Nonequilibrium steady states in a model for prebiotic evolution

A. Wynveen, I. Fedorov, and J. W. Halley*
School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA
(Received 30 October 2013; published 27 February 2014)

Some statistical features of steady states of a Kauffman-like model for prebiotic evolution are reported from computational studies. We postulate that the interesting "lifelike" states will be characterized by a nonequilibrium distribution of species and a time variable species self-correlation function. Selecting only such states from the population of final states produced by the model yields the probability of the appearance of such states as a function of a parameter p of the model. p is defined as the probability that a possible reaction in the the artificial chemistry actually appears in the network of chemical reactions. Small p corresponds to sparse networks utilizing a small fraction of the available reactions. We find that the probability of the appearance of such lifelike states exhibits a maximum as a function of p: at large p, most final states are in chemical equilibrium and hence are excluded by our criterion. At very small p, the sparseness of the network makes the probability of formation of any nontrivial dynamic final state low, yielding a low probability of production of lifelike states in this limit as well. We also report results on the diversity of the lifelike states (as defined here) that are produced. Repeated starts of the model evolution with different random number seeds in a given reaction network lead to final lifelike states which have a greater than random likelihood of resembling one another. Thus a form of "convergence" is observed. On the other hand, in different reaction networks with the same p, lifelike final states are statistically uncorrelated. In summary, the main results are (1) there is an optimal p or "sparseness" for production of lifelike states in our model—neither very dense nor very sparse networks are optimal—and (2) for a given p or sparseness, the resulting lifelike states can be extremely different. We discuss some possible implications for studies of the origin of life.

DOI: 10.1103/PhysRevE.89.022725 PACS number(s): 82.39.−k, 87.15.R−, 82.20.Wt, 87.23.Kg

I. INTRODUCTION

A central theoretical question associated with the origin of life [1] is sometimes termed the "evolutionary paradox." As described by many authors and with particular clarity by Eigen [2], the problem is sometimes also termed "Eigen's paradox." Briefly, the paradox arises because, if it is assumed that prebiotic chemistry must randomly construct a starter "naked gene" of length N, then the number of possible biopolymers which must be randomly sampled to find a biologically viable one is of order B^N where B is the number of available types of monomer (B = 4 for DNA). If N is of the order of the length of the shortest gene known in terrestrial biology, then the time to randomly hit upon a viable starting gene is easily estimated to be hundreds of orders of magnitude longer than the age of our universe. The argument is seen to be an entropic one, associated with the number of ways to assemble a polymer of length N.

The still undetermined solution to this problem is absolutely central to a meaningful quantitative estimate of the frequency of life in the universe. In fact, if the naked gene model just described were the correct one for describing the origin of life, then Eigen's paradox would clearly imply that no other life forms will ever be found on other planets. It should be emphasized that the argument depends on very few details of the nature of terrestrial biochemistry. The only assumption is that a unique high information content polymer or small class of polymers of length comparable to that in biological DNA on Earth must be formed at the initial stage.

While one group of scientists [3,4] basically accepts the argument and concludes that life is likely to be extremely rare in the universe, others [5–12] including Eigen himself have proposed a variety of models in which life is more likely to arise than in the naked gene model.

In the present paper we describe results from such a model. However, our version of this type of model differs from previous ones in its definition of a lifelike state as we will describe in more detail below. Briefly, and unlike previous versions, we require that lifelike states not be in chemical equilibrium and that they are not temporally static. All these models use highly abstracted mathematical descriptions of the biopolymer chemistry. There are two reasons for this abstraction. First, it is totally impractical by computational or analytical mathematical means to draw rigorous conclusions from a complete mathematical description of the relevant polymer chemistry in the forseeable future. But secondly and more importantly, if the abstraction of the description of the chemistry retains the basic entropic conundrum which leads to the paradox, then the models can hope to provide insight into the way, if one exists, to avoid the paradox and predict the existence of other biospheres with an observably high probability.

These models which suggest a way around the paradox all basically take a similar approach, though details vary: it is assumed that the rate limiting prebiotic step is not the formation of one of a small number of initially viable naked genes, but is instead the formation of a population of interacting biopolymers interacting autocatalytically, dynamically, and metastably, far from equilibrium. Even with the simplifications provided by the abstraction of the chemical description, the mathematics of these models is very difficult to solve, and

*woods@woods1.spa.umn.edu

approximations, sometimes uncontrolled, are often used. With that caveat, results are reported which predict the appearance of such populations, arising from a randomly interacting "soup" of monomers, with observably high probabilities.

In the present paper we describe results from such a model. We make the same hypothesis that the initial step in the origin of life was the formation of an autocatalytic set of interacting polymers and not the formation of a single "gene." Though there is no direct experimental evidence for this initial step (or for any other initial step) in the origin of life, this hypothesis has the advantage of yielding these observably high probabilities, whereas naked gene models do not. However, our version of this type of model differs from previous ones in its definition of a lifelike state as we will describe in more detail below. Briefly, and unlike previous versions, we require that lifelike states not be in chemical equilibrium and that they are not temporally static.

Three issues arise concerning these models and these issues motivate our modifications in the model in this paper as described below. The three issues are as follows.

First, it appears that in some such models, the final states of the simulated prebiotic evolution are autocatalytic, as desired for a lifelike state, but also may be in chemical equilibrium within the simplified artificial laws of chemistry set up by the model. (In other words, the time-averaged polymer populations are in the ratios which maximize the entropy.) But intuitive notions of "lifelike" usually exclude systems in chemical equilibrium and most formal lists of requirements (see, e.g., Ref. [13]) for a system to be regarded as lifelike also explicitly or implicitly exclude such chemical equilibrium states.

Second, some such models (see, e.g., Ref. [11]) include dynamically inert final fixed points of the simulated dynamics within the definition of lifelike. While such states can certainly be out of chemical equilibrium (as in a glass), such inert states would not meet most intuitive notions of lifelike.

Third, whether the models exclude chemical equilibrium states and dynamically inert states from the definition of lifelike or not, they accept final states of the prebiotic evolution as lifelike as long as they satisfy a short list of very abstract generic properties. Thus another question which arises is, given a set of laws of chemistry, and an interacting set of chemical entities (as such models do), does the model dynamics lead to a unique lifelike chemistry according to the definition of lifelike adopted, or are many very different final-state chemistries generated which meet the generic criteria of lifelike? Within the models, this may be termed the issue of the "convergence" of the models for prebiotic evolution. If the models were realistic and the answer to the question posed was that they generate many very different lifelike states, then that answer would have implications for the kind of searches of other planets for lifelike systems which would be most likely to yield success. If prebiotic evolution is not convergent, then a search for lifelike biochemistry much like our own is likely to be fruitless, whereas a search for more generic lifelike chemical qualities such as nonequilibrium population ratios in steady states might be more successful.

