


Some generic measures of the extent of chemical disequilibrium applied to living and abiotic systems

B. F. Intoy and J. W. Halley

School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA (Received 23 August 2018; revised manuscript received 21 May 2019; published 28 June 2019)

We report results of evaluation of several measures of chemical disequilibrium in living and abiotic systems. The previously defined measures include R_T and R_L , which are Euclidean distances of a coarse grained polymer length distribution from two different chemical equilibrium states associated with equilibration to an external temperature bath and with isolated equilibration to a distribution determined by the bond energy of the system, respectively. The determination uses a simplified description of the energetics of the constituent molecules. We evaluated the measures for data from the ribosome of *E. coli*, a variety of yeast, and the proteomes (with certain assumptions) of a large family of prokaryotes, and for mass spectrometric data from the atmosphere of the Saturn satellite Titan and for nonliving commercial copolymers. We find with surprising consistency that R_L is much smaller than R_T for all these systems. The living (protein) systems have a well defined value of R_T that is sharply defined and distinct from that obtained from the nonliving Titan and copolymer systems. The living systems are also distinguishably characterized by larger values of R_L than most of the nonliving systems, but R_L values vary more from one living system to another than the R_T values do. These data suggest that the measures R_L and R_T can distinguish living from nonliving systems.

DOI: [10.1103/PhysRevE.99.062419](https://doi.org/10.1103/PhysRevE.99.062419)**I. INTRODUCTION**

To define the mission of the search for extraterrestrial biology with quantitative precision, it would be desirable to specify measures of molecular systems that are more generic than searches for particular molecules or combinations of molecules like those in the earth's biosphere. That is because it is not known that all nonequilibrium systems having lifelike characteristics will turn out to be at all chemically similar to the terrestrial biosphere. Here we test the appropriateness of two such possible measures by comparing their values for known living and nonliving systems in order to determine how appropriate they would be for making a preliminary identification of lifelike chemistry in extraterrestrial contexts.

The systems we study are the proteomes of 4555 prokaryotes, the ribosomes in *E. coli*, a budding yeast, five artificial copolymers, and the atmosphere of Titan. The choice of systems was made in order to provide a sampling of living systems and some engineered system (the copolymers). Titan's atmosphere was selected because Titan's atmosphere is reported [1] to contain an abundance of organic molecules and may be characteristic of a prebiotic environment [2].

We evaluate the measures for this set of experimentally known systems, including some that are in the biosphere and others that come from nonliving systems. Similar measures were used earlier by us to identify the extent to which steady states resulting from simulations of a Kauffman-like model were out of equilibrium [3]. We do not need to assume that the real systems considered here have all the detailed properties of the model used in Ref. [3], and in particular, no detailed model of reaction rates or mechanism of molecule formation from monomers is required. Some simplifying assumptions about the bonding energy of the system considered are used as discussed below.

The experimental inputs to the analysis presented here are the molecular weight density distributions, the numbers of types of monomers available for forming molecules, two parameters characterizing the size of a polymer coil containing L monomers, and the average free energy of monomer-monomer bonds. The outputs are values of two dimensionless parameters termed R_T and R_L , which describe how far the observed system is from chemical equilibrium with an external thermal bath and from chemical equilibrium if isolated, respectively. Detailed definitions are provided in Sec. II.

We do not contend that the quantities R_T and R_L are the only quantities that might be used to characterize lifelike systems. We are testing the hypothesis, not making the assumption, that they are among the quantities that might be used in that way. It should be emphasized that we do not expect any of the experimental systems considered to be in chemical equilibrium so we do not expect the measures R_T and R_L to be small. We are not attempting to optimize an equilibrium model by minimizing R_T or R_L with respect to some model parameters. Instead we are using R_T and R_L as determined from experimental data to determine, within the assumptions made, how far each system is from chemical equilibrium. The results are used to test, not assume, the hypothesis that the values of R_T and R_L found for living systems may be sufficiently different from those for the nonliving systems (engineered copolymers and the atmosphere of Titan) to make them useful as determinants of whether an unfamiliar chemical system (such as might be found on an exoplanet) might be "lifelike" or not.

Our results, as presented at the end of the paper, show that all the systems considered, both living (proteomes) and nonliving (the atmosphere of Titan and five engineered copolymer systems), are closer to isolated equilibrium than they are to chemical equilibrium with the ambient temperature of their

surroundings. However, the living systems have a sharply defined R_T value different from the values found for the other systems. It appears that a small R_L and a large well defined R_T distinguish the living from the nonliving systems. We briefly discuss detailed molecular weight distributions to contrast what is observed experimentally with the corresponding equilibria.

In the next section we define the essential features and assumptions that allow definition and evaluation of the measures R_L , R_T and other generic characteristics. We focus on the case in which all the bond energies are the same followed by a brief discussion of the case in which there is more than one bond energy. Section III reports the processes by which we extracted the needed information from the raw experimental data in each of the cases considered. In Sec. IV we report the results, and Sec. V contains a discussion and conclusions.

II. ANALYSIS

We first describe the case in which all the polymeric bond energies are the same, as we have assumed for all systems except the Titan atmosphere. Polymers are assumed to consist of strings of monomers with b types possible for each monomer. We took $b = 20$ for the ribosomes in *E. coli*, the prokaryotes, and yeast, and $b = 2$ for the Titan data and the copolymer data. To any polymer of length L we attribute an energy $-(L-1)\Delta$, where Δ is a real number that is the bonding energy between two monomers. It is taken to be positive for the copolymers and the Titan atmosphere but negative for the peptide bonds in the biological systems as discussed in the next section. The total energy E of any population $\{n_m\}$ of polymers in which n_m is the number of polymers of type m is $E = -\sum_{L=1}^{l_{\max}} (L-1)N_L\Delta$. Here the $N_L = \sum_{m \text{ of length } L} n_m$ is the same set of macrovariables used in [4] and [5]. We denote the total number of polymers in a sample by $N = \sum_{L=1}^{l_{\max}} N_L$. However, in contrast to the situation in the dynamic simulations described in [3], the input data for calculation of equilibrium distributions are not N and E but the volumetric polymer concentration $\rho = N/V$, where V is the solution volume and the volumetric energy density $e = E/V$. To take entropic account of the dilution of the experimental sample we introduce a microscopic length R_0L^ν , where R_0 is a length related to the polymer persistence length, ν is an index that would be 1/2 for a random walk, and L is the number of monomer units in the polymer, as above. $v_pL^{3\nu}$ approximates the volume of a polymer of monomer length L . We report the numerical values for ν and R_0 used for the various systems considered in the next section. We modify the expression for the entropy used in [3] to take account of the number of ways to distribute N polymers in a volume V as follows: $S/k_B = \ln W$ with

$$W = \prod_L \frac{(N_L + G_L - 1)!}{N_L!(G_L - 1)!}$$

and $G_L = b^L V / v_p L^{3\nu}$ and $v_p = R_0^3$. The expression is identical to the one used in [3] except for the factor $V/v_p L^{3\nu}$ in the degeneracy G_L . A similar configurational entropy factor was used by us in an earlier paper [5].

