Title

Origins of Life: The Protein Folding Problem all over again?

Permalink

https://escholarship.org/uc/item/5tv606f3

Journal

Proceedings of the National Academy of Sciences, 121(34)

Authors

Kocher, Charles

Dill, Kenneth

Publication Date

2024-08-20

DOI

10.1073/pnas.2315000121

Peer reviewed

PNAS PERSPECTIVE

Origins of Life: The Protein Folding Problem all over again?

Charles D. Kocher[a,b] and Ken A. Dill[a,b,c,1]

Edited by Ulyana Shimanovich, Department of Materials and Interfaces, Weizmann Institute of Science, Rehovot, Israel; received September 19, 2023; accepted February 4, 2024, by Editorial Board Member Ulrich Hartl

How did specific useful protein sequences arise from simpler molecules at the origin of life? This seemingly needle-in-a-haystack problem has remarkably close resemblance to the old Protein Folding Problem, for which the solution is now known from statistical physics. Based on the logic that Origins must have come only after there was an operative evolution mechanism—which selects on phenotype, not genotype—we give a perspective that proteins and their folding processes are likely to have been the primary driver of the early stages of the origin of life.

Protein Folding Problem | origin of life | foldcats | disorder-to-order transition

No one knows how life originated—presumably on earth—about 3.5 to 3.8 billion years ago (1–3). No experiment has rediscovered it. In that breach, modeling can give some guidance. For instance, there are insightful speculations on "What type of biomolecule came first?" (4–11). This question can be narrowed further to a focus on nucleic acids vs. proteins, because those are uniquely the two types of sequence → structure → function molecules at the beating heart of biology. A popular view is that RNA came first because it can serve the dual purposes of both information storage and catalysis, leading to self-replication (5, 12, 13). Here, we summarize a recent alternative perspective that protein folding and function was among the first steps.

Why proteins, rather than RNA? In short, it follows from a different logic. The primacy of RNA has followed from supposing that self-replication is the central issue. In contrast, the primacy of proteins follows from reasoning that the central question instead is What is the driving force toward biology? Why was there any force at all? What molecular process starts from disordered states—and through some sort of needle-in-a-haystack search through a huge sequence space (haystack)—finds a few functional biomolecular sequences (needles)? We describe how the needle-in-a-haystack nature of the origin of life (OOL, or Origins) closely resembles—and can be resolved by—what we now know to be essentially the same physics that underpinned the solution to another apparent needle-in-a-haystack conundrum, the Protein Folding Problem (PFP) (14–21).

Basic Questions of Origins

How Did Evolution Begin? According to NASA's definition, "life is a self-sustaining chemical system capable of Darwinian Evolution" (22). The emphasis is ours, to emphasize the implication that because life cannot be defined in the absence of its adaptation dynamics, then some form of that dynamics must have been operating at or before the OOL. In short, life cannot originate until it can propagate. Evolution must have had a beginning. Darwinian processes among molecules were apparently not acting before Origins approximately 3.5 billion years ago, when earth's processes were purely governed by physics and chemistry. Before living systems could pick and choose molecules, there must have been a sustainable process for doing so. What was it? Here are some of the key questions:

How Did Polymer Sequences Come to Encode Molecular Functions? The heart of biology is sequence-based heteropolymers (RNA, DNA, proteins) that are the cell's catalysts, machines, and memory. Origins is sometimes expressed as finding a needle in a haystack, or as a Blind Watchmaker (23), or as monkeys on typewriters that create Shakespearean plays, because the specific sequences that give biopolymers their functions must be found from the huge space of mostly useless alternative sequences. How did order arise from disorder in polymer sequences?

What Was the Tipping Point from Degradation and Hydrolysis to Long-Term Persistence? Prebiotic chemical reactions tend toward hydrolysis, degradation, and dilution, as expressed by the second law of thermodynamics of equilibria. But, living systems are not tendencies toward equilibrium. They are driven by resource intake. How did prebiotic molecules rise up to persistent dynamics that overcame decay forces?

What Was Fitness Before There Were Cells? Biology drives toward self-servingness, by winning and losing in competitions for finite resources. There is no evident equivalent of simple prebiotic molecules being self-serving. What selection principle among molecules preceded the climbing of fitness landscapes observed in cells?

Author affiliations: [a]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794; [b]Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794; and [c]Department of Chemistry, Stony Brook University, Stony Brook, NY 11794

Author contributions: C.D.K. and K.A.D. designed research; performed research; and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. U.S. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: dill@laufercenter.org.

Published August 12, 2024.

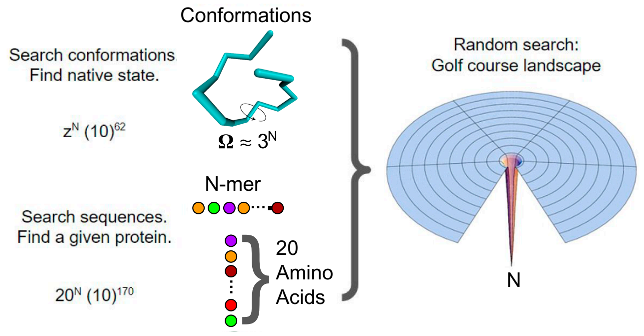

Fig. 1. Searches using a golf course landscape are random and slow. (*Left*, *Top*) Protein folding searches conformations to find the native structure. (*Left*, *Bottom*) Origins searches sequence space to find given functional proteins. (*Right*) A golf course landscape representing a completely random search process. It is shown with a single minimum; however, sometimes protein folding funnels can have multiple minima, and we expect that sequence funnels will too.

