solution. Let X_t be the population at time $t \in \mathbb{R}_{\geq 0}$ with initial population X_0. X_t is a multiset, that is, it is a set containing elements possibly with

repeats. We assume that the system is well mixed and has access to an infinite source of activated nucleotides.

The complement of $x \in E$ is denoted $x^c \in E$, attained using the base-pairing A with U and C with G. Let

$$\mathfrak{h}(x, y) \in \{0, 1, \ldots, n\} \quad \text{for} \quad x, y \in E \tag{1}$$

be the Hamming distance between $x, y \in E$ as the number of positions in x and y where the nucleotides differ. We have $\mathfrak{h}(x, x) = 0$ and $\mathfrak{h}(x, x^c) = n$.

2.1. Core Model

We describe the reaction network of the system below. We simulate trajectories of the system using the stochastic simulation algorithm (SSA) [39], to simulate exact trajectories for the evolution of stochastic reaction networks here. SSA forms a Markovian process, where the arrival of reactions follows a Poisson (point) process, and assumes that the reaction volume is well mixed and homogeneous, with all parts of the system accessible for reactions. Reactions across various disjoint volume elements of the system are dependent. We do not consider the effects of finite diffusion, which effects a length scale above which disjoint volume elements are effectively independent [38].

2.1.1. System

We model the population of sequences which can form double-stranded helices, dissociate, and replicate with mutation with replicator fitness and sequence specificity. We interpret each system element as a set, either containing one element—a single-stranded sequence—indicated as $\{x\}$—or two elements, single and complementary stranded sequences, indicated as $\{x\} \cup \{x^c\} = \{x, x^c\}$, where $x, x^c \in E$ (double bracket notation indicates a collection, or set, of sets, i.e., $\{\{x\}, \{y\}, \{z\}, \cdots\}$). We define system elements as sets

$$\begin{aligned} \bar{E} &\equiv \{\{x\} : x \in E\} \\ \bar{F} &\equiv \{\{x, x^c\} : x \in E\} \\ \bar{G} &\equiv \bar{E} \cup \bar{F} \end{aligned}$$

with respective σ-algebras, $\bar{\mathcal{E}} = 2^{\bar{E}}$, $\bar{\mathcal{F}} = 2^{\bar{F}}$ and $\bar{\mathcal{G}} = 2^{\bar{G}}$. The sizes are $|\bar{E}| = 4^n$ and $|\bar{F}| = 4^n/2$, and $|\bar{G}| = 3 \times 4^n/2$. The set $\bar{E}$ contains single-stranded sequences, whereas the set $\bar{F}$ contains double-stranded sequences; the set $\bar{G}$ is the union of $\bar{E}$ and $\bar{F}$, containing both single and double stranded sequences. These sets are necessary to track the various sequences (single and double stranded). The reaction network of the system is given by

$$x + x^c \xrightarrow{k_{ds}} x \cup x^c \tag{2}$$

$$x \cup x^c \xrightarrow{k_{ss}} x + x^c \tag{3}$$

$$x + y \xrightarrow{k_{rep}(x,y)} x + y + y^c_*. \tag{4}$$

and expressed in terms of sets

$$\begin{aligned} \{x\}, \{x^c\} &\xrightarrow{k_{ds}} \{x, x^c\} \\ \{x, x^c\} &\xrightarrow{k_{ss}} \{x\}, \{x^c\} \\ \{x\}, \{y\} &\xrightarrow{k_{rep}(x,y)} \{x\}, \{y\}, \{y^c_*\}. \end{aligned}$$

Reaction (2) is double-strand formation from complementary single-strands with reaction rate k_{ds}. Reaction (3) is the dissociation of double-strands into single-strands, caused by a heat source, with reaction rate k_{ss}. Reaction (4) is template-directed polymerization of a single-strand (the template) by another single-strand (the polymerase), producing a

single-strand complementary to the template with some fidelity, with reaction rate k_{rep} (which functionally depends on the polymerase and template sequences).

The polymerization reaction rate is defined by

$$k_{rep}(x,y) = af(x)s(x,y) \in (0,a] \quad \text{for} \quad x,y \in E \tag{5}$$

where $a > 0$ is a positive constant, $f : E \mapsto (0,1]$ is the replicative *fitness* of x (as a polymerase) and $s : E \times E \mapsto (0,1]$ is the *similarity* between x and y, a symmetric function. Finally, x replicates y, outputting a version y^c_* with mutations, where each nucleotide position has fidelity probability $p : E \mapsto (0,1]$. Note that the similarity function can be either trivial/constant, i.e., $s(x,y) = 1$, or non-trivial. For example, if we assume sequence similarity to correlate to spatial proximity of sequences, as assumed below, then $s(x,y)$ is non-trivial.

2.1.2. High-Fidelity Set

To define a high-fidelity set, pick an arbitrary subset of sequences $\mathfrak{R} \subset E$ as high-fidelity replicators of size $r = |\mathfrak{R}|$. We define $\mathfrak{R}$ two ways.

We define $\mathfrak{R}$ using a product of non-empty random nucleotide subsets $\{A_i \subseteq M : i = 1,\ldots,n\}$ for each nucleotide position

$$\mathfrak{R} = A_1 \times \cdots \times A_n$$

so that $r = \prod_{i=1}^{n} |A_i| = 1^{r_1}2^{r_2}3^{r_3}4^{r_4}$ where $r_2 = |\{A_i : |A_i| = 2\}|$, etc. and $r_1 + r_2 + r_3 + r_4 = n$. For simplicity, we assume $|A_i| \in \{1,4\}$ with fraction 4 being $q \in (0,1)$. Thus, $\mathfrak{R}$ is a subset of E defined as a product space.

For another construction of R, we define a finite union of m random sequences $\mathfrak{R} = \{x_1,\ldots,x_m\}$.

2.1.3. Distance

We define the Hamming distance $\mathfrak{H}$ between sequence x and high-fidelity sequence set $\mathfrak{R}$ as

$$\mathfrak{H}(x,\mathfrak{R}) = \min\{\mathfrak{h}(x,y) : y \in \mathfrak{R}\} \in \{0,1,\ldots,n\} \quad \text{for} \quad x \in E. \tag{6}$$

2.1.4. Fitness

We define "tent-pole" fitness f_k of sequence $x \in E$ and high-fidelity sequence set $\mathfrak{R}$ for curvature parameter $k \in \mathbb{R}_{\geq 0}$ as

$$f_k(x,\mathfrak{R}) = \exp[-k\,\mathfrak{H}(x,\mathfrak{R})] \in (0,1] \quad \text{for} \quad x \in E. \tag{7}$$

The maximums are the sequences of the high-fidelity sequence set $\mathfrak{R}$, which are the "points" or "poles" of the surface, with exponential decay into the remainder of the space in string distance. The strength of the decay is governed by parameter k, called the curvature parameter, which can be specified through the value of fitness at $\mathfrak{H}(x,\mathfrak{R}) = n$ (sequence dimension),

$$k = -\frac{\log(f_k(x,\mathfrak{R}))}{n} \tag{8}$$

Appendix B describes other fitness functions.

2.1.5. Similarity

We define two cases for the similarity function appearing in the reaction rate of template-directed polymerization. The first case is the trivial (constant) case where the

$$s_b(x,y) = b \in (0,1] \quad \text{for} \quad (x,y) \in E \times E.$$

This assumes that there is no mechanism by which sequence specificity is selected for, such as in the case that polymerases should evolve to generically well replicate sequences,

including their own. This means that they will spend much of their time replicating other sequences.

For the second case of a non-trivial similarity function, we note that RNA origins of life are thought to be a spatially localized stochastic transition, where high-fidelity replicators are found concentrated in foci, following from the increased replicative mass of the replicators. Hence, we implicitly encode spatial information through a non-trivial similarity function, based on a distance function, which increases the replicative system mass for similar (and here nearby) high-fidelity replicators, that is, if they're similar, then they're likely proximal. In the following definition, we use the notation $\wedge$ for the minimum of two numbers, $x \wedge y = \min\{x, y\}$. Distance S is defined between two sequences $x, y \in E$ as

$$S(x,y) = \mathfrak{h}(x,y) \wedge \mathfrak{h}(x,y^c) \wedge \mathfrak{h}(x^c,y) \wedge \mathfrak{h}(x^c,y^c) \in \{0,1,\ldots,n\} \quad \text{for} \quad x,y \in E. \tag{9}$$

Similarity s_k of sequences $x, y \in E$ for curvature parameter $k \in \mathbb{R}_{\geq 0}$ is defined in terms of exponential decay as

$$s_k(x,y) = \exp[-k\,S(x,y))] \in (0,1] \quad \text{for} \quad x,y \in E. \tag{10}$$

Presently, replicators can replicate other sequences well but not their own [34]. There may exist RdRps that are excellent polymerases and, in conjunction with RNA hammerhead ribozyme, engage in rolling circle amplification of the polymerase-hammerhead sequence (genome) so that the amplification process is self-cleaving. This results in a large increase in replicative mass due to the super-exponential growth in the population of the high-fidelity replicators within a small volume. In the context of SSA, the similarity function here is an ansatz for spatial locality.

2.1.6. Fidelity

Finally, polymerization fidelity probability for curvature $k \in \mathbb{R}_{\geq 0}$ is defined as

$$p_k(x,R) = f_k(x,R) \in (0,1] \quad \text{for} \quad x \in E. \tag{11}$$

Note that $f_k(x,R) = p_k(x,R) = 1$ for high-fidelity sequences $x \in \mathfrak{R}$.

Note that fitness, similarity, and fidelity are defined for single-stranded sequences $(E, \mathcal{E})$.

2.1.7. Counting Representation

The process $X = (X_t)_{t \in \mathbb{R}_{\geq 0}}$ is the time-evolution of the system. Recall that X_t is a multiset. X_t contains the individual single stranded molecules in the set $\bar{E}$, i.e., sequences $\{x\} \in \bar{E}$ and double stranded molecules in the set $\{x, x^c\} \in \bar{F}$, with overall set $\bar{G} = \bar{E} \cup \bar{F}$ having size $m = |\bar{G}| = 3|\bar{E}|/2$. Note that, in the set representation, there is symmetry $\{x, x^c\} = \{x^c, x\}$, so that the size of the double-stranded set is equal to $|\bar{F}| = |\bar{E}|/2$. The system evolution X_t induces a random counting measure N_t on the overall space of single and double stranded sequences $(\bar{G}, \bar{\mathcal{G}})$ as

$$N_t(A) = \sum_{x \in X_t} \mathbb{I}_A(x) \quad \text{for} \quad A \in \bar{\mathcal{G}} \tag{12}$$

The total count, that is, the total number of molecules, is $K_t \equiv |X_t| = N_t(\bar{G})$. We assume that the counter N_t is maintained for all times $t \in \mathbb{R}_{\geq 0}$.

2.1.8. Reaction Rates

The total reaction rate is given by the sum of the individual reaction rates

$$\check{k}(t) = \check{k}_{ds}(t) + \check{k}_{ss}(t) + \check{k}_{rep}(t) \quad \text{for} \quad t \in \mathbb{R}_{\geq 0}$$

where the reaction rate of double-strand formation from complementary single-strands is given by

$$\check{k}_{ds}(t) = \frac{1}{2} \sum_{\{x\} \in \bar{E}} k_{ds} N_t(\{x\}) N_t(\{x^c\}),$$

the reaction rate of dissociation of double-strands into single-strands is given by

$$\check{k}_{ss}(t) = \sum_{\{x, x^c\} \in \bar{F}} k_{ss} N_t(\{x, x^c\}),$$

and the reaction rate for template-directed polymerization of a single-strand by another single-strand (the polymerase) is given by

$$\check{k}_{rep}(t) = \sum_{(\{x\}, \{y\}) \in \bar{E}^2} k_{rep}(x, y) N_t(\{x\}) (N_t(\{y\}) - \mathbb{I}(x = y))$$

The first reaction rate is a sum over the single-strands of $\bar{E}$ with size 4^n. The second is a sum over double-strands $\bar{F}$ with size $4^n/2$. The third reaction is a sum over the product space of single stranded sequences, $\bar{E}^2$, with $4^{2n} = 16^n$ number of elements. Therefore, $\check{k}(t)$ requires $4^n(\frac{3}{2} + 4^n)$ elements to be evaluated for every reaction. Clearly, direct representation on the full space is very expensive and impractical for even modest n. One obvious way to improve efficiency is not summing over the zero elements. We define sets

$$\mathcal{X}_1 = \{x \in \text{set}(X_t) : |x| = 1\} \tag{13}$$

and

$$\mathcal{X}_2 = \{x \in \text{set}(X_t) : |x| = 2\} \tag{14}$$

as the unique single and double stranded sequences of the system. Then, direct calculations of the reaction rates are

$$\check{k}_{ds}(t) = \frac{1}{2} \sum_{\{x\} \in \mathcal{X}_1} k_{ds} N_t(\{x\}) N_t(\{x^c\})$$

and

$$\check{k}_{ss}(t) = \sum_{\{x, x^c\} \in \mathcal{X}_2} k_{ss} N_t(\{x, x^c\})$$

and

$$\check{k}_{rep}(t) = \sum_{(\{x\}, \{y\}) \in \mathcal{X}_1^2} k_{rep}(x, y) N_t(\{x\}) (N_t(\{y\}) - \mathbb{I}(x = y)).$$

For this approach, the replication rate has quadratic dependence on $|\mathcal{X}_1|$. Using the reaction rates, the system may be exactly simulated using SSA. The reaction at time t with rate $\check{k}(t)$ occurs over time interval $\triangle t \sim \text{Exponential}(1/\check{k}(t))$. As the reaction rate $\check{k}(t)$ increases with increasing number of molecules $K_t = N_t(\bar{G})$, the reaction rate increases and reaction duration $\triangle t$ decreases over time. The natural consequence of increasing process intensity is that the system speeds up.

The quadratic dependence may still be too expensive for large simulations. Appendix E describes Monte Carlo approximation of the reaction rates.

2.2. Hitting Times

We define some hitting times. The initial population consists of I single-stranded sequences X_0, i.e., $|X_0| = I$. We define the hitting time τ for the time of the first replication event

$$\tau_{\text{rep}} = \inf\{t \in \mathbb{R}_{\geq 0} : K_t > I\}. \tag{15}$$

We define the hitting time τ for the appearance of sequences in the high-fidelity sequence set $\mathfrak{R}$,

$$\tau_{\mathfrak{R}} = \inf\{t \in \mathbb{R}_{\geq 0} : \mathbb{I}_{\{1,2,\cdots\}}(|\mathfrak{R} \cap X_t|) = 1\}. \tag{16}$$

Put $X_{\mathfrak{R}} = \mathfrak{R} \cup \{\{x, x^c\} : x \in \mathfrak{R}\}$ and define the volume fraction of high-fidelity sequences $\mathfrak{R}$ at time t as

$$V(t) = \frac{N_t(X_{\mathfrak{R}})}{K_t}. \tag{17}$$

We define the hitting time τ where high-fidelity sequences of $\mathfrak{R}$ emerge and reach a minimum volume fraction,

$$\tau_{\min} = \inf\{t \in \mathbb{R}_{\geq 0} : t \geq \tau_{\mathfrak{R}}, V(t) \leq V(s) \text{ for } \tau_{\mathfrak{R}} \leq s \leq t\}. \tag{18}$$

This hitting time reflects that period wherein a high-fidelity replicator has been identified yet there exists no complementary high-fidelity sequence for amplification, hence the system diversity continues to increase, decreasing the concentrations of all extant sequences as more sequences are discovered. The minimum hitting time captures the duration of time the high-fidelity replicator exists by itself. We define the hitting time τ for the time high-fidelity sequences in $\mathfrak{R}$ constitutes some volume fraction $v \in (0, 1]$ of the population,

$$\tau_v = \inf\{t \in \mathbb{R}_{\geq 0} : V(t) \geq v\}. \tag{19}$$

In practice for simulations, τ_v is censored based on some total number of reactions, that is, if the volume fraction is not achieved by n reactions, $\tau_v = \infty$ because there is no arrival time.

For SSA, we specify a maximum number of reactions N to simulate. We have parameters $\theta \in \Theta$ for τ, such as landscape curvature k, sequence dimension n, etc. Therefore, $\tau(\theta)$ is right-censored with value ∞ at simulation time a, as some simulations will stop at time a with no arrival time. These are censoring events. For fixed θ, the $\tau(\theta)$ is a random variable, due to the stochastic nature of SSA. Hence, for each parameter vector θ, we attain a set of M realizations of hitting time τ as

$$\mathcal{T}(\theta) = \{\tau_i(\theta) : i = 1, \ldots, M\}. \tag{20}$$

For convenience, we assume that the realizations $\mathcal{T}(\theta)$ are ordered by non-censored followed by censored.

2.2.1. Functional Structure

For each parameter vector θ, we record two values: the number of hitting events in the hitting time set $\mathcal{T}$

$$g(\theta) = |\{x \in \mathcal{T}(\theta) : x < \infty\}| \in \{0, 1, \ldots, M\} \tag{21}$$

and the average of the hitting time τ

$$f(\theta) = \begin{cases} \frac{1}{g(\theta)} \sum_{i=1}^{g(\theta)} \tau_i(\theta) & \text{if } g(\theta) > 0 \\ 0 & \text{if } g(\theta) = 0 \end{cases} \in \mathbb{R}_{\geq 0} \tag{22}$$

To describe the functional structure of the average hitting time $f(\theta)$, we require a classifier which determines whether or not there are zero hittings $g(\theta) = 0$ and a regressor for the value of $f(\theta)$ for hittings $g(\theta) > 0$. We assume that the parameters $\theta = (\theta_1, \cdots, \theta_n)$ are randomly sampled according to distribution $\nu = \prod_i \nu_i$ and the hitting times recorded. High dimensional model representation (HDMR) may be attained for the classifier (as a

probabilistic discriminative model) and the regressor of $f(\theta)$. For the regressor, we have HDMR expansion

$$f(\theta_1, \ldots, \theta_n) = f_0 + \sum_i f_i(\theta_i) + \sum_{i<j} f_{ij}(\theta_i, \theta_j) + \cdots + f_{1\cdots n}(\theta_1, \ldots, \theta_1)$$

The HDMR component functions $\{f_u\}$ convey a global sensitivity analysis, where, defining variance term

$$\sigma_u^2 = \mathbb{V}ar f_u = \int_\Theta f_u^2(\theta_u)\nu(\mathrm{d}\theta),$$

we have a decomposition of variance

$$\sigma_f^2 = \mathbb{V}ar f = \sum_{u\subseteq\{1,\cdots,n\}:|u|>0} \sigma_u^2$$

The normalized terms $\mathbb{S}_u = \sigma_u^2/\sigma_f^2$ are called sensitivity indices. Appendix F gives a brief description of global sensitivity analysis via HDMR.