With this modification we apply Stirling's approximation and maximize the entropy subject to the density and energy constraints associated with the experimental data discussed in the next section. We have

$$S/k_B = \sum_L \{\ln[(G_L + N_L - 1)!] - \ln(N_L!) - \ln[(G_L - 1)!\]. \quad (1)$$

Proceeding in the standard way to maximize the entropy under these constraints, we have, when both energy density $e = E/V$ and polymer number density $\rho = N/V$ are fixed, that the values \bar{N}_L of the populations that maximize this entropy are

$$\bar{N}_L = \frac{G_L - 1}{\exp[-\beta(e, \rho)\mu(e, \rho) - \beta(e, \rho)\Delta(L-1)] - 1}. \quad (2)$$

Here the parameters $\beta(e, \rho)$ and $\mu(e, \rho)$ are determined from the total energy density e and polymer number density $\rho = N/V$ by the implicit equations [with (2)]

$$e = -(1/V) \sum_{L=1}^{l_{\max}} (L-1)\bar{N}_L\Delta \quad (3)$$

and

$$\rho = (1/V) \sum_{L=1}^{l_{\max}} \bar{N}_L. \quad (4)$$

We use the definitions of G_L and v_p to write these relations as

$$\bar{N}_L v_p / V = \frac{b^L / L^{3\nu} - (v_p / V)}{\exp[-\beta(e, \rho)\mu(e, \rho) - \beta(e, \rho)\Delta(L-1)] - 1}, \quad (5)$$

and

$$e v_p = - \sum_{L=1}^{l_{\max}} (L-1) \bar{N}_L v_p / V \Delta \quad (6)$$

and

$$\rho v_p = \sum_{L=1}^{l_{\max}} \bar{N}_L v_p / V. \quad (7)$$

These are in dimensionless form, convenient for solving for $\beta(e, \rho)\mu(e, \rho)$ and $\beta(e, \rho)\Delta$ numerically because they do not involve macroscopically large numbers. The term v_p/V on the right hand side of (5) is in all cases much less than $b^L/L^{3\nu}$ and is dropped in the numerical analysis. As before [3] we refer to this equilibrium as "isolated." There is no reference to an external temperature bath. Note that in isolation, the system energy is fixed, whereas there are small fluctuations in the energy predicted by the distribution derived. However, as is well known in statistical mechanics, when the number of molecules is large, the fluctuations are very small relative to the average energy [inversely proportional to the square root of the number of molecules, which is tiny ($\approx 10^{-8}$) for molar quantities]. This is the "equivalence of ensembles" discussed in many textbooks (e.g., [6]). Isolated systems achieve this equilibrium as long as they are not kinetically blocked from doing so. Such blocking has occurred in all the systems considered here.

As in [3] we also determine a “thermal” equilibrium distribution by solving (7) with a fixed value of β using reported approximate values of the ambient temperature of the environment in the experiments considered and making no use of e . In each case we can then use (5) to evaluate the polymer length density distributions expected in those two equilibrium states.

The systems of interest are not expected (and in fact are found not) to be in either kind of chemical equilibrium. The hypothesis that we are exploring in this paper is that the quantitative extent to which they are out of equilibrium may be sufficient to distinguish living from nonliving systems. To measure that extent quantitatively, we have chosen to use the Euclidean distance in the space of values of sets $\{N_L v_p/V\}$ between the actual population sets $\{N_L v_p/V\}$ and the ones corresponding to the two kinds of equilibria given by (5) with $\beta(e, \rho)$, $\mu(e, \rho)$ in the isolated case and with a fixed ambient β and $\mu(\rho)$ in the case in which the system is equilibrated to an external bath.

The choice of Euclidean distances for these measures was motivated by the facts that they can be determined from population data alone without a detailed model of the reaction network and they can be normalized to their maximum values so that the values found can be quantitatively and meaningfully compared for very disparate systems. Any such measure that allows quantitative comparison of very disparate systems will have to abstract some general features, as these do, and may seem to be a somewhat arid characterization to individuals familiar with the detailed structure and dynamics of the individual systems compared. A similar Euclidean measure to describe how far a chemical system is from chemical equilibrium in the context of prebiotic evolution has been suggested by others [7].

The use of a Euclidean measure is simpler than the use of direct entropic measures used by us in an earlier study [4] of an ensemble of steady states generated computationally using a specific and very simplified model of the reaction dynamics of a prebiotic system. Later, we used a Euclidean measure similar to the one used here but different in some details, to study ensembles of steady state systems generated by another specific and abstract dynamic model of prebiotic chemistry [5]. However in the present paper we make no use of the dynamic details of either of those models and evaluate the Euclidean measures entirely from the experimental population data.

Thus we define two Euclidean distances R_L and R_T in the l_{\max} dimensional space of sets $\{N_L v_p/V\}$ that characterize how far the system of interest is from the two kinds of equilibria described above:

$$R_L = \sqrt{\sum_L (v_p/V)^2 [N_L - \overline{N_L}(\beta(e, \rho), \mu(e, \rho))]^2 / (\sqrt{2} v_p \rho)} \quad (8)$$

for distance from the locally equilibrated state and

$$R_T = \sqrt{\sum_L (v_p/V)^2 [N_L - \overline{N_L}(\beta, \mu(\beta, \rho))]^2 / (\sqrt{2} v_p \rho)} \quad (9)$$

for distance from the thermally equilibrated state.

In the cases of prokaryotes, *E. coli* ribosomes, yeast, and the nonliving copolymers that we evaluate, we use the assumptions as just described. However, in the analysis of the data from the Titan atmosphere, we take account of the large difference between the energies of CC and CN bonds and NN bonds by reformulating the description with two bond strengths as described in Appendix A and Ref. [8]. We show there that an accurate treatment of that case gives results very close to those obtained by using the same model as the one described above, but with an average bond strength $\overline{\Delta}$ of

$$\overline{\Delta}(p_c) = \Delta_{CC} p_c^2 + 2\Delta_{CN} p_c(1 - p_c) + \Delta_{NN}(1 - p_c)^2. \quad (10)$$

p_c is the atomic fraction of the atmosphere that is carbon. (We assume $\Delta_{CC} = \Delta_{CN}$ as discussed further in the next section and Appendix B.)

III. EXTRACTION OF POPULATION DISTRIBUTIONS FROM DATA

Here we describe how data were extracted from data available on a budding yeast, 4555 prokaryotes, the ribosomes in *E. coli*, the five nonliving commercial copolymers, and mass spectrographic data on the atmosphere of Titan in order to compute the characteristic quantities R_L , R_T , $\beta\mu$, and $\beta\Delta$ defined in the preceding section. We report a comparison of the results in the next section in order to determine the extent to which these quantities may distinguish living or lifelike systems from nonliving ones.

