Nonequilibrium abundances for the building blocks of life

Elan Stopnitzky[1,*] and Susanne Still[2,1]
[1]*Department of Physics and Astronomy, University of Hawai'i at Mānoa, Honolulu, Hawai'i 96822, USA*
[2]*Department of Information and Computer Sciences, University of Hawai'i at Mānoa, Honolulu, Hawai'i 96822, USA*

(Received 20 August 2018; revised manuscript received 4 April 2019; published 2 May 2019)

The difficulty of obtaining appreciable quantities of biologically important molecules in thermodynamic equilibrium has long been identified as an obstacle to life's emergence, and determining the specific nonequilibrium conditions that might have given rise to life is challenging. To address these issues, we investigate how the concentrations of life's building blocks change as a function of the distance from equilibrium *on average*, in two example settings: (i) the synthesis of heavy amino acids and (ii) their polymerization into peptides. We find that relative concentrations of the heaviest amino acids can be boosted by four orders of magnitude, and concentrations of the longest peptide chains can be increased by hundreds of orders of magnitude. The average nonequilibrium distribution does not depend on the details of how the system was driven from equilibrium, indicating that environments might not have to be fine-tuned to support life.

DOI: 10.1103/PhysRevE.99.052101

I. INTRODUCTION

Biology requires the coordination of many complex molecules to store and copy genetic information, harness energy from the environment, and maintain homeostasis. The spontaneous emergence of life thus hinges upon the abundances of such molecules in an abiotic environment. At first glance, statistical mechanics seems to pose a serious barrier: the high molecular mass and structural specificity of many biomolecules severely limit their abundances in thermodynamic equilibrium and thus make the emergence of life implausible [1–5]. Many biomolecules require considerable free energy to form, and this leads to an exponential suppression of their equilibrium concentrations.

The apparent severity of this problem, which appears under rather general considerations, has motivated researchers to search for special environments, either extant or belonging to the early Earth, which would be ideally suited for producing the necessary molecules in significant quantities. Due to the free-energy requirement, an essential feature of these environments is that they include nonequilibrium driving of some kind [1,3–7]. Some proposed sources of this driving on prebiotic Earth are radiation [6,7], temperature and ion gradients [6–9], concentration fluxes [10,11], and electrical discharge [12]. Examples of such environments include hydrothermal vent systems [13–16] and the surfaces of minerals [3]. Yet it remains an open question to what extent environmental conditions must be fine-tuned to give rise to life, and there is considerable uncertainty about the chemistry of the early Earth [5,17,18].

Here, we use a relatively new approach proposed by Crooks [19] which allows us to explore how the abundances of life's building blocks change away from thermodynamic equilibrium *on average*, where the average is taken over all the possible ways the system could be driven from equilibrium and depends only on a simple parametric measure of the distance from equilibrium.

Our calculation does not hinge upon specific assumptions about the conditions that might have created life and therefore does not require significant knowledge about the early Earth. The question we answer is more general: can we quantify how much nonequilibrium conditions *typically* change the abundances of the complex molecules that life relies on? We study this dependence for two simple models describing, respectively, the concentrations of heavy amino acids and their polymerization into peptides. The result is that away from equilibrium, the abundances of rare molecules become, on average, increasingly favorable, potentially boosted by many orders of magnitude. The specific forms of nonequilibrium driving previously considered can thus be recognized as part of a much more general phenomenon, whereby driving is expected on average to increase the probabilities of rare states as one moves further from equilibrium. By dramatically augmenting the concentrations of biologically important molecules without fine-tuning conditions, this effect makes the appearance of life on Earth a much more plausible event.

II. THE NONEQUILIBRIUM MODEL

Statistical mechanics tells us that we do not need to describe the full microscopic state of a system in order to predict macroscopic characteristics, as those are understood as expectation values, or ensemble averages. Therefore, all we need to infer is the *probability*, ρ_i, of every state, $i = 1, \ldots N$. This is a hard problem, as we have only a handful of constraints, namely measured average quantities, together with normalization of probability. Say we have M constraints. Then we are still lacking $N - M$ equations to determine the ρ_i. These probabilities can be assigned by choosing the probability distribution with the largest entropy, $S[\rho] \equiv -\sum_{i=1}^{N} \rho_i \ln(\rho_i)$, subject to the constraints imposed by the

*elanstop@hawaii.edu

system's bulk properties [20]. This maximization of entropy can be interpreted as choosing a model that makes use of only the information provided by the measured quantities [20–22], ensuring that information that we do not actually have is not falsely being ascribed to the system. This powerful inference tool has been applied successfully to many other problems in a diverse range of fields from ecology to neuroscience, and is commonly known under the name of MaxEnt [23,24]. In statistical physics, we find that under the constraint that only the average energy is known, the Boltzmann distribution, describing thermodynamic equilibrium states, $\rho = \frac{1}{Z}e^{-E/kT}$, is recovered by this MaxEnt inference method [20]. Here, the temperature is denoted by T, Bolzmann's constant by k, and normalization is ensured by the partition function, Z, which is related to the equilibrium free energy, F, by $Z = e^{-F/kT}$.