Two Stories of Needles-in-Haystacks

There is a remarkable similarity between the two puzzles of origins of life and the old PFP—both have apparent needle-in-a-haystack unlikelihood combinatorics and both pertain to protein sequences. The Origins problem can be regarded as a search of a large sequence space to pick out particular sequences. What is the probability that a present-day amino acid sequence, say of lysozyme, arose from random selection out of the vast sea of all possible sequences? That probability is often taken to be infinitesimal. Needle-in-haystack problems can be expressed as a landscape shaped like a golf course (*Right* side of Fig. 1) that is randomly searched by a ball rolling on it. Finding the hole by rolling a ball randomly would be impossibly slow.

Likewise, previously seen as a needle-in-a-haystack search was the PFP (16, 17, 19–21): how does a protein molecule search its vast conformational space to find its unique native structure (*Top Left* of Fig. 1)? This, too, was expressed in terms of golf-course-shaped landscapes. But, the solution to the physical folding problem, of the driving forces and kinetic routes, is now well-known[*] (14–18), and it gives useful insights for the OOL. In short, needles-in-haystacks and golf courses are now seen as just incorrect conceptualizations. The problem is not one of random independent steps; the problem is to find what types of physical cooperativity cause the snowballing, or bootstrapping, of one state of small probability to another state of higher probability. Fig. 2 shows three lessons from the PFP: 1) That a physical code, based on hydrophobic (H) and polar (P) patterning reduces the haystack by more than 100 orders of magnitude, as confirmed by experiments (26–30). 2) That the landscape is relatively funnel-shaped, not golf-course-like, because of physical cooperativities in secondary and tertiary drivers (16, 17, 31). 3) That the kinetics is local first, global later; helices and turns early, tertiary structure later (32); reducing an NP-completeness challenge to an often very fast process. In the Foldon Funnel Model, the early fast steps (helices and turns) are not stable (i.e., not downhill); they are just less unstable, continuing up a landscape of diminishing steepness until reaching a tipping point at which full stability is achieved by the native structure (33).

Finding the sequence of a particular protein, say lysozyme, by randomly searching the space of all 20^N sequences of length N, entails the infinitesimal probability of 20^{-N}. But, searching the space of particular folds and functions requires only a search of 2^N sequences because of the binary HP folding code (26, 27); see Fig. 2, *Left*. This space is vastly smaller, by a factor of 10^N.

Learning from the Logic of Evolution

Insights into the beginnings of evolutionary dynamics can come from knowing how cellular Darwinian evolution (DE) currently operates (34, 35). DE is a process of mutational search, competition, and fitness ratcheting. What does today's process tell us about the singularity 3.5 billion years ago at the beginnings of that process?

Evolution's Grist Is Molecules Making Molecules. To be self-sustaining and persistent, evolution is rooted in the autocatalysis (positive feedback) of replication. We call it "moms making moms." Evolution selects on phenotypes, not genotypes. Selection cannot see a gene if it is not

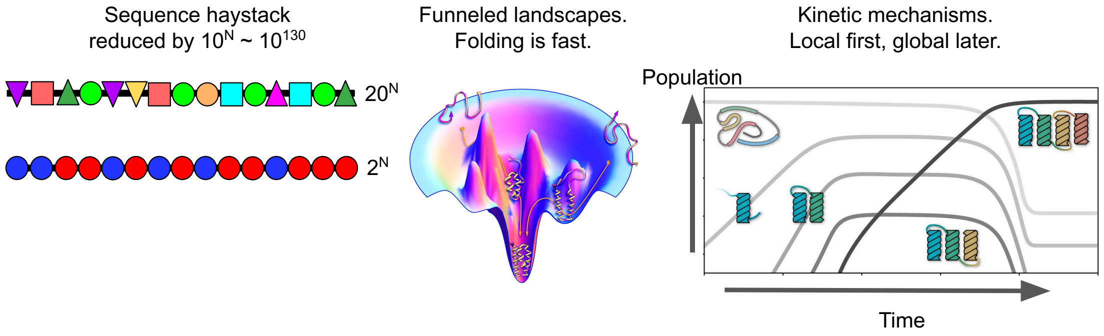

Fig. 2. Three lessons from the PFP (16, 17, 31). (*Left*) Hydrophobic/polar patterning determines the conformation space, not the exact amino acid sequence, making the search space smaller (26–29). (*Center*) Folding landscapes are funneled, representing a driven search, not a random one. (*Right*) When proteins fold, local structures collapse first, which then allow for the global structure to form. Populations of chains with varying degrees of collapse are shown, with darker curves indicating more folded structure. Figure from ref. 33.

[*] The physical folding problem is different than the computational protein-structure-prediction problem that machine learning methods are now used for (24, 25).

expressed. Selection in simple organisms is based on differential growth rates, and the main mass production that constitutes growth in cells is protein. Moreover, the makers and catalysts that produce that growth are also mainly proteins. For prebiotic origins, the equivalent autocatalytic process was "molecules making molecules." Of course, some RNA molecules can be functional, in self-catalysis like splicing of mRNAs, in ribosomes, and others, but today's functionality is mostly proteins.