2.2.2. Statistical Structure

A second analysis can be conducted on the hitting times $\mathcal{T}(\theta)$ for the parameter vector θ using reliability theory. Put random hitting set $\mathfrak{T}(\Theta) \equiv \{\mathcal{T}(\theta) : \theta \in \Theta\}$ where $\Theta = \{\theta_i\}$ is an independency of parameter values. For each parameter vector $\theta \in \Theta$, we partition the hitting times $\mathcal{T}(\theta)$ into C censored values with censor times $\mathcal{C}(\theta) = \{a_i(\theta)\}$ and $M - C$ non-censored (hitting) values $\mathcal{N}(\theta) = \{x \in \mathcal{T}(\theta) : x < \infty\}$. The likelihood is given by

$$L(\mathfrak{T}(\Theta)|\vartheta) = \prod_{\theta\in\Theta} \prod_{x\in\mathcal{N}(\theta)} f(x|\vartheta) \prod_{x\in\mathcal{C}(\theta)} R(x|\vartheta),$$

where f is the hitting time probability density function ('failure density') and R is censoring time distribution ('reliability distribution'), and ϑ are the parameters of the density and distribution functions. Note that f and R each specify each other, so ϑ are the common parameters. Reliability definitions are given in Appendix G.

We show reliability quantities in Table 1 for the two-parameter $(\alpha, \beta) \in (0, \infty)^2$ Weibull(α, β) distribution and Cox proportional hazard's model where γ is a vector of coefficients for the θ. We use the Python software *lifelines* for estimation of ϑ for the Weibull–Cox model from data [53]. The mean failure time $\mathbb{E}\tau(\theta)$ is given by

$$\mathbb{E}\tau(\theta) = \int_0^\infty t\, f(t, \theta|\alpha, \beta, \gamma) = \beta e^{-\gamma\cdot\theta/\alpha}\Gamma(1 + \tfrac{1}{\alpha})$$

We have second moment

$$\int_0^\infty t^2\, f(t, \theta|\alpha, \beta, \gamma) = \beta^2 e^{-2\gamma\cdot\theta/\alpha}\Gamma(1 + \tfrac{2}{\alpha})$$

giving variance

$$\mathbb{V}ar\tau(\theta) = \beta^2 e^{-2\gamma\cdot\theta/\alpha}(\Gamma(1 + \tfrac{2}{\alpha}) - \Gamma^2(1 + \tfrac{1}{\alpha}))$$

Thus, if $\gamma < 0$, then $\mathbb{E}\tau(\theta)$ and $\mathbb{V}ar\tau(\theta)$ exponentially increase in θ.

Table 1. Weibull reliability model.

Name	Quantity	Baseline	Quantity	Proportional Hazards
Failure density	$f(t\|\alpha,\beta)$	$\frac{\alpha}{t}(\frac{t}{\beta})^{\alpha}e^{-(\frac{t}{\beta})^{\alpha}}$	$f(t,\theta\|\alpha,\beta,\gamma)$	$\frac{\alpha}{t}(\frac{t}{\beta})^{\alpha}e^{\gamma\cdot\theta}e^{-(\frac{t}{\beta})^{\alpha}e^{\gamma\cdot\theta}}$
Failure distribution	$F(t\|\alpha,\beta)$	$1-e^{-(\frac{t}{\beta})^{\alpha}}$	$F(t,\theta\|\alpha,\beta,\gamma)$	$1-e^{-(\frac{t}{\beta})^{\alpha}e^{\gamma\cdot\theta}}$
Reliability distribution	$R(t\|\alpha,\beta)$	$e^{-(\frac{t}{\beta})^{\alpha}}$	$R(t,\theta\|\alpha,\beta,\gamma)$	$e^{-(\frac{t}{\beta})^{\alpha}e^{\gamma\cdot\theta}}$
Cumulative hazard	$H(t\|\alpha,\beta)$	$(\frac{t}{\beta})^{\alpha}$	$H(t,\theta\|\alpha,\beta,\gamma)$	$(\frac{t}{\beta})^{\alpha}e^{\gamma\cdot\theta}$
Hazard rate	$h(t\|\alpha,\beta)$	$\frac{\alpha}{t}(\frac{t}{\beta})^{\alpha}$	$h(t,\theta\|\alpha,\beta,\gamma)$	$\frac{\alpha}{t}(\frac{t}{\beta})^{\alpha}e^{\gamma\cdot\theta}$

2.3. Surface Chemistries

The system given by (2) of polymerization with mutation requires two separate hitting events, one sequence in the high-fidelity set $x \in \mathfrak{R}$ and either another sequence in the high-fidelity set $x \in \mathfrak{R}$ or its complement $x^c \in E$, in order for high-fidelity replicators to maximally engage in templated-directed polymerization and achieve some fraction of the population. This setup of RNA polymerase action, requiring two such events for the polymerase and template, makes the hitting times long. Basically, the same information must be discovered twice before it can be used, which is unsatisfactory. We idealize polymerase activity conveyed by a non-RNA species, here clay, with the parameter k_{clay-p}, as clay itself is not thought to be capable of polymerization but is capable of oligomerization of RNA. The reactions are given by

$$\varnothing \xrightarrow{k_{clay-o}} x \tag{23}$$

$$x \xrightarrow{k_{clay-p}} x + x_*^c \tag{24}$$

where non-RNA polymerization has mutation with fidelity probability $p \in (0,1]$ and x_*^c is the complement of x with mutation. The reaction rates are given by

$$\check{k}_{clay-o}(t) = k_{clay-o}$$

and

$$\check{k}_{clay-p}(t) = k_{clay-p}N_t(\mathcal{X}_1)$$

Therefore, upon the first hitting of the high-fidelity replicators $\mathfrak{R}$ with sequence $x \in \mathfrak{R}$ through (4) or (23), x gives two high-similarity single-stranded sequences x and x_*^c through (24), which then may participate in template-directed RNA polymerization (4).

2.4. Reactions as Measure-Kernel-Functions

All the reactions $x \mapsto y$ which involve substrate x may be represented using transition kernels, which form linear operators. At each iteration of SSA, a reaction type is chosen, followed by a transition to a particular domain $(X,\mathcal{X})$ with distribution ν_t, followed by mapping into a codomain $(Y,\mathcal{Y})$ using transition probability kernel Q with distribution $\mu_t = \nu_t Q$. The notions of ν_t and Q involve measure-kernel-functions. The probability of transition of y into $B \in \mathcal{Y}$ given x is given by Q

$$Q(x,B) = \int_B Q(x,\mathrm{d}y) = \mathbb{P}(y \in B : x)$$

Appendix C recalls some facts about Q.

We define kernels Q for RNA and non-RNA polymerization to provide insight into the reactions. Consider X_t for some $t \in \mathbb{R}_{\geq 0}$. Recall that N_t is the random counting measure of X_t on single and double-stranded sequences $(E \cup F, 2^{E \cup F})$. For RNA and non-RNA polymerization, we take ν_t as a (random) probability measure for x in domain $(X,\mathcal{X})$ and describe a transition probability kernel Q from x into y in codomain $(Y,\mathcal{Y})$.

2.4.1. RNA Polymerization

For RNA polymerization, we have that

$$\{x\},\{y\} \mapsto \{x\},\{y\},\{y_*^c\}$$

which results in the creation of the single-stranded sequence $\{y_*^c\}$. The first dimension is the polymerase and the second is the template. We give a definition of the probability measure on the product space of sequences. Recall that $(\mathcal{E}\otimes\mathcal{E})_{\geq 0}$ denotes the collection of non-negative $\mathcal{E}\otimes\mathcal{E}$-measurable functions.

Definition 1 (Measure on domain ν_t)**.** *Let ν_t be a random probability measure on $(E\times E, \mathcal{E}\otimes\mathcal{E})$ formed by random counting measure N_t (12)*

$$\nu_t\{x,y\} = \frac{k_{rep}(x,y)N_t(\{x\})(N_t(\{y\}) - \mathbb{I}(x=y))}{\check{k}_{rep}(t)} \quad \textit{for} \quad (x,y)\in E\times E$$

with

$$\nu_t(f) = \sum_{(x,y)\in E\times E} \nu_t\{x,y\} f\circ(x,y) \quad \textit{for} \quad f\in(\mathcal{E}\otimes\mathcal{E})_{\geq 0} \tag{25}$$

We write $\nu_t(A) = \nu_t\mathbb{I}_A$ for $A\in\mathcal{E}\otimes\mathcal{E}$.

Because the first two coordinates are preserved under the mapping, we focus on the new dimension as a transition from $(E\times E, \mathcal{E}\otimes\mathcal{E})$ into $(E,\mathcal{E})$ using transition probability kernel Q. In this case, Q is defined by a $16^n\times 4^n$ matrix whose rows vectors (dimension 4^n) are probability vectors. The structure of Q follows from the polymerase replication with mutation, whereby each nucleotide position has fidelity probability $p: E\mapsto(0,1]$, which depends on the first dimension of $E\times E$. We put $p_x = p(x)$ for sequence $x\in E$. Now, we state a simple fact on the binomial structure of the number of mutations made by a polymerase.

Theorem 1 (Mutation distribution)**.** *The number of mutations by polymerase $x\in E$ on template $y\in E$ is distributed*

$$\mathfrak{h}(y^c, y_*^c) \sim \textit{Binomial}(n, 1-p_x) \quad \textit{for} \quad p_x\in(0,1)$$

with mean $n(1-p_x)$ and variance $np_x(1-p_x)$ and

$$\mathfrak{h}(y^c, y_*^c) \sim \textit{Dirac}(0) \quad \textit{for} \quad p_x = 1.$$

Now, we partition E into level sets $(\mathfrak{H}_i(y))$ by Hamming distance to the template complement y^c,

$$\mathfrak{H}_i(y) = \{x\in E : \mathfrak{h}(y^c, x) = i\} \quad \text{for} \quad i\in\{0,\ldots,n\}. \tag{26}$$

We define the transition kernel Q for RNA polymerization, where Q completely encodes RNA polymerization using Theorem 1.

Corollary 1 (Transition probability kernel Q)**.** *We have that the transition probability kernel Q for RNA polymerization is defined by*

$$Q((x,y),\mathfrak{H}_i(y)) = \binom{n}{i}(1-p_x)^i p_x^{n-i} \quad for \quad (x,y) \in E \times E, \quad i \in \{0,\ldots,n\}, \quad p_x \in (0,1)$$

and

$$Q((x,y),\{z\}) = \frac{1}{|\mathfrak{H}_i(y)|}\binom{n}{i}(1-p_x)^i p_x^{n-i} \quad for \quad (x,y) \in E \times E, \quad i = \{0,\ldots,n\}$$
$$z \in \mathfrak{H}_i(y), \quad p_x \in (0,1)$$

and

$$Q((x,y),\{y^c\}) = 1 \quad for \quad (x,y) \in E \times E, \quad p_x = 1.$$

RNA polymerization is defined by Q using the binomial structure of polymerase mutation. A more sophisticated model could be defined as a sum of Bernoulli random variables with varying success probabilities in the Poisson binomial distribution. This could be used to take into account polymerase mutation that varies with nucleotide position. Another idea is taking into account schemata such as repeats which destabilize the polymerase [54].

Proposition 1 (Measure on codomain μ_t)**.** *$\mu_t = \nu_t Q$ is a probability measure on $(E,\mathcal{E})$ defined by*

$$\mu_t(f) = \int_{E\times E} \nu_t(\mathrm{d}x,\mathrm{d}y)\int_E Q((x,y),\mathrm{d}z)f(z) \quad for \quad f \in \mathcal{E}_{\geq 0} \tag{27}$$

It is multiplication of ν_t as a 16^n dimension row vector with $16^n \times 4^n$ dimension matrix Q, giving a 4^n dimension row vector $\nu_t Q$. We write $\mu_t(A) = \mu_t \mathbb{I}_A$ for $A \in \mathcal{E}$.

Define the partition $(\mathfrak{H}_i)$ of E as

$$\mathfrak{H}_i \equiv \{x \in E : \min\{\mathfrak{H}(x,\mathfrak{R}),\mathfrak{H}(x^c,\mathfrak{R})\} = i\} \quad \text{for} \quad i \in \{0,\ldots,n\}. \tag{28}$$

Then, $\mu_t(\mathfrak{H}_i)$ for $i \in \{0,\ldots,n\}$ is the distribution on sequences by distance to R, i.e.,

$$\mu_t(\mathfrak{H}_i) = \sum_{x\in E} \mu_t\{x\}\mathbb{I}_{\mathfrak{H}_i}(x) \quad \text{for} \quad i \in \{0,\ldots,n\}$$

contains the instantaneous information of RNA polymerization.

A more general model for replication is where polymerase activity is tied to geometry, i.e., compartmentalization/spatial confinement, with state space $(C,\mathcal{C})$ and to metabolic state $(M,\mathcal{M})$. In this telling, the polymerase reaction rate could be tied to the degree of spatial confinement and the source of the activated nucleotides from metabolic precursors. Then, the polymerase state-space is $(C\times M\times E\times E, \mathcal{C}\otimes\mathcal{M}\otimes\mathcal{E}\otimes\mathcal{E})$ with law ν_t and the polymerase transition kernel Q_{cme} is defined as the mapping from $(C\times M\times E\times E, \mathcal{C}\otimes\mathcal{M}\otimes\mathcal{E}\otimes\mathcal{E})$ into $(E,\mathcal{E})$. Thus, the law on the input–output space-state $(C\times M\times E\times E\times E, \mathcal{C}\otimes\mathcal{M}\otimes\mathcal{E}\otimes\mathcal{E}\otimes\mathcal{E})$ is given by $\mu_t = \nu_t \times Q_{cem}$, or in differential notation,

$$\mu_t(\mathrm{d}c,\mathrm{d}m,\mathrm{d}x,\mathrm{d}y,\mathrm{d}z) = \nu_t(\mathrm{d}c,\mathrm{d}m,\mathrm{d}x,\mathrm{d}y)Q_{cem}((c,m,x,y),\mathrm{d}z)$$

2.4.2. Non-RNA Polymerization

If there exists some kind of non-RNA polymerase activity, we have that the mapping

$$\{x\} \mapsto \{x\},\{x_*^c\}$$

which we regard as a mapping from $(E,\mathcal{E})$ into $(E,\mathcal{E})$. Let ν_t be a probability measure on $(E,\mathcal{E})$ defined by

$$\nu_t\{x\} = \frac{N_t(\{x\})}{N_t(E)} \quad \text{for} \quad x \in E$$

Similar to RNA polymerization, for fidelity probability $p \in (0,1]$, we have Q as the $4^n \times 4^n$ matrix defined by

$$Q(x, \mathfrak{H}_i(x)) = \binom{n}{i}(1-p)^i p^{n-i} \quad \text{for} \quad x \in E, \quad i \in \{0,\ldots,n\}, \quad p \in (0,1)$$

and

$$Q(x,\{z\}) = \frac{1}{|\mathfrak{H}_i(x)|}\binom{n}{i}(1-p)^i p^{n-i} \quad \text{for} \quad x \in E, \quad i = \{0,\ldots,n\}, \quad z \in \mathfrak{H}_i(x), \quad p \in (0,1)$$

and

$$Q(x,\{x^c\}) = 1 \quad \text{for} \quad x \in E, \quad p = 1.$$

$\mu_t = \nu_t Q$ is a probability measure on $(E, \mathcal{E})$ defined by

$$\mu_t(f) = \int_E \nu_t(\mathrm{d}x) \int_E Q(x, \mathrm{d}y) f(y) \quad \text{for} \quad f \in \mathcal{E}_{\geq 0}$$

Note that, for SSA, Q is fixed over the simulation, whereas the probability measure ν_t depends on time. That is, the reactions are chosen according to the reaction rates, and the reactions each use respective Q. The ν_t is formed using a random counting measure, so ν_t is random. This approach generalizes in the obvious way to all the reactions.

2.5. Decay

The RNA sequences have finite lifetimes in reality. This comes from a variety of sources, including radiation, pH, intrinsic molecular stability, etc. We assume double-stranded RNA is stable, whereas single-stranded RNA is not. Therefore, we create a reaction for decay of single-stranded RNA into constitutive nucleotides

$$x \xrightarrow{k_\varnothing} \varnothing \tag{29}$$

with reaction rate

$$\check{k}_\varnothing(t) = k_\varnothing N_t(\mathcal{X}_1)$$

2.6. Compartmentalization

It is thought that compartmentalization plays a role in RNA origins of life, giving foci of reproducing sequences [23,55]. This is somewhat anticipated by the similarity function $s : E \times E \mapsto (0,1]$, where sequences are more likely to copy similar sequences than less similar ones, due to an underlying spatial localization. Explicit spatial effects may be modeled by assuming each $x \in E$ is marked with a position on a bounded subset of the real line $([-T,T], \mathcal{B}_{[-T,T]}) \subset (\mathbb{R}, \mathcal{B}_\mathbb{R})$. We think of this as a one-dimensional projection of the three-dimensional system. Additional species can be introduced, such as lipids, with reactions forming a vesicle M (vesiculation), which encloses some $A = [r,s] \subset [-T,T]$. We assume the lipids interact with the single stranded sequences in A to form vesicles as

$$x \xrightarrow{k_{mic}} M(A) \tag{30}$$

with reaction rate

$$k_{mic}(t) = k_{mic} N_t(\mathcal{X}_1).$$

Hence, vesiculation is coupled to the population of sequences by design so that it evolves on roughly the same time-scale as sequence activities. Note that vesicles can enclose one another, i.e., $M(A)$ and $M(B)$ where $A \subset B$ or $B \subset A$, but cannot cross, i.e., for all vesicles at locations $A,\ldots,B$ we have that $A \cap B \in \{A, B, \varnothing\}$. For example, suppose one vesicle $A = [0,1]$ encloses another two, $B = [\frac{1}{3}, \frac{1}{2}]$ and $C = [\frac{2}{3}, \frac{3}{4}]$. Then, $A \setminus (B \cup C) = [0, \frac{1}{3}) \cup (\frac{1}{2}, \frac{2}{3}) \cup (\frac{3}{4}, 1]$. Although $A \setminus (B \cup C)$ is disconnected in one dimension, the intervals

are physically connected in three dimensions, where vesicles are spheres. We identify each vesicle to a union of disjoint intervals, disjoint across the vesicles.