On the early Earth, conditions governing the processes preceding life were *not* consistently in thermodynamic equilibrium. It is much harder to infer the distribution, θ, of a system that is away from thermodynamic equilibrium without detailed information. The distribution can no longer be inferred straight from a MaxEnt argument, and information is lacking to make up for the missing equations. Without specific knowledge about some particular process generating biomolecules on early Earth, little can be done.

Here, we propose to calculate instead the *average* nonequilibrium distribution. The idea is that there are many diverse environments on Earth and a large variety of energy sources that act as nonequilibrium drives. If all we are interested in are the expected abundances we would get somewhere on Earth, then we can average out details of the nonequilibrium driving. We do so, following [19], by giving probability distributions a weight, i.e., we will assume that there is a *distribution over distributions*, $P(\theta)$, and compute the average: $\langle\theta\rangle = \int \theta P(\theta) d\theta$.

For our purposes, we need only consider distributions on a discrete state space. We will compare the probability of finding the building blocks of life as computed from this average nonequilibrium distribution to that computed from the equilibrium distribution for two biologically relevant model systems in the following sections. Clearly, the answer will depend on the probabilities assigned to different nonequilibrium probability distributions, $P(\theta)$. Crooks suggested [19] to find $P(\theta)$ by maximizing the entropy of the distribution over distributions subject to physical constraints, in analogy to what is done in equilibrium [20]. In the absence of additional information, this maximum entropy approach ought to best describe the ensemble of nonequilibrium distributions, as it ensures that only available information is included in the description.

We elaborate on the details of Crooks' approach in the Appendix and mention here only the resulting formula:

$$\langle\theta\rangle = \frac{1}{\mathcal{Z}(\beta,\lambda)} \int \theta e^{-\lambda D(\theta\|\rho)} d\theta. \tag{1}$$

The normalization constant, $\mathcal{Z}(\beta,\lambda)$ depends on the inverse temperature, $\beta = 1/kT$, where k denotes the Boltzmann constant. The factor $e^{-\lambda D(\theta\|\rho)}$ determines the weight given to each distribution θ. It is controlled by the product of the distribution-independent parameter $\lambda \geqslant 0$, and the relative entropy $D(\theta||\rho)$ between the nonequilibrium distribution, θ, in question, and the corresponding equilibrium distribution ρ:

$$D(\theta||\rho) = \sum_i \theta_i \ln\left[\frac{\theta_i}{\rho_i}\right]. \tag{2}$$

A system away from thermodynamic equilibrium contains free energy in excess of the corresponding equilibrium system. This additional free energy is given by $kTD(\theta||\rho)$ [25–28]. The second law of thermodynamics implies that the work input to a system is always greater than or equal to the corresponding change in free energy, and so this formalism assigns higher probabilities to distributions that require a lower minimum amount of work to create. At a fixed value of the parameter λ, a nonequilibrium distribution is thus more likely to occur, if less work is needed to produce it.

Relative entropy also measures the coding cost encountered when the canonical distribution ρ is used as a model for θ [29,30]. Relative entropy is thus both a physically and an information-theoretically meaningful measure for deviation from equilibrium.

In equilibrium, it is the free-energy difference between reactants and products alone that sets their relative abundances. Thus, a natural measure for the difficulty of creating a molecule is its free energy of formation. The hyperensemble extends this notion, in a sense, to the situation away from equilibrium.

We stress that using this approach does not imply that possible path dependencies of the nonequilibrium states are being neglected; they may very well retain some memory of their history. Each system in the ensemble is driven to a nonequilibrium distribution in a path-dependent way, as the arrival at a distribution, in general, depends on the trajectory generated by the drive.

In the limit $\lambda \to \infty$, all nonequilibrium distributions will have negligible probability, and the average nonequilibrium distribution converges to the equilibrium distribution: $\langle\theta\rangle = \rho$. For finite values of λ, the distribution $\langle\theta\rangle$ is in general flatter than its equilibrium counterpart, thereby augmenting the probabilities of states that would otherwise be rare [19]. This is most apparent in the limit $\lambda \to 0$. In that case, all distributions become equally likely. In that sense, λ encodes the extent to which driving conditions can push the system out of equilibrium. In the most extreme out of equilibrium limit ($\lambda \to 0$), the average distribution over a finite state-space has, by symmetry, equal probabilities for every state. The overall flattening effect would persist if we were to replace the measure $D(\theta||\rho)$ with any other function (see the Appendix for details). Conclusions we draw for the extreme nonequilibrium limit ($\lambda \to 0$) in Sec. III are therefore invariant with respect to how distance from equilibrium is measured.

The extreme nonequilibrium limit is different from the high-temperature limit of an equilibrium distribution, because the free energies of the molecules are themselves temperature dependent, and so the high-temperature limit would not assign equal weight to every possible distribution of molecules. For example, polymerization of amino acids into long chains would generally be disfavored in the high-temperature limit (compare Sec. III B).

Probabilities of rare states are only augmented on average. There are individual nonequilibrium systems that give

rise to worse-than-equilibrium odds for forming the desired molecules. The nonequilibrium distributions describing those systems are included in the average. Individual distributions that exhibit large numbers of rare molecules are less probable at all finite values of λ, due to the exponential dependence on $D(\theta||\rho)$ [see Eq. (A1)]. In what follows, it is the average nonequilibrium distribution $\langle\theta\rangle$, and not any particular nonequilibrium distribution θ, that we use for our analysis.