Evolution Is Implemented by Sequence → Function Molecules. On the whole, proteins are good at function and catalysis, while RNA is a good vehicle for information. This difference can be rationalized by their chain physics. 1) Proteins are versatile catalysts over a broad range of reactions, in part because of their 20 side-chain moieties covering the chemical spectrum, vs. the smaller repertoire of interaction types in nucleic acids (36, 37). 2) Proteins make good catalysts because they fold into single solid-like miniature stable surfaces. RNAs are stringier: they are less compact and more structurally heterogeneous because RNAs are dominated by secondary structure forces while proteins are dominated by tertiary forces (38, 39). And even those RNAs that are structured are assisted, in ribosomes, by protein–RNA interactions (40); or in tRNA molecules, by modified nucleotides (41). 3) Proteins give a more direct and unique mapping of sequence to structure because of funnel-shaped energy landscapes. While some RNA molecules do have unique folds, when taken over RNA sequence space as a whole, landscapes are generally rugged with multiple minima (42), because of multiple hydrogen bonds per base pair and because of the greater degrees of freedom around the backbone [eight in RNA vs. two in proteins (43)].

Evolution Is a "Feynman Ratchet," Not a Copy Machine. Evolutionary replication is not perfect autocatalysis: It is heritable variation, i.e., descent with modification. If moms made identical copies of themselves, evolution would have died out, having been too brittle in the face of environmental variations and unruliness, particularly in life's fragile early stages (34). The winners that take all would die out when the environmental "winds" shift. Rather, evolution's replication, through a process of search/compete/select, is of autocatalytic sets, wherein one element of the set can produce others (44–47). Evolution is a Feynman Ratchet, like a Brownian ratchet, where a random noisy input drives a directed output. Mutational searching generates much random junk that becomes selected through competition.

Evolution Is Driven by an Out-of-Equilibrium Environment. Cells are driven by intake of resources—food, energy, and water. They are open systems, not driven by a second law tendency to equilibrium. Open systems are of two types: 1) "equilibrium," of the system with a bath, having no net flow either way, having only fluctuations—These tend to equilibria—and 2) "nonequilibrium," where there is a net unbalanced flow between system and environment. In biology, only death is explained by (1); life must be explained by (2). Think about a TV set. Its action is not a tendency toward equilibrium until it is unplugged; as long as a TV set is plugged in, its action is persistent complex flows of electrical currents that enact its functionality. Origins too must have entailed some form of persistent environmental driver that coupled to molecule-making. What type of molecule-making could have innovated through positive feedback cooperativity?

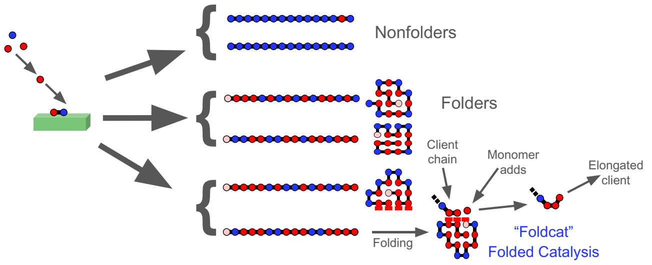

Fig. 3. Bootstrapping foldcats from stationary catalysis. The Founding Rock (green) assembles polymers from hydrophobic (red) and polar (blue) monomers. Some of these polymers fold, and some of the folders catalyze elongation of other chains on hydrophobic patches of their surface. The two foldcat structures at the bottom are from a computation showing that they are the unique native conformations of those two sequences in the two-dimensional HP lattice model. The elongated client chain shown could be a piece of any foldcat containing the sequence PHHH, contributing to the foldcat autocatalytic set.

Hypothesis: Evolution Started with Proteins

The HP Foldcat Mechanism. Now, we discuss the link between prebiotic evolution and the protein folding process, which is encapsulated by the HP Foldcat (HPF) mechanism. The HP Foldcat mechanism is presented in detail elsewhere (35, 48); here, is a summary. The "HP" stands for hydrophobic and polar, the two types of (amino acid) monomers that can get polymerized into a chain. Peptides are assumed to be continuously synthesized from these monomers through catalysis on some prebiotic catalyst, which we call the Founding Rock[†]; see Fig. 3. At first, the peptides are short and random. A small fraction of those chains are longer, collapsing (folding) into compact conformations with a hydrophobic core. Some folders have stable surfaces. Indicated here as having hydrophobic sticky "landing pads," these foldcat sequences are catalysts that accelerate elongation of a client chain by bringing the next monomer to be added into juxtaposition. What follows below is first the big-picture conclusions from the model, followed by a more quantitative description of it.

This Mechanism Has Emergent Properties. The foldcat process has emergent properties—i.e., properties that are not anticipated from simple noncooperative random short-chain synthesis alone. a) Chains grow longer. b) There are autocatalytic sets, where some sequences become preferentially populated, explaining the beginnings of sequence–structure relationships (48, 49). c) A fitness property emerges—namely the folding stability (and ultimately the catalytic effectiveness)—through which some sequences survive and win while other sequences degrade and recycle. And whereas fitness starts as simple folding stability, once this form of privileging takes hold persistently, any other factors

[†] Simply as a shorthand notation for some possibly macroscopic mineral surface, volcanic vent, or air–water interface, constrained to be located at particular spatial location(s).

that can stabilize proteins or their communities can also support further evolutionary change. d) Accordingly, an evolution-like process emerges, namely of sequence searching and fitness selection via competition for monomers; see also ref. 6. The emergence of autocats and function here is not dependent on a nucleic acid template.

e) Moreover, this mechanism has long-term persistence for the following reasons. 1) It is an open system, not driven by a second law tendency to equilibrium. It is driven by a nonequilibrium input of monomers and a Founding Rock that initially facilitates the otherwise nonspontaneous polymerization of these monomers. 2) At some point in the process, when chains become sufficiently good foldcats, there would be an untethering transformation, a point in time at which catalysis is mostly carried out by foldcat proteins and is no longer reliant on the Founding Rock. This would be a major evolutionary event, because no longer is Origins localized, i.e., stuck in some "small pond in Nebraska." Now catalysts are mobile and can go anywhere; now catalysts are programmable (different proteins can catalyze different reactions or work in different condition); and now catalysis becomes miniaturized and capturable within cells (50).