In this paper we report simulations of a version of one of these models [6] with the specific objective of determining the following.

(i) First, how does a definition of lifelike which excludes chemical equilibrium states and dynamically inert states affect estimates of the probability of generating lifelike states within the model under various conditions?

(ii) Second, within a definition of lifelike which includes the nonequilibrium and dynamic constraints, how diverse are the resulting "biochemistries"?

In this way we seek to provide partial answers to the questions posed in the preceding paragraphs within the model. Previewing the answers qualitatively, we find that the constraint to nonequilibrium systems very significantly reduces the probability of producing lifelike systems and provides interesting and understandable constraints on the parameter space which allows such systems to appear with the largest probability.

With regard to convergence, we find that, within a given artificial chemistry, some convergence of the dynamics to similar final lifelike states is observed. However, if one compares final lifelike states generated by chemical reaction networks which are different but have the same statistical characteristics (parametrized by the quantity p defined below), then there is no convergence. In that latter case the final states are almost as diverse as they could possibly be.

In the next section we describe the construction of the model for the artificial chemistry, which, following Ref. [6], proceeds in two distinct steps. First a network of reactions is selected. Second, a model is adopted for the kinetic rates of the selected reactions. A set of small polymers termed a food set is maintained and the total population is limited during the dynamics (though, as we will discuss, for the states of interest the population is self-limiting and the imposition of a maximum population does not result in finite-size effects). This leads to final states from which we identify those which are lifelike according the definition of lifelike that we have qualitatively discussed. The next section describes how the nonequilibrium and dynamical constraints are implemented. Then we describe results of repeatedly using the resulting chemical networks to generate ensembles of final states which are analyzed to determine the probability that they are lifelike according to our definition. The last section contains a discussion, conclusions, and suggestions for further work. Algorithms are described in more detail in Appendixes A and B.

II. ESTABLISHING THE NETWORK

The "chemical species" in the model are described as numerical strings interpreted as polymers (e.g., nucleic acids or proteins). In the realization used here each digit is 0 or 1 so the chemical species are binary strings. Previous work on similar models [6,7] indicates that increasing the number of possible digits per place in the strings (e.g., by using 20 for proteins or 4 for nucleic acids) does not have large effects on the qualitative nature of the results. The allowed reactions are scission and ligation corresponding to the separation of a polymer into two parts and the connection of two polymers at their ends. As in Ref. [6], not all possible reactions are included in the model network. We include all possible ligation and scission reactions of all possible polymers of length $\leqslant L_{\max}$ with probability p. The parameter p characterizes the extent of connectivity of the resulting network of modeled chemical reactions. For each

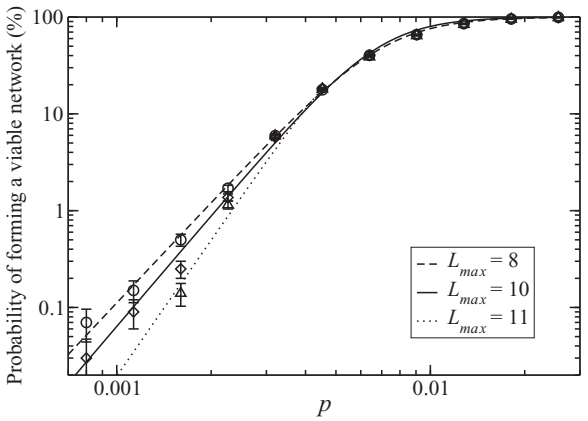

FIG. 1. Probability of forming a "viable network" as defined in the text as a function of the parameter p. The fitted curves are intended to be guides to the eye.

reaction an "enzyme" or catalyst is assigned, without which the reaction cannot take place.

The algorithm for selection of reactions included in a realization of the model is described in Appendix A. It is designed to construct an artificial chemistry of reaction networks and to determine which of these networks might be viable, i.e., satisfying those criteria outlined in the Introduction upon simulating the dynamics of those networks. First, to construct the reaction network, each possible reaction is admitted to the network with probability p. Then, in order to determine if a given network might be "viable," the connectivity of the network is probed. Only those networks that contain reaction paths connecting members of the small polymer "food set" defined below to at least one reaction involving a polymer of the maximum size $L_{\max}$ are accepted for further dynamics studies. If a network is not found to be viable, the attempt to form a network is still included in a count of such attempts (successful and unsuccessful) for use in computation of the probability of forming a viable network at that value of p. Our algorithm has eliminated an implicit dependence on the run time which seems to have been present in some previous such algorithms, because we exhaust the list of possible reactions.

We implemented this algorithm for a wide range of the parameters p, and for values $L_{\max} = 8$, 10, and 11. The probability of forming such a network as a function of p as determined from this process is shown in Fig. 1. The curves connecting the data points are a fit to the functional form $Ap^{\eta}/(1 + Ap^{\eta})$ and are only intended as guides to the eye and for interpolation between the calculated values. We attribute no theoretical significance to the functional form used for fitting. The dependence on $L_{\max}$ evident in Fig. 1 suggests that viable networks are less likely to be formed at a given p as $L_{\max}$ gets larger. This is qualitatively similar to results for similar, but simpler, networks [14,15] in which the probability for forming a spanning cluster at fixed p shrinks as the size of the system grows.