We posit that compartmentalization precedes the hitting of a high-fidelity replicator through ensuring necessary concentration of RNA and a stable environment. Generally, we identify compartmentalization state to $(C, \mathcal{C})$ with probability measure ν. Let Q_c be a transition probability kernel from $(C, \mathcal{C})$ into $(E, \mathcal{E})$, encoding the transition from compartmentalization coordinates to RNA sequences. The product space $(C \times E, \mathcal{C} \otimes \mathcal{E})$ has law $\mu = \nu \times Q_c$. Upon hitting a high-fidelity replicator and achieving Darwinian evolution to acquire information, e.g., genes, the sequences are assumed to become adapted to compartmentalization coordinates $(C, \mathcal{C})$ through the transition probability kernel Q_c' from $(C \times E, \mathcal{C} \otimes \mathcal{E})$ into $(C, \mathcal{C})$, so that $\mu = \nu \times Q_c \times Q_c'$ is the law on the full space $(C \times E \times C, \mathcal{C} \otimes \mathcal{E} \otimes \mathcal{C})$. In this telling, compartmentalization precedes RNA activity, and, upon hitting high-fidelity replicators that can maintain their information, is followed by genomic adaptation.

2.7. Metabolism

We identify metabolism reaction-state to the measurable space $(M, \mathcal{M})$ with probability measure ν. Let Q_m be a transition kernel from $(M, \mathcal{M})$ into $(E, \mathcal{E})$, positing that metabolism precedes replication. For example, certain metabolic state may be precursors to the synthesis of RNA. Consider product space $(M \times E, \mathcal{M} \otimes \mathcal{E})$ with measure $\mu = \nu \times Q_m$. Now we suppose that, upon achieving Darwinian evolution in replicators, the replicators will eventually become adapted to $(M, \mathcal{M})$. Hence, we interpret $(M, \mathcal{M})$ as a mark-space of $(M \times E, \mathcal{M} \otimes \mathcal{E})$, representing genomic adaptation. Let Q_m' be a transition kernel from $(M \times E, \mathcal{M} \otimes \mathcal{E})$ into $(M, \mathcal{M})$. Then, $\mu = \nu \times Q_m \times Q_m'$ is a probability measure on $(M \times E \times M, \mathcal{M} \otimes \mathcal{E} \otimes \mathcal{M})$, where

$$\mu(f) = \int_M \nu(\mathrm{d}x) \int_E Q_m(x, \mathrm{d}y) \int_M Q_m'((x, y), \mathrm{d}z) f(x, y, z) \quad \text{for} \quad f \in (\mathcal{M} \otimes \mathcal{E} \otimes \mathcal{M})_{\geq 0}$$

or

$$\mu(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) = \nu(\mathrm{d}x) Q_m(x, \mathrm{d}y) Q_m'((x, y), \mathrm{d}z).$$

Therefore, metabolism-first followed by replication and genomic adaptation is encoded by the structures of Q_m and Q_m'. We do not specify these transition kernels in this article but mention that they are richly textured.

2.8. Reaction Overview

The reactions of the system having decay and clay and their reaction orders are shown in Table 2. There is one zero-order reaction, three first-order reactions, and two second-order reactions. Additionally, we show reactions and orders for compartmentalization and metabolism.

Table 2. Reactions and orders.

Reaction	Order
RNA double strand formation	2
RNA double strand dissociation	1
RNA polymerization	2
RNA decay	1
Clay polymerization	1
Clay oligomerization	0
Compartmentalization	1
Metabolism to replication	1
Metabolism & replication to metabolism	2

3. Results

Consider some initial population of I random sequences X_0. The population over time is given by X_t with associated random counting measure N_t on $(\bar{G}, 2^{\bar{G}})$. Recall parameters

$$\theta = (n, q, k, l, m, p, k_{\varnothing}, k_{ss}, k_{ds}, k_{rep}, k_{clay-o}, k_{clay-p})$$

for sequence dimension n, high-fidelity sequence set size q, fitness degree k, similarity degree l, fidelity degree m, clay fidelity probability p, RNA decay rate $k_{\varnothing}$, double-strand dissociation rate k_{ss}, double-strand formation rate k_{ds}, RNA replication rate k_{rep}, and clay oligomerization and polymerization rates k_{clay-o} and k_{clay-p}. These parameters are summarized in Table 3.

The following is the description of how the parameter values were specified and to what they biologically correspond. The sequence dimension n is chosen from $\{3, 4, 5\}$. The fitness and similarity functions are chosen by setting the value of the range of the curvature parameters k and l from one (inside the high-fidelity manifold) to some small values, such as over an exponential grid. For example, when $i = 0.1$, the fitness of sequences that are maximally dissimilar have 10% of the fitness of the high-fidelity sequences. We range the grid from 0.1 to 0.001 for fitness and similarity. The RNA fidelity parameter $m = 0.25$ is chosen such that the high-fidelity sequences have value one and the lowest fidelity sequences have value 0.25, equal to random chance. The clay fidelity parameter is set to an optimistically high value of 0.9 for clay studies. The double-strand dissociation and formation rates k_{ss} and k_{ds} are set to unity as a baseline. In comparison, the RNA replication rate is set to a large value, 10, whereby replication is the dominant reaction. The decay parameter $k_{\varnothing}$ is set to some uniform random value in $(0, 1)$. The clay RNA oligomerization rate is set to unit, and 'clay' RNA polymerization rate is set to a uniform random value in $(0, 20)$.

Table 3. Model parameters.

θ	Name	Domain	Value(s)
n	sequence dimension	$\mathbb{N}_{>0}$	$\{3, 4, 5\}$
k	RNA fitness parameter	$\mathbb{R}_{\geq 0}$	$\{-\log(i)/n : i = 0.1, 0.05, 0.01, 0.005, 0.001\}$
l	RNA similarity parameter	$\mathbb{R}_{\geq 0}$	$\{-\log(i)/n : i = 0.1, 0.05, 0.01, 0.005, 0.001\}$
m	RNA fidelity parameter	$\mathbb{R}_{\geq 0}$	$-\log(0.25)/n$
p	clay fidelity probability	$(0, 1]$	0.9
k_{ss}	double-strand dissociation rate	$\mathbb{R}_{\geq 0}$	1
k_{ds}	double-strand formation rate	$\mathbb{R}_{\geq 0}$	1
k_{rep}	RNA replication rate	$\mathbb{R}_{\geq 0}$	10
$k_{\varnothing}$	RNA decay rate	$\mathbb{R}_{\geq 0}$	$(0, 1)$
k_{clay-o}	clay RNA oligomerization rate	$\mathbb{R}_{\geq 0}$	1
k_{clay-p}	clay RNA polymerization rate	$\mathbb{R}_{\geq 0}$	$(0, 20)$

With the parameters governing the reaction rates, different values of these parameters confer different regimes for the system.

3.1. Stability: ODEs

We characterize the zeros of the vector field f from ODE system (A1) and use the eigenvalues of the Jacobian (A2) to determine their stability.

Theorem 2. *The ODE system* (A1) *for* $R = \{x\}$, $x \in E$, *has a single unstable fixed-point at* $[x] = 1$ *and* $[y] = 0$ *for* $y \in G \setminus x$.

Proof. Solving $f = 0$ gives a single solution $[x] = 1$ and $[y] = 0$ for $y \in G \setminus x$. For this solution, the eigenvalues of the Jacobian contain no zero values and positive values. Therefore, the solution is unstable. □

It follows from Theorem 2 that, for all other initial conditions, the system has no equilibria.

Corollary 2 (Unbounded). *For all initial conditions X_0 such that $I = |X_0| > 1$, the system is unbounded.*

This confirms the obvious: the system, a replicating network with no death, is almost always an increasing system.

3.2. Simulation Reaction State

We are interested in the behavior of temporal probability measures, ν_t (25) on the sequence product space and $\mu_t = \nu_t Q$ (27) on the sequence space, for RNA polymerization. These reveal the instantaneous information of the system. The structure of μ_t reveals the state of polymerization and is a leading indicator of the population concentrations over time.

3.2.1. Core Model with "Tent" Functions, Probable Hitting $\mathbb{P}(\tau_v(\theta) < \infty) \sim 1$

Take sequence dimension $n = 3$ and fitness and similarity curvature parameters $k = l = -\log(0.01)/n$ and fidelity parameter $m = -\log(0.25)/n$. Set rates for double-strand dissociation and formation $k_{ss} = k_{ds} = 1$ and polymerization rate $k_{rep} = 10$ and use the "tent" function for fitness, similarity, and fidelity. Take random initial population X_0 with initial population size $I = |X_0| = 10$ and random singleton $R = \{\{x\}\}$ ($q = 0$). We simulate 5000 reactions, simulation censored at hitting time τ_v for volume fraction $v = 0.25$. Take partition of the sequence space by Hamming distance to the high-fidelity manifold $(\mathfrak{H}_i)$ (28) of sequence space E. In Figure 1, we plot measures of a typical realization of the system X_t on the partition $(\mathfrak{H}_i)$ of sequence concentration (Figure 1a), growth (Figure 1b), and polymerase sequence output μ_t (Figure 1c). Some quantities are plotted on log-log scale, whereas others are plotted on a linear-log scale. These results show that the concentrations are relatively stable for most time, until the high-fidelity manifold is hit. Then, the concentration of high-fidelity replicators rapidly increases to exceed 25%. Similarly, Figure 1b shows the growth curves on a log-log scale, where the high-fidelity manifold rapidly increases near the end of the simulation. Figure 1c shows the structure of the RNA sequence polymerization output temporal probability measure μ_t. Low probability is assigned to polymerization of high-fidelity replicators for most of the reaction time, followed by a large increase near the end of the simulation, where high-fidelity replicators dominate with 56% probability. *Therefore, the RNA sequence polymerization output temporal probability measure μ_t is a leading indicator of the concentration curve, i.e., at simulation end-time, concentration of high-fidelity replicators is 25% and polymerization output is 56%.*

3.2.2. Core Model with "Tent" Functions, Improbable Hitting $\mathbb{P}(\tau_v(\theta) < \infty) \sim 0$

We use the same configuration as Section 3.2.1 except for setting fitness and similarity curvature parameters to $k = l = -\log(0.1)/n$. In Figure 2, we plot measures of a typical realization of the system X_t on sequence partition by Hamming distance to the high-fidelity manifold $(\mathfrak{H}_i)$ of concentration (Figure 2a), growth (Figure 2b), and μ_t (Figure 2c). The behavior has completely changed: the high-fidelity group ends the simulation with around 6% concentration, only steadily increasing, and never hits. The polymerase output μ_t shows 6%. This indicates that the concentration of high-fidelity replicators is unlikely to increase further, as the population is generally in equilibrium with the polymerase output.

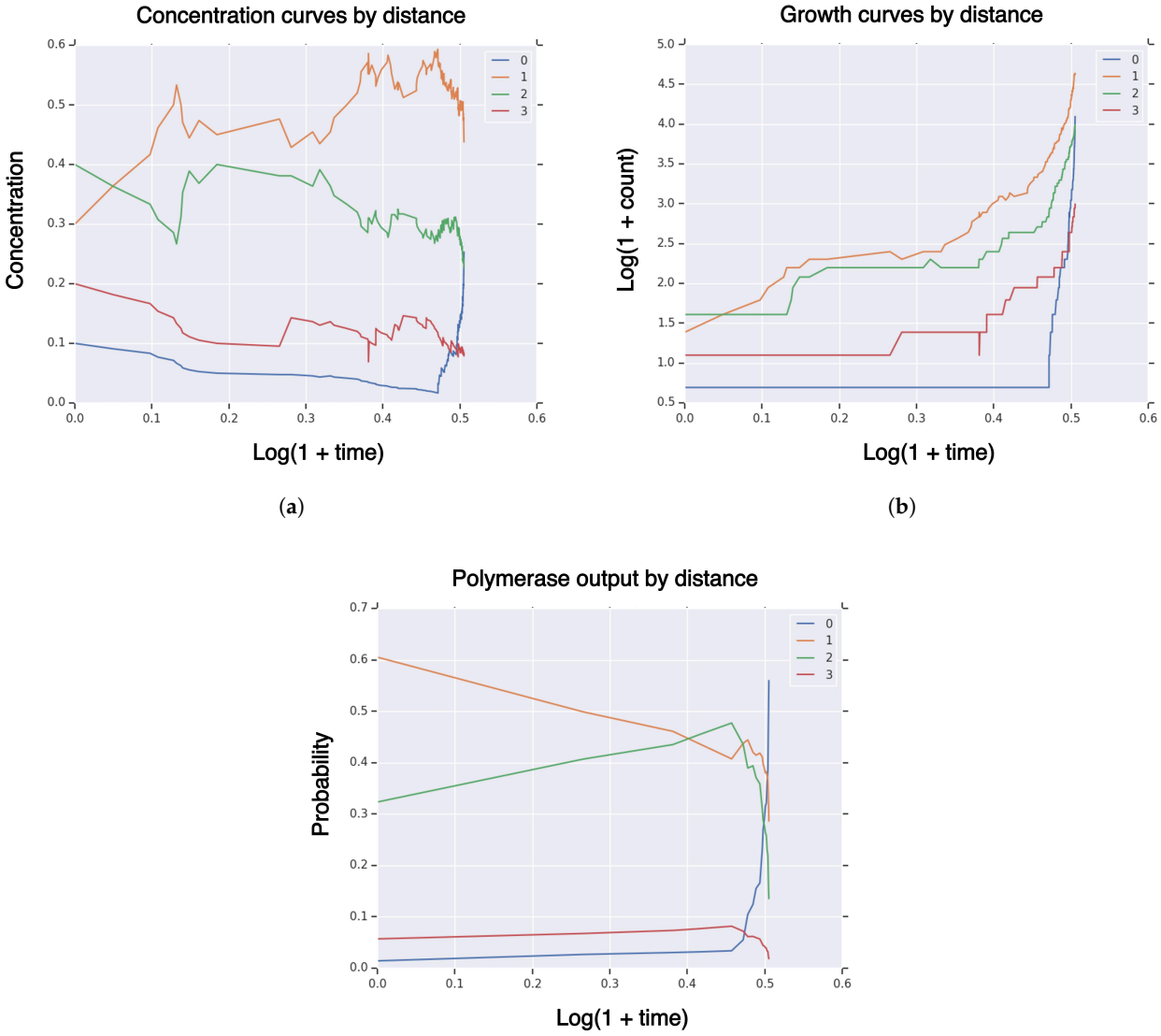


Figure 1. Measures of system population X_t until hitting time τ_v for high-fidelity replicator volume fraction $v = 0.25$ with sequence dimension $n = 3$, fitness/similarity curvature $l = k = -\log(0.01)/n$, initial population size $I = |X_0| = 10$, singleton high-fidelity replicator $R = \{\{x\}\}$, with "tent" fitness and similarity functions. (**a**) concentration of RNA sequences by Hamming distance to high-fidelity replicator; (**b**) population size of RNA sequences by Hamming distance to high-fidelity replicator; (**c**) polymerase RNA sequence output by Hamming distance to high-fidelity replicator.

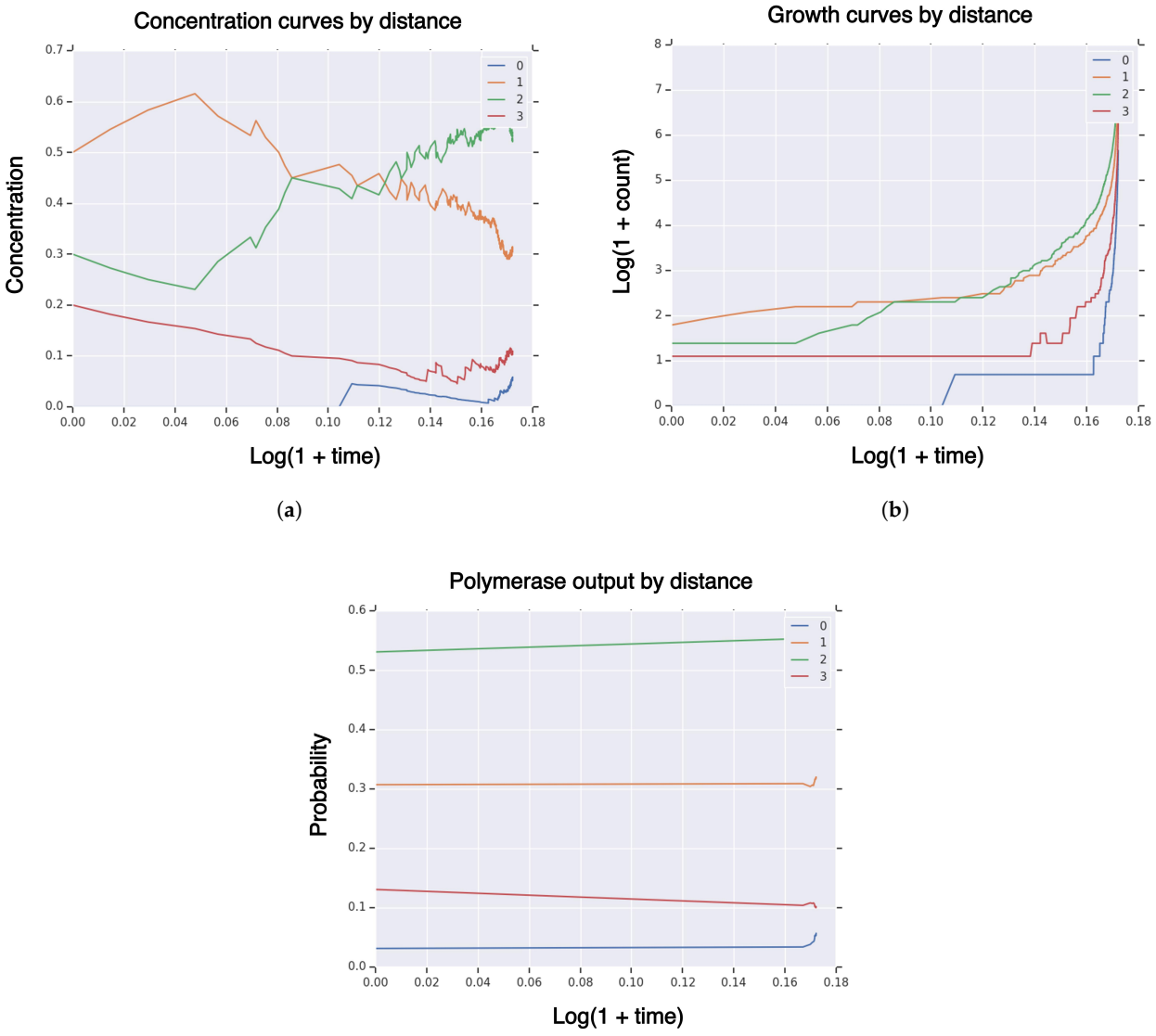


Figure 2. Measures of system population X_t until hitting time τ_v for high-fidelity replicator volume fraction $v = 0.25$ with sequence dimension $n = 3$, fitness/similarity curvature $l = k = -\log(0.1)/n$, initial population size $I = |X_0| = 10$, singleton high-fidelity replicator $R = \{\{x\}\}$, with "tent" fitness and similarity functions. (**a**) concentration of RNA sequences by Hamming distance to high-fidelity replicator; (**b**) population size of RNA sequences by Hamming distance to high-fidelity replicator; (**c**) polymerase RNA sequence output by Hamming distance to high-fidelity replicator.