We interpret the average nonequilibrium distribution as describing the result that would be obtained if one took samples from a diverse collection of nonequilibrium environments, and averaged the concentrations of the various molecules found. The average nonequilibrium distribution provides the expected value, or best guess, for what we would find in a single sample, taken anywhere on the planet. In the context of molecules relevant for forming living structures, using the average nonequilibrium distribution to make an inference about relative abundances should be more appropriate than using the equilibrium distribution, because we know that conditions on early earth were not consistently in thermodynamic equilibrium.

The average nonequilibrium distribution does not depend on the details of any particular driving protocol, but rather on the *set* of driving protocols that generate the nonequilibrium systems in question. The set of local processes that could drive a system out of equilibrium on the Earth is extremely large and diverse, to the degree that the entropy over the set of possible distributions might be, to a good approximation, maximal. This would not necessarily be the case if, for example, the only process driving various systems on early Earth out of equilibrium was the rising and setting of the sun. That restriction would then impose additional constraints on our ensemble that would need to be taken into consideration, and we could not expect the maximization of entropy to sidestep those details. However, environments on the Earth permit a diversity of local processes. This inhomogeneity of conditions on early Earth supports the use of the maximum entropy hyperensemble, which allows us to compute averages without requiring any information beyond that captured by the temperature T and the nonequilibrium parameter λ.

Let us now explore how the concentrations of large and complex molecules change as a function of the distance from equilibrium.

III. RESULTS

A. Amino acid abundances and functional proteins

The possibility of prebiotic synthesis of amino acids was established in the landmark experiment by Miller and Urey [12]. They have since been detected in meteors [31], and produced in other experiments seeking to model the conditions of the early Earth [17,32]. However, the abundances with which the amino acids appear in abiotic settings do not match their biotic abundances [33]. In particular, functional proteins tend to employ the various amino acids in roughly equal proportions [33,34], whereas in abiotic sources there is an exponential suppression in the abundances of the larger amino acids, and none heavier than threonine have yet been found [35]. The apparent inability of the environment to produce heavier amino acids in sufficient quantities has been identified by several authors as a barrier to the emergence of life [5,34,35].

The difficulty of synthesizing the heavier amino acids in a prebiotic setting is usually ascribed to them having a larger Gibbs free energy of formation, ΔG [35]. The free energies of formation of the amino acids were calculated in Ref. [16], assuming synthesis from CO_2, NH_4^+, and H_2 in surface seawater at a temperature of 18 °C. The concentrations of amino acids relative to glycine, taken from nine different data sets, were fit using an exponential function [35]:

$$C_{\rm rel} = 15.8\exp[-\Delta G/31.3]. \tag{3}$$

We rescale these values so that they may be interpreted as probabilities (i.e., fraction of the total amino acid concentration occupied by amino acid x):

$$P(x) = \frac{C_{\rm rel}(x)}{\sum_{i=1}^{N} C_{\rm rel}(i)}, \tag{4}$$

where $C_{\rm rel}(x)$ is the relative concentration of amino acid x, and the index $i = 1, \ldots, N$ runs over all measured amino acids. The exponential dependence of the probabilities on the free energy of formation ΔG is consistent with an equilibrium distribution [35], although we caution that there are difficulties with this interpretation [31]. Nevertheless, we take Eq. (4) as our best approximation to the equilibrium distribution. We furthermore assume that this function correctly predicts the equilibrium abundances of the heavier amino acids which have not yet been found in abiotic sources, consistent with the fact that it predicts abundances too low to observe for these heavy amino acids [35].

We compare the distribution calculated from Eqs. (3) and (4) to the average nonequilibrium distribution, calculated numerically from Eq. (1). We assume that amino acids are the most thermodynamically costly molecules that can be formed in the system. This ought to be the case if the system is physically confined to a small volume (e.g., a mineral pore), or the reactants are very diluted. Such a restriction on the available state space is needed because in the extreme nonequilibrium limit, all states become equally probable. This means that if more costly molecules can be formed than amino acids, then the probabilities of forming any amino acids could go down relative to these more costly molecules. Yet, even without this restriction, the distribution of amino acids would become more uniform out of equilibrium. In the last section we will relax this assumption on the maximum cost of molecules, as we look at the asymptotic behavior of amino acids polymerizing into arbitrarily long chains.