III. DYNAMICAL MODEL

A dynamical simulation is carried out on the ensemble of networks representing artificial chemistries described in the previous section. The system is assumed, as in previous similar models, to be "well mixed" spatially and no explicit representation of the position of species is made. The dynamical laws are stochastic. That is to say, if one were to carry out the simulation many times on the same network, one would find that the average number of polymers $n_l(t)$ of species l obeyed a master equation of the form

$$dn_l/dt = \sum_{l',m,e} \left[v_{l,l',m,e}\left(-k_d n_l n_{l'} n_e + k_d^{-1} n_m n_e\right) + v_{m,l',l,e}\left(+k_d n_m n_{l'} n_e - k_d^{-1} n_l n_e\right)\right]. \quad (1)$$

Here $v_{l,l',m,e}$ is proportional to the rate of the reaction $l + l' \xrightarrow{e} m$. e denotes the catalyst, l and l' denote the polymer species combined during ligation or produced during cleavage, and m denotes the product of ligation or the reactant during cleavage. If the ligation reaction in the network is chosen to be "forward" as indicated, then $k_d = k_f > 1$, where k_f is a global model parameter roughly characterizing the temperature. [In the case of the first two terms of the summand in Eq. (1), the ratio of the ligation rate to the scission rate if all the populations were 1 would be k_d^2 and would have value $e^{\Delta/k_B T}$, where Δ was the free-energy difference between the reactants and products in the ligation reaction. This motivates the choice of parametrization, but a realistic temperature dependence would be much more involved.) On the other hand, if the reaction has been chosen to be forward in the other, cleavage, direction, then $k_d = 1/k_f$. The parameter k_f is a rough proxy for the effects of temperature in the model. In the simulation results reported here we set $k_f = 1$ corresponding to infinite "temperature." The v factors (times the length of the time step) are chosen at random from the interval [0,1] for each reaction in the network before the dynamical simulation starts and are held fixed thereafter.

During the dynamic simulation, a set of small polymers (we use 0, 1, 00, 10, 01, and 11) are designated the food set. The food set is assigned a fixed population which is maintained after each time step by addition or subtraction. We also assign a fixed maximum polymer population $N_{\max}$ and maintain the total number of polymers near or below it by subtracting polymers from the total population at random after each time step if the population exceeds $N_{\max}$. In many of the simulation results below, we found that steady lifelike states (according to our criteria) were maintained at populations below $N_{\max}$ without the need for activation of this removal. The number of polymers present in the population, which is $\leqslant N_{\max}$, is denoted N. The dynamical algorithm is described in more detail in Appendix B.

The algorithm is repeated until the set of populations $\{n_l\}$ which we term the system point reaches a stable final state as defined in more detail below. In our simulations the system point usually converges to such a final state quite quickly. A key point is that this dynamical algorithm is stochastic in the sense that a given reaction occurs with a certain fixed probability, but it need not occur in any given realization of the dynamics even though all the parameters (k_f, $N_{\max}$, v's, and the reaction network itself) have been fixed. Thus the results can be different each time the algorithm is run with a different initial seed for the numerical random number generator. Comparing results obtained by using different initial random seeds to

simulate reactions allowed by a given network of reactions corresponds to comparing results carried out within the same chemistry repeatedly, whereas comparing results with different networks generated with the same parameter p corresponds to comparing the results from chemistries which are different but with a similarly sparse number of allowed reactions.

IV. IMPLEMENTATION OF ENTROPIC AND DYNAMICAL CONSTRAINTS

To analyze the results of the dynamical simulation, we must define the criteria for identifying lifelike systems. There is a large amount of literature debating the issue of what characterizes lifelike states [13]. As discussed in the Introduction, we select final states to be lifelike by criteria which require that they not be in chemical equilibrium and that they be dynamical as described in more detail below.

Each state of the system is characterized by the set of populations n_l of each polymer species l. We computed coarse-grained entropies for each such state, characterized by a set of occupancies $\{n_l\}$ by defining the numbers $N_L = \sum_{\{\text{polymers } l \text{ of length } L\}} n_l$ of polymers of length L. Since there are 2^L possible polymers of length L in the model, it is straightforward, for example following Ref. [16] (see Appendix C), to determine that the entropy is

$$S/k_B = \sum_L \{\ln[(2^L + N_L - 1)!] - \ln(N_L!) - \ln[(2^L - 1)!]\}. \tag{2}$$

Using Stirling's approximation for the factorials gives

$$S/k_B = \sum_L 2^L[(1 + \bar{N}_L)\ln(1 + \bar{N}_L) - \bar{N}_L \ln \bar{N}_L], \tag{3}$$

in which $\bar{N}_L = N_L/2^L$. The maximum configurational entropy S_{eq} for a given number N of polymers is then determined by standard methods and is provided in Appendix C [Eq. (C3)]. We impose the condition that lifelike states be nonequilibrium states by requiring that those final states which are accepted as lifelike have values of $S/S_{\text{eq}} \leqslant S_{\max}/S_{\text{eq}} < 1$. The value chosen for the cutoff $S_{\max}/S_{\text{eq}}$ determining exactly how small S/S_{eq} should be is somewhat arbitrary and should be regarded as a parameter of the model. We report results for a spectrum of values of the cutoff $S_{\max}/S_{\text{eq}}$.

In Fig. 2, we show results for this entropy as a function of simulation time in a dynamic simulation using the same chemical network with different random number seeds at the start ($L_{\max} = 10$). In the cases shown, the system reached a final steady state containing a number of polymers N which was less than $N_{\max}$. The values of $S_{\text{eq}}(N)/k_B$, at the steady-state values of N, for different dynamical runs are shown as horizontal straight lines in the figure. At larger values of p most of the dynamics simulations lead to final states in which $N = N_{\max}$ and $S/k_B = S_{\text{eq}}/k_B$. Such final states do not satisfy our lifelike criteria. In the results described in the next section we report results for different values of the cutoff parameter $S_{\max}/S_{\text{eq}}$, below which the final state was deemed to meet the nonequilibrium criterion for a lifelike state. We did not observe qualitatively significant dependence of the results on this threshold.

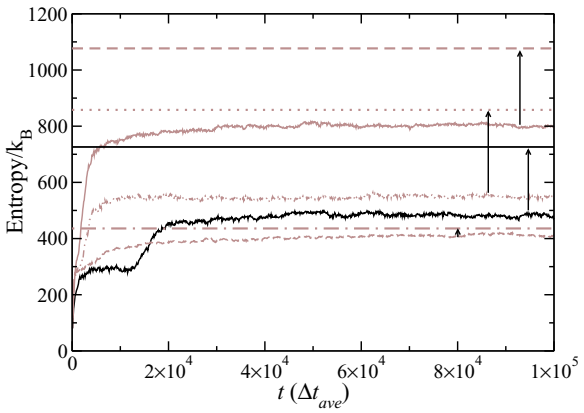

FIG. 2. (Color online) Entropy as a function of time is shown for four different dynamics runs on the same network. The maximum entropy, determined from the total number of polymers at the last time step, is drawn as a straight line with an arrow pointing from the calculated entropy to its maximum value for a given run. The same chemical network of reactions, for the same value of $p = 0.0032$ but with different values of the random number seed, was used to generate each curve.