A. Yeast

We used data on protein population distributions in the budding yeast *Saccharomyces cerevisiae* from Ref. [9]. From the small angle x-ray scattering experiments determining the root mean square size of many denatured proteins as reported in [10] we took the experimental values $R_0 = 1.927 \text{ \AA}$ and $\nu = 0.588$ for this and all the other protein systems considered in solving Eqs. (6) and (7). For determining the chemical potential in the case of equilibrium with an external temperature bath [Eq. (7)] we used a temperature of 293 K and a bond energy [11] of -2.2 kcal/mol . It is important to note that for this and the other protein systems Δ is taken to be negative, meaning that it costs energy to make a peptide bond. The value of -2.2 kcal/mol taken from Ref. [11] is the Gibbs free energy of reaction in aqueous solution near neutral pH and can be seen from [11] to be the appropriate energy to identify with Δ within our description of the equilibrium state. That Gibbs free energy difference includes effects of the difference in entropy of the solvating water in the bonded and nonbonded states as discussed in [11].

B. Ribosomes in *E. coli*

We used protein distribution data from [12]. There are approximately 50 000 ribosomes per cubic micron in an *E. coli* cell [13], giving a value for Nv_p/V of about 1.86×10^{-5} assuming that, in equilibrium, the proteins would be denatured and evenly distributed throughout the cell. We took $R_0 = 1.927 \text{ \AA}$ and $\nu = 0.588$ as for the yeast and prokaryotes.

TABLE I. Parameters used for the five nonliving copolymer systems considered here. N_{mon} is the total number of monomer blocks in units of moles, $\langle L \rangle$ is the average polymer length, and V is the volume in liters. R_0 values are calculated by taking the average of the two monomer types from [18].

Data reference	Monomer types	Figure in this paper	N_{mon} (moles)	$\langle L \rangle$	V (liters)	R_0 (Å)	Density (Nv_p/V)
Fig. 7 of [16]	Isoprene and styrene	3	8.6	28.9	6	5.10	0.00395
Fig. 6-13b of [17]	Isoprene and styrene	3	1.21	32.9	2	5.10	0.00147
Fig. 6-13c of [17]	Isoprene and styrene	3	1.21	41.3	2	5.10	0.00117
Fig. 6-13d of [17]	Isoprene and styrene	3	1.21	49.9	2	5.10	0.000968
Fig. 6-15 (left) of [17]	Styrene and butadiene	3	0.966	36.6	2	5.27	0.00116

C. Prokaryote proteins

We used data from the Kyoto Encyclopedia of Genes and Genomes [14] (KEGG) (the faa, fasta amino acid file) to extract approximate length distributions for 4555 prokaryotes assuming that there is only one of each protein in each prokaryotic cell. The KEGG data gave amino acid sequences for all the proteins, from which the coarse grained number N_L of all proteins of each length L was extracted. By averaging data in [15] we obtained an approximate average volumetric protein density of 2.5 million proteins per cubic micron that was used for all the prokaryotes when solving Eqs. (6) and (7). As for yeast, the form [10] $v_L = R_0^3 L^{3\nu}$ with $R_0 = 1.927$ Å and $\nu = 0.588$ was used in the solutions to Eqs. (6) and (7). For thermal equilibrium calculations a temperature of 20 °C and a bond energy of -2.2 kcal/mol were used.

D. Copolymers

We used data from Refs. [16,17], which report polymer length distributions using a special mass spectrometry technique. The experimental results are summarized in Table I. The data references given in Table I give the relative abundances of polymers as a function of their composition. The table also reports the monomers used in each of copolymer systems. They were either polystyrene-polyisoprene or polystyrene-polybutadiene. For example the data from [16] give the relative abundances $f(l_s, l_i)$ of polymers in a solution of a polystyrene-polyisoprene system as a function of their styrene l_s and isoprene l_i compositions, as extracted from mass spectrometry data. To get values of f_L we sum all the values of $f(l_s, l_i)$ for which $l_s + l_i = L$. From [16], the total number of monomer blocks in the copolymer system was 8.6 mol, consisting of 0.7 mol of styrene and 7.9 mol of isoprene. From the length distribution $\{f_L\}$ the average polymer length is calculated to be approximately 28.9 monomer units. Dividing the total number of monomer blocks by the average polymer length gives the total number of polymers to be approximately $N = 1.79 \times 10^{23}$. The experiment took place in a $V = 6$ liter reactor vessel giving N/V . For solving Eqs. (6) and (7) we obtained v_p for this system using listed values for the persistence lengths l_p , statistical segment lengths \tilde{b} , and characteristic ratios C_∞ cited in [18] to obtain the value $R_0 = \tilde{b}^{1-\nu} l_{\text{eff}}^\nu$, with $l_{\text{eff}} = 2l_p/\sqrt{C_\infty}$ the effective bond length as defined in [18]. For the copolymers we assumed that $\nu = 0.5$ and used the average of the R_0 values calculated from the parameters \tilde{b} , l_p , and C_∞ for isoprene and styrene in [18]. For calculations

of the chemical potential μ from (7) when the system is in thermal equilibrium with an external thermal bath, a bath temperature of 293 K and an average carbon-carbon bond energy of 347 kJ/mol were used [19]. (\tilde{b} is not to be confused with b . For copolymers, we took $b = 2$.)

E. Titan data extraction and analysis

We used mass spectrographic data on the atmosphere of Titan taken by the Cassini spacecraft as reported in [1]. The data on mass distributions are reported in [1] in units of charge detected per second. In this paper we only report analysis of the Cassini mass spectroscopy data on detected negative ions. The mass distribution of negative ions contains the widest mass distribution reported, including molecules as massive as 10^4 daltons. Data are available for detected neutral and positively charged molecules [20] and may be analyzed later. We assume in our analysis that all the molecules detected had unit charge (in units of the magnitude of the electron charge). Masses were reported in daltons and converted approximately to monomer units by dividing by an assumed average monomer mass of 13 daltons because the monomers are believed to be predominantly single nitrogen or carbon entities. (We are neglecting the contribution of the hydrogen masses.) We converted the data reported in [1] in log scale bins to linear scale bins as described in Appendix A. To extract volumetric densities N/V we used the kinetic relation $\dot{N} = \epsilon A(N/V)v$, where \dot{N} is the rate at which particles are detected per second, $\epsilon = 0.05$ is the detector efficiency [1], A is the detector area, reported to be 0.33 cm^2 [1], and v is the velocity of the spacecraft relative to the Titan atmosphere for which we used $v = 6.3 \text{ km/sec}$ from Ref. [21]. This procedure gave volume densities for data taken at different altitudes above the Titan surface as summarized in Table II. We assumed

TABLE II. Titan atmospheric densities for various altitudes as detected during the 40th Titan encounter of Cassini, which occurred on 05 January 2008.