Figure 1 shows the probability of obtaining the rarest amino acid, tryptophan, as a function of λ. On average, the relative concentrations of the rarest amino acid can be boosted by four orders of magnitude in the nonequilibrium regime. In the extreme nonequilibrium limit, $\lambda \to 0$, the hyperensemble becomes a symmetric Dirichlet distribution, which, in a state space of dimension d, has an expectation value for each outcome of $1/d$ and a variance of $\frac{1}{d^2}\frac{(d-1)}{(d+1)}$ [36], meaning that the standard deviation is of the same order as the mean. For a state space of dimension $d = 20$, the relative concentration of tryptophan in the extreme nonequilibrium limit is then

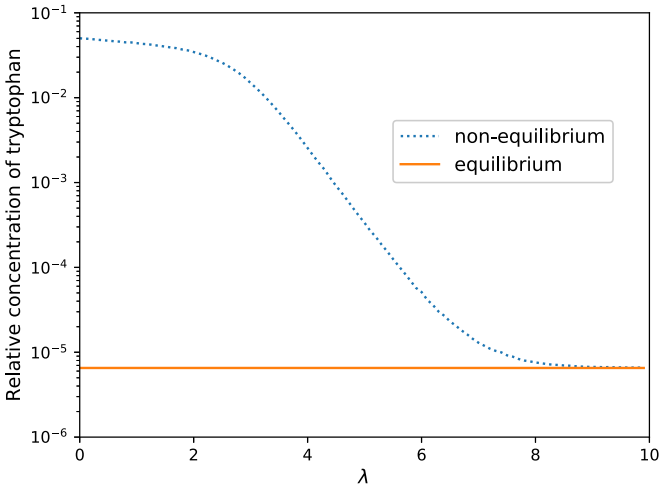

FIG. 1. Tryptophan requires the largest free energy to form of the protein amino acids and has not yet been found in an abiotic setting. Here we show how the relative concentration of tryptophan changes as one moves away from equilibrium, with the distance from equilibrium controlled by the parameter λ. The equilibrium relative concentration is plotted with an orange solid line. The average nonequilibrium relative concentration is plotted with a blue dotted line. Values are computed numerically from Eq. (1). We see that in the extreme nonequilibrium limit $\lambda \to 0$, the relative concentration of tryptophan can be increased up to four orders of magnitude.

$5.0 \times 10^{-2} \pm 4.8 \times 10^{-2}$, while its equilibrium relative concentration is $\sim 6 \times 10^{-6}$. Figure 2 shows a normalized histogram of amino acid samples in this limit. The distribution of relative concentrations is the same for any amino acid in this extreme nonequilibrium limit. For tryptophan, we observe that while a significant fraction of samples end up close to their

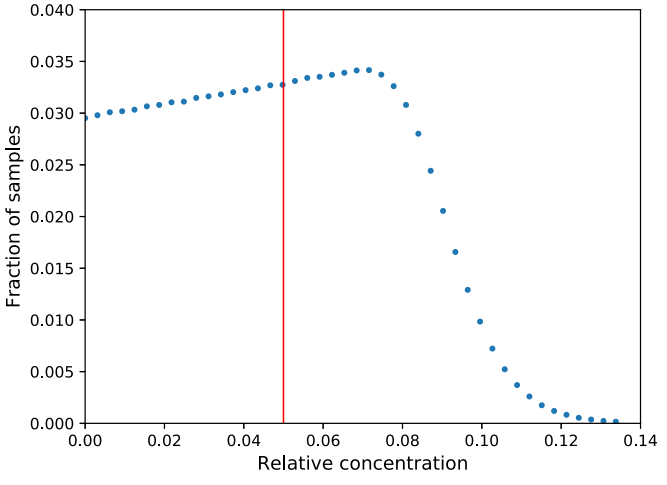

FIG. 2. A normalized histogram of 10^7 samples of the relative concentration of tryptophan from the hyperensemble in the extreme nonequilibrium limit. The red vertical line indicates the mean value. We have confirmed numerically that there are an equal number of samples above and below the mean. On this scale, the equilibrium relative concentration of tryptophan, at $\sim 6 \times 10^{-6}$, would not be distinguishable from the y axis.

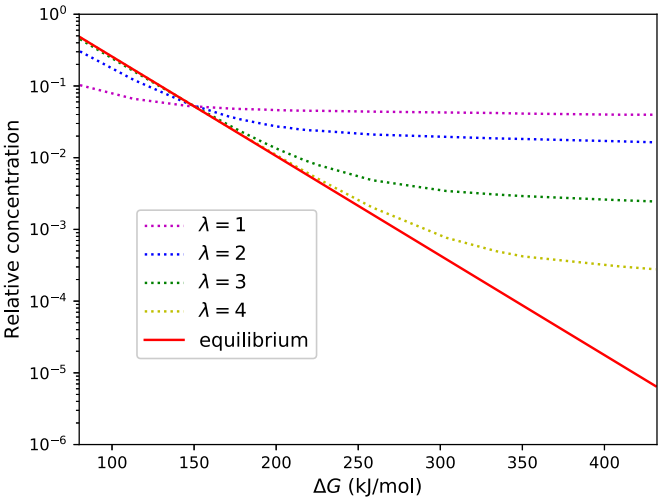

FIG. 3. The distribution of amino acids, arranged on the x axis in order of increasing Gibbs free energy, ΔG. The relative concentration in thermodynamic equilibrium is given by Eq. (4) and plotted with a solid red line. The other curves are the average nonequilibrium distribution, computed numerically from Eq. (1), at different distances from equilibrium (i.e., different values of λ). Note that as the distance from equilibrium increases, i.e., λ gets smaller, the distribution becomes flatter, and the relative concentrations of the rarest amino acids increase by several orders of magnitude. The flatter distribution observed out of equilibrium is consistent with the fact that roughly equal numbers of amino acids are found in functional proteins, and thus boosts the odds of forming them.

equilibrium values, the distribution as a whole gives radically more favorable odds of drawing a high concentration.