We applied a second filter to the final states which were found to be stable nonequilibrium states in order to assure that they displayed the dynamical behavior minimally expected of lifelike states, thus excluding states of a glasslike nature. To do this, we computed the following time-dependent self-correlation function for each nonequilibrium asymptotic state α:

$$C(\tau) = (1/N_{\text{st}}) \sum_t \sum_j n_j^{\{\alpha\}}(t) n_j^{\{\alpha\}}(t + \tau), \tag{4}$$

where N_{st} corresponds to the number of discrete time steps over which the sum over the time steps had been averaged. The times over which the average is taken were chosen to be well into the asymptotic region in which the computed configurational entropy had stopped changing significantly. As indicated in Fig. 2, the asympototic region defined in this way turns out to be numerically well defined.

We explored several algorithms for selecting states which might be regarded as sufficiently dynamic to be lifelike from the time dependence of $C(\tau)$. Here we report results from the following characterization of the dynamics of $C(\tau)$: $C(\tau)$ is Fourier transformed and its power spectrum $P(\omega)$ is integrated about the origin out to a characteristic frequency ω_m defined implicitly by the relation

$$\frac{\int_{-\omega_m}^{\omega_m} P(\omega)d\omega}{\int_{-\infty}^{\infty} P(\omega)d\omega} = 1/2.$$

Then states are regarded as lifelike if $\omega_m > \omega_c$, where ω_c is a parameter of the model, thereby requiring that the populations of individual species are still temporally varying on time scales of order $2\pi/\omega_c$ in states characterized as lifelike. (In practice the transforms and integrals are discretized as described in Appendix D.)

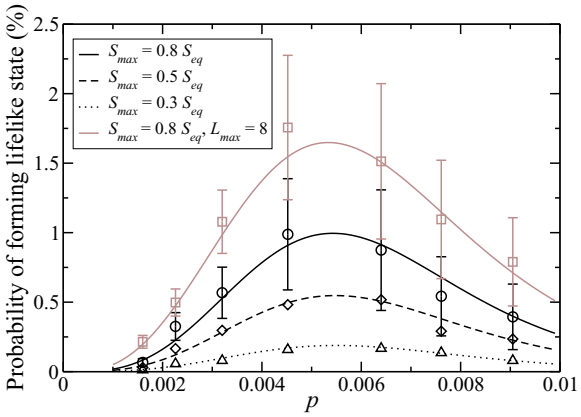

FIG. 3. (Color online) Probability of obtaining a lifelike state in the model. Results are show for $L_{\max} = 10$ unless otherwise noted. Results are shown for the choice of the dynamics cut $\omega_c = 0.01(2\pi/\Delta t_{\rm av})$, where $\Delta t_{\rm av}$ is the average time step in a given dynamics simulation. The points are the products of the data in Figs. 1, 4, and 5, and the lines are the products of the fits to that data as described in the captions to those figures.

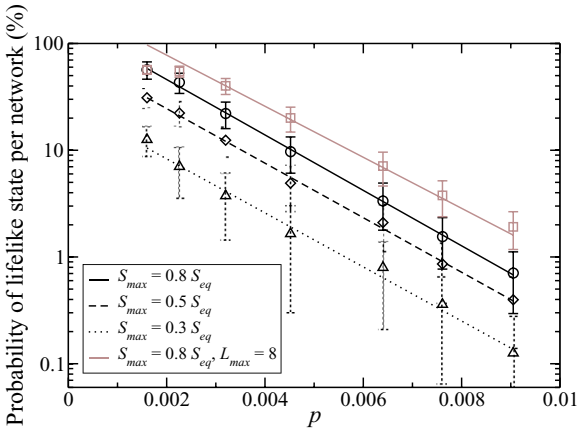

FIG. 4. (Color online) Probability of forming a steady state with entropy lower than the indicated cutoff, given that the network is "viable."Results are shown for $L_{\max} = 10$ unless otherwise noted. $S_{\max}$ is the cutoff for the calculated entropy of the system, whereas $S_{\rm eq}$ is the equilibrium entropy for the number of polymers in the steady state.

V. RESULTS

We repeated such calculations for a series of networks, characterized by various p. For decreasing p, the number of nonequilibrium final states increases to a maximum and then decreases. Results obtained for $L_{\max} = 10$ and for three values of the cutoff $S/S_{\rm eq}$ (and for one value of the cutoff for $L_{\max} = 8$) are shown in Fig. 3. The dynamical cutoff parameter used here was $\omega_c = 0.01(2\pi/\Delta t_{\rm av})$, where $\Delta t_{\rm av}$ is the average value of the time step during the dynamics simlulation. The value chosen for ω_c assures that significant temporal variations in the self-correlation function occur at least on time scales down to 10 time steps. Dependence of the results on ω_c was explored as described below (Fig. 5). The error bars were determined as described in Appendix D.

The qualitative results are not strongly dependent on the choice of the entropy cutoff, and the value of p for which the probability of forming a final state below the cutoff is maximum is quite stable at a value of around $p = 0.005$. The probability of forming a nonequilibrium final state at that value of p is around a percent in all cases, varying between 0.3% and 1.0% as the choice of the entropy cutoff is increased from 0.3 to 0.8. We find that the nonequilibrium states are generated predominantly in very sparse networks, and that, even in them, the likelihood of trapping a nonequilibrium state is quite small.

We understand the nonmonotonicity diagramed in Fig. 3 as follows. At large values of p, there are many reactions in the network, making it relatively easy to achieve equilibrium, so equilibrium is usually achieved and the number of nonequilibrium fixed points is small. As p becomes smaller, the likelihood of kinetic blocking of the paths to equilibrium grows and, with it, the number of nonequilibrium final states. However, at very small p, the likelihood of forming a network with a fixed number of species itself gets small (see Fig. 1). As a result, the average number of nonequilibrium lifelike states starts to shrink as p gets very small.