Altitude (km)	Figure in this paper	Density (Nv_p/V)
1013	4	4.56×10^{-20}
1032	4	3.79×10^{-20}
1078	4	1.72×10^{-20}
1148	4	6.54×10^{-21}
1244	4	7.89×10^{-22}

TABLE III. The average bond energies for carbon and nitrogen [19].

Bond	Average Bond Energy (kJ/mol)
C-C	347
C-N	305
N-N	160

that the species detected in the mass spectrometer were singly charged. Uncertainties resulting from this procedure are discussed in [1]. For the value of $v_p = R_0^3$ in Eqs. (6) and (7) we used the $R_0 = 4.97 \text{ \AA}$ extracted from data for polyethylene in [18] in the same way that was done for the copolymers as explained above (with $\nu = 0.5$).

For determining the equilibrium distribution in the presence of an ambient thermal bath using Eq. (7) we used an ambient temperature of 120 K [22].

In applying Eqs. (7) and (6) to the Titan data, we are assuming that all the molecules detected are linear chains, which is certainly not expected to be true [2]. Also, CC and CN bond energies are quite closely similar, as described in Table III, but the N-N bond is much weaker. In principle, one should therefore take account of the different bond energies in the model used to obtain Eqs. (6) and (7), which we solve to determine the equilibrium distributions. We have done that, with results described elsewhere [8]. However, it turned out that a simplifying “mean field” approximation works quite well to describe the results, and we use that here: We used an “average” bond energy $\Delta = \bar{\Delta}$, where

$$\bar{\Delta}(p_c) = \Delta_{CC}p_c^2 + 2\Delta_{CN}p_c(1 - p_c) + \Delta_{NN}(1 - p_c)^2 \quad (11)$$

with $\Delta_{CC} = \Delta_{CN} = 325 \text{ kJ/mol}$ and $\Delta_{NN} = 160 \text{ kJ/mol}$. p_c is the atomic fraction of the atmosphere that is carbon. We use the value $p_c = 0.02$ [23]. As explained in Appendix B, we also obtained results using the maximum and minimum values that the bond strength could have at the observed p_c and found that the results were quite insensitive to the change and consistent with the results of [8].

IV. RESULTS

We summarize the data found for R_T and R_L for all the systems studied in Fig. 1, where one sees that the proteomes of living systems have values of R_T and R_L that are clearly distinguishable from those of the nonliving copolymers and the Titan atmosphere.

The population distributions for three of the 4555 prokaryotes analyzed are shown in Fig. 2. Similar results were obtained for the ribosome and yeast proteomes. Length distributions for five copolymer systems are shown in Fig. 3. Figure 4 compares the observed length distributions with the calculated equilibrium for the Titan data at five elevations. Values of R_L and R_T for the data exhibited in the figures appear in Table IV.

V. DISCUSSION AND CONCLUSIONS

A cursory inspection of Fig. 1 makes it clear that for this collection of molecular population data on living and

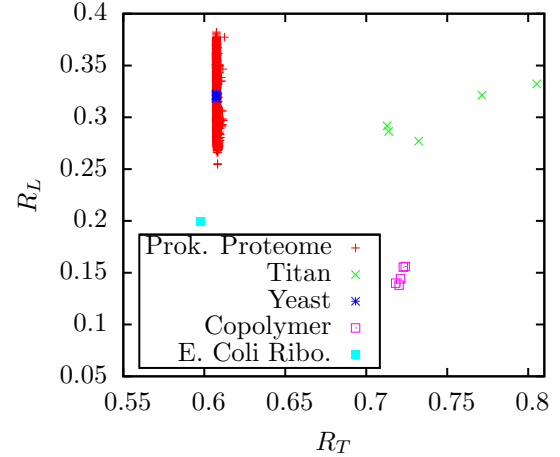


FIG. 1. R_T and R_L summary scatter plot for all the systems considered. There are 4555 points (red ‘+’) associated with the 4555 different prokaryotes for which we obtained KEGG data as well as 5 points (green ‘x’) for the Titan data, 5 points (pink open squares) for the copolymer systems, 1 point (filled blue square) for the E. Coli ribosome proteome and 1 point (blue ‘*’) for the proteome of the yeast.

nonliving systems, the measures R_T and R_L quite clearly distinguish the living (proteome) systems from the nonliving Titan atmosphere and engineered copolymer systems. One sees that the prokaryotic and yeast proteomes are far from chemical equilibrium with the ambient environment as measured by nearly identical values of about $R_T \approx 0.61$ found

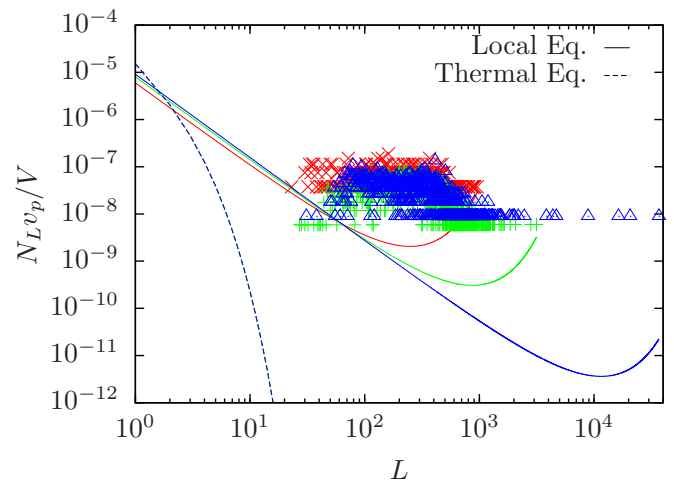


FIG. 2. Normalized length density ($N_L v_p / V$) distributions for three of the 4555 proteomes included in the data set together with the corresponding local and thermal equilibrium distributions. Red symbols (‘x’) give data for the proteome of the prokaryote *Buchnera aphidicola* JF98, an endosymbiont of *Acyrtosiphon pisum* (KEGG code baw; $R_L = 0.25$ and $R_T = 0.61$); blue symbols (‘+’) indicate data from prokaryote *Chlorobium chlorochromatii* CaD3 (KEGG code cch; $R_L = 0.38$ and $R_T = 0.61$), and green symbols (triangles) give data from prokaryote *Corynebacterium variable* DSM 44702 (KEGG code cva; $R_L = 0.33$ and $R_T = 0.61$). Local equilibrium curves corresponding to the 3 proteomes are red (dashes) for baw, blue (dots) for cch and green (dot-dashes) for cva. Solid line is the thermal equilibrium curve (same for all three proteomes.)