3.2.3. Core Model with Linear Functions, Improbable Hitting $\mathbb{P}(\tau_v(\theta) < \infty) \sim 0$

The same configurations for Section 3.2.1 are used, except the fitness, similarity, and fidelity functions are linear. Similar to the "tent" functions, we specify the terminus landscape curvature for fitness and similarity $k = l$. Then, the fitness function for RNA polymerization is given by

$$f_k(x, \mathfrak{R}) = 1 + \left(\frac{k-1}{n}\right)\mathfrak{H}(x, \mathfrak{R}) \quad \text{for} \quad x \in E$$

and

$$s_k(x, y) = 1 + \left(\frac{k-1}{n}\right)S(x, y) \quad \text{for} \quad x, y \in E$$

We put fitness and similarity landscape curvature $k = l = 0.01$ for fitness and similarity and $v = 0.25$ for hitting volume fraction of high-fidelity replicators. We simulate X_t for 5000 reactions. We find that the probability of hitting is near zero $\mathbb{P}(\tau_{0.25}(\theta) < \infty) \sim 0$. In Figure A1, we plot measures of a typical realization of X_t on $(\mathfrak{H}_i)$ of concentration (Figure A1a), growth (Figure A1b), and μ_t (Figure A1c). The simulation ends with high-fidelity concentration of ∼5% and polymerase output of ∼4%. Therefore, the concentration of high-fidelity replicators will continue to decrease. *Linear surfaces are not sufficient to achieve hitting times $\tau_v(\theta) < \infty$ for high-fidelity replicator volume fraction $v = 0.25$, in contrast to the nonlinear "tent" functions.*

3.2.4. Expanded Model with "Tent" Functions, Probable Hitting $\mathbb{P}(\tau_v(\theta) < \infty) \sim 1$

We consider a similar model to the previous subsections and expand it with clay oligomerization rate (of RNA) k_{clay-o}, clay polymerization rate (of RNA) k_{clay-p}, and clay polymerization fidelity p. Therefore, the full set of variables is given by $\theta = (n, k_{ss}, k_{ds}, k, k_{clay-o}, k_{clay-p}, p)$. The value of fitness/similarity landscape curvature k, l and clay RNA polymerization raet k_{clay-p} are set such that the replicative mass of each is initialized to 10. This means that RNA and clay polymerization have the same reaction mass at the beginning of the simulation. We set sequence dimension $n = 3$, fitness/similarity landscape curvature $k = l = -\log(0.01)/n$, clay RNA polymerization fidelity $p = 0.9$, and double-strand dissociation and formation rates $k_{ss} = k_{ds} = 1$. This is a high hitting regime, i.e., the probability of hitting is close to one $\mathbb{P}(\tau_v(\theta) < \infty) \sim 1$. In Figure A2, we plot measures of a typical realization of X_t on $(\mathfrak{H}_i)$ and additionally the probability of reactions over time. High-fidelity replicators ended the simulation with 25% concentration (Figure A2a) and RNA polymerase output ∼62% (Figure A2c), indicating that the concentration of high-fidelity replicators will continue to increase. All species exhibit superexponential growth (Figure A2b). Clay polymerization decreases in contribution over time, whereas RNA polymerization increases substantially over time, and RNA double-strand reactions are small and stable (Figure A2d).

3.3. Hitting Times: Functional and Survival Analysis

We study various models in order of increasing complexity. We examine the hitting time surface $\tau_v(\theta)$ in the parameters $\theta \in \Theta$, including probability of hitting $\mathbb{P}(\tau_v(\theta) < \infty)$. We begin with the core model with no decay or clay.

3.3.1. Core Model, $\tau_v(\theta)$ for $v = 0.1$ with $\theta = (n, k)$ and "Tent" Functions

We are interested in the structure of the hitting time $\tau_v(\theta)$ of (19) as a function of the parameter vector θ. We use the Weibull–Cox proportional hazard's model of Table 1 for the hitting time τ_v for volume fraction $v = 0.1$. Let $\theta = (n, k)$ with sequence dimension $n \in \{3, 4\}$ and fitness and similarity parameters $k = l = -\log(i)/n$ for $i \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$. Set $m = -\log(0.25)/n$ for fidelity probability parameters. For each value of sequence dimension n, take random initial population X_0 with initial population size $I = |X_0| = 10$ and random singleton $\mathfrak{R} = \{\{x\}\}$ for the high-fidelity manifold and fix these for the fitness/similarity landscape curvature parmeters $k = l$. We fix the double-strand dissociation and formation rate parameters $k_{ss} = k_{ds} = 1$ and set RNA polymerization rate k_{rep} such that the overall RNA polymerization rate is given by $\check{k}_{rep}(0) = 10$ and use the "tent" function for fitness, similarity, and fidelity. We take 10 realizations of hitting time $\tau_v(\theta)$ for each parameter vector $\theta \in \Theta$ and allocate 5000 reactions. This gives 100 independent hitting times and up to 500,000 reactions. The times are comparable because the system is initialized to the same replication mass.

For the simulations, 66 hitting times are finite. The coefficients positively contribute to hitting, where $\gamma_n \approx 0.97$ and $\gamma_k \approx 13.29$, both with p-values less than 0.005. Therefore, hittings are strongly positively influenced by the parameters of the fitness and similarity functions and less so by the dimension. Plots of the coefficients and survival and cumulative

hazard curves are given in Figure A3. *High survival is found for k large and low survival for k small. Cumulative hazard is highest for k small.*

We estimate HDMR of the classifier (whether or not hitting time is finite) using all 100 samples. The results are shown in Figure A5 and Table 4. The HDMR explains roughly 80% of variance. $\mathbb{S}_k \approx 0.69$ and $\mathbb{S}_n \approx 0.06$, so *hitting probability is strongly influenced by fitness landscape curvature k and less so by sequence length n*. The component functions f_k and f_n for fitness landscape curvature and sequence dimension are strictly decreasing, where larger fitness landscape curvature parameter k results in decreasing hitting probability. These results are consistent with the survival analysis.

Table 4. HDMR sensitivity indices of hitting probability $\mathbb{P}(\tau_v(\theta) < \infty)$ for the core model.

$\boldsymbol{\theta}$	$\mathbb{S}_{\boldsymbol{\theta}}$
Sequence length n	0.06
Curvature k	0.69
$\sum$	0.75

We estimate HDMR of the regressor (hitting time) using the 66 simulations with finite hitting time. The results are shown in Figure A4 and Table 5. The HDMR explains roughly 60% of variance. $\mathbb{S}_n \approx 0.57$ and $\mathbb{S}_k \approx 0.04$, so *sequence dimension dominates the hitting time*. Both HDMR component functions f_n and f_k for sequence dimension and fitness landscape curvature are increasing. The HDMR results reveal that conditioning on hitting reverses the roles of sequence dimension n and fitness landscape curvature k.

Table 5. HDMR sensitivity indices of hitting time $\tau_v(\theta)$ for the core model.

$\boldsymbol{\theta}$	$\mathbb{S}_{\boldsymbol{\theta}}$
Sequence length n	0.57
Curvature k	0.04
$\sum$	0.61

3.3.2. Clay and Decay Model, $\tau_v(\theta)$ for $v = 0.1$ with $\theta = (n, k, k_\varnothing, k_{clay-p}, p)$ and "Tent" Functions

We expand the model to include clay and decay. We take parameter vector

$$\theta = (n, k, k_\varnothing, f_{clay}, p_{clay}) \in \Theta$$
$$\Theta = \{3, 4\} \times \{-\log(i)/n : i = 0.1, 0.05, 0.01, 0.005, 0.001\} \times (0,1) \times (0,1) \times (0,1)$$

with double-strand dissociation and formation and clay oligomerization reaction rates $k_{ss} = k_{ds} = k_{clay-o} = 1$. For each parameter vector $\theta \in \Theta$, (i) we choose singleton high-fidelity replicator manifold $\mathfrak{R} = \{x\}$ for some RNA sequence $x \in E$ and choose random initial population of RNA molecules X_0 such that the initial population size is 10, $I = |X_0| = 10$, and where the initial population does not intersect the high-fidelity manifold $X_0 \cap \mathfrak{R} = \varnothing$, that is, the initial population does not reside on the high-fidelity manifold; (ii) we initialize the replicative mass of the system such that the initial overall RNA polymerization reaction rate is given by $\check{k}_{rep}(0) = (1 - f_{clay})20$ and the initial overall clay RNA polymerization reaction rate $\check{k}_{clay-p}(0) = f_{clay}20$; (iii) we sample the hitting times $\tau_v(\theta)$ for volume fraction of high-fidelity replicators $v = 0.10$ a total of $M = 10$ times, each censored by 5000 reactions, giving hitting time set $\mathcal{T}(\theta)$ of (20). We attain input–output data set as $\mathfrak{D} = \{(\theta_i, \mathcal{T}(\theta_i)) : i = 1, \ldots, 240\}$. This gives a total of 2400 simulations.

For the simulations, 1546 hitting times are finite. The results of fitting the Weibull–Cox model are shown below in Table 6 and Figure A6. The curvature parameter k again

significantly dominates with a large positive value. All the remaining parameters have values less than one. Sequence dimension n again is a relatively small positive contributor. The clay fraction f_{clay} is small and positive and replication fidelity parameter p is not significant.

Table 6. Weibull–Cox model parameters for hitting times of clay and decay model.

θ	Name	Coefficient γ_θ	p-Value
n	sequence dimension	0.54	$<$0.005
$k = l$	RNA fitness parameter	27.87	$<$0.005
p	clay fidelity probability	$-$0.08	0.35
$k_\varnothing$	RNA decay rate	0.80	$<$0.005
f_{clay}	fraction clay RNA polymerization rate	0.96	$<$0.005

We estimate HDMR of the classifier (whether or not hitting time is finite) using all 2400 samples. Component functions and sensitivity indices are shown below in Figure A7 and Table 7. First-order HDMR captures 74% of explained variance, and second-order captures 4%. *Curvature dominates hitting probability with large sensitivity index* $\mathbb{S}_k \approx 67\%$. The HDMR component function in landscape curvature f_k is a decreasing function, where small values increase and large values decrease hitting probability. Sequence dimension n has sensitivity index $\mathbb{S}_n \approx 2\%$, and the HDMR component function f_n is decreasing, where high dimension decreases the probability of hitting. Clay parameter sensitivity index is small $\mathbb{S}_{f_{clay}} \approx 2\%$, and the HDMR component function for fractional clay RNA polymerization rate, $f_{f_{clay}}$, is decreasing, where low-to-medium clay fractions increase and high-clay fractions decrease probability of hitting. The HDMR results are consistent with the Weibull–Cox model.

Table 7. HDMR sensitivity indices of hitting probability $\mathbb{P}(\tau_v(\theta) < \infty)$ for expanded model (clay and decay).

θ	$\mathbb{S}_\theta$
Sequence length n	0.0213
Curvature k	0.6732
Decay rate $k_\varnothing$	0.0120
Clay fidelity p	0.0114
Fraction clay RNA polymerization rate f_{clay}	0.0219
$\sum$	0.7399

We estimate HDMR of the regressor (hitting time) using the 1546 simulations with finite hitting time. Component functions and sensitivity indices are shown below in Figure A8 and Table 8. First-order HDMR captures 33% of explained variance, and second-order captures 7%. *In stark contrast to the contributions to the classifier, the parameters k and n are insignificant to hitting time. Instead, the largest sensitivity index is* $\mathbb{S}_{f_{clay}} \approx 20\%$. The HDMR component function for fractional clay RNA polymerization rate, $f_{f_{clay}}$, is an increasing function, where small f_{clay} decreases and large f_{clay} increases the hitting time. This suggests that high clay-fractions representing first-order reactions increase the hitting time, as clay polymerization has less replicative mass than RNA polymerization, i.e., things go faster with RNA polymerization. The second largest sensitivity index is decay $\mathbb{S}_{k_\varnothing} \approx 11\%$. Decay is an increasing function, with sharp increase in hitting times nearby one, i.e., things go slower with large decay resulting in increased hitting time.

Table 8. HDMR sensitivity indices of $\tau_v(\theta) < \infty$ for expanded model (clay and decay).

$\boldsymbol{\theta}$	$\mathbb{S}_{\boldsymbol{\theta}}$
Sequence dimension n	0.0013
Curvature k	0.0030
Decay $k_{\varnothing}$	0.1129
Clay fidelity p	0.0152
Fraction clay RNA polymerization rate f_{clay}	0.2014
$\sum$	0.3339

3.4. Compartmentalization

Compartmentalization has a direct effect on the calculation of the reaction rates, specifically replication, by computing only a subset of the reactions in $\mathcal{X}_1^2$. Put

$$X_t(A) = \{x \in X_t : l(x) \in A\}.$$

For vesicle region $A \in \mathfrak{M}$, we have that

$$\begin{aligned}\check{k}_{rep}(t,A) &= \sum_{(x,y)\in\mathcal{X}_1^2} k_{rep}(x,y)M_t(\{x\}\times A)(M_t(\{y\}\times A) - \mathbb{I}(x=y)) \quad \text{for} \quad t\in\mathbb{R}_{\geq 0}, \quad A\subset[-T,T] \\ &= \sum_{(x,y)\in X_t^2(A)} k_{rep}(x,y)M_t(\{x\}\times A)(M_t(\{y\}\times A) - \mathbb{I}(x=y))\end{aligned}$$

and total replicative mass

$$\check{k}_{rep}(t) = \sum_{A\in\mathfrak{M}} \check{k}_{rep}(t,A) \quad \text{for} \quad t\in\mathbb{R}_{\geq 0}$$

As $\mathfrak{M}$ increases in size over time, the number of partitions grows, and

$$\sum_{A\in\mathfrak{M}} |X_t^2(A)| \ll |\mathcal{X}_1^2|.$$

Therefore, the replicative mass will be reduced with $\mathfrak{M}$, and the system evolves less quickly. *This suggests that there is a trade-off between the degree of compartmentalization and the replicative mass of the system.*

4. Discussion and Conclusions

Origins of life is a fascinating problem. The wonderful complexity of extant life follows from origins. The distribution of life in the universe is tied to origins.

In this article, we have attempted to peek into the problem by concentrating on the RNA world hypothesis, studying hitting times of high-fidelity replicators. We develop fitness, similarity, and fidelity functions as landscapes for a mathematical model of replicating RNA molecules at the sequence level and observe hitting times through simulation studies. We draw attention to the distinction between the probability of hitting $\mathbb{P}(\tau(\theta) < \infty)$ and the hitting time $\tau(\theta) < \infty$.

In terms of mathematical set-up, we interpret the reactions as measure-kernel-functions. Each reaction is identified to and fully encoded by a probability transition kernel. The reactions take place in some domain, whereby all molecules may interact. We note that, in reality, molecules have limited diffusion, and this effectively breaks the reaction domain into independent subdomains above some length scale, i.e., molecules are more likely to react with their neighbors. Therefore, we assume our reaction volume is sufficiently small such that all molecules may participate in the reactions. We use for modeling purposes the ansatz that sequence distance is correlated to spatial proximity, where similar sequences are proximal, using a non-trivial similarity function $s : E \times E \mapsto (0,1]$.

Theorem 2 and its Corollary 2 show that the system (without decay) is unbounded and strictly increases. This formally shows the system to be a growth process. Next, we illustrate findings about the core system with probable hitting (Section 3.2.1). In particular, we see that the temporal image measure $\mu_t = \nu_t Q$, which describes the polymerization output, is a leading indicator of high-fidelity sequence concentration. Polymerization output and high-fidelity replicators super-exponentially increase near the end of the simulation. Next, the fitness and sequence curvature parameters are set at a higher value which confers reduced fidelity (Section 3.2.2). This reveals that hitting is never achieved and that the polymerization output is in equilibrium with population composition. Hence, the probability of hitting is strongly influenced by landscape curvatures. Next, linear curvature is utilized for fitness and similarity and results in no hitting (Section 3.2.3). This reveals that nonlinear curvature is necessary to achieve hitting of high-fidelity replicators. Next, we expand the model with non-RNA ('clay') based polymerization and find that such activity decreases over time, in contrast to RNA polymerization, which greatly increases and dominates other reactions over time the system (Section 3.2.4).