The results shown in Figs. 1, 4, and 5 are consistent with this understanding. The probability of forming a network rises sharply at a threshold around $p = 0.002$ in a percolationlike transition, much like the one discussed by Kauffman and co-workers [5–7], as shown in Fig. 1. However, as the probability of forming a viable network rises, the network of reactions becomes more connected and the likelihood of forming an equilibrium final dynamical state increases, with a corresponding decrease in the likelihood of the formation of nonequilibrium dynamical final states of the sort which we characterize as lifelike. This effect is shown in Fig. 4. Given a viable network, the likelihood of forming a nonequilibrium state drops with increasing values for p, as shown there, becoming negligible when $p \geqslant 0.01$.

Finally, we show the likelihood that, given a viable network and a nonequilibrium final steady state, the steady state has lifelike dynamics (as a function of p) in Fig. 5. Although we observe that the systems tend to be "more dynamic," i.e., there are greater variations in the populations during the simulation runs, with increasing values of p (due to

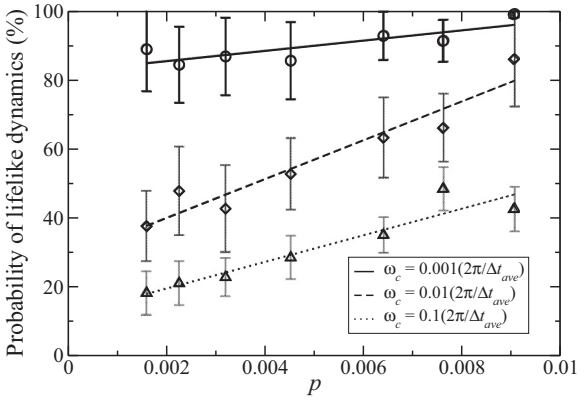

FIG. 5. Probability of "lifelike dynamics" given that the network is viable and the entropy is below the cutoff $S_{\max} = 0.8S_{\rm eq}$. Results are shown for $L_{\max} = 10$. Again, $\Delta t_{\rm av}$ in the legend corresponds to the average time step of the dynamics simulation.

the networks being larger), no definitive dependence on the frequency cutoff appears to be emerging from the simulations. We take the probability, within the model, of forming a lifelike state according to our definition to be the product of [probability of forming a network (Fig. 1)] × [probability of lifelike state per network (Fig. 4)] × [probability of lifelike dynamics (Fig. 5)], as shown in Fig. 3. As noted above, a peak appears around $p = 0.005$. If we were to interpret this as a model for the formation of biospheres on planets, our model would suggest that under optimal conditions (corresponding to chemical networks with $p = 0.005$), the probability of forming a lifelike biosphere would be around 1%.

Next, we analyzed the dispersion of the dynamical nonequilibrium states which we found. The motivation is to understand, within the model, whether the selection principle we have chosen to characterize lifelike states leads in any sense to chemical convergence. That is, do the requirements that the final state be nonequilibrium and dynamic select out any particular types of combinations of polymer species in the model? In the real world of the search for life, this would translate into the question of whether all biospheres are expected to be chemically similar. To test this, we characterized the final states of the "successful" runs by a vector $\{n_l(t)\}$ of species populations in the space of system points. In the final states which our criteria accept as lifelike these vectors move through this space, but they are not found to move over large distances through the space. We accordingly characterize each lifelike state by the time-average value of its state vector, with the average taken over times long with respect to the period of the characteristic time scale $(2\pi/\omega_m)$ of its steady-state dynamics, and well after the steady state has been realized. We characterize the angular difference between two lifelike states $\{\langle n_i^\alpha(t)\rangle\}$ and $\{\langle n_i^\beta(t)\rangle\}$ by the angle

$$\Theta_{\alpha\beta} = \cos^{-1} \frac{\sum_i \langle n_i^\alpha(t)\rangle\langle n_i^\beta(t)\rangle}{\sqrt{\sum_i \langle n_i^\alpha(t)\rangle^2 \sum_j \langle n_i^\beta(t)\rangle^2}} \tag{5}$$

between the time-averaged vectors of the two states in this space. If the "evolution" within our model is "convergent," we expect to get a narrow distribution of the $\Theta_{\alpha\beta}$ angles about the value $\Theta = 0$ $(\cos\Theta = 1)$, whereas we will interpret a broad distribution of Θ values as indicative of nonconvergent evolution. Results are shown for $L_{\max} = 10$ and for two different values for p, $p = 0.0032$, and 0.004 52 in Fig. 6.

As noted, for repeated simulations using the same chemical network, there is evidence of convergence, as indicated in the curves labeled "within network" in Fig. 6. In contrast, the curves labeled "outside network" were obtained by evaluating the distribution over all networks with a fixed value of p and show no correlation. The curves labeled "random distribution" in Fig. 6 were obtained from a sample of randomly selected vectors in six dimensions subject to the constraint that a number of vectors corresponding to the food set be maintained, as we do in the model simulation. The randomly generated sample, using the minimum number of species for a given network (the number of species defined in the food set), provides a distribution of the angle most skewed toward convergence $(\Theta_{\alpha,\beta} = 0)$ to which we can make a comparison with the model simulation. The shift of the distribution of

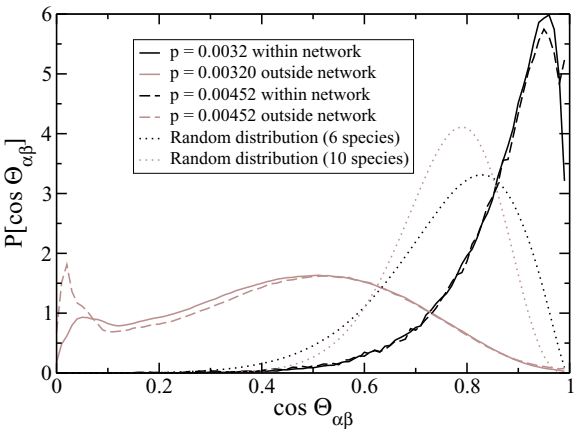

FIG. 6. (Color online) Distribution of the cosine of the angle between those final states of dynamical simulations of the model which are lifelike as defined in the text. The curves to the right of the diagram indicate some convergence of states generated from a given chemical network. The wider curves indicate no convergence between states generated by chemical networks which are different but characterized by the same sparseness, parametrized by p. Results are provided for $L_{\max} = 10$, $S_{\max} = 0.8S_{\text{eq}}$, $\omega_c = 0.01(2\pi/\Delta t_{\text{av}})$, and $p = 0.0032$ and 0.004 52, as labeled. Distributions for random six-dimensional and ten-dimensional vectors, corresponding to the number of species in the food set and the average number of species in lifelike steady states, are provided for comparison.