In Fig. 3, the average nonequilibrium distribution of amino acids is plotted as a function of the free energy of formation ΔG and compared to the equilibrium distribution, for various values of the nonequilibrium parameter λ, showing how the distribution becomes flatter as λ decreases. Importantly, the roughly uniform distribution of amino acids employed in functional proteins is exactly what the average nonequilibrium distribution gives in the extreme nonequilibrium regime (for values of λ close to zero). Thus, far away from equilibrium, the distribution of amino acids moves closer to its biotic distribution, thereby greatly enhancing the chances of spontaneously assembling functional proteins [2,34].

B. Polymerization of amino acids

Amino acids may be linked with one another via the peptide bond to form long chains. These chains then fold into proteins, with a typical protein containing ∼500 amino acids. However, the free energy, ΔG, for the peptide bond is on the order of several thousand kJ/mole [37], making the formation of long chains extremely improbable in thermodynamic equilibrium. It has been estimated that a solution containing 1 molar concentrations of each of the amino acids would require a volume 10^{50} times the size of the Earth to produce a single molecule of protein in equilibrium [1].

The thermodynamics of polymerization of amino acids were explored in Ref. [37], where, for simplicity, the chains were assumed to consist entirely of glycine. It was found that

dimerization of two glycine molecules requires the greatest amount of free energy per bond ($\Delta G = 3.6$ kcal/mole), being about eight times more difficult to form than subsequent additions to the chain. The relative concentration $[GG]/[G]$ is predicted to be about $1/400$ in equilibrium, and each subsequent addition of a glycine to the peptide results in a decrease by a factor of $1/50$ [37]. The probability of getting a chain of length $l \geqslant 2$ then follows a power law,

$$P_{\rm eq}(l) \propto \left(\frac{1}{50}\right)^{l-2}, \tag{5}$$

with the proportionality constant set by normalization of the probability. We examine the change in this distribution for nonequilibrium systems. To proceed, we identify each macrostate of a solution containing N glycine molecules with a partition of the number N into a sum of positive integers. For example, in a solution containing $N = 3$ glycine molecules there are three possibilities: the solution could contain three monomers (corresponding to 1+1+1), one monomer and one dimer (1+2), or one trimer. In number theory, the partition function, which we denote here by $Q(N)$, counts the number of distinct ways that a positive integer N can be decomposed into a sum of positive integers. For example, $Q(N = 3) = 3$. For tractability, we consider in this section only the extreme nonequilibrium limit $\lambda \to 0$, where all partitions of N become equally likely. First, we examine the probability of the rarest state, in which all N glycine molecules become bound into one chain of length $l = N$. The probability of observing this state is $P(l = N) = 1/Q(N)$. For large N, we can estimate $P(l = N)$ using the Hardy-Ramanujan asymptotic expression for $Q(N)$ [38], giving us

$$P_{\rm neq}(l = N) \approx 4N\sqrt{3}e^{-\pi\sqrt{\frac{2N}{3}}}. \tag{6}$$

Far away from equilibrium, the maximum probability of the rarest state is a decreasing function of N. Yet the odds of finding all N particles bound into a single chain decrease much more rapidly in equilibrium [refer to Eq. (5)], meaning that as the system gets larger, the factor by which nonequilibrium driving enhances probabilities of the rarest states grows without bound. This effect radically augments the chances of forming proteins in an abiotic setting. We display the ratio $P_{\rm neq}(l)/P_{\rm eq}(l)$ in Fig. 4, computed from Eqs. (5) and (6) using an exact expression for $P_{\rm neq}(l)$ obtained from SageMath's built-in Partitions function. With only 100 glycine molecules, the chance of finding them all bound into a single chain is found to be more than 100 orders of magnitude greater out of equilibrium than in equilibrium, and this effect continues to become more dramatic as the number of molecules in the system increases.

Of interest is also the number of chains of each possible length l, which we denote by m_l. When every partition is equally likely, the average number of chains of length l is given by [39,40]

$$\langle m_l \rangle = \frac{1}{Q(N)} \sum_{n=1}^{\text{floor(N/l)}} Q(N - nl). \tag{7}$$

This distribution was previously studied in the context of a fragmentation process, e.g., where a nucleus is broken apart and each partition is equally likely [39–44]. We calculate the expected number of chains of length l, $\langle m_l \rangle$, numerically for a system of size $N = 100$ and compare to that computed from the equilibrium distribution, Eq. (5). The results are displayed in Fig. 5. The numbers of chains of all lengths are increased dramatically, whereas in equilibrium most molecules would remain unbound to one another.