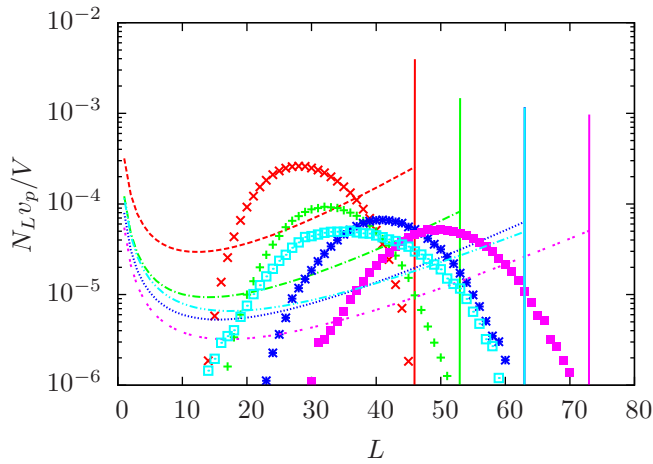


FIG. 3. $N_L v_p / V$ distributions for the data from the five nonliving copolymer systems with the corresponding equilibrium distributions. Isoprene and styrene copolymers: red (\times 's and dashes for local equilibrium), "JCA 2010," Fig. 17 of [16]; green (+'s and dot-dashes for local equilibrium), "Staal 13b," Fig. 6-13b of [17], $R_L = 0.16$ and $R_T = 0.72$; blue (*'s and dots for local equilibrium), "Staal 13c," Fig. 6-13c of [17], $R_L = 0.14$ and $R_T = 0.72$; magenta (filled squares and double-dash line for local equilibrium), "Staal 13d," Fig. 6-13d of [17], $R_L = 0.14$ and $R_T = 0.72$. Styrene and butadiene: cyan (open squares and double-dash-dot for local equilibrium). The vertical lines at the immediate right of each local equilibrium curve indicate the corresponding thermal equilibrium (almost all polymers of maximum length), "Staal 15," Fig. 7-15 (left) of [17], $R_L = 0.14$ and $R_T = 0.72$.

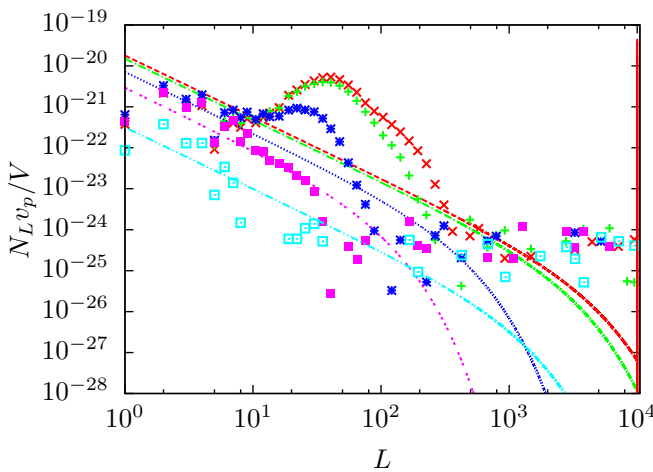


FIG. 4. $N_L v_p / V$ distribution values for the Titan atmosphere at 1013 km [red (\times 's and dashes for local equilibrium), $R_L = 0.29$ and $R_T = 0.71$], 1032 km [green (+'s and dot-dashes for local equilibrium), $R_L = 0.29$ and $R_T = 0.71$], 1078 km [blue (*'s and dots for local equilibrium), $R_L = 0.28$ and $R_T = 0.73$], 1148 km [magenta (filled squares and double-dash line for local equilibrium), $R_L = 0.33$ and $R_T = 0.81$], and 1244 km [cyan (open squares and double-dash-dot for local equilibrium) $R_L = 0.33$ and $R_T = 0.81$]. The vertical lines at the right in the figure indicate the corresponding thermal equilibria (almost all polymers of maximum length).

for all of them. The reason for this is that the large number ($b = 20$) of available monomers (amino acids) plays a significant role in determining the thermal equilibrium population distribution. [For all cases considered here, the terms -1 in the numerator and denominator of Eq. (2) can be ignored (the "Gibbs limit").] The dominant factor determining the equilibrium distribution is $e^{(\ln b + \beta \Delta)L}$. When $\Delta < 0$ as it is for polypeptides the exponent can be negative. If the exponent is positive the thermal equilibrium is predominantly long polymers whereas if it is negative, the distribution is predominantly short polymers. This is a manifestation of the competition between entropy, which in the case of peptides drives the system toward long polymers, and energy, which drives it toward monomers. When $b = 20$ the external temperature (or the bond energy) must be finely tuned to avoid one of those extremes and for most reasonable values of Δ and β the fine tuning does not occur and the thermal equilibrium state is described by one of the extremes. The R_T value is then the distance in the space of coarse grained populations between one of those extreme cases (all monomers or all polymers of maximum length) and the actual, nonequilibrium population distribution. Because the actual polymer distributions of the living systems are all similar and the thermal equilibrium distribution is at one of the extremes, the R_T values tend to be the same. However, because of the sensitivity of the thermal equilibrium population to the value of $\beta \Delta$ one finds one or the other of the two extremes with values of Δ that are within the range of measured peptide bond energies. We illustrate this in Fig. 5 where we show data for a prokaryote analyzed using the average (-2.2 kcal/mol) of the values of the peptide bond energy reported in Ref. [11] and the results of the same analysis using the average plus one standard deviation and the average minus one deviation. For the least negative value of $\beta \Delta$ the thermal distribution jumps from all monomers to all polymers of maximum length.

We conclude that within our model, with ambient temperatures in the range of hundreds of degrees, a system of polypeptides with a large b (which is 20 for terrestrial biochemistry) will be very sensitive to the external temperature and will be driven toward long polymers at higher temperatures and monomers at lower ones with a sharp transition in between. The exact transition will be determined by the value of Δ (< 0). In real living systems, that value of the peptide bond energy varies from one amino acid pair to another and the corresponding sharp transition will be smeared by the distribution of bond energies. We illustrate these points in another way in Fig. 6, where we show the calculated value of R_T for a series of $\beta \Delta$ values for one of the data sets. There is a sharp minimum in R_T when the thermal distribution is at the tipping point between all monomers and maximum length polymers. The minimum is close to but different from the local value of $\beta \Delta$ that was calculated from the protein energy and polymer number.

The implication seems to be that the observed protein distributions in the living systems are closest to a local distribution, which is quite unlikely since it requires a fine-tuned value of $\beta \Delta$ to be produced in a thermal environment. One may construct the following plausibility argument to speculatively explain this: Generally, a large number of possible chemical configurations must be explored in order to find those special

TABLE IV. Values of parameters found here characterizing the observed population distributions of some of the systems considered. (Only representative prokaryote data are listed here. R_T and R_L values for all 4555 proteomes are shown in Fig. 1.) Equations (7) and (6) were solved with six-figure precision but only three significant figures are reported here.