For functional and survival analysis of the hitting times, we study the core model, whereby hitting times are strongly positively influenced by the fitness and similar functions yet are not impacted significantly by sequence dimension (Section 3.3.1). In particular, survival analysis reveals low fitness curvature confers low survival (high hitting), whereas high fitness curvature confers high survival (low hitting). HDMR analysis shows that hitting probability is strongly influenced by fitness curvature and much less so by sequence dimension, supporting the survival analysis. HDMR analysis of hitting time shows reversed roles for sequence dimension and landscape curvature, where sequence dimension dominates hitting time, with curvature playing a far less significant role. This gives the finding that hitting probability is driven by curvature, whereas hitting time is driven by sequence dimension. Next, we perform functional and survival analysis of the core model augmented with 'clay and decay' dynamics (Section 3.3.2). Survival analysis shows similar results to the core model, where curvature dominates survival (no hitting), with sequence dimension playing a significantly reduced role. HDMR analysis of hitting probability shows that curvature dominates hitting probability, similar to the core model, whereas sequence dimension again plays a significantly reduced role. HDMR analysis of hitting time reveals that the presence of 'clay and decay' significantly increase hitting time, with curvature and sequence dimension playing insignificant roles. These results are consistent in that clay polymerization has less replicative mass than RNA polymerization, where RNA polymerization is a faster reaction.

Overall, we find that *nonlinear landscapes* are *necessary* for hitting: linear landscapes are insufficient. For nonlinear landscapes, we find that the *probability of hitting* is dominated by *curvature* and that *hitting times* are dominated by *sequence dimension*. These results suggest that the landscapes in nature are nonlinear with high curvature, and that the hitting time for high-fidelity replicators is an increasing function of sequence dimension. When clay and decay are added to the model, hitting probability is again dominated by curvature, and clay and decay are relatively insignificant. This reflects that clay and decay are low order reactions. They increase hitting times.

For replication and compartmentalization, we suggest that compartmentalization, while a necessary condition, slows overall system dynamics with increasing vesiculation rate. Essentially, as compartmentalization increases, there is a corresponding reduction in absolute replicative system mass, as certain reactions among elements are no longer possible, being physically sequestered. While the timescale of a simulation is tied to the replicative system mass of the system, there is variability in replicative mass across compartments. Some compartments contain large genomic and metabolic populations. It favors the search for the high-fidelity replicator by there being a distribution on compartmental 'fitness' such as resource concentrations so that the high-fitness compartments drive replicative system mass. Compartmentalization is identified to the measure ν on coordinates in

$(C, \mathcal{C})$, for which coordinates are "marked" by sequences through the transition probability kernel Q_c, and followed by genomic adaptation via the transition probability kernel Q'_c.

Metabolism is thought to be identified to production of precursors to RNA synthesis, leading to replication identification, followed by genomic adaptation to metabolic state. Metabolism is thus defined through the transition kernels Q_m and Q'_m.

The independence of the transition kernels can be scrutinized, and it is possible that general transition kernels on the full product spaces across location, genomic, and metabolic states are necessary to satisfactorily explain RNA origins, i.e., all three functions may have co-evolved. This notion is suggested in the hot springs hypothesis for origins, where compartmentalization is hypothesized to furnish necessary conditions to genomic and metabolic state [26–28]. In this telling, the base measurable space of interest is $(F, \mathcal{F}) = (C \times E \times M, \mathcal{C} \otimes \mathcal{E} \otimes \mathcal{M})$ with measure ν. Then, identification of a high-fidelity replicator is described through the transition probability kernel Q_{cem} from $(F, \mathcal{F})$ into $(F, \mathcal{F})$ in "marking" the base measurable space with state for genomic replication; finally, genomic adaptation is conveyed through the kernel Q'_{cem} from $(F \times F, \mathcal{F} \otimes \mathcal{F})$ into $(F, \mathcal{F})$. Hence, RNA origins of life has law $\nu \times Q_{cem} \times Q'_{cem}$ on the product space $(F \times F \times F, \mathcal{F} \otimes \mathcal{F} \otimes \mathcal{F})$, reflecting the steps of replicator identification and adaptation through the definitions transition kernels Q_{cem} and Q'_{cem}.

A putative "genesis machine" here is a mapping from the base (initial) measurable space $(F, \mathcal{F})$ into the product space of identified high-fidelity replicators undergoing adaptation, i.e., $(F \times F \times F, \mathcal{F} \otimes \mathcal{F} \otimes \mathcal{F})$. More generally the base space could additionally contain amino acid sequence space $(P, \mathcal{P})$. Such a machine is fully specified through the definitions of the distribution ν on the base space and the transition kernels Q_{cemp} and Q'_{cemp} (pre and post genes, respectively). Because the stages of transition occur purely through random drift, an experiment performed by such a machine would take an unacceptably long period of time to complete. Experimental demonstration can be contemplated by augmenting the base measurable space with a control space $(X, \mathcal{X})$ to accelerate dynamics, using for instance closed-loop shaped radiation to address molecular degrees of freedom in their appropriate frequency domains, (open-loop) catalysts, temperature, geometry, selection, concentration through centrifugal force, etc., resulting in the new (four-dimensional) base space $(\widetilde{F}, \widetilde{\mathcal{F}}) = (C \times E \times M \times P \times X, \mathcal{C} \otimes \mathcal{E} \otimes \mathcal{M} \otimes \mathcal{P} \otimes \mathcal{X})$. Then, the transition kernels $\widetilde{Q}_{cemp}$ and $\widetilde{Q}'_{cemp}$ become mappings from $(\widetilde{F}, \widetilde{\mathcal{F}})$ into $(\widetilde{F}, \widetilde{\mathcal{F}})$ and from $(\widetilde{F} \times \widetilde{F}, \widetilde{\mathcal{F}} \times \widetilde{F})$ into $(\widetilde{F}, \widetilde{\mathcal{F}})$, respectively. The general design of a genesis machine is the definitions of the augmented base space $(\widetilde{F}, \widetilde{\mathcal{F}})$, its distribution $\widetilde{\nu}$, and the augmented transition kernels $\widetilde{Q}_{cemp}$ and $\widetilde{Q}'_{cemp}$, giving law $\widetilde{\mu} = \widetilde{\nu} \times \widetilde{Q}_{cemp} \times \widetilde{Q}'_{cemp}$ on the full 15-dimensional product space $(\widetilde{F} \times \widetilde{F} \times \widetilde{F}, \widetilde{\mathcal{F}} \otimes \widetilde{\mathcal{F}} \otimes \widetilde{\mathcal{F}})$, written in differential notation

$$\widetilde{\mu}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) = \widetilde{\nu}(\mathrm{d}x)\widetilde{Q}_{cemp}(x, \mathrm{d}y)\widetilde{Q}'_{cemp}((x, y), \mathrm{d}z)$$

Origins could be experimentally demonstrated using a sequence of adaptive control fields in $(X, \mathcal{X})$, cycling through the transitions, and a detection system for online identification of the system whose elements belong to the product measurable space. The full design and estimated operating timescale for such a machine needs further research to assess practical feasibility. We call the creation of primordial life (*primordia*) by the continuous causal efforts of a genesis machine given initial prebiotic conditions *artebiogenesis*, where *arte-* is Latin and means "from skill." The primordia are not necessarily those that occurred in nature. Primordia and their genesis represent non-trivial system trajectories across the transition to the earliest life in sterile environments and belong to a manifold of primordial lifeforms, each having characteristic geochemical setting.

The probability measure $\mu_t = \nu_t Q$ has additional utility to integrate 'test' functions or queries about the system. If we let $f \in \mathcal{E}_{\geq 0}$ be a fitness function, then the fitness value $J(\mu_t) = \mu_t(f)$ is the expected value of the fitness function with respect to the probability measure μ_t. In OptiEvo theory, $J(\mu_t)$ is studied as a function of the population X_t on $(E, \mathcal{E})$ [52]. OptiEvo assumes that the set of all probability measures $\{\mu_t\}$ is convex and

that X_t has sufficient flexibility such that $J(\mu_t)$ may be explored around μ_t. Then, OptiEvo predicts that $J(\mu_t)$ has global maxima on $(E, \mathcal{E})$ and that these form a connected level-set of sequences with the same fitness value. Both predictions are consistent with our model. The first prediction is consistent with zero distance in fitness and similarity functions for high-fidelity sequences. The second prediction is consistent with the high-fidelity set being a singleton or a product-space construction. A contention is whether X_t has sufficient flexibility in exploring $J(\mu_t)$ around μ_t. This has direct bearing on the structure of Q: if X_t is inflexible, then Q is constrained to certain subspaces of $(E, \mathcal{E})$, i.e., not all transitions are possible.

In future work, the model could be extended to the space of sequences of lengths up to n, $(E^*, \mathcal{E}^*)$ or even the space of sequences of all lengths, where distance and similarity functions would utilize a more general string distance metric, e.g., Levenshtein distance. We note that the size of $(E^*, \mathcal{E}^*)$ is not much larger than $(E, \mathcal{E})$. Alternative similarity functions could be explored, such as the trivial case of constant similarity, e.g., $s(x, y) = 1$ for $(x, y) \in E \times E$. A limitation of this article is the restriction to short sequences due to computational efficiency. The numerical size of the sequences is mathematically low-dimensional and does not correspond to actual functions of RNA molecules. Other parameter sets can be explored for example using experimentally derived values for reaction rates, so that the timescales are calibrated. Future research could see the simulation software rewritten for a high-performance computing environment, enabling much longer, e.g., length 10–1,000, sequences to be studied. Polymerase fitness can be made empirical using known RdRp sequences as members of the high-fidelity manifold. Another area of future work could be studying the aforementioned transition kernels Q_c, Q_c', Q_m, and Q_m'. More general models for polymerization transition kernel based on the structure of the Poisson-binomial distribution could be employed. It would be interesting to study lipid-RNA and metabolism-RNA interactions and equip the system with the ability to append nucleotides to their sequences to form functional genes, such as storing useful information for the replication channel, perhaps a Hammerhead ribozyme to convey rolling-circle amplification. We note that transition kernels here generally lack amino acid state and are pure-RNA. An area to explore is the notion of the transition kernel into the space of high-fidelity replicators to depend on amino acid sequences and then to elaborate the system to contain a primitive translation system and examine various hitting times. Additional reactions can be introduced as operations on pairs of sequences, such as concatenation, and others for sequence splitting, and so on, with corresponding transition kernels enabling RNA networking and recombination dynamics.

Author Contributions: All authors contributed to the study conception and design. Material preparation, prepared software, data generation and analysis were performed by C.D.B. The first draft of the manuscript was written by C.D.B. and all authors commented on previous versions of the manuscript. Grant funding was provided through H.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Foundation Grant No. CHE-1763198.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Please see https://github.com/calebbastian/originoflife (accessed on 31 October 2021) for the Python software and example script of usage.

Acknowledgments: The authors thank the late Freeman Dyson for the discussions in 2018 of these ideas at the Institute for Advanced Study in Princeton, New Jersey, as well as anonymous reviewers who gave critical comments that substantially improved the quality of this article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Discrete Probability Space

We define some concepts related to the space $(E, \mathcal{E})$. The discrete probability measure ν on $(E, \mathcal{E})$ is defined by

$$\nu(A) = \nu \mathbb{I}_A = \sum_{x \in E} \nu\{x\} \mathbb{I}_A(x) \quad \text{for} \quad A \in \mathcal{E}$$

where $\nu\{x\}$ is the probability mass at the point $x \in E$ and

$$\mathbb{I}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

For the collection of non-negative $\mathcal{E}$-measurable functions $\mathcal{E}_{\geq 0}$, we have

$$\nu(f) = \sum_{x \in E} \nu\{x\} f(x) \quad \text{for} \quad f \in \mathcal{E}_{\geq 0}.$$

Appendix B. Other Fitness Functions

Another fitness function can be defined using polynomials, such as lines, quadratics, etc, in terms of $k \in \mathbb{N}_{>0}$

$$f_k(x, R) = \left(1 - \frac{\mathfrak{H}(x, \mathfrak{R})}{n+1}\right)^k \in (0, 1] \quad \text{for} \quad x \in E$$

or

$$f_k(x, R) = 1 - \left(\frac{\mathfrak{H}(x, \mathfrak{R})}{n+1}\right)^k \in (0, 1] \quad \text{for} \quad x \in E$$

However, another surface is using a sigmoid function. Put

$$E(x) = \frac{1}{1 + \exp[-x]} \quad \text{for} \quad x \in \mathbb{R}$$

We have fitness for $k \in (0, \infty)$

$$f_k(x, \mathfrak{R}) = \frac{1 - E\left(\frac{\mathfrak{H}(x, \mathfrak{R}) - \frac{n}{2}}{k}\right)}{1 - E\left(-\frac{n}{2k}\right)} \in (0, 1] \quad \text{for} \quad x \in E$$

Appendix C. Measure-Kernel-Function

We recall a few facts about transition kernel Q. Q defines a function

$$Qf(x) = \int_Y Q(x, \mathrm{d}y) f(y) \quad \text{for} \quad x \in E$$

that is in $\mathcal{X}_{\geq 0}$ for every function $f \in \mathcal{Y}_{\geq 0}$.

For every probability measure ν_t on $(E, \mathcal{E})$ and at time $t \geq 0$, the quantity $\mu_t = v_t Q$ defines a probability measure on $(Y, \mathcal{Y})$ as

$$\mu_t(A) = \int_X \nu_t(\mathrm{d}x) Q(x, A) \quad \text{for} \quad A \in \mathcal{Y}.$$

For every probability measure ν_t on $(X, \mathcal{X})$ and function $f \in \mathcal{Y}_{\geq 0}$, we have that

$$\mu_t(f) = (\nu_t Q) f = \int_X \nu_t(\mathrm{d}x) \int_Y Q(x, \mathrm{d}y) f(y).$$

Here, the spaces are discrete, i.e.,

$$\nu_t(A) \equiv \nu_t(\mathbb{I}_A) = \sum_{x \in X} \nu_t\{x\}\mathbb{I}_A(x) \quad \text{for} \quad A \in \mathcal{X}$$

where $\nu_t\{x\}$ is the probability mass at the point $x \in X$ at time $t \geq 0$.

Appendix C.1. Reactions as Measure-Kernel-Functions

We index the reaction types on $(Z, \mathcal{Z}) = (\mathbb{N}_{>0}, 2^{\mathbb{N}_{>0}})$. Let η_t be the probability measure on $(Z, \mathcal{Z})$ formed from the normalized reaction rates. Let Q_* be the transition kernel from $(Z, \mathcal{Z})$ into $(X, \mathcal{X})$. Then, $\eta_t Q_* = \nu_t$ is the distribution on $(X, \mathcal{X})$ and $\mu_t = \nu_t Q$ is the distribution on $(Y, \mathcal{Y})$. Then, a reaction is the mapping $(Z, \mathcal{Z}) \mapsto (X, \mathcal{X}) \mapsto (Y, \mathcal{Y})$.

Appendix C.2. Deterministic Model

Consider the core model defined in (2)–(4) with reaction rates $\check{k}_{ds}(t)$, $\check{k}_{ss}(t)$, and $\check{k}_{rep}(t)$. We are neglecting clay and decay for the moment. All the reactions impact $(E, \mathcal{E})$. For double-strand formation, the input space is $(X, \mathcal{X}) = (E \times E, \mathcal{E} \otimes \mathcal{E})$ and the output space is $(Y, \mathcal{Y}) = (F, \mathcal{F})$. For double-strand dissociation, the input and output spaces are swapped. For polymerization, $(X, \mathcal{X}) = (E\times, \mathcal{E} \otimes \mathcal{E})$ and $(Y, \mathcal{Y}) = (E, \mathcal{E})$. Hence, $(E, \mathcal{E})$ is positively impacted by polymerization, positively impacted by double-strand dissociation, and negatively impacted by double-strand formation. Put $[x] = N_t(\{\{x\}\})$ and $[x, x^c] = N_t(\{\{x, x^c\}\})$. Recall the ν_t, Q, and $\mu_t = \nu_t Q$ for the reactions, e.g., $\nu_t^{rep}, \nu_t^{ds}, Q^{rep}$, etc.

For $x \in E$, we have the system of $m = \frac{3}{2}4^n$ deterministic nonlinear ordinary differential equations (ODEs) in m variables as mean-field equations

$$\begin{aligned}
\frac{\mathrm{d}[x]}{\mathrm{d}t} &= f_x = \check{k}_{rep}(t)\mu_t^{rep}\{\{x\}\} + \check{k}_{ss}(t)\mu_t^{ss}\{\{x\}, \{x^c\}\} - \check{k}_{ds}(t)\mu_t^{ds}\{\{x, x^c\}\} \qquad \text{(A1)}\\
&= \check{k}_{rep}(t) \sum_{(y,z) \in E \times E} \nu_t^{rep}\{\{y\}, \{z\}\} Q^{rep}((y, z), \{\{x\}\}) + k_{ss}[x, x^c] - k_{ds}[x][x^c]\\
&= \sum_{(y,z) \in E \times E} k_{rep}(y, z)[y]([z] - \mathbb{I}(y = z)) Q^{rep}((y, z), \{\{x\}\}) + k_{ss}[x, x^c] - k_{ds}[x][x^c]\\
\frac{\mathrm{d}[x, x^c]}{\mathrm{d}t} &= f_{xx^c} = k_{ds}[x][x^c] - k_{ss}[x, x^c]
\end{aligned}$$

The fixed points of f are the equilibria of the system, i.e., $f(x) = \mathbf{0}$ for $x \in \mathbb{R}^m_{\geq 0}$. The Jacobian of the system is

$$J = \begin{bmatrix} \frac{\mathrm{d}f_1}{\mathrm{d}x_1} & \cdots & \frac{\mathrm{d}f_1}{\mathrm{d}x_m} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}f_m}{\mathrm{d}x_1} & \cdots & \frac{\mathrm{d}f_m}{\mathrm{d}x_m} \end{bmatrix} \qquad \text{(A2)}$$

The eigenvalues of the Jacobian reveal the stability of the fixed points. If all the eigenvalues of the Jacobian evaluated at the fixed point have negative real parts, then the fixed point is stable. If none of the eigenvalues are zero and at least one of the eigenvalues has a positive real part, then the fixed point is unstable. If at least one eigenvalue is zero, then the fixed point can be either stable or unstable.

Appendix D. Hitting Cardinality

We index the N reactions of X_t with arrival times $\{T_i\}$ in $(\bar{\mathbb{R}}_{\geq 0}, \mathcal{B}_{\bar{\mathbb{R}}_{\geq 0}})$, where $\bar{\mathbb{R}} = \mathbb{R}_{\geq 0} \cup \{\infty\}$. Define the hitting reaction $\varpi(\theta) \in \mathbb{N}$ in terms of hitting time $\tau(\theta) \in \bar{\mathbb{R}}_{\geq 0}$ as

$$\varpi(\theta) = \sum_{i=1}^{N} \mathbb{I}_{[0,\tau(\theta)]}(T_i)$$

$\varpi(\theta)$ is right-censored at N reactions. If $\tau(\theta) = \infty$, then $\varpi(\theta) = N$. If $\tau(\theta) < \infty$, then $\varpi(\theta) < N$.