the random sample from a symmetrical distribution about 0.5 arises from this requirement that the food set be maintained. Thus part, but not all, of the convergence observed in the model appears to arise just from that requirement. The additional convergence (the more peaked curves to the right of the randomly generated sample distribution) is presumably due to a combination of the sparseness of the network and of the entropy constraint. Effects of the sparseness of the network are further elucidated by the curves providing the distribution of the angles between lifelike states generated in different chemical networks with the same value of p. There is no convergence at all between such states. This shows that the convergence observed is for particular chemical networks and not for any chemical network generated with a given value of p. We understand this qualitatively as follows: because the networks generating lifelike states are all very sparse, a relatively small number of paths through the population space are allowed by each network. Thus the paths to lifelike states in each such sparse network are likely to sample a different part of the space, leading to no correlation. On the other hand, for repeated runs with different random number seeds on the same network, the small number of reaction paths to lifelike states leads to the expectation that the lifelike states found will be similar leading to correlation. We discuss the possible significance of these results for the search for lifelike systems on other planets or elsewhere in the next section.

VI. DISCUSSION

We have explored the consequences of a prebiotic evolutionary model using criteria for the acceptance of lifelike states in the asymptotic dynamics which are different from

corresponding criteria used by others [5–7]. In particular, many authors explicitly or implicitly assume that final dynamical states which are "autocatalytic" are lifelike. The meaning of autocatalytic has been made mathematically precise, and such criteria lead to a threshold on the value p, which defines the connectivity of a chemical network, for forming lifelike states much like the curves shown in Fig. 1. With such results, authors sometimes conclude that one should be "optimistic." That is to say, one should expect that a wide range of chemical networks will lead to lifelike states. On the other hand, most states in chemical equilibrium will be autocatalytic, even though most biologists (and others) do not regard systems in chemical equilibrium as lifelike. Accordingly we have applied the constraint that lifelike states must be kinetically trapped out of chemical equilibrium. This constraint results in a low probability of production of lifelike states, which are not in chemical equilibrium when p is large, and, accordingly, greatly limits the range of networks in which it is at all probable to find a lifelike state as seen in Figs. 3 and 4. As noted, in our numerical realization, the maximum probability of producing a lifelike state by our criteria occurred at $p \approx 0.005$ and resulted in a probability of finding a lifelike state in this optimally tuned network of about a percent, as summarized in Fig. 3.

The peak in the probability at small p arises from the requirement that the final states of the dynamic evolution be kinetically trapped out of equilibrium. Sparse networks of reactions provide few paths to equilibrium and increase the likelihood of such trapping. This conclusion that sparse networks are favorable is different from the conclusions of Kauffman and co-workers [5–7] who used similar models but did not apply the entropy constraint. It suggests qualitatively that a "desert" might be more favorable than a strongly interacting Darwinian "pond" for starting lifelike evolution. In laboratory attempts to generate lifelike states, the reported struggles [17] to avoid "asphalt" are a version of the problem of generating lifelike states in systems in which chemical equilibrium is easily reached.

We suggest that it may be possible to realize laboratory systems in which sparse networks of reactions lead to dynamically stable, out of equilibrium, states thus demonstrating this effect. For example, dynamical molecular networks [18,19] with a set of dithiol "building blocks" as a food set have been shown to result in nonequilibrium, kinetically trapped states under some conditions. Whether such nonequilibrium states involve molecular populations which fluctuate or oscillate in time is not reported. This could possibly be determined optically if the fluctuating constituents have distinguishable absorption characteristics (as in the Belousov-Zhabotinsky interaction [20]). The rates of reaction in the dithiol system are reportedly controlled by the pH, which might then approximate the effects of p in our model.

The other criterion for a lifelike state which we have applied is that the final asymptotic state of the dynamics be "dynamic" and not a glasslike inert collection of populations. Though this again is consistent with the biological intuition of most biologists and other scientists, such dynamically inert states are included in the final states of interest in some models of prebiotic evolution [11]. However, this second change in the criteria for selecting a lifelike state has less dramatic qualitative effects on the results than the constraint to nonequilibrium final states. The dynamic character of the lifelike states that we find in our model has not been fully explored. There are anecdotal indications from samples of the functions $C(\tau)$ [Eq. (4)] that a wide variety of dynamical types is selected by the low entropy condition, and this will be studied further.

The other question which we addressed here is whether our criterion for lifelike states selects any particular combinations of chemical species. This question is a crude form of the issue of convergence in this model of prebiotic evolution. By determining the distribution of angles between time-averaged final states in the model we found evidence that there is convergence in a given sparse reaction network of reactions, but that there is no convergence when comparing final states associated with different sparse reaction networks with the same degree of sparseness as measured by the parameter p. However, our results show approximately the same probability of finding lifelike states for all ensembles of networks generated with a given value of p. If lifelike states were observed (for example, on exoplanets) under conditions in which the chemical network was similarly sparse in each case, but environmentally constrained to favor different sets of reactions, this then suggests that such lifelike states might be completely different from each other.

It should be noted that taking more realistic account of the details of allowed reactions might result in more convergence than is manifest here. For example, if one confines one's attention to planets orbiting stars, then the addition of criteria associated with the need to efficiently utilize radiant energy from the star could be conjectured to favor some chemistries over others. Nevertheless, our results suggest that nonequilibrium dynamical limit cycles of many divergent chemical types can form in networks of the type we consider, and we believe that they therefore suggest some caution concerning the assumption that lifelike states on other planets will be chemically similar to our own.

An unexpected feature of our results is that the steady nonequilibrium dynamic states which we find are not saturating the bound on the total number of polymers which we imposed. This means that, without this bound, the population would nevertheless remain approximately fixed. An advantage of this result is that it means that the steady states we find are not artificially affected by the "finite-size effect" imposed by the population bound. On the other hand, evolution to larger polymer lengths and more complexity is apparently not occurring in these steady states (though we have not yet explored the time dependence of the individual polymer populations and this conclusion is tentative). The model may require environmental "shocks" or spatial heterogeneity in order to exhibit such evolution. Both effects are known to be important in macroscopic Darwinian evolution. The effects of environmental shocks of various kinds and of spatial heterogeneity are quite easy to add to the model as discussed briefly below, and we have begun some such studies.