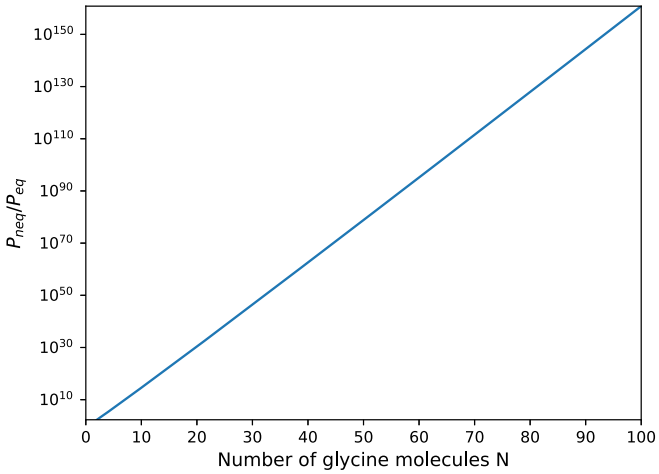

FIG. 4. Glycine molecules can be linked together via a peptide bond to form chains. Due to the large amount of free energy required per bond, the concentrations of longer chains drop precipitously in thermodynamic equilibrium [Eq. (5)]. Here we consider a system of N glycine molecules, and compare the probability of finding all of them bound into a single long chain, in thermodynamic equilibrium ($P_{\rm eq}$) to that far away from thermodynamic equilibrium ($P_{\rm neq}$), in the extreme nonequilibrium limit [given approximately by Eq. (6) but using exact values here]. We plot the ratio $P_{\rm neq}/P_{\rm eq}$ as a function of N, and see an exponential increase.

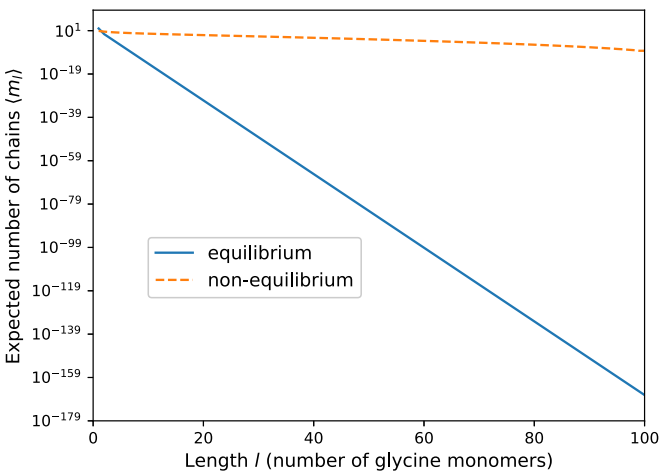

FIG. 5. The expected number of chains of length l in the extreme nonequilibrium limit is given by Eq. (7) and plotted with an orange dashed line for a system of size $N = 100$. We compare it to the values computed from the equilibrium distribution, given by Eq. (5) (plotted with blue solid line). In equilibrium, long chains are suppressed exponentially. This is not the case far away from equilibrium, where concentrations of the longest chains are increased by hundreds of orders of magnitude.

When N is large and the chains are not too long relative to N, Eq. (7) is well approximated by [39]

$$\langle m_l \rangle \approx \frac{1}{\exp\left[\sqrt{\frac{\pi^2}{6N}}l\right] - 1}, \tag{8}$$

which again will drop off much more slowly than the equilibrium distribution. Overall, this means that in the extreme nonequilibrium limit, the abundances of long peptide chains, and therefore proteins, can be increased by hundreds of orders of magnitude. This shows that obtaining appreciable quantities of proteins on the early Earth, which is all but excluded in equilibrium statistical mechanics, is a viable possibility considering the average odds out of equilibrium.

IV. DISCUSSION

Using two examples for which equilibrium thermodynamics seems to prohibit the spontaneous emergence of biologically important molecules, we have demonstrated that, under very modest assumptions, the concentrations of these molecules might be significantly larger (by many orders of magnitude) when odds are calculated from an average nonequilibrium distribution instead of the equilibrium distribution.

It is well known that nonequilibrium conditions of some kind are necessary for life. The degree by which the abundances are improved depends, of course, on how far from equilibrium the system has been driven. Since this is not known, we can not determine the parameter λ in our model, and hence can not provide a definitive number for the concentrations of life's building blocks. But what this study reveals is that, on average, nonequilibrium systems exhibit significantly more favorable conditions, provided that the distance from equilibrium is large enough. Importantly, this approach does not rely on specific knowledge about the conditions on the early Earth.

Another model-independent approach to assessing the odds of life's formation was presented in Ref. [45]. The chance of life's emergence on other worlds was calculated from estimating parameters in a Drake-type equation. One of the parameters appearing in this equation is the abiogenesis probability, which estimates the chances of life forming per unit time within a set of building blocks. An implication of our conclusions is that this parameter ought to be increased on planets where conditions are far from equilibrium, as for example on planets with rich weather phenomena, tectonic activity, or tidal interactions [7]. The necessity for chemical disequilibrium on a planetary scale for the emergence of life has been identified by several authors [6,7,46]. The average nonequilibrium distribution provides a concrete way of quantifying this effect as a function of how far conditions are from equilibrium.

Explaining the presence of heavy amino acids and peptides is, of course, far from a complete account of life's origins. But we wish to emphasize that the average nonequilibrium distribution's increased odds for attaining otherwise rare states should be independent of the details of any particular reaction. Thus, the same effect is likely to play an important role in other situations where equilibrium thermodynamics appear to create barriers to the emergence of life, e.g., the polymerization of nucleotides in RNA and DNA [10]. It is also possible that the effect might be compounded. This could happen, for example, if a more favorable distribution of amino acids, resulting from a nonequilibrium process is input to another nonequilibrium system that assembles the amino acids into peptides.