Reaction Cardinality

In this section, we study, instead of the hitting time $\tau_v(\theta)$, the hitting reaction number $\varpi_v(\theta)$ for $v = 0.1$. We study the core model with parameter vector $\theta = (n, k)$. We uniformly sample sequence length $n \in \{3, 4, 5\}$ and fitness/similarity landscape curvature $k = l \in \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$ and form input–output data $\mathfrak{D} = \{(\theta_i, \varpi(\theta_i)) : i = 1, \ldots, 1200\}$. We set double-strand dissociation and formation rates $k_{ss} = k_{ds} = 1$ and initialize the overall RNA polymerization rate k_{rep} such that the initial replicative mass is 10. For each parameter vector $\theta \in \Theta$, we randomly choose the initial population of sequences X_0 with initial population size $I = |X_0| = 10$ and singleton high-fidelity sequence manifold $\mathfrak{R}$ such that the initial population of sequences does not intersect the high-fidelity replicator manifold, $X_0 \cap \mathfrak{R} = \varnothing$.

$\mathfrak{D}$ contains 781 hitting events. We denote these $\mathfrak{D}^*$. We form a first-order HDMR on $\varpi(\theta)$ using $\mathfrak{D}^*$. The HDMR truth-plot, component functions, and sensitivity indices are shown below in Figure A9. First-order HDMR captures approximately 31% of variance. Both component functions f_n and f_k have similar sizes, with f_n somewhat larger than f_k. The HDMR component function for sequence dimension f_n is essentially an increasing linear function of sequence length n, and component function for landscape curvature f_k is generally an increasing function of the fitness/similarity landscape curvature. *These results suggest that sequence dimension and curvature influence the hitting reaction.* Larger sequences and flatter curvature increase the hitting reaction.

Table A1. HDMR sensitivity indices of $\varpi_v(\theta) < \infty$ for core model and $v = 0.1$.

$\boldsymbol{\theta}$	$\mathbb{S}_{\boldsymbol{\theta}}$
Sequence dimension n	0.1619
Curvature k	0.1464
$\sum$	0.3083

Appendix E. Approximate Reaction Rates

One approach to reducing the computational complexity of $\check{k}(t)$ is to approximate the sums using Monte Carlo. Define random variables $\mathfrak{X}_1 \sim \text{Uniform}(\mathcal{X}_1)$ and $\mathfrak{X}_2 \sim \text{Uniform}(\mathcal{X}_2)$. Let $\{\mathfrak{X}_{1i}\}$ and $\{\mathfrak{X}_{2i}\}$ be independencies of such random variables. Given N random samples of $\mathfrak{X}_1$, the first reaction rate becomes

$$\check{k}_{ds}(t|N) = \frac{|\mathcal{X}_1|}{2N} \sum_{i=1}^{N} k_{ds} N_t(\mathfrak{X}_{1i}) N_t(\mathfrak{X}_{1i}^c)$$

whose expected value is approximated using M realizations,

$$\mathbb{E}\check{k}_{ds}(t|M, N) \simeq \frac{|\mathcal{X}_1|}{2MN} \sum_{j=1}^{M} \sum_{i=1}^{N} k_{ds} N_t(\mathfrak{X}_{1ij}) N_t(\mathfrak{X}_{1ij}^c)$$

requiring a total of MN evaluations. In a similar manner, the second reaction rate is

$$\mathbb{E}\check{k}_{ss}(t|M, N) \simeq \frac{|\mathcal{X}_2|}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} k_{ss} N_t(\mathfrak{X}_{2ij} \cup \mathfrak{X}_{2ij}^c)$$

and, putting $\mathfrak{X}_1^* \sim \text{Uniform}(\mathcal{X}_1)$, we have the third reaction

$$\mathbb{E}\check{k}_{rep}(t|M,N) \simeq \frac{|\mathcal{X}_1|^2}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N} k_{rep}(\mathfrak{X}_{1ij},\mathfrak{X}_{1ij}^*)N_t(\mathfrak{X}_{1ij})(N_t(\mathfrak{X}_{1ij}^*)-\mathbb{I}(\mathfrak{X}_{1ij}=\mathfrak{X}_{1ij}^*)) \quad \text{(A3)}$$

We refer to SSA simulation with Monte Carlo approximate reaction rates as Monte Carlo Approximate SSA, or MCASSA.

Appendix F. High Dimensional Model Representation

Suppose we have a real-valued square-integrable function $f(x) \in L^2(E,\mathcal{E},\nu)$ with $E = \mathbb{R}^n$, $\mathcal{E} = \mathcal{B}_{\mathbb{R}^n}$, $\nu = \prod_i \nu_i$, and $x = (x_1,\ldots,x_n) \in E$. The $\{\nu_i\}$ may be diffuse (continuous) and/or discrete. Put $B = \{1,\ldots,n\}$. We would like to decompose f into orthogonal function subspaces $\{f_u : u \subseteq B\}$ (projections) in such a way that each projection on an input subspace f_u maximizes variance and across subspaces retrieves total variance, i.e., $f = \sum_{u\subseteq B} f_u$ and $\mathbb{V}arf = \sum_{u\subseteq B}\mathbb{V}arf_u$. The solution to this problem in the retrieval of $\{f_u : u \subseteq B\}$ is known as *high dimensional model representation* (HDMR) or functional ANOVA expansion and for f is written as

$$f(x_1,\ldots,x_n) = f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i,x_j) + \ldots + f_{1\ldots n}(x_1,\ldots,x_n)$$

where the $\{f_u : u \subseteq B\}$ are called *component functions*. For independent inputs, the component functions are mutually orthogonal and, aside from the constant component function $f_0 = \mathbb{E}f$ (order zero), have zero mean $\mathbb{E}f_u = 0$ for all non-empty $(2^n - 1)$ subspaces, where

$$\mathbb{V}arf_u = \int_E f_u^2(x_u)\nu(\mathrm{d}x) < \infty \quad \text{for} \quad u \subseteq B$$

and

$$\mathbb{V}arf = \int_E (f(x)-f_0)^2\nu(\mathrm{d}x) = \sum_{u\subseteq B}\mathbb{V}arf_u < \infty.$$

A key principle of HDMR is that the expansion for most f may be truncated at low order $T \ll n$ in a T-order HDMR,

$$f(x) \simeq f^T(x) = \sum_{u\subseteq B:|u|\le T} f_u(x_u) \quad \text{for} \quad T \ll n.$$

HDMR is often used in *global sensitivity analysis* to assess input–output correlations at various orders, where the variances are normalized to define *sensitivity indices*

$$\mathbb{S}_u = \frac{\mathbb{V}arf_u}{\mathbb{V}arf} \quad \text{for} \quad u \subseteq B.$$

When the inputs are correlated $\nu \neq \prod_i \nu_i$, then the component functions may still be uniquely recovered under hierarchical orthogonality, the variance decomposes

$$\mathbb{V}arf = \sum_{u,v\subseteq B}\mathbb{C}ov(f_u,f_v),$$

where

$$\mathbb{C}ov(f_u,f_v) = \int_E f_u(x_u)f_v(x_v)\nu(\mathrm{d}x) \quad \text{for} \quad u,v \subseteq B$$

and the sensitivity indices generalize to *structural* and *correlative* sensitivity indices [56], defined respectively as

$$\mathbb{S}_u^a = \frac{\mathbb{V}arf_u}{\mathbb{V}arf} \quad \text{for} \quad u \subseteq B$$

and

$$\mathbb{S}_u^b = \frac{\sum_{v \subseteq B: u \neq v} \mathbb{C}ov(f_u, f_v)}{\mathbb{V}arf} \quad \text{for} \quad u \subseteq B$$

with *total* sensitivity index

$$\mathbb{S}_u = \mathbb{S}_u^a + \mathbb{S}_u^b \quad \text{for} \quad u \subseteq B$$

where $\sum_{u \subseteq B} \mathbb{S}_u = 1$. We use the total sensitivity index as a measure of variable importance and the component functions as profiles of output dependence on the input subspaces.

Appendix G. Reliability Definitions

Given a failure distribution f and reliability (survival) distribution R, we give some relations: the cumulative failure distribution is defined as

$$F(t|\vartheta) = \int_0^t f(s|\vartheta)\mathrm{d}s$$

where

$$R(t|\vartheta) + F(t|\vartheta) = 1,$$

the hazard rate $h(t|\vartheta)$ is defined as

$$h(t|\vartheta) = \frac{f(t|\vartheta)}{R(t|\vartheta)},$$

the cumulative hazard is defined as

$$H(t|\vartheta) = \int_0^t h(s|\vartheta)\mathrm{d}s,$$

and we have reliability expressed in terms of the cumulative hazard

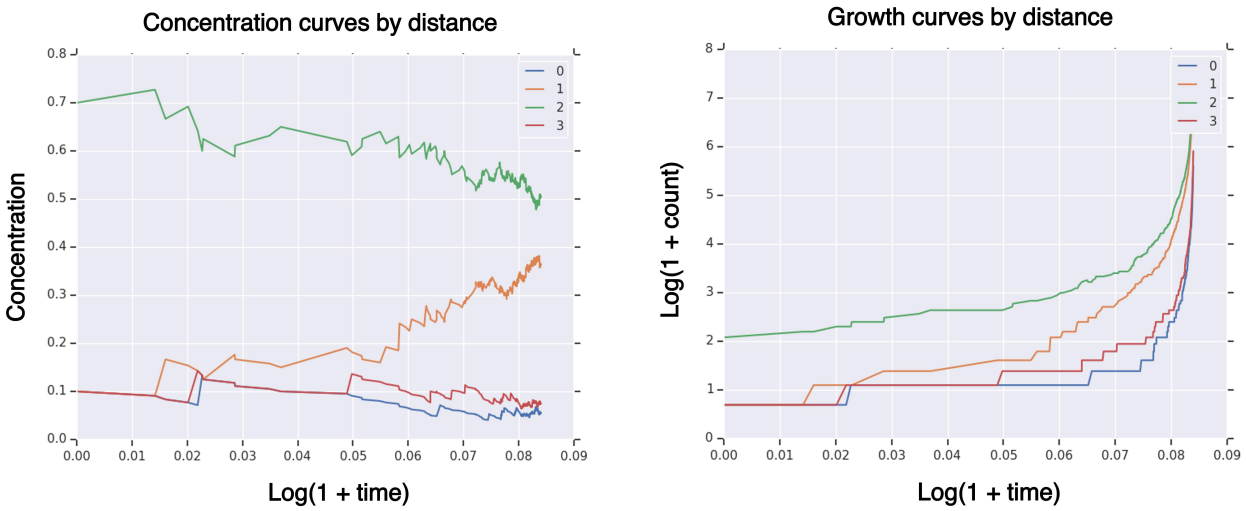
$$R(t|\vartheta) = e^{-H(t|\vartheta)}.$$

Another useful quantity is the mean residual life

$$\mu(t|\vartheta) = \frac{\int_t^\infty R(s|\vartheta)\mathrm{d}s}{R(t|\vartheta)}.$$

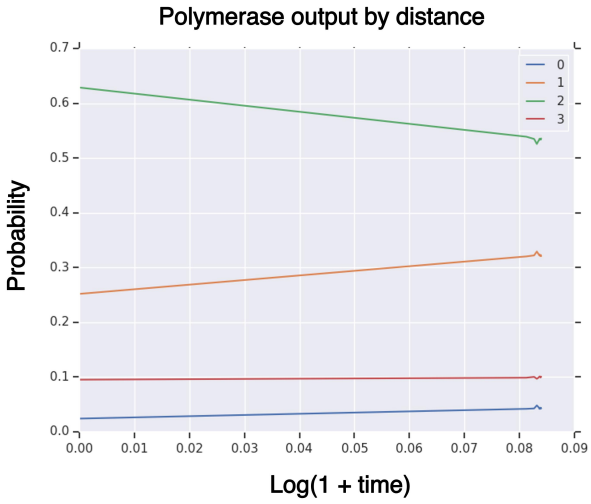
Appendix H. Additional Figures

Appendix H.1. Linear Landscape



(**a**)

(**b**)



(**c**)

Figure A1. Linear landscape: Measures of system population X_t until hitting time τ_v for high-fidelity replicator volume fraction $v = 0.25$ with sequence dimension $n = 3$, fitness/similarity curvature $k = l = -\log(0.01)/n$, initial population size, $I = |X_0| = 10$, singleton high-fidelity replicator $R = \{\{x\}\}$, with linear fitness and similarity functions. (**a**) Concentration of RNA sequences by Hamming distance to high-fidelity replicator; (**b**) population size of RNA sequences by Hamming distance to high-fidelity replicator; (**c**) polymerase RNA sequence output by Hamming distance to high-fidelity replicator.

Appendix H.2. Core Model with Clay

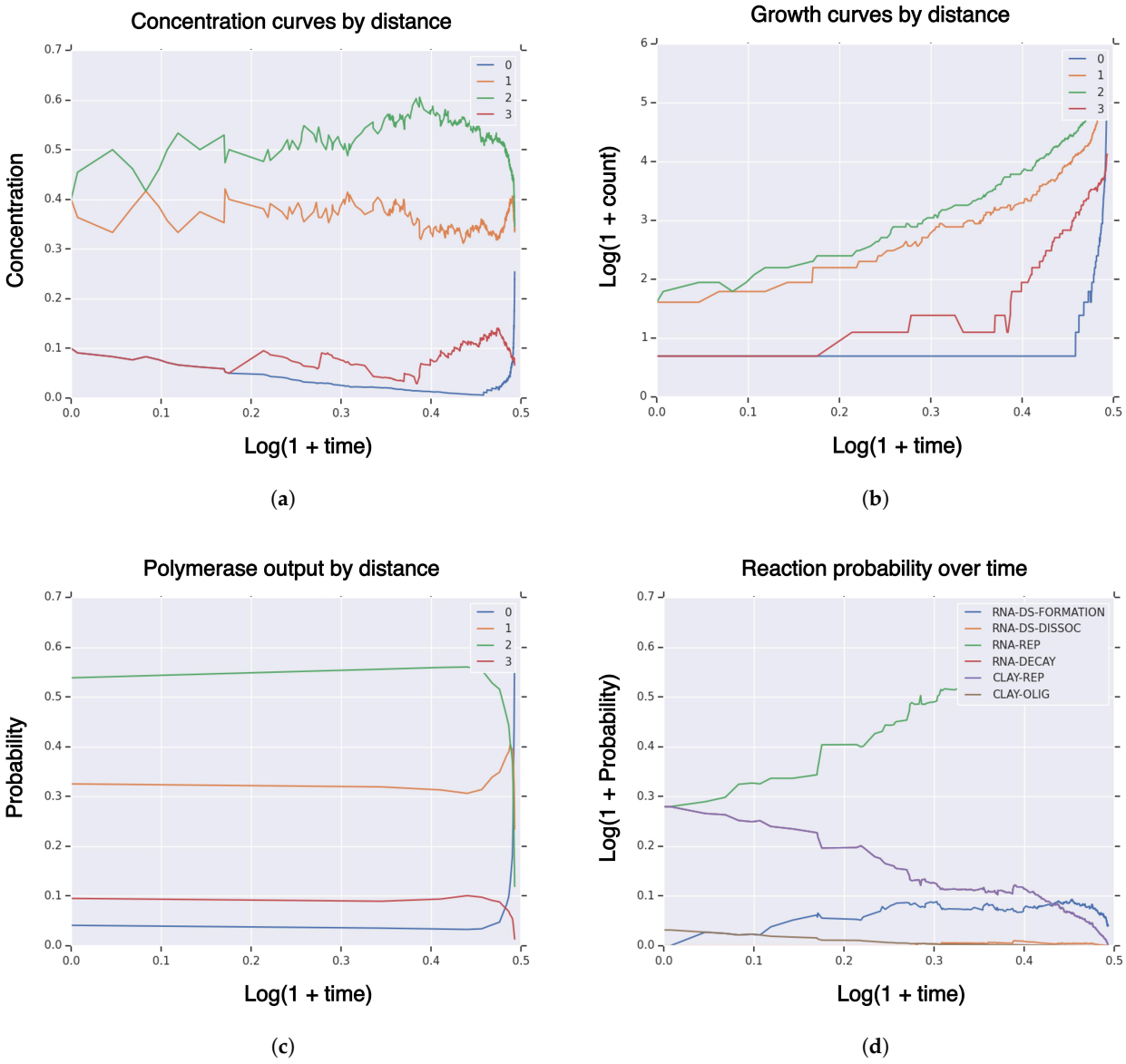


Figure A2. Core model with clay: Measures of system population X_t until hitting time $\tau_v(\theta)$ for high-fidelity replicator volume fraction $v = 0.25$ with sequence dimension $n = 3$, fitness/similarity curvature $k = -\log(0.01)/n$, initial population size $I = |X_0| = 10$, singleton high-fidelity replicator $R = \{\{x\}\}$, double strand separation and formation rates reaction rate $k_{ss} = k_{ds} = 1$, clay replication fidelity probability $p = 0.9$, and RNA polymerization rate k_{rep} and clay polymerization rate k_{clay-p} chosen such that the replicative mass of each is 10. (**a**) Concentration of RNA sequences by Hamming distance to high-fidelity replicator; (**b**) population size of RNA sequences by Hamming distance to high-fidelity replicator; (**c**) polymerase RNA sequence output by Hamming distance to high-fidelity replicator; (**d**) probability of reactions over time.