These emergent properties come from two physical cooperativities: i) Chains that are long enough have folded cores with enhanced protection against degradation, and ii) the collection of chains that help elongate other chains forms an autocatalytic set. Like a snowflake that accretes more snow, that turns into a snowball and ultimately into an avalanche, cooperativities are key to explaining how the improbable first steps rise up to dominate the macroscale. The HP Foldcat mechanism snowballs toward longer, more folded, and more catalytic molecules.

This Mechanism Is Prebiotically Plausible. While it has not been observed directly, the HP Foldcat mechanism could plausibly have arisen on the early earth. First, amino acids and peptides have been produced in prebiotically plausible ways through terrestrial processes (9, 51–54) or from space (55–58).[‡] Founding-Rock-like catalysis of peptide elongation through dehydration reactions has been demonstrated on mineral surfaces (60–62) and at air–water interfaces (sea spray) (63–65), or even by unknown extraterrestrial processes (66). Plausible nonequilibrium drivers of peptide bond formation could have been wet-dry cycles (67, 68) or hot-cold cycles (69).

Second, polymers of hydrophobic and polar monomers can fold and catalyze, even if the chains are random and/or short. Proteins are known to be driven by a binary HP code (16, 29, 30, 70, 71). Because today's 20 amino acids are found in roughly equal hydrophobic and polar proportions in the PDB, it means that most sequences, of any sufficient chain length, will collapse to compactness in water (26, 27, 72–76) and thus have cores that are protected from access to the external solvent (77, 78). While speculations suggest that early alphabets may have had fewer than 20 amino acids (79, 80), all that matters here is just a binary code. Also, short proteins are ubiquitous in biology: Humans have thousands of microproteins (81), which are proteins that are less than 100 amino acids long. Although modern microproteins may have a distinct, later evolutionary origin, they demonstrate that short proteins can be interesting and functional as required for the HP Foldcat mechanism. Many microproteins perform biological functions including catalyzing reactions (82–90).

Third, folding and catalysis are simple physical properties of HP chains. They are found in prebiotic mixtures of amino acids (79, 91–96), possibly assisted by available small molecules (97). Moreover, even just cysteine alone, a single hydrophobic amino acid, has been suggested to be both prebiotically available and capable of performing peptide ligation reactions under plausible prebiotic conditions (98). In addition, catalysis has also been observed in amyloids (99, 100). And, while today's enzyme catalysts often utilize high levels of atomically detailed chemical and spatial specificity, simpler spatial proximity effects, as envisioned here, are capable of giving orders of magnitude speed-ups to reactions (101–103). The HP foldcat mechanism predicts that persistent generation of HP chains could lead to some longer folded chains, a fraction of which could catalyze other actions.

The Origins Problem Resembles the Folding Problem

Now, compare the origins of life problem to the PFP. In the Foldcat conception, both problems are centered on proteins—one in conformational space and the other in sequence space; one driven by equilibrium forces and the other by nonequilibrium forces. Nevertheless, both have funnel-shaped landscapes; both have dynamical epochs for how order arises from disorder; both entail cooperative interactions through which random steps lead from local to global order; and both of them solve apparent needle-in-a-haystack combinatorics problems through the protein folding code.

Both Have Funnel Landscapes. See Fig. 4. In the end, protein folding turned out not to be a needle-in-a-haystack problem. Although it entails a search through a huge space to find the single native structure, it is not random. Energetic preferences favor compact hydrogen-bonded states with hydrophobic contacts. Steps downhill in free energy lead to more stability and facilitate additional steps. Even though individual steps are stochastic, the net result is directed. It matters not if a state is highly unlikely based on the count of other options; it only matters if one advance can lead to another, the way one snowflake can start a snowball, and an avalanche, downhill. The reason for the large width at the top (high entropy) and smooth reduction to the low-entropy native state is because of excluded volume (104). The denser the chain gets as it grows more compact and native-like, the fewer configurations it has that remain available. Protein folding entropies are huge: $T\Delta S \approx -100$ kcal/mol for 100-mer sized proteins, about half of which is due to the backbone and half to the sidechains (105, 106). In short, the PFP was not about aimless searching, but about the accumulation of local advantages and the cooperativities of one step leading to the next (33).

[‡] In contrast, the production of RNA under prebiotic conditions has been more challenging (9, 59).