Figure in this paper	Description	R_L	Local $\beta\mu$	Local $\beta\Delta$	R_T	Thermal $\beta\mu$	Thermal $\beta\Delta$
Not shown	Yeast	0.319	-15.2	-2.99	0.607	-14.5	-3.78
2	Prokaryote baw	0.254	-15.0	-2.99	0.608	-14.1	-3.78
2	Prokaryote cch	0.382	-14.6	-3.00	0.608	-14.1	-3.78
2	Prokaryote cva	0.333	-14.7	-2.99	0.607	-14.1	-3.78
Not shown	<i>E. coli</i> ribosomes	0.199	-15.4	-2.98	0.597	-14.1	-3.78
3	Copolymer JCA 2010	0.156	-8.57	-0.570	0.724	-6.44×10^3	142
3	Copolymer Staal 13b	0.155	-9.53	-0.586	0.723	-7.45×10^3	142
3	Copolymer Staal 13c	0.144	-9.96	-0.597	0.721	-8.88×10^3	142
3	Copolymer Staal 13d	0.138	-10.3	-0.605	0.720	-1.03×10^4	142
3	Copolymer Staal 15	0.140	-9.52	-0.607	0.718	-8.88×10^3	142
4	Titan 1013 km	0.292	-45.8	-0.693	0.713	-1.68×10^5	167
4	Titan 1032 km	0.286	-45.9	-0.694	0.714	-1.68×10^5	167
4	Titan 1078 km	0.277	-46.7	-0.697	0.732	-1.68×10^5	167
4	Titan 1148 km	0.321	-47.6	-0.708	0.771	-1.68×10^5	167
4	Titan 1244 km	0.332	-49.8	-0.694	0.805	-1.68×10^5	167

ones that result in lifelike properties such as autocatalysis. If a system is in a thermal environment that strongly favors all monomers or all long polymers, not many configurations will be explored. Hence the systems from which a lifelike system is most likely to emerge will be those that are thermally fine-tuned to be in the intermediate region in which the system has a spread of polymer lengths and, we expect, large fluctuations in the polymer length distribution. Once

autocatalysis begins, the system may become self-stabilizing and its local distribution will differ from the one imposed by its thermal environment, as we observe for the prokaryotes, the yeast, and the ribosome. Hence we suggest that the “local” chemical equilibrium associated with the polymer density and bond energy density of these living systems may be a relic of the early thermal environment in which those living systems originally evolved. The needed value of temperature is easily estimated from the relation $\ln b = -\beta\Delta$ from which $T = |\Delta|/(k_B \ln b)$. With the average bond energy we have used and $b = 20$, this is within a few degrees of the boiling point of water at 1 atmosphere.

These speculations suggest experiments with peptide systems in which the value of the effective $\beta\Delta$ is varied over a very fine scale to find the tipping point and discover whether any interesting behavior emerges near it. For a more careful estimate of the needed temperature one could take the expected (or measured) temperature dependence of the bond

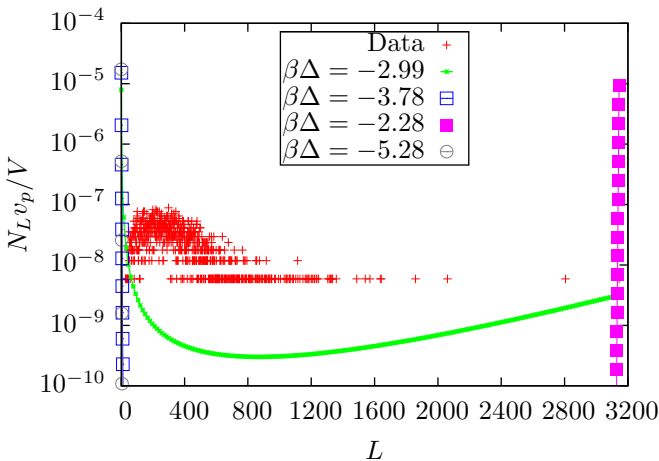


FIG. 5. Data on prokaryote cva from Fig. 2 compared to equilibria at different values of $\beta\Delta$. $\beta\Delta = -2.99$ corresponds to the local equilibrated distribution, $\beta\Delta = -3.78$ is the thermally equilibrated distribution using a bond energy of -2.2 kcal/mol and a temperature of 293 K, $\beta\Delta = -2.28$ is the distribution using a bond energy of $(-2.2 + 0.875)$ kcal/mol and a temperature of 293 K, and $\beta\Delta = -5.28$ is the distribution using an energy of $(-2.2 - 0.875)$ kcal/mol and a temperature of 293 K. Here -2.2 kcal/mol and 0.875 kcal/mol are the average and standard deviation, respectively, of the protein bond energies found in [11].

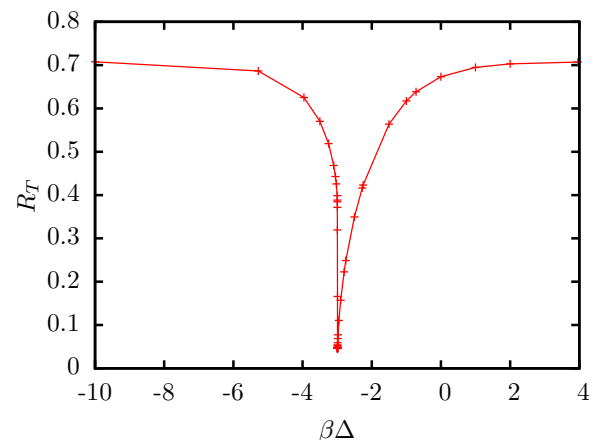


FIG. 6. R_T as a function of $\beta\Delta$ for a fixed data set.

energy into account and solve the resulting implicit equation for T .

The living systems have values of the parameters R_T and R_L that are robustly distinct from those for the nonliving copolymer ($R_T \approx 0.725$, $R_L \approx 0.15$) and the atmosphere of Titan $0.71 < R_T < 0.82$, $0.28 < R_L < 0.33$. In these two measures, Titan is mainly distinguished from the living systems by the values of R_T . Its values of R_L fall within the same range as those of the living systems. Of course many other features of the upper atmosphere of Titan differ from those of the living systems. We regard it as a strength of the present method of analysis that these dimensionless measures of disequilibrium permit a dimensionless quantitative comparison of such dissimilar systems with regard to this feature.

Note that the engineered, nonliving copolymer systems are far from equilibrium by design: Most combinatorial combinations are excluded by the preparation procedure, which assures that the molecules will all consist of a string of one type of monomer attached by one bond to a string of the other type of monomer, whereas the corresponding approximate equilibria that we find take account of the possibility of mixing the two types of monomers in all possible ways along the chain.