There is a wide variety of possible ways to model environmental shocks in this model, all involving some kind of rare event which changes the parametrization of the chemical network. For example, one could change all the parameters $v_{l,l',m,e}$ in Eq. (1) by $\epsilon r v_{l,l',m,e}$ at some point during the steady-state evolution, where r is evenly distributed between -0.5 and 0.5 and ϵ is a positive real parameter controlling the magnitude

of the shock. We have made preliminary explorations of the effects of spatial diffusion in the model. These will be reported later. Introducing space also opens the possibility of spatially heterogeneous diffusion rates and distributions of reaction rates which are doubtless also present in real contexts and which may be explored later. The effects of temperature have yet to be explored, even at the primitive level of varying the parameter k_f from unity. Sudden changes in the temperature are also a form of shock which can be explored. The limited results on the dependency of the conclusions on $L_{\max}$ suggest that no new qualitative changes will result, but, on the other hand, the dependency is quantitatively significant. In the study of convergence, correlations functions other than $\Theta_{\alpha,\beta}$ could obviously be studied. For example, one could look at the "Hamming distance" between the vectors. However, along a "ray" from the origin to a particular state in the population space, the Hamming distances between points can be large, though the "chemistry" is essentially the same. Nevertheless, the study of other correlations in the structure of the space of lifelike states in the model may yield further insight.

ACKNOWLEDGMENT

This work was supported by the Minnesota Supercomputing Institute.

APPENDIX A: ALGORITHM FOR GENERATION OF THE NETWORK

(1) Select parameters p, and $L_{\max}$ characterizing the network. p is the probability that a possible reaction is actually included in the network (as described in more detail below), and $L_{\max}$ is the maximum allowed length of the polymer strings of integers (here restricted to 0's and 1's) which are allowed in the network.

(2) Provisionally accept each possible reaction [$2(2^{L_{\max}+1}-2)(L_{\max}2^{L_{\max}}-2^{L_{\max}+1}+2)$ total] involving the $(2^{L_{\max}+1}-2)$ different species, as a candidate for the network with a probability p.

(3) Select a "firing disk" of initial short polymer species (binary strings of polymers) to include in the network.

(4) For the species within the firing disk, determine what candidate reactions accepted in step (2) consist solely of reactants and enzymes that are species within the firing disk. If they do, include these reactions and the resulting product species in the network.

(5) For all candidate reactions (both scission and ligation) determine if the reactants and enzymes of each reaction are species that have already been included in the network. For those reactions that do satisfy this criterion, include the reaction and its product species, unless the product species have already been included, in the network.

(6) Repeat step (5), iterating through all those candidate reactions that have not yet been included in the network, until no more candidate reactions satisfy the criterion that the reactants and enzymes are species in the network.

(7) Once the network is constructed, determine if any of the species included in the network have a length of $L_{\max}$. If so, this network is considered viable. If not, this is still counted as an attempt to form a network.

(8) This procedure [steps (2)–(7)] is performed repeatedly, using different random number seeds, in order to determine the probability of forming a viable network (the ratio of successes of forming a viable network to the total number of attempts to form a network), which is shown in Fig. 1. Dynamics simulations (see the next section) are then performed on those networks that are deemed viable.

APPENDIX B: DYNAMICS ALGORITHM

The dynamics simulation proceeds as follows. Note that states in the dynamical simulation are characterized by sets $\{n_l\}$ of polymer populations where l runs over species in the network.

(1) Choose a food set of initial polymer populations. In our simulations, the set of species in the food set is the same as the set in the firing disk of the network generating algorithm though in general it does not need to be. We have not yet explored the effects of varying the choice of food set on the results numerically but plan to do so in the future. We do not anticipate large qualitative effects of such a variation, as long as both the food set and the firing disk consist of polymers of length much less than $L_{\max}$. The populations $\{n_l\}$ for these polymers in the food set are initialized.

(2) From the reactions in the network with reactants and enzymes that consist of polymer species that have finite populations, select one without bias. Carry out the reaction with probability indicated in the master equation and adjust the numbers $\{n_l\}$ accordingly. The time elapsed is computed as described in Ref. [21].

(3) Calculate the total number of polymers. If it exceeds a target $N_{\max}$ (fixed at the outset), then choose polymers at random from the set and remove them until the number is less than or equal to $N_{\max}$.

(4) Return to (2).

APPENDIX C: CONFIGURATIONAL ENTROPY

We consider the configurational entropy associated with a coarse-grained prescription of the state given by the number of polymers N_L for each length L between $L=1$ and $L=L_{\max}$. Using the fact that, in the model, there are 2^L possible polymers of length L, counting possible states for a given state specification $\{N_L\}$ is the same problem that occurs in the boson statistics problem [16] (though of course this is not to imply that this model has any quantum features). The result is

$$S/k_B = \ln\left[\sum_{L=1}^{L_{\max}} \frac{(2^L+N_L-1)!}{N_L!(2^L-1)!}\right]. \tag{C1}$$

Denoting $\bar{N}_L = N_L/2^L$ and using Stirling's approximation for the factorials, one finds the familiar form

$$S/k_B = \sum_{L=1}^{L_{\max}} 2^L[(1+\bar{N}_L)\ln(\bar{N}_L+1) - \bar{N}_L \ln \bar{N}_L]. \tag{C2}$$

This form was used to calculate the entropy of the polymer systems during the dynamics simulations.

To find the equilibrium configurational entropy for a system (which depends only on the total number of polymers N at a given time during the dynamics simulation), we maximize

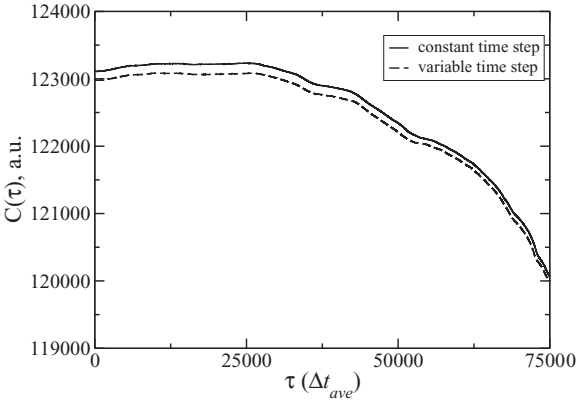

FIG. 7. Self-correlation function [Eq. (4)] for one trajectory, calculated by the two methods described in Appendix D.