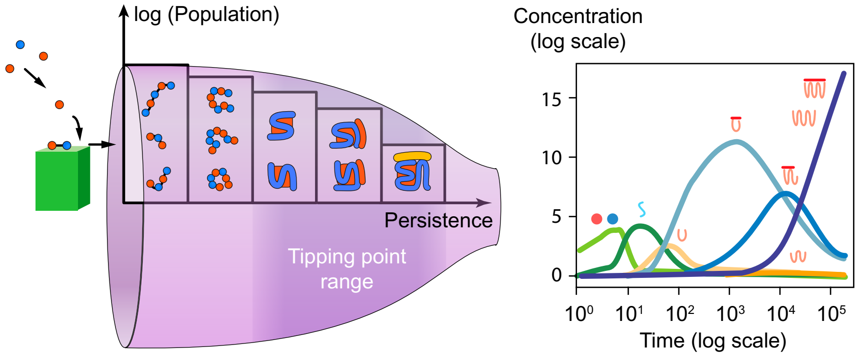

Fig. 4. Foldcats cause a funneling-like exploration toward a particular region of sequence space by leveraging local advantages. (*Left*) The Founding Rock makes random chains, from which stable and catalytically active ones are selected. The few discovered foldcats untether the protein synthesis process from the Founding Rock and preferentially make more of themselves. (*Right*) A simplified model of the foldcat mechanism, introduced in ref. 107, shows how monomers (light green) initially form random, useless chains (dark green), slowly develop folding (orange), then catalysis (light blue), which enables longer, more functional chains to be created (darker blue curves).

Understanding the basis for cooperativity is important for the origins problem, as it was for the folding problem. It is funnel-like. The *Left* side of Fig. 4 has a large space of random short sequences, leading to a later stage of a much smaller space of longer folders and foldcats.[§] The *Right* side of Fig. 4 gives an example of the time-course of this funneling in terms of populations of different types of HP chains (nonfolding, folding, foldcat, of various lengths) (107). Proteins fold because conformational space is shaped like a funnel. The HPF model shows how origins, too, may result from funneling, in this case in sequence space.

Briefly, here is the model more quantitatively (107). As noted above, any origins model must explain cooperativities, i.e., a physical basis for nonrandom "snowballing." For the two types of cooperativity embodied here, the present model is minimal insofar as it lumps together microstates into mesostates in a way that requires only 3 rate parameters after simplification. Let M be the total amount of monomer, r the total number of nonfoldcat chains, and A the total number of foldcats. Monomer is supplied to the system at a rate α_M and decays at a rate $d_M M$, nonfoldcat chains are created (from monomers, by both the foldcats and the Founding Rock) with rate $\alpha_r(A, M)$ and decay at a rate $d_r r$, and foldcats decay at a rate DA. The specific way in which foldcats cooperatively speed up their communal formation by making their precursors from monomers in the function $\alpha_r(A, M)$—the feature that we want to study—is importantly still present in this simplified model even though other details of the foldcat mechanism have not been explicitly tracked. Finally, the elongation reactions, which are catalyzed both by the Founding Rock and by the foldcats A, are as follows: 1) $r + M \rightarrow A$ (nonfoldcat is elongated into a foldcat), 2) $r + M \rightarrow r$ (nonfoldcat is elongated and still is not a foldcat), 3) $A + M \rightarrow A$ (foldcat is elongated and is still a foldcat), and 4) $A + M \rightarrow r$ (foldcat is elongated and is no longer a foldcat). Elongation reaction i has mass-action rate constant $K_i(A)$, to which the Founding Rock and foldcats both contribute. These equations give a set of ODEs for foldcat cooperativity:

$$\begin{aligned}\frac{dr}{dt} &= \alpha_r - d_r r + K_4(A)AM - K_1(A)rM\,,\\ \frac{dM}{dt} &= \alpha_M - [d_M + r(K_1(A) + K_2(A)) +\\ &\quad + A(K_3(A) + K_4(A))]M\,,\\ \frac{dA}{dt} &= K_1(A)rM - K_4(A)AM - DA. \end{aligned} \qquad [1]$$

We now reduce these to a single rate equation. By eliminating M and r through steady-state arguments, and by switching to dimensionless variables (for details, see ref. 107), we get

$$\frac{dA}{dt} = \frac{k_1 A}{1 + k_1 A} + \frac{k_1 k_2 A^2}{(1 + k_1 A)(1 + A/A_s)} - A. \qquad [2]$$

where k_1 characterizes the rate at which foldcats make new foldcats (related to $K_1(A)$ above), k_2 characterizes the rate at which foldcats make their precursors (related to $\alpha_r(A, M)$ above), and A_s is the number of foldcats at which the latter reaction starts to saturate. The rate k_2 provides the cooperativity: The ability of foldcats to make their precursors allows them to accelerate their collective production nonlinearly. Specifically, when the number of foldcats is small, i.e., in the prebiotic stage, $A \rightarrow 0$ and we find

$$\frac{dA}{dt} \approx k_1(1 + k_2 A - k_1 A)A - A. \qquad [3]$$

The two terms in Eq. **3** give growth and decay rates, respectively. For a noncooperative autocatalyst, the growth rate is gA, where g is a constant. In this case, $g - 1$ is either positive or negative for all values of A, meaning that the foldcats either grow or decay just depending on the constant value of g. Cooperativity occurs when g is itself a function of A, such as $g = k_1(1 + k_2 A - k_1 A)$. When $k_2 < k_1$, the cooperativity is negative and inhibits further growth. However, when $k_2 > k_1$, the cooperativity is positive and can encourage further growth. Now, the sign of dA/dt also depends on the value of A, because g itself depends on the value of A. Instead of having only-growth or only-decay behavior as in the noncooperative case, the population of foldcats can have a bistable behavior: When A is small, $dA/dt < 0$, but when A increases, eventually, $dA/dt < 0$. Positive cooperativity allows for an initially unfavorable

[§] In simplest approximation, the walls of this funnel are linear on a log scale because funneling follows 20^{N-m}, where N is total target chain length and m is the particular sequence length.