Moreover, the biological relevance of this effect need not be limited to the origin of life. Indeed, it is possible that early metabolic processes drove intracellular molecular distributions even further from equilibrium, creating a feedback process whereby the state-space of useful molecules could be more effectively sampled. A similar effect can be observed in kinetic proofreading, where energy is expended to drive reactions out of equilibrium and reduce the rate at which disadvantageous molecules are formed [47].

Altogether, the approach we presented here raises the possibility that the formation of life does not require a particular environment that has been fine-tuned for life. Rather, it may be sufficient to have a set of environments that have been driven far enough away from thermodynamic equilibrium. Not only is nonequilibrium driving a prerequisite for life, but nonequilibrium driving may thus, in this very general way, be a catalyst for life's emergence.

ACKNOWLEDGMENTS

We thank Lee Altenberg, Gavin Crooks, Joshua Deutsch, Norman Packard, Sebastián Pardo, Rob Shaw, and Eric Smith for helpful discussions. We are grateful for support from the Foundational Questions Institute and Fetzer Franklin Fund (a donor advised fund of Silicon Valley Community Foundation, Grant No. FQXi -RFP 1820).

APPENDIX

Crooks' approach [19] finds the $P(\theta)$ that maximizes the entropy $S[P(\theta)] = -\int P(\theta)\ln P(\theta)d\theta$, subject to:

(1) Normalization of probability: $\int P(\theta)d\theta = 1$.

(2) $\langle \bar{E}[\theta] \rangle = \int P(\theta)\bar{E}(\theta)d\theta$, a constraint on the average energy. Here, $\bar{E}[\theta] = \sum_i E_i\theta_i$ denotes the average energy, averaged over an individual nonequilibrium distribution, θ.

(3) $\langle S \rangle = \int P(\theta)S[\theta]d\theta$, a constraint on the average entropy. The entropy of a nonequilibrium distribution is given by $S[\theta] = -\sum_i \theta_i \ln(\theta_i)$. The Lagrange multiplier, λ, used to enforce this constraint then parameterizes the deviation from the equilibrium distribution. While this constraint is necessary to distinguish equilibrium systems from nonequilibrium ones, it also implicitly introduces the quantitative measure of deviation from equilibrium.

Solving the above constrained optimization problem results in the distribution [19]

$$P(\theta) = \frac{1}{\mathcal{Z}(\beta,\lambda)}\exp[-\lambda D(\theta\|\rho)], \tag{A1}$$

where $\mathcal{Z}(\beta,\lambda)$ is a normalization constant, and $D(\theta||\rho)$ is the relative entropy between the nonequilibrium distribution, θ, and the corresponding equilibrium distribution ρ. The average nonequilibrium distribution is then found by integrating:

$$\langle\theta\rangle = \int \theta P(\theta)d\theta = \frac{1}{\mathcal{Z}(\beta,\lambda)}\int \theta e^{-\lambda D(\theta\|\rho)}d\theta. \tag{A2}$$

The flattening effect on the average distribution, which is observed as $\lambda \to 0$, is invariant with respect to the choice of distance measure used in constraint number 3. If this constraint was replaced by a generic constraint on the average distance from equilibrium, $\int P(\theta)d(\theta, \rho)d\theta$, for any distance measure $d(\theta, \rho)$, then $P(\theta) \propto e^{-\lambda d(\theta,\rho)}$. This would change the exact form of $\langle\theta\rangle$, but the limit $\lambda \to 0$ would nonetheless give a flat average distribution.

Numerical calculations were performed in SageMath. To calculate $\langle\theta\rangle$ in Figs. 1 and 3, we generated 20 000 random distributions, calculated the relative entropy of each one [Eq. (2)] using the corresponding equilibrium distribution, then weighted them using Eq. (A1) and took the average using Eq. (A2). We also added a sample of the equilibrium distribution to the set of random distributions, in order to correct for the possibility that no samples would be generated close enough to the equilibrium distribution to obtain appreciable weight, when λ was high. For Fig. 2, each possible nonequilibrium distribution was generated from a list of 20 random numbers. Each entry in the list was sampled uniformly from the interval [0,1]. The list was then normalized. We generated 10^7 such distributions, and examined the distribution of a single element in the list, which corresponds with the relative concentration of an amino acid. Due to symmetry, the distribution is the same for each amino acid in the limit $\lambda \to 0$. For Figs. 4 and 5 we were only interested in the extreme nonequilibrium limit $\lambda \to 0$ where all states become equally likely, in which case the probability of each state is just the inverse of the number of states, and the number of states is given by the partition function. The partition function was calculated exactly, using Sage's built in Partitions function.