All the molecules detected in the Titan atmosphere are almost certainly not linear, as we have assumed in calculating the equilibrium distributions. Bonding rules assure that the energies will not be greatly affected by that error, but the degeneracies G_L that describe the numbers of molecules containing L monomers may be, particularly for small L . If, as we anticipate, the error were larger for smaller L , then that would shift the equilibrium distributions toward small L , and it appears that this might increase the values of R_L . The analysis could be repeated with more detailed models of the small molecule chemistry of the Titan atmosphere, but the idea here was to use only the empirical data available, and those data do not currently provide the information on the distribution of molecular types that is required.

There are some issues with the R_T values for the Titan atmosphere: We used a reported value [22] for the temperature of the Titan atmosphere in the analysis. But that is not really consistent because we are finding that the Titan atmosphere is not in equilibrium and therefore cannot itself be characterized by a temperature. Second, the temperature used for the Titan atmosphere in the determination of R_T is much lower ($T = 120$ K) than the characteristic terrestrial temperature ($T = 293$ K) used in the corresponding analysis of the living systems. However, we have recalculated R_T for all the Titan data assuming an ambient temperature of 293 K and find negligible changes.

We note that the molecular density of the organic molecules in the Titan upper atmosphere, both in absolute terms (particles per unit volume) and in dimensionless form as reported in Fig. 4, is much less than that of the living systems. One point that we are making in this work is that we can nevertheless compare its degree of equilibration with that for the living systems by a common measure. It has been a primary goal of our program to find such measures, which can be used to analyze and compare systems found elsewhere to determine whether they have lifelike properties without imposing excessive terracentric bias. We do not have data on

the atmosphere of Titan near its surface, where the pressure exceeds that of the earth's surface atmosphere. However, as discuss briefly below, we do not anticipate the chemistry on the surface to be as far from equilibrium (and thus to have such large values of R_T and R_L) as that in the upper atmosphere, unless of course there is actually a biosphere on or below the solid surface of Titan.

The selection of the Titan data (as well as data from the nonliving copolymer systems) as one of the nonliving test cases for applying these measures was based in part on our earlier work on Kauffman models, which suggest that very dilute chemical systems might have a better chance than denser ones of not falling into chemical equilibrium during their temporal evolution, and that appears to be a minimal requirement for the spontaneous development of lifelike molecular dynamics. It has been known for a long time [24] that chemistry in the dilute environment of space, both in the upper atmosphere of earth and other planets and in the interplanetary medium, does in fact avoid falling into chemical equilibrium while sustaining substantial chemical activity. That condition is much closer to that of the small p Kauffman models that we previously simulated and that led to nonequilibrium, possibly lifelike, dynamics, than to the environments of ponds and oceanic trenches that are often speculated to be possible sites of prebiotic evolution. Astrophysical energy sources for sustaining prebiotic chemistry in such environments have been extensively studied (e.g., [25]) and many of the chemical constituents believed to be required are present (e.g., [2,26]).

ACKNOWLEDGMENTS

This work was supported by the United States National Aeronautics and Space Administration (NASA) through Grant No. NNX14AQ05G and used the computational resources of the Minnesota Supercomputing Institute, the Open Science Grid, the University of Minnesota School of Physics and Astronomy Condor cluster, and the NASA Advanced Supercomputing division Pleiades supercomputer. We thank Aaron Wynveen for helpful discussions and Ravindra Desai, Joao Paulo, Bastian Staal, Gabriel Vivo Truyola, and Niels Fischer for answering questions about their work and supplying us with their data. The Titan data are available on NASA's Planetary Database System, as well as in summary form, in Ref. [1].

APPENDIX A: CONVERSION OF TITAN DATA

The reported data are of the form $(\log L_i, y_{L_i})$, where y_{L_i} is the number of counts in bin L_i centered at $\log L_i$ and L is the number of carbon or nitrogen atoms in a molecule. (The Cassini mass spectrometer did not have sufficient resolution to distinguish carbon from nitrogen masses.) To convert the reported distributions to a distribution as a function of L we suppose that the number y_{L_i} in the L_i th bin represents a uniform distribution over the range $(1/2)[\log(L_{i-1}) + \log(L_i)]$ and $(1/2)[\log(L_{i+1}) + \log(L_i)]$. On a linear scale the corresponding range for L is from $\sqrt{L_i L_{i-1}}$ to $\sqrt{L_i L_{i+1}}$. We round the lower limit to the nearest larger integer and the upper limit to the nearest lower integer and attribute a count value

of $y_{L_i}/(\sqrt{L_i L_{i+1}} - \sqrt{L_i L_{i-1}} + 1)$ to each integer within the interval.

APPENDIX B: BOND ENERGIES FOR TITAN DATA

Because the N-N bond energy is much lower than the C-N and C-C bond energy as reviewed in Table III, the model for the energy expressed in the relation $E = -\sum_{L=1}^{L_{\max}} (L-1)N_L \Delta$ cannot be used and the calculation of the equilibrium distributions involves, in principle, a more complicated statistical mechanical calculation that is described elsewhere [8]. Here we describe a kind of mean field approximation in which we use an average bond strength $\bar{\Delta}(p_c)$ in the expression for the energy: $E = -\sum_{L=1}^{L_{\max}} (L-1)N_L \bar{\Delta}(p_c)$, where p_c is the fraction of the monomers in the system that are carbon. The average value that we use is

$$\bar{\Delta}(p_c) = \Delta_1 p_c^2 + 2\Delta_1 p_c(1-p_c) + \Delta_2(1-p_c)^2, \quad (\text{B1})$$

where Δ_1 is the bond energy of the C-C and C-N bonds (assumed to be the same) and Δ_2 is the bond energy of the N-N bond. The expression implicitly assumes that the probability of finding a carbon at a site is p_c independent of its environment and that is not expected to be a very good assumption. To test the effects of the error, we consider the maximum and minimum energies per bond that a polymer of long length L could have given p_c and show that the results would not be much affected by using these extremal values.

To find the average bond strength that gives the minimum energy consider a polymer of length L and $L-1$ total bonds and that $L_c = p_c L$ of the monomers are carbon. The answer depends on whether $L_c \leq L/2$ or not. In the former case, the total energy is minimized by starting with a nitrogen atom and then alternating between carbon and nitrogen until all of the carbon atoms are used giving a sequence NCNCNCNCN...N. (The NCNCNCNCNCN sequence can be

placed anywhere in the chain, giving a degeneracy, but we do not consider that here.) For each carbon atom there are two carbon-nitrogen bonds; the rest are nitrogen-nitrogen bonds. This gives a total energy of $-2\Delta_1 L_c - \Delta_2(L-1-2L_c)$. For the other case $L_c > L/2$ start with all C atoms and replace less than half of them with N atoms. This can be done without introducing any N-N bonds and, since we assume that the C-N and C-C bond energies are equal, we have a total minimum energy of $-(L-1)\Delta_1$. Dividing these energies by L and taking the large L limit with $p_c = L_c/L$ fixed we obtain an average bond energy giving a minimum polymer energy of

$$\Delta_{\min}(p_c) = \begin{cases} 2\Delta_1 p_c + \Delta_2(1-2p_c), & p_c \leq 0.5, \\ \Delta_1, & p_c > 0. \end{cases} \quad (\text{B2})$$