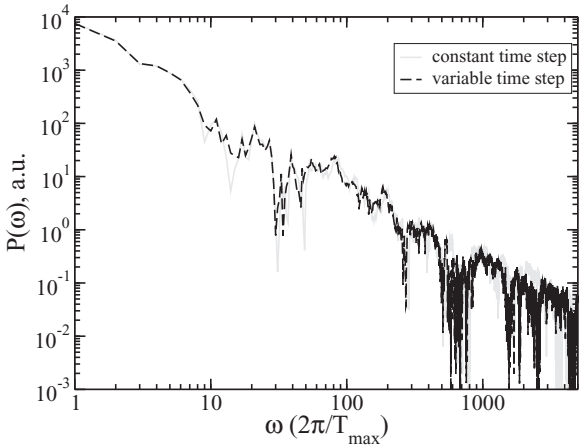

FIG. 8. Fourier transform of the functions shown in Fig. 7.

the free energy $S/k_B + N\mu/k_BT$, where $N = \sum_{L=1}^{L_{\max}} N_L$ is the total number of polymers within the system. We find (with $\beta = 1/k_BT$) that the extremum value for N_L is $N_L^{\rm eq} = 2^L/(e^{-\mu\beta} - 1)$. Substituting this extremum value back into $N = \sum_{L=1}^{L_{\max}} N_L$, for fixed N, we may eliminate the $(e^{-\mu\beta} - 1)$ term and solve for the equilibrium value $N_L^{\rm eq}$ explicitly in terms of N. This yields $\bar{N}_L^{\rm eq}(=N_L^{\rm eq}/2^L) = N/[2(2^{l\max} - 1)]$. Substituting this into the above equation for entropy yields the maximal equilibrium configurational entropy:

$$S_{\rm eq}(N) = \sum_{L}^{L_{\max}} 2^L\left[\left(1 + \bar{N}_L^{\rm eq}\right)\ln\left(\bar{N}_L^{\rm eq} + 1\right) - \bar{N}_L^{\rm eq}\ln\bar{N}_L^{\rm eq}\right], \tag{C3}$$

which was used in the calculations to evaluate the value of $S/S_{\rm eq}(N)$. The validity of the Stirling approximation for the N's of interest here was evaluated by direct computation and was adequate.

APPENDIX D: DETAILS OF DYNAMICS CUTOFF AND ERROR ESTIMATION

The error bars in Fig. 3 were obtained as follows: the standard deviations associated with the averaging giving the data in Figs. 1, 4, and 5 were determined. Denoting them $\Delta P_{\rm net}$, ΔP_S, and $\Delta P_{\rm dyn}$, and the corresponding averages as $P_{\rm net}$, P_S, and $P_{\rm dyn}$, the error bars in Fig. 3 were (conservatively) determined as $\pm[P_{\rm net}(\Delta P_S + \Delta P_{\rm dyn}) + P_S(\Delta P_{\rm net} + \Delta P_{\rm dyn}) + P_{\rm dyn}(\Delta P_{\rm net} + \Delta P_S)]$.

The time dependence of the populations in our simulations is determined at variable time intervals because of our use of the Gillespie algorithm [21]. To avoid numerical issues associated with calculation of the Fourier transform of data provided at variable times, we explored two approaches. The least computationally expensive, but also least accurate, approach was to treat each data point as if it were equally spaced in time with time intervals between successive data points equal to the average time step in the simulation. In a few cases, to test if this method gave a good approximation to the self-correlation function, we also found an approximation to the self-correlation function by linearly interpolating between the unequally spaced data points in the time-dependent population record and evaluating the populations at equally spaced times using the interpolations. We compare results of the two approaches for one case in Figs. 7 and 8. The results are nearly identical so we used the less expensive method in our screening algorithm.

[1] D. Lynn, C. Burrows, J. Goodwin, and A. Mehta (eds.), Acc. Chem. Res. **45**, 2023 (2012).
[2] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[3] P. Davies, *The Eerie Silence* (Houghton Mifflin Harcourt, Boston, 2010).
[4] M. Hart, in *Extraterrestrials, Where are They?*, 2nd ed., edited by Zuckerman and M. H. Hart (Cambridge University Press, Cambridge, UK, 1995), Chap. 22.
[5] J. Doyne Farmer, S. A. Kauffman, and N. H. Packard, Physica D **22**, 50 (1986).
[6] R. Bagley and J. D. Farmer, in *Artificial Life II*, edited by C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen (Addison-Wesley, Redwood City, CA, 1991), p. 93.
[7] S. A. Kauffman, *The Origins of Order–Self-Organization and Selection in Evolution* (Oxford University Press, New York, 1993), Chap. 7.
[8] M. Eigen and P. Schuster, *The Hypercycle: A Principle of Self-Organization* (Springer, Berlin, 1979).
[9] P. W. Anderson, Proc Natl. Acad. Sci. USA **80**, 3386 (1983).
[10] F. Dyson, J. Mol. Evol. **18**, 344 (1982).
[11] S. Jain and S. Krishna, Phys. Rev. Lett. **81**, 5684 (1998); Comput. Phys. Commun. **121**, 116 (1999); Proc. Natl. Acad. Sci. USA **98**, 543 (2001); Phys. Rev. E **65**, 026103 (2002); Proc. Natl. Acad. Sci. USA **99**, 2055 (2002); S. Krishna, Ph.D. thesis, Indian Institute of Science, Bangalore, 2003.

[12] I. A. Chen and M. A. Nowak, Acc. Chem. Res. **45**, 2088 (2012).
[13] R. Popa, *Between Necessity and Probability: Searching for the Definition and Origin of Life* (Springer-Verlag, Berlin, 2004).
[14] R. Albert and Q.-L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002).
[15] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[16] L. Landau and E. Lifshitz, *Statistical Mechanics* (Addison-Wesley, Reading, MA, 1958), p. 116.
[17] S. Benner, H.-H. Kim, and M. A. Carrigan, Acc. Chem. Res. **45**, 2025 (2012).
[18] S. Otto, Acc. Chem. Res. **45**, 2200 (2012).
[19] F. B. Cougnon and J. K. M. Sanders, Acc. Chem. Res. **45**, 2211 (2012).
[20] I. R. Epstein, V. K. Vanak, A. C. Balazs, O. Kuksenok, P. Dayal, and A. Bhattacharya, Acc. Chem. Res. **45**, 2160 (2012).
[21] D. Gillespie, J. Comput. Phys. **22**, 403 (1976).