[1] M. Dixon and E. Webb, *Enzymes* (Academic Press, Cambridge, 1964), p. 667.
[2] S. Walker, P. Davies, and G. Ellis, *From Matter to Life: Information and Causality* (Cambridge University Press, Cambridge, 2017).
[3] J.-F. Lambert, Orig. Life Evol. Biosph. **38**, 211 (2008).
[4] H. Cleaves, A. Aubrey, and J. Bada, Orig. Life Evol. Biosph. **39**, 109 (2009).
[5] A. Brack, Chem. Biodivers. **4**, 665 (2007).
[6] H. Morowitz and E. Smith, Complexity **13**, 51 (2007).
[7] L. Barge, E. Branscomb, J. Brucato, S. Cardoso, J. Cartwright, S. Danielache, D. Galante, T. Kee, Y. Miguel, S. Mojzsis *et al.*, Orig. Life Evol. Biosph. **47**, 39 (2017).
[8] C. B. Mast, S. Schink, U. Gerland, and D. Braun, Proc. Natl. Acad. Sci. USA **110**, 8030 (2013).
[9] M. Kreysing, L. Keil, S. Lanzmich, and D. Braun, Nat. Chem. **7**, 203 (2015).
[10] D. Andrieux and P. Gaspard, Proc. Natl. Acad. Sci. USA **105**, 9516 (2008).
[11] R. S. Shaw, N. Packard, M. Schroter, and H. L. Swinney, Proc. Natl. Acad. Sci. USA **104**, 9580 (2007).
[12] S. L. Miller and H. C. Urey, Science **130**, 245 (1959).
[13] E. L. Shock and M. D. Schulte, J. Geophys. Res.: Planets **103**, 28513 (1998).
[14] W. Martin, J. Baross, D. Kelley, and M. J. Russell, Nat. Rev. Microbiol. **6**, 805 (2008).
[15] B. Herschy, A. Whicher, E. Camprubi, C. Watson, L. Dartnell, J. Ward, J. R. G. Evans, and N. Lane, J. Mol. Evol. **79**, 213 (2014).
[16] J. Amend and E. Shock, Science **281**, 1659 (1998).
[17] J. L. Bada, Chem. Soc. Rev. **42**, 2186 (2013).
[18] S. Chang, in *The Chemistry of Life's Origins* (Springer, Berlin, 1993), pp. 259–299.
[19] G. E. Crooks, Phys. Rev. E **75**, 041119 (2007).
[20] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[21] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).
[22] J. W. Gibbs, *Elementary Principles in Statistical Mechanics* (Courier Corporation, Mineola, 2014).
[23] J. Skilling, *Maximum Entropy and Bayesian Methods* (Springer Science & Business Media, Berlin, 2013), Vol. 36.
[24] J. N. Kapur, *Maximum-Entropy Models in Science and Engineering* (John Wiley & Sons, Hoboken, 1989).
[25] R. Shaw, *The Dripping Faucet as a Model Chaotic System* (Aerial Press, Santa Cruz, 1984).
[26] K. Takara, H.-H. Hasegawa, and D. Driebe, Phys. Lett. A **375**, 88 (2010).
[27] M. Esposito and C. Van den Broeck, Europhys. Lett. **95**, 40004 (2011).
[28] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, Phys. Rev. Lett. **109**, 120604 (2012).
[29] S. Kullback, *Statistics and Information Theory* (John Wiley & Sons, Hoboken, 1959).
[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, Hoboken, 2012).
[31] S. Pizzarello, Y. Huang, and M. Fuller, Geochim. Cosmochim. Acta **68**, 4963 (2004).
[32] T. M. McCollom, Annu. Rev. Earth Planet. Sci. **41**, 207 (2013).
[33] E. D. Dorn, K. H. Nealson, and C. Adami, J. Mol. Evol. **72**, 283 (2011).
[34] C. Adami, Orig. Life Evol. Biosph. **45**, 309 (2015).
[35] P. G. Higgs and R. E. Pudritz, Astrobiology **9**, 483 (2009).
[36] N. Balakrishnan and V. B. Nevzorov, *A Primer on Statistical Distributions* (John Wiley & Sons, Hoboken, 2004).
[37] R. B. Martin, Biopolymers **45**, 351 (1998).
[38] G. E. Andrews, *The Theory of Partitions* (Cambridge University Press, Cambridge, 1998), Vol. 2.
[39] K. C. Chase and A. Z. Mekjian, Phys. Rev. C **49**, 2164 (1994).
[40] J. R. Iafrate, S. J. Miller, and F. W. Strauch, Phys. Rev. E **91**, 062138 (2015).
[41] L. G. Moretto and G. J. Wozniak, Prog. Part. Nucl. Phys. **21**, 401 (1988).
[42] A. Z. Mekjian, Phys. Rev. Lett. **64**, 2125 (1990).
[43] S. J. Lee and A. Z. Mekjian, Phys. Rev. C **45**, 1284 (1992).
[44] A. S. Botvina, A. D. Jackson, and I. N. Mishustin, Phys. Rev. E **62**, R64 (2000).
[45] C. Scharf and L. Cronin, Proc. Natl. Acad. Sci. USA **113**, 8127 (2016).
[46] M. J. Russell, L. M. Barge, R. Bhartia, D. Bocanegra, P. J. Bracher, E. Branscomb, R. Kidd, S. McGlynn, D. H. Meier, W. Nitschke *et al.*, Astrobiology **14**, 308 (2014).
[47] J. J. Hopfield, Proc. Natl. Acad. Sci. USA **71**, 4135 (1974).