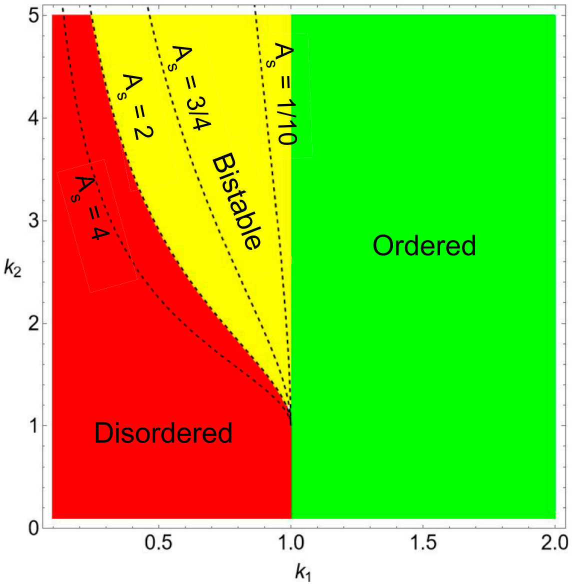

Fig. 5. The dynamical phase diagram of foldcat origins, from Eq. **2**. Increasing the cooperative reproduction rate k_2 above the noncooperative reproduction rate k_1 allows for a bistable region where a stochastic jump can take foldcats from a disordered state (mostly small useless chains) to an ordered state (folders, foldcats, long chains enriched).

environment (small $A \rightarrow dA/dt < 0$) to be overcome by fluctuating to a higher population where the growth rate is positive (107).

Disorder to Order: From Many to One. Because foldcats can exhibit positive cooperativity, they can transition from a world where degradation dominates, with only random short chains, to a world of persistent growth, with longer-chain folders and foldcats that propagate stably with evolution-like dynamics. Other studies explore similar disorder-to-order transitions of other Origins models (108–111).

Fig. 5 shows the kinetic phase diagram of this model Eq. **2**. In the red region, foldcats just die off and cannot survive. The environment is too unfavorable; it produces only short random chains, the disordered state. In the green region, foldcats grow deterministically into a persistent population because both growth rate parameters are favorable. In the yellow region, the environment is unfavorable, according to the mass-action dynamics of Eq. **2**, but stochastic fluctuations can drive the system over the barrier from disorder to order. This stochastic behavior is because of the nonlinearity (cooperativity) in Eq. **3**: there are two stable steady-states, one with $A = 0$ (disordered) and one with $A > 0$ (ordered), which are separated by a "kinetic barrier" that must be hopped over.

The goal of this modeling is to ask whether foldcat cooperativities admit of any possible window of viable parameters that could lead to a persistent evolutionary process toward further complexity and biology. It is only a minimal model, and of necessity neglects many things, including repurposing, other forms of noncatalytic functionalities, and/or protein assemblies. The main conclusion here is that the yellow region is a viable tipping point route from prebiotic degradation to kinetic persistence of foldcats in an unfavorable initial environment. As described below, it leads to an evolutionary funnel in sequence space.

From Disorder to Order. Models of protein folding show epochs of steps that follow a hierarchy of local-first, global later; see the *Right* side of Fig. 2. Microscopically, the steps are stochastic. But "mindless" local advantages add up to globally optimal and ordered structures. First to appear are local interactions in helices and turns; later are nonlocal helix–helix interactions. In folding, most early steps are undirected and unproductive, but ultimately the native state (ordered) arises from the denatured states (disordered). The foldcat mechanism of origins reflects similar hierarchical epochs, as shown on the *Right* side of Fig. 4. In the foldcat model, first come short unfunctional molecules, followed by systematically longer and functional molecules. In both evolution and folding processes, small incremental advantages are found among a sea of options, and then further advantages accumulate, leading ultimately to a greater global advantage.

The Rest of the Story: From Evolution to Origins

Our foldcat mechanism is not a full story of Origins. It is missing major components of life, including cell encapsulation, nucleic acids, lineages and inheritance through a genetic code, and the complex biochemical pathways needed to implement them. What is the fitness ratchet that is preserving value among prebiotic molecules? Here, it is simply persistence, i.e., the folding stability of a chain—longer chains are more stably folded, so they persist longer in a fluctuating environment. Once this simple evolutionary dynamics is stable, this machinery can then further discover other forms of persistence and fitness. Various such discoveries have been proposed: selection of amino acid type (79, 91, 95, 112); use of an energy currency such as ATP (109); a better protein chain elongator [ribosome, (113)]; or other features (114).

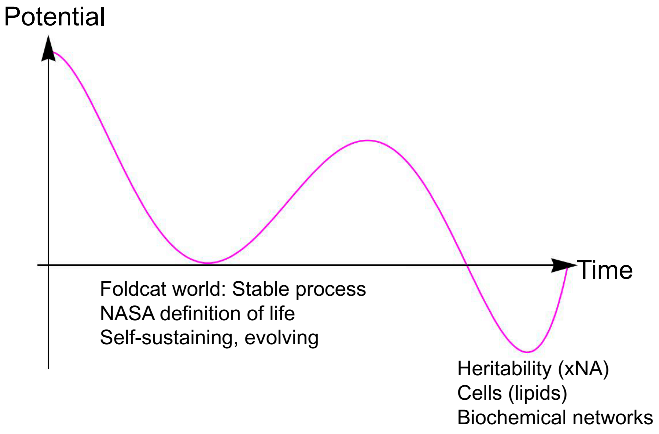

Fig. 6. The potential-energy-like valleys (stable states) on the path to cellular life, starting with foldcats. Foldcats represent a stable, persistent state, from which evolutionary dynamics can jump to the next minimum closer to biological life.