The calculation of the average bond energy that maximizes the total energy (Δ_{\max}) given L atoms is simpler: One must maximize the number of N-N bonds and that is done by connecting all the carbon atoms together, connecting all the nitrogen atoms together, and then connecting the two with a single carbon-nitrogen bond (C...CCNN...N). The total energy is given by $-\Delta_1 L_c - \Delta_2(L-1-L_c)$. Dividing this total energy by L and taking the limit $L \rightarrow \infty$ yields the average bond energy that maximizes the total energy: Using each of these three bond energies for all the bonds, we computed the values of $\beta\mu$, $\beta\Delta$ for the Titan data for some test cases and found that the results were not sensitive to the value used. Using the three bond energies $\bar{\Delta}$, Δ_{\min} , and Δ_{\max} for calculating the thermal equilibrium length distribution for the Titan 1078 km data, it was found that the value of $\beta\mu$ changed by only 1% and there was no discernible change in the value of R_T . We also have preliminary results for the full model as described in [8] and find that the full model gives results quite close to those found by the approximate approach described above and used to obtain the results in the rest of the Titan data considered here.

-
- [1] R. Desai, A. Coates, A. Wellbrock, V. Vuitton, F. Cray, D. González-Caniulef, O. Shebanits, G. Jones, G. Lewis, J. Waite *et al.*, Carbon chain anions and the growth of complex organic molecules in Titan's ionosphere, *Astrophys. J. Lett.* **844**, L18 (2017).
- [2] M. Rahm, J. I. Lunine, D. A. Usher, and D. Shalloway, Polymorphism and electronic structure of polyimine and its potential significance for prebiotic chemistry on Titan, *Proc. Natl. Acad. Sci. USA* **113**, 8121 (2016).
- [3] B. F. Intoy and J. W. Halley, Energetics in a model of prebiotic evolution, *Phys. Rev. E* **96**, 062402 (2017).
- [4] A. Wynveen, I. Fedorov, and J. W. Halley, Nonequilibrium steady states in a model for prebiotic evolution, *Phys. Rev. E* **89**, 022725 (2014).
- [5] B. F. Intoy, A. Wynveen, and J. W. Halley, Effects of spatial diffusion on nonequilibrium steady states in a model for prebiotic evolution, *Phys. Rev. E* **94**, 042424 (2016).
- [6] R. C. Tolman, *The Principles of Statistical Mechanics* (Oxford University Press, Oxford, 1938), pp. 58–59.
- [7] D. Baum, The origin and early evolution of life in chemical composition space, *J. Theor. Biol.* **456**, 295 (2018).
- [8] B. F. M. Intoy and J. W. Halley, Quantitative estimates of chemical disequilibrium in Titan's atmosphere, [arXiv:1811.10760](https://arxiv.org/abs/1811.10760).
- [9] J. A. Paulo, J. D. O'Connell, R. A. Everley, J. O'Brien, M. A. Gygi, and S. P. Gygi, Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 *S. cerevisiae* proteins across 10 carbon sources, *J. Proteomics* **148**, 85 (2016).
- [10] J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach, and K. W. Plaxco, Random-coil behavior and the dimensions of chemically unfolded proteins, *Proc. Natl. Acad. Sci. USA* **101**, 12491 (2004); **102**, 14475(E) (2005).
- [11] R. B. Martin, Free energies and equilibria of peptide bond hydrolysis and formation, *Biopolymers* **45**, 351 (1998).
- [12] N. Fischer, P. Neumann, A. L. Kovevega, L. V. Bock, R. Ficner, M. V. Rodnina, and H. Stark, Structure of the *E. coli* ribosome-EF-Tu complex at $<3 \text{ \AA}$ resolution by C_s-corrected cryo-EM, *Nature (London)* **520**, 567 (2015).
- [13] R. Milo and R. Phillips, *Cell Biology by the Numbers* (Garland Science, New York, 2015), Chap. 2.

- [14] M. Kanehisa and S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.* **28**, 27 (2000).
- [15] R. Milo, What is the total number of protein molecules per cell volume? A call to rethink some published values, *Bioessays* **35**, 1050 (2013).
- [16] G. Vivó-Truyols, B. Staal, and P. J. Schoenmakers, Strip-based regression: A method to obtain comprehensive Co-polymer architectures from matrix-assisted laser desorption ionisation-mass spectrometry data, *J. Chromatogr., A* **1217**, 4150 (2010).
- [17] B. B. P. Staal, *Characterization of (Co)polymers by MALDI-TOF-MS* (Technische Universiteit Eindhoven, Netherlands, 2005).
- [18] P. C. Hiemenz and T. P. Lodge, *Polymer Chemistry* (CRC Press, Boca Raton, 2007), Table 6.1.
- [19] S. S. Zumdahl and S. A. Zumdahl, *Chemistry*, 7th ed. (Houghton Mifflin Company, Boston, MA, 2007), Table 8.4.
- [20] J. H. Waite Jr., D. T. Young, T. E. Cravens, A. J. Coates, F. J. Crary, B. Magee, and J. Westlake, The process of tholin formation in Titan's upper atmosphere, *Science* **316**, 870 (2007).
- [21] P. Valerino and B. Buffington, Titan Event Summary Table, Reference Trajectory 110818v4, NASA Planetary Data System, 2013.
- [22] F. Crary, B. Magee, K. Mandt, J. Waite Jr., J. Westlake, and D. Young, Heavy ions, temperatures and winds in Titan's ionosphere: Combined Cassini CAPS and INMS observations, *Planet. Space Sci.* **57**, 1847 (2009).
- [23] T. Cravens, R. Yelle, J.-E. Wahlund, D. Shemansky, and A. Nagy, Composition and structure of the ionosphere and thermosphere, in *Titan from Cassini-Huygens* (Springer, Dordrecht, Heidelberg, London, NY, 2009), pp. 259–295.
- [24] J. P. Ferris, L. Becker, K. Boering, G. D. Cody, G. B. Ellison, J. M. Hayes, R. E. Johnson, W. Klemperer, K. J. Meech, K. S. Noll, and M. Saunders, *Exploring Organic Environments in the Solar System* (National Academies Press, Washington, DC, 2007).
- [25] M. Lingam, C. Dong, X. Fang, Bruce M. Jakosky, and A. Loeb, The propitious role of solar energetic particles in the origin of life, *Astrophys. J.* **853**, 10 (2018).
- [26] L. Majumdar *et al.*, Methyl isocyanate (CH₃NCO): An important missing organic in current astrochemical networks, *Mon. Not. R. Astron. Soc.: Lett.* **473**, L59 (2018).