Appendix H.3. Hitting/Survival Analysis

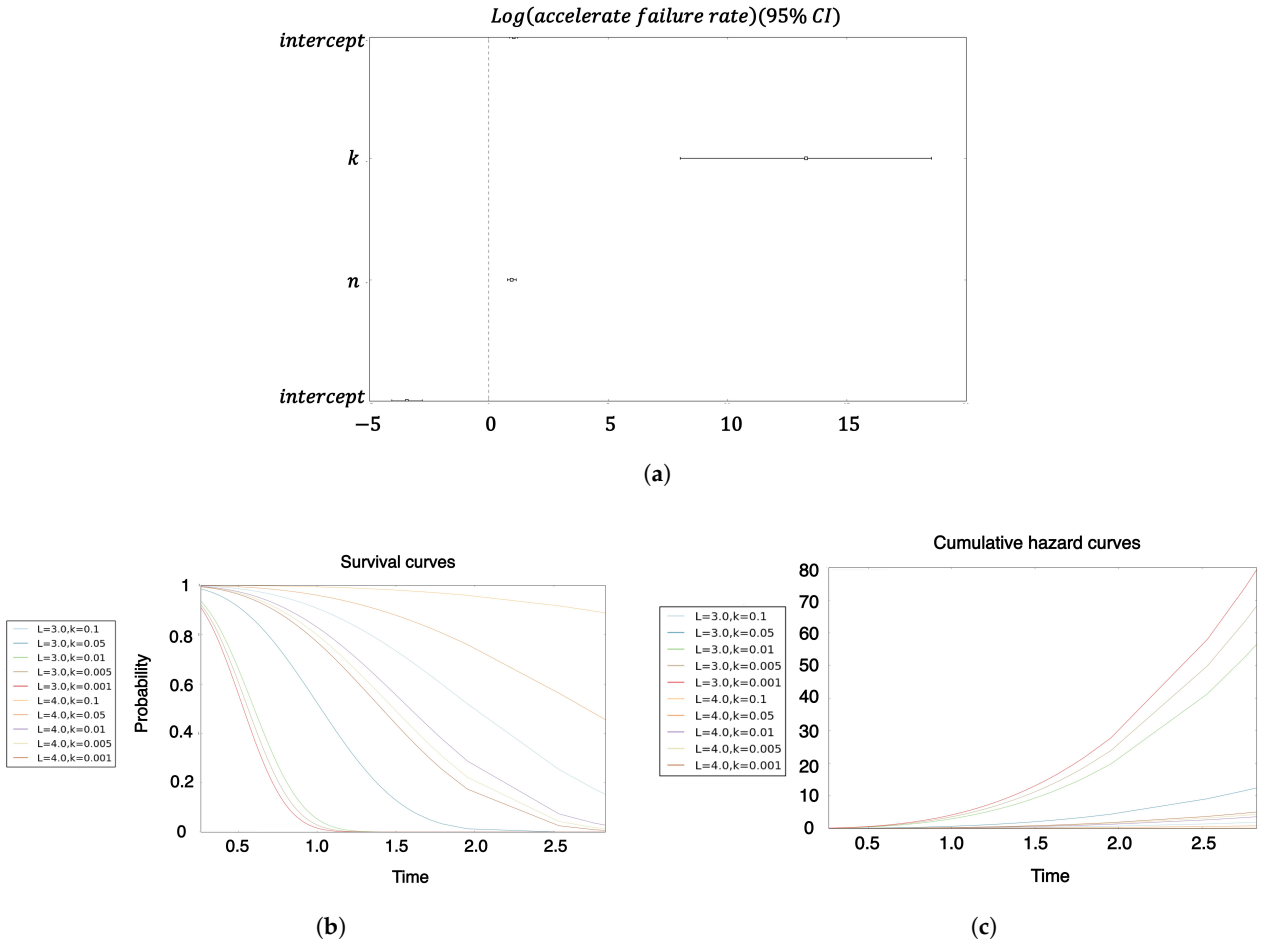
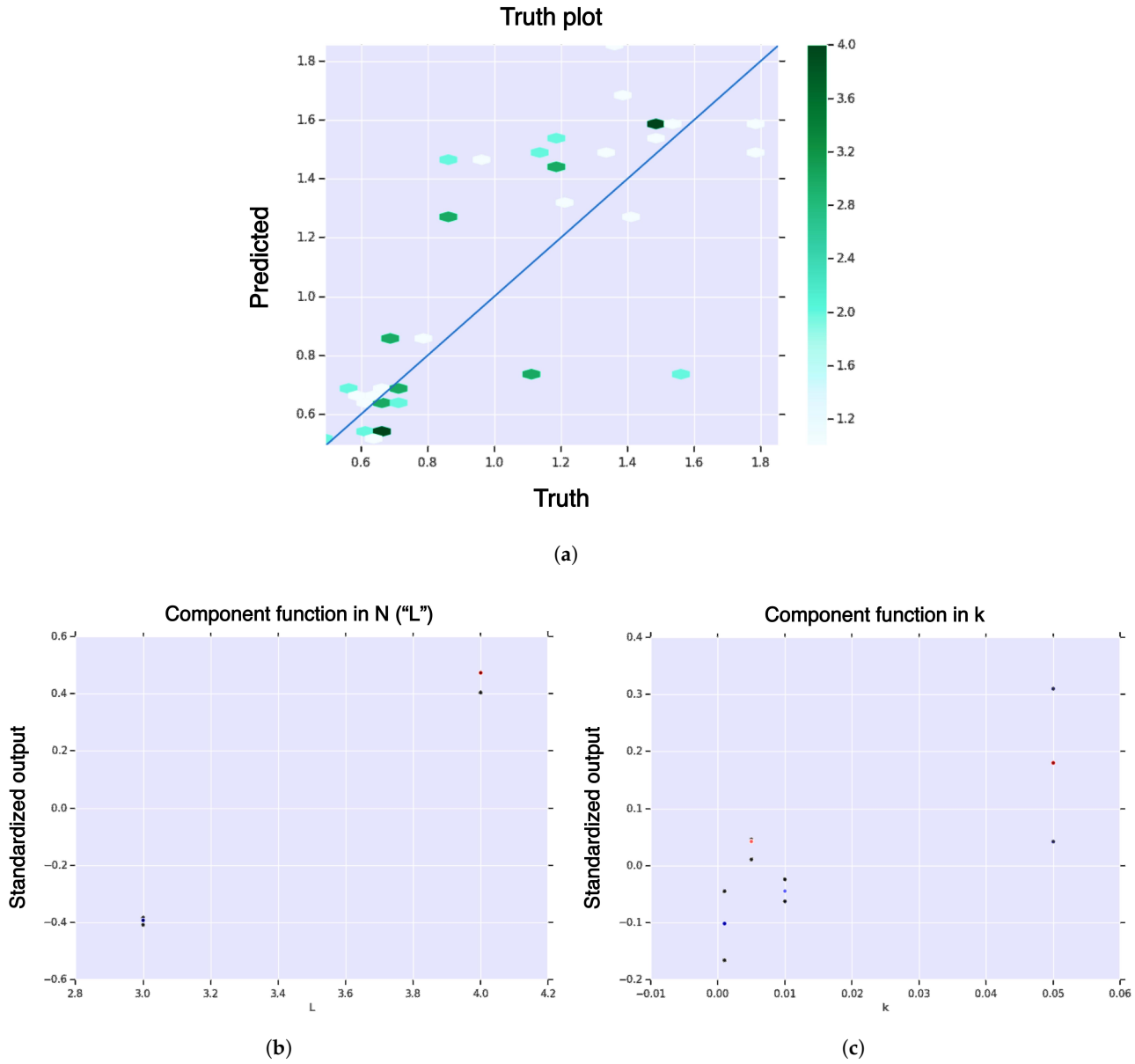


Figure A3. Survival analysis of hitting time τ_v for high-fidelity replicator volume fraction $v = 0.1$ for core model. (**a**) Coefficients of the Cox proportional hazard survival model; (**b**) survival curves in sequence dimension $n = L$ and landscape curvature $k = l$; (**c**) cumulative hazard in sequence dimension $n = L$ and landscape curvature $k = l$.



(**a**)

(**b**) (**c**)

Figure A4. First-order HDMR analysis of hitting time $\tau_v(\theta) < \infty$ for core model. (**a**) Hexagonal-bin truth plot; (**b**) HDMR component function for sequence length $n = L$, $f_n(n)$, in sequence length for hitting time, of hitting time; (**c**) HDMR component function for landscape curvature, $f_k(k)$, in landscape curvature, of hitting time. The color function is from blue (negative) to white (zero) to red (positive). The black dots represent standard deviation of the error.

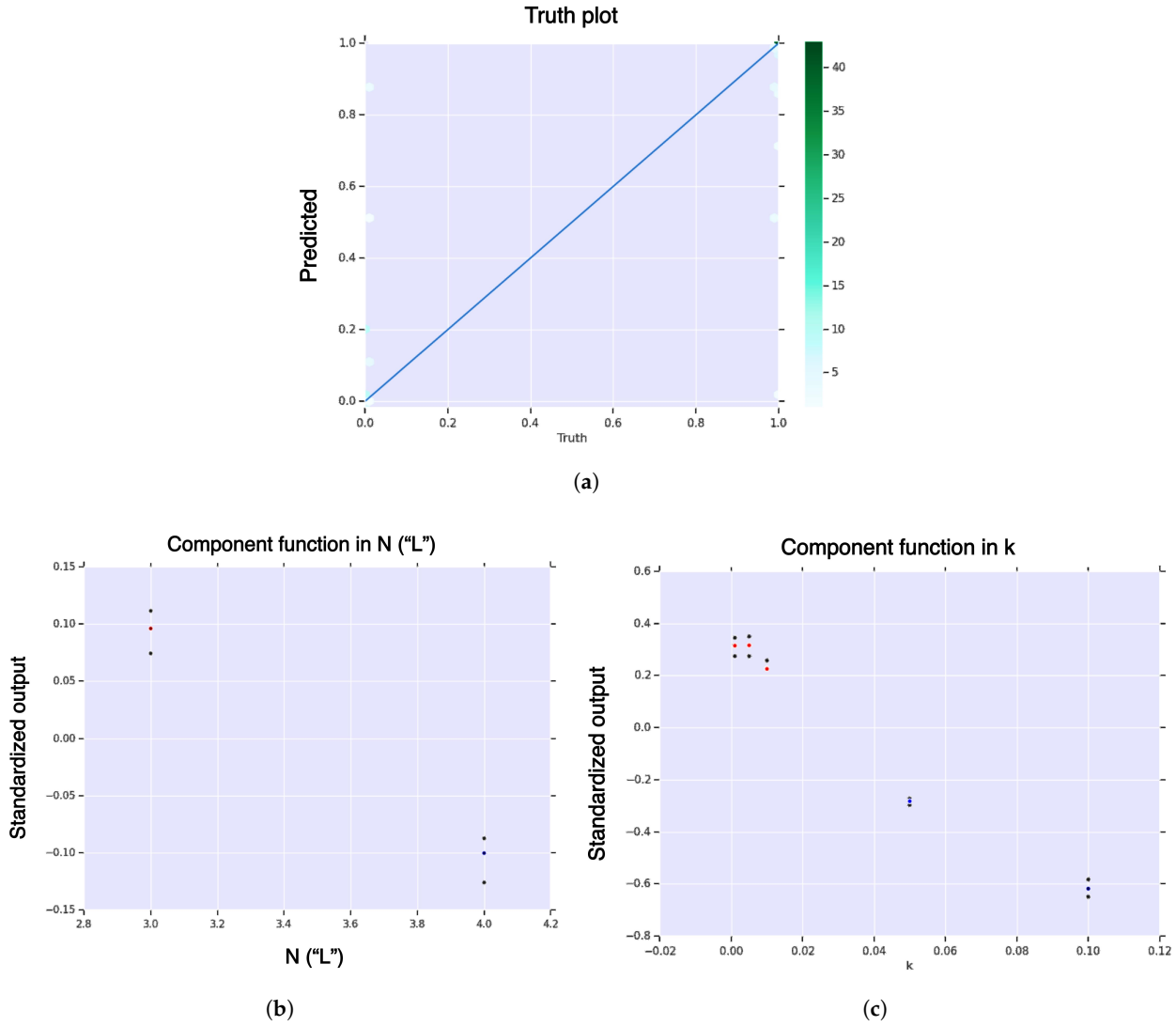


Figure A5. First-order HDMR analysis of hitting probability $\mathbb{P}(\tau_v(\theta) < \infty)$ for core model. (**a**) Hexagonal-bin truth plot; (**b**) HDMR component function for sequence length, $f_n(n)$, in sequence length, of hitting probability; (**c**) HDMR component function for landscape curvature, $f_k(k)$, in landscape curvature, of hitting probability. The color function is from blue (negative) to white (zero) to red (positive). The black dots represent standard deviation of the error.

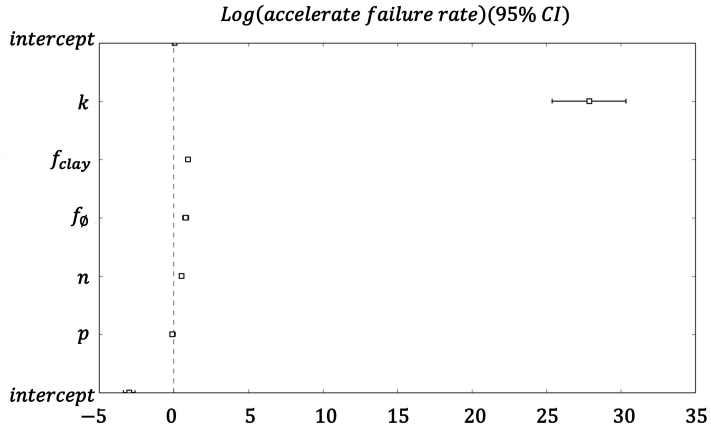


Figure A6. Survival analysis of hitting time τ_v for volume fraction $v = 0.1$ for expanded model (clay and decay). Coefficients of the Cox proportional hazards survival model.

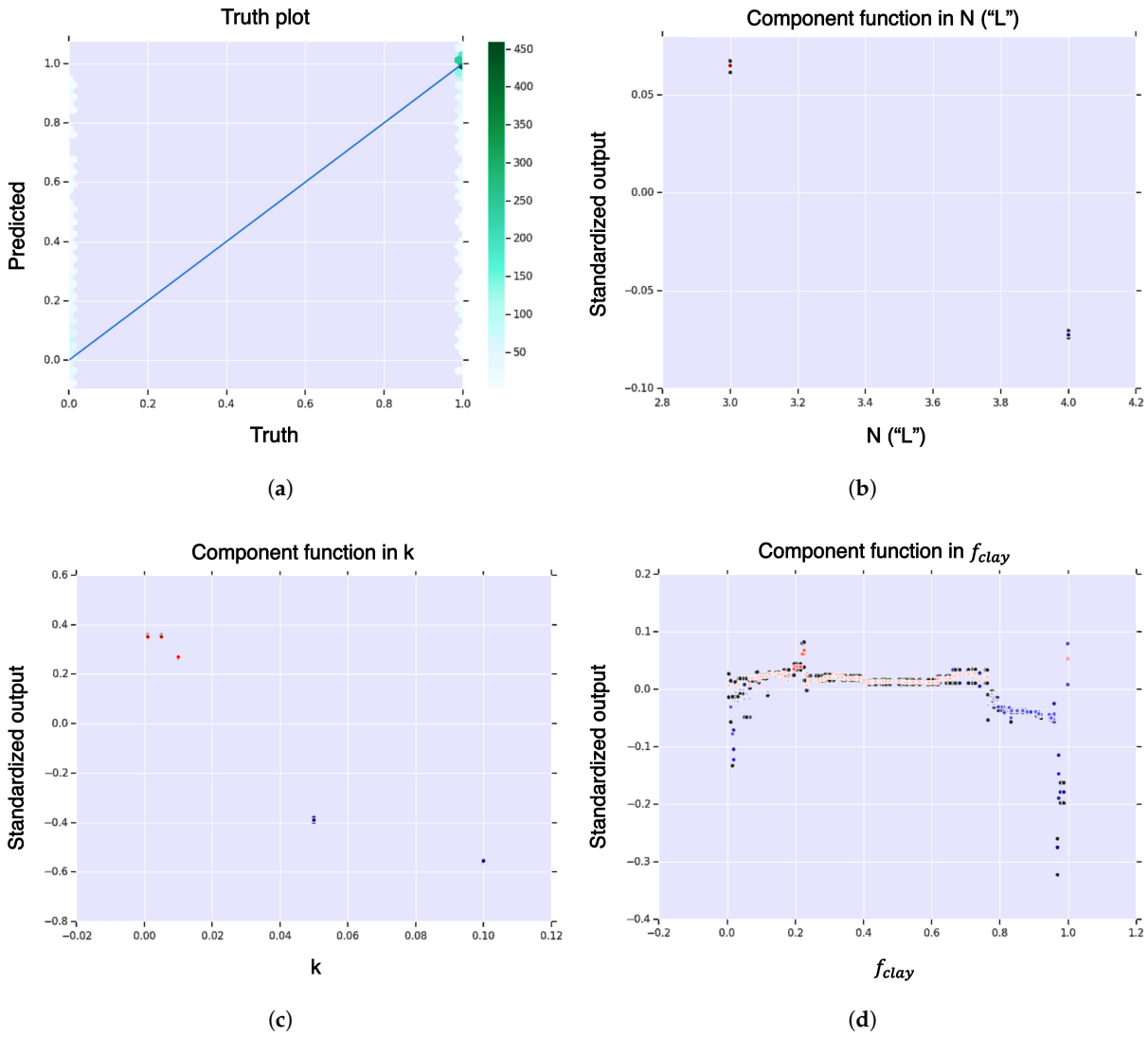


Figure A7. First-order HDMR analysis of hitting probability $\mathbb{P}(\tau_v(\theta) < \infty)$ for expanded model (clay and decay). (**a**) Hexagonal-bin truth plot; (**b**) HDMR component function for sequence length $n = L$, $f_n(n)$, in sequence length, of hitting probability; (**c**) HDMR component function for curvature, $f_k(k)$, in curvature, of hitting probability; (**d**) HDMR component function for clay fitness, $f_{f_{clay}}(f_{clay})$, in clay fitness, of hitting probability. The color function is from blue (negative) to white (zero) to red (positive). The black dots represent standard deviation of the error.