On the one hand, the present model assigns primacy to proteins insofar as proteins alone are the minimal system that can explain arguably the first step—which is an evolution-like process toward origins—without requiring any other biomolecules. On the other hand, that does not have temporal implications about other components. It does not mean that other molecules, including RNA, were not present concurrently or even undergoing interesting dynamics themselves.

Fig. 6 shows a potential-energy-like diagram with two steps to Origins. First are evolutionary dynamics such as the foldcat mechanism; second are the further ingredients we just listed. Before the advent of lineages—i.e., nucleic acids and cells—persistence is only on the time scales of molecule processes. The advent of lineages gives extraordinary extension of the time scale of persistence, all the better for handling larger environmental unruliness. This model brings the perspective that prebiotic chemistry was not "aiming to become biology," but that it ratcheted up chemical persistence and biomolecules were the best way to achieve it.

Conclusions

The origins of life must have been preceded by a stable evolution-like propagation mechanism. We review how evolution could arise from a random generator of peptide sequences that could ultimately function and catalyze reactions. Two types of proteins' cooperativities, in their assembly and reduction of degradation rates, lead to the emergence of longer chains, autocatalytic sets, increasing persistence and function through a narrowing sequence space funnel, and a tipping point from disorder to order. This origins process in sequence space resembles—and originates partly from—the folding process in conformational space.

Data, Materials, and Software Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We are grateful to the Laufer Center for Physical and Quantitative Biology at Stony Brook and to the John Templeton Foundation for financial support (Grant ID 62564).