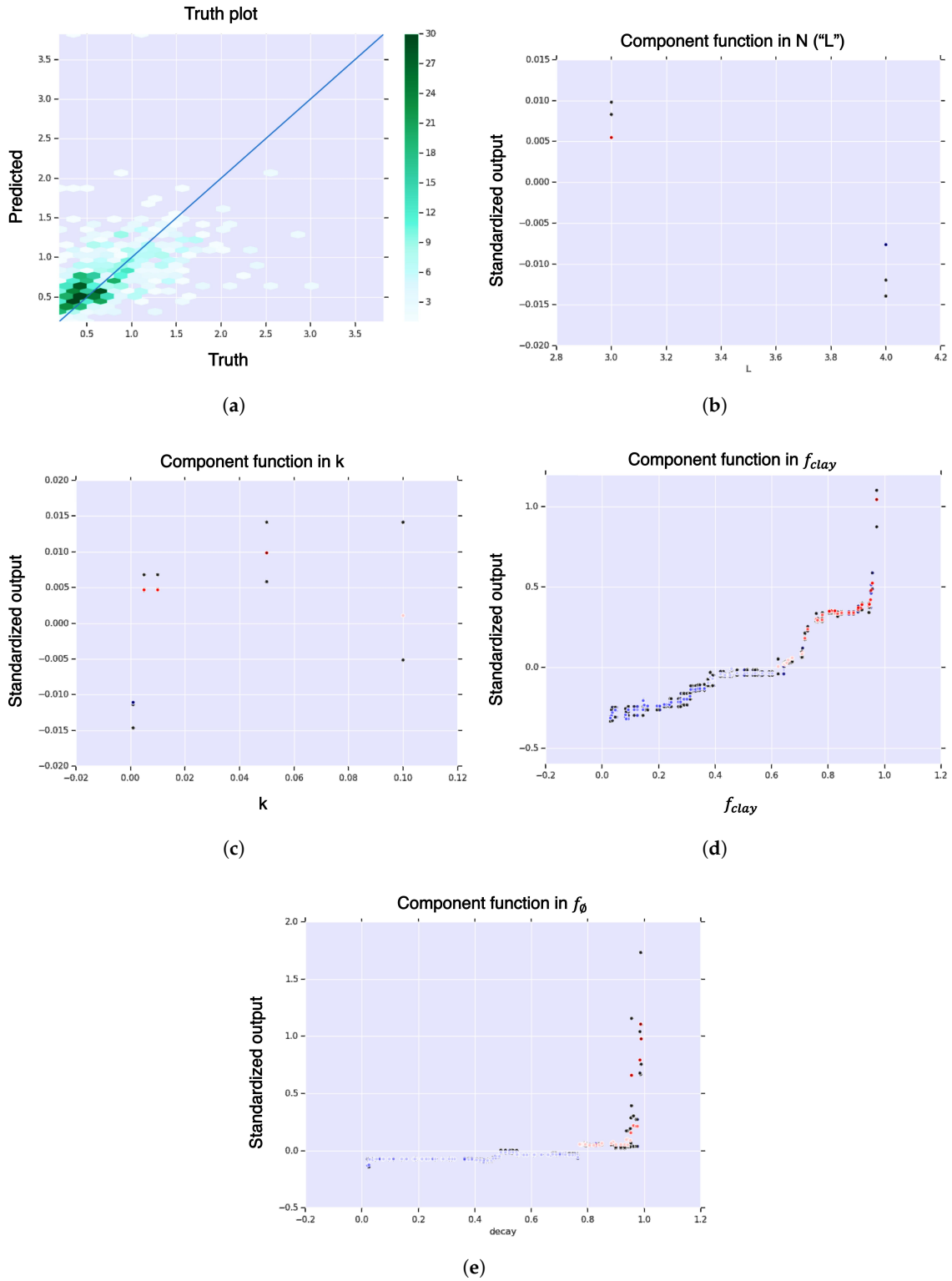
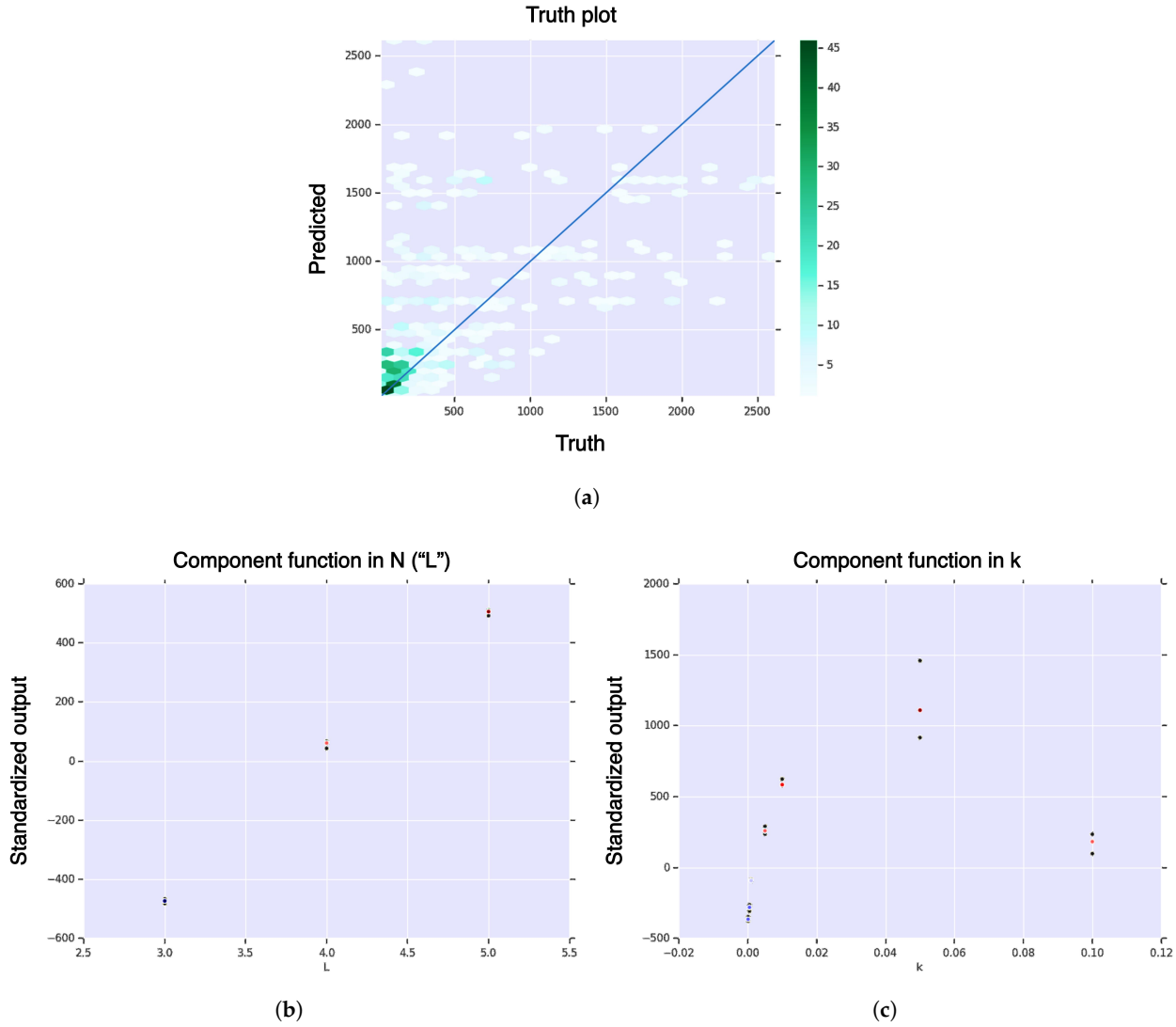


Figure A8. First-order HDMR analysis of hitting time $\tau_v(\theta) < \infty$ for expanded model (clay and decay). (**a**) Hexagonal-bin truth plot; (**b**) HDMR component function for sequence length, $f_n(n)$, in sequence length, of hitting time; (**c**) HDMR component function for curvature, $f_k(k)$, in curvature, of hitting time; (**d**) HDMR component function for clay-fraction, $f_{f_{clay}}(f_{clay})$, in clay fraction, of hitting time; (**e**) HDMR component function for decay rate, $f_{k_\varnothing}(k_\varnothing)$, in decay rate, of hitting time. The color function is from blue (negative) to white (zero) to red (positive). The black dots represent standard deviation of the error.



(**a**)

(**b**) (**c**)

Figure A9. First-order HDMR analysis of hitting cardinality $\varpi_v(\theta) < \infty$ for core model and volume fraction $v = 0.1$. (**a**) Hexagonal-bin truth plot; (**b**) HDMR component function for sequence dimension, $f_n(n)$, in sequence dimension, of hitting cardinality; (**c**) HDMR component function for curvature, $f_k(k)$, in curvature, of hitting cardinality. The color function is from blue (negative) to white (zero) to red (positive). The black dots represent standard deviation of the error.

References

1. Ganti, T. *The Principles of Life*; Oxford University Press: Oxford, UK, 2003.
2. Ganti, T. *Chemoton Theory*; Kluwer Academic/Plenum Publishers: Dordrecht, The Netherlands, 2003.
3. Orgel, L.E. Evolution of the genetic apparatus. *J. Mol. Biol.* **1968**, *38*, 381–393. [CrossRef]
4. Gilbert, W. Origin of life: The RNA world. *Nature* **1986**, *319*, 618–618. [CrossRef]
5. Joyce, G.F. The antiquity of RNA-based evolution. *Nature* **2002**, *418*, 214–221. [CrossRef] [PubMed]
6. White, H.B. Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **1976**, *7*, 101–104. [CrossRef]
7. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **1971**, *58*, 465–523. [CrossRef] [PubMed]
8. Eigen, M.; Schuster, P. A principle of natural self-organization. *Naturwissenschaften* **1977**, *64*, 541–565. [CrossRef]
9. Cech, T.R. The Ribosome Is a Ribozyme. *Science* **2000**, *289*, 878. [CrossRef]
10. Diener, T.O. Potato spindle tuber "virus". IV. A replicating, low molecular weight RNA. *Virology* **1971**, *45*, 411–428. [CrossRef]
11. 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]
12. 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] [PubMed]

13. de Farias, S.T.; dos Santos Junior, A.P.; Rêgo, T.G.; José, M.V. Origin and Evolution of RNA-Dependent RNA Polymerase. *Front. Genet.* **2017**, *8*, 125. [CrossRef] [PubMed]
14. Koonin, E.V.; Krupovic, M.; Ishino, S.; Ishino, Y. The replication machinery of LUCA: Common origin of DNA replication and transcription. *BMC Biol.* **2020**, *18*, 61. [CrossRef]
15. Ghadessy, F.J.; Ong, J.L.; Holliger, P. Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 4552. [CrossRef] [PubMed]
16. Tjhung, K.F.; Shokhirev, M.N.; Horning, D.P.; Joyce, G.F. An RNA polymerase ribozyme that synthesizes its own ancestor. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 2906. [CrossRef]
17. Ertem, G.; Ferris, J.P. Template-Directed Synthesis Using the Heterogeneous Templates Produced by Montmorillonite Catalysis. A Possible Bridge Between the Prebiotic and RNA Worlds. *J. Am. Chem. Soc.* **1997**, *119*, 7197–7201. [CrossRef]
18. Acevedo, O.L.; Orgel, L.E. Non-enzymatic transcription of an oligodeoxynucleotide 14 residues long. *J. Mol. Biol.* **1987**, *197*, 187–193. [CrossRef]
19. Szostak, J.W. The eightfold path to non-enzymatic RNA replication. *J. Syst. Chem.* **2012**, *3*, 2. [CrossRef]
20. Cairns-Smith, A.G. *Genetic Takeover and the Mineral Origins of Life*; Cambridge University Press: Cambridge, UK, 1987.
21. Dyson, F. *Origins of Life*; Cambridge University Press: Cambridge, UK, 1999. [CrossRef]
22. Sakuma, Y.; Imai, M. From vesicles to protocells: The roles of amphiphilic molecules. *Life* **2015**, *5*, 651–675. [CrossRef]
23. Szostak, J.W.; Bartel, D.P.; Luisi, P.L. Synthesizing life. *Nature* **2001**, *409*, 387–390. [CrossRef]
24. Segré, D.; Ben-Eli, D.; Deamer, D.W.; Lancet, D. The Lipid World. *Orig. Life Evol. Biosph.* **2001**, *31*, 119–145. [CrossRef]
25. Martin, W.; Baross, J.; Kelley, D.; Russell, M.J. Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **2008**, *6*, 805–814. [CrossRef] [PubMed]
26. Damer, B.; Deamer, D. The Hot Spring Hypothesis for an Origin of Life. *Astrobiology* **2019**, *20*, 429–452. [CrossRef]
27. Damer, B.; Deamer, D. Coupled phases and combinatorial selection in fluctuating hydrothermal pools: A scenario to guide experimental approaches to the origin of cellular life. *Life* **2015**, *5*, 872–887. [CrossRef] [PubMed]
28. Deamer, D.; Damer, B.; Kompanichenko, V. Hydrothermal Chemistry and the Origin of Cellular Life. *Astrobiology* **2019**, *19*, 1523–1537. [CrossRef]
29. Kvenvolden, K.; Lawless, J.; Pering, K.; Peterson, E.; Flores, J.; Ponnamperuma, C.; Kaplan, I.R.; Moore, C. Evidence for Extraterrestrial Amino-acids and Hydrocarbons in the Murchison Meteorite. *Nature* **1970**, *228*, 923–926. [CrossRef] [PubMed]
30. Miller, S.L.; Urey, H.C. Organic Compound Synthesis on the Primitive Earth. *Science* **1959**, *130*, 245. [CrossRef]
31. Pino, S.; Sponer, J.E.; Costanzo, G.; Saladino, R.; Mauro, E.D. From formamide to RNA, the path is tenuous but continuous. *Life* **2015**, *5*, 372–384. [CrossRef]
32. Powner, M.W.; Sutherland, J.D. Prebiotic chemistry: A new modus operandi. *Phil. Trans. R. Soc. B Biol. Sci.* **2011**, *366*, 2870–2877. [CrossRef]
33. Powner, M.W.; Gerland, B.; Sutherland, J.D. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature* **2009**, *459*, 239–242. [CrossRef]
34. Attwater, J.; Wochner, A.; Holliger, P. In-ice evolution of RNA polymerase ribozyme activity. *Nat. Chem.* **2013**, *5*, 1011–1018. [CrossRef]
35. Hays, L. *NASA Astrobiology Strategy*; Technical report; National Aeronautics and Space Administration: Washington, DC, USA, 2015.
36. Coveney, P.V.; Swadling, J.B.; Wattis, J.A.D.; Greenwell, H.C. Theory, modelling and simulation in origins of life studies. *Chem. Soc. Rev.* **2012**, *41*, 5430–5446. [CrossRef]
37. Lanier, K.A.; Williams, L.D. The Origin of Life: Models and Data. *J. Mol. Evol.* **2017**, *84*, 85–92. [CrossRef]
38. Wu, M.; Higgs, P.G. The origin of life is a spatially localized stochastic transition. *Biol. Direct* **2012**, *7*, 42. [CrossRef]
39. Gillespie, D.T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **1977**, *81*, 2340–2361. [CrossRef]
40. Hordijk, W.; Steel, M. Detecting autocatalytic, self-sustaining sets in chemical reaction systems. *J. Theor. Biol.* **2004**, *227*, 451–461. [CrossRef]
41. Walker, S.I. Origins of life: A problem for physics, a key issues review. *Rep. Prog. Phys.* **2017**, *80*, 092601. [CrossRef]
42. Wattis, J.A.D.; Coveney, P.V. The Origin of the RNA World: A Kinetic Model. *J. Phys. Chem. B* **1999**, *103*, 4231–4250. [CrossRef]
43. Kun, Á.; Szilágyi, A.; Könnyű, B.; Boza, G.; Zachar, I.; Szathmáry, E. The dynamics of the RNA world: Insights and challenges. *Ann. N. Y. Acad. Sci.* **2015**, *1341*, 75–95. [CrossRef] [PubMed]
44. Szilágyi, A.; Zachar, I.; Scheuring, I.; Kun, Á.; Könnyű, B.; Czárán, T. Ecology and Evolution in the RNA World Dynamics and Stability of Prebiotic Replicator Systems. *Life* **2017**, *7*, 48. [CrossRef] [PubMed]
45. Scheuring, I.; Szilágyi, A. Diversity, stability, and evolvability in models of early evolution. *Curr. Opin. Syst. Biol.* **2019**, *13*, 115–121. [CrossRef]
46. Takeuchi, N.; Hogeweg, P. Evolutionary dynamics of RNA-like replicator systems: A bioinformatic approach to the origin of life. *Phys. Life Rev.* **2012**, *9*, 219–263. [CrossRef]
47. Orgel, L. A Simpler Nucleic Acid. *Science* **2000**, *290*, 1306. [CrossRef] [PubMed]
48. Maury, C.P.J. Origin of life. Primordial genetics: Information transfer in a pre-RNA world based on self-replicating beta-sheet amyloid conformers. *J. Theor. Biol.* **2015**, *382*, 292–297. [CrossRef]

49. Ehrenfreund, P.; Rasmussen, S.; Cleaves, J.; Chen, L. Experimentally Tracing the Key Steps in the Origin of Life: The Aromatic World. *Astrobiology* **2006**, *6*, 490–520. [CrossRef]
50. Kunin, V. A System of Two Polymerases—A Model for the Origin of Life. *Orig. Life Evol. Biosph.* **2000**, *30*, 459–466. [CrossRef] [PubMed]
51. Wright, S. *Evolution and the Genetics of Populations, Volume 1*; The University of Chicago Press: Chicago, IL, USA, 1984.
52. Feng, X.; Pechen, A.; Jha, A.; Wu, R.; Rabitz, H. Global optimality of fitness landscapes in evolution. *Chem. Sci.* **2012**, *3*, 900–906. [CrossRef]
53. Davidson-Pilon, C. lifelines: Survival analysis in Python. *J. Open Source Softw.* **2019**, *4*, 1317. [CrossRef]
54. Bornholt, J.; Lopez, R.; Carmean, D.M.; Ceze, L.; Seelig, G.; Strauss, K. A DNA-Based Archival Storage System. In Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, Association for Computing Machinery, New York, NY, USA, 19–23 April 2016; pp. 637–649. [CrossRef]
55. Kitadai, N.; Maruyama, S. Origins of building blocks of life: A review. *Geosci. Front.* **2018**, *9*, 1117–1153. [CrossRef]
56. Li, G.; Rabitz, H. General formulation of HDMR component functions with independent and correlated variables. *J. Math. Chem.* **2012**, *50*, 99–130. [CrossRef]