1. J. W. Schopf, A. B. Kudryavtsev, A. D. Czaja, A. B. Tripathi, Evidence of Archean life: Stromatolites and microfossils. *Precamb. Res.* **158**, 141–155 (2007).
2. Y. Ohtomo, T. Kakegawa, A. Ishida, T. Nagase, M. T. Rosing, Evidence for biogenic graphite in early Archaean Isua metasedimentary rocks. *Nat. Geosci.* **7**, 25–28 (2014).
3. M. S. Dodd *et al.*, Evidence for early life in earth's oldest hydrothermal vent precipitates. *Nature* **543**, 60–64 (2017).
4. G. Wachtershauser, Before enzymes and templates: Theory of surface metabolism. *Microbiol. Rev.* **52**, 452–484 (1988).
5. W. Gilbert, Origin of life: The RNA world. *Nature* **319**, 618 (1986).
6. G. F. Joyce, J. W. Szostak, Protocells and RNA self-replication. *Cold Spring Harb. Perspect. Biol.* **10**, a034801 (2018).
7. C. de Duve, The beginnings of life on earth. *Am. Sci.* **83**, 428–437 (1995).
8. D. Segré, D. Ben-Eli, D. W. Deamer, D. Lancet, The lipid world. *Origins Life Evol. Biosp.* **31**, 119–145 (2001).
9. S. D. Fried, K. Fujishima, M. Makarov, I. Cherepashuk, K. Hlouchova, Peptides before and during the nucleotide world: An origins story emphasizing cooperation between proteins and nucleic acids. *J. R. Soc. Interface* **19**, 20210641 (2022).
10. C. P. J. Maury, Self-propagating β-sheet polypeptide structures as prebiotic informational molecular entities: The amyloid world. *Origins Life Evol. Biosp.* **39**, 141–150 (2009).
11. C. P. J. Maury, Origin of life., Primordial genetics: Information transfer in a pre-RNA world based on self-replicating beta-sheet amyloid conformers. *J. Theor. Biol.* **382**, 292–297 (2015).
12. M. P. Robertson, G. F. Joyce, The origins of the RNA world. *Cold Spring Harb. Perspect. Biol.* **4**, a003608 (2012).
13. J. F. Atkins, R. F. Gesteland, T. Cech, *RNA Worlds: From Life's Origins to Diversity in Gene Regulation* (Cold Spring Harbor Laboratory Press, 2011).
14. K. A. Dill, S. B. Ozkan, T. R. Weikl, J. D. Chodera, V. A. Voelz, The protein folding problem: When will it be solved? *Curr. Opin. Struct. Biol.* **17**, 342–346 (2007).
15. K. A. Dill, S. B. Ozkan, M. S. Shell, T. R. Weikl, The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316 (2008).
16. K. A. Dill, J. L. MacCallum, The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
17. R. Nassar, G. L. Dignon, R. M. Razban, K. A. Dill, The protein folding problem: The role of theory. *J. Mol. Biol.* **433**, 167126 (2021).
18. J. N. Onuchic, Z. Luthey-Schulten, P. G. Wolynes, Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).
19. P. G. Wolynes, J. N. Onuchic, D. Thirumalai, Navigating the folding routes. *Science* **267**, 1619–1620 (1995).
20. A. V. Finkelstein, N. S. Bogatyreva, D. N. Ivankov, S. O. Garbuzynskiy, Protein folding problem: Enigma, paradox, solution. *Biophys. Rev.* **14**, 1255–1272 (2022).
21. D. Thirumalai, E. P. O'Brien, G. Morrison, C. Hyeon, Theoretical perspectives on protein folding. *Annu. Rev. Biophys.* **39**, 159–183 (2010).
22. G. F. Joyce, D. W. Deamer, G. Fleischaker, *Forward to Origins of life: The Central Concepts in Origins of Life: The Central Concepts* (Jones and Bartlett Publishers, 1994).
23. R. Dawkins, *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design* (W. W. Norton & Company, 1996).
24. J. Jumper *et al.*, Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
25. A. Madani *et al.*, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1–8 (2023).
26. K. F. Lau, K. A. Dill, A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* **22**, 3986–3997 (1989).
27. K. F. Lau, K. A. Dill, Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 638–642 (1990).
28. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, Deciphering the message in protein sequences: Tolerance to amino acid substitutions. *Science* **247**, 1306–1310 (1990).
29. S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680–1685 (1993).
30. W. A. Lim, R. T. Sauer, Alternative packing arrangements in the hydrophobic core of Λrepresser. *Nature* **339**, 31–36 (1989).
31. K. Dill, R. Jernigan, I. Bahar, *Protein Actions: Principles and Modeling* (CRC Press, 2017).
32. S. W. Englander, L. Mayne, Z. Y. Kan, W. Hu, Protein folding-how and why: By hydrogen exchange, fragment separation, and mass spectrometry. *Annu. Rev. Biophys.* **45**, 135–152 (2016).
33. G. C. Rollins, K. A. Dill, General mechanism of two-state protein folding kinetics. *J. Am. Chem. Soc.* **136**, 11420–11427 (2014).
34. C. D. Kocher, K. A. Dill, Darwinian evolution as a dynamical principle. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218390120 (2023).
35. C. Kocher, K. A. Dill, Origins of life: First came evolutionary dynamics. *QRB Discov.* **4**, e4 (2023).
36. K. W. Plaxco, M. Gross, *Astrobiology: An Introduction* (JHU Press, 2021).
37. G. J. Narlikar, D. Herschlag, Mechanistic aspects of enzymatic catalysis: Lessons from comparison of RNA and protein enzymes. *Annu. Rev. Biochem.* **66**, 19–59 (1997).
38. J. A. Cruz, E. Westhof, The dynamic landscapes of RNA architecture. *Cell* **136**, 604–609 (2009).
39. S. E. Butcher, A. M. Pyle, The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc. Chem. Res.* **44**, 1302–1311 (2011).
40. D. Klein, P. Moore, T. Steitz, The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.* **340**, 141–177 (2004).
41. T. Biedenbänder *et al.*, RNA modifications stabilize the tertiary structure of tRNAfMet by locally increasing conformational dynamics. *Nucl. Acids Res.* **50**, 2334–2349 (2022).
42. S. J. Chen, K. A. Dill, RNA folding energy landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 646–651 (2000).
43. K. S. Keating, E. L. Humphris, A. M. Pyle, A new way to see RNA. *Q. Rev. Biophys.* **44**, 433–466 (2011).
44. W. Hordijk, A history of autocatalytic sets. *Biol. Theory* **14**, 224–246 (2019).
45. S. A. Kauffman, Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. *J. Cybernet.* **1**, 71–96 (1971).
46. S. A. Kauffman, *The Origins of Order: Self-organization and Selection in Evolution* (Oxford University Press, 1993).
47. W. Hordijk, M. Steel, Conditions for evolvability of autocatalytic sets: A formal example and analysis. *Origins Life Evol. Biosph.* **44**, 111–124 (2014).
48. E. Guseva, R. N. Zuckermann, K. A. Dill, Foldamer hypothesis for the growth and sequence differentiation of prebiotic polymers. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7460–E7468 (2017).
49. T. Farquharson, L. Agozzino, K. Dill, The bootstrap model of prebiotic networks of proteins and nucleic acids. *Life* **12**, 724 (2022).
50. C. Kocher, L. Agozzino, K. Dill, Nanoscale catalyst chemotaxis can drive the assembly of functional pathways. *J. Phys. Chem. B* **125**, 8781–8786 (2021).
51. S. L. Miller, A production of amino acids under possible primitive earth conditions. *Science* **117**, 528–529 (1953).
52. S. L. Miller, H. C. Urey, Organic compound synthesis on the primitive earth. *Science* **130**, 245–251 (1959).
53. E. T. Parker *et al.*, Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5526–5531 (2011).
54. A. P. Johnson *et al.*, The Miller volcanic spark discharge experiment. *Science* **322**, 404 (2008).

55. J. R. Cronin, S. Pizzarello, Amino acids in meteorites. *Adv. Space Res.* **3**, 5–18 (1983).
56. D. P. Glavin *et al*., Extraterrestrial amino acids and L-enantiomeric excesses in the CM2 carbonaceous chondrites Aguas Zarcas and Murchison. *Meteorit. Planet. Sci.* **56**, 148–173 (2021).
57. Y. Kebukawa, S. Asano, A. Tani, I. Yoda, K. Kobayashi, Gamma-ray-induced amino acid formation in aqueous small bodies in the early solar system. *ACS Cent. Sci.* **8**, 1664–1671 (2022).
58. O. Botta, J. L. Bada, Extraterrestrial organic compounds in meteorites. *Surv. Geophys.* **23**, 411–467 (2002).
59. N. Kitadai, S. Maruyama, Origins of building blocks of life: A review. *Geosci. Front.* **9**, 1117–1153 (2018).
60. Y. Furukawa, T. Otake, T. Ishiguro, H. Nakazawa, T. Kakegawa, Abiotic formation of valine peptides under conditions of high temperature and high pressure. *Origins Life Evol. Biosph.* **42**, 519–531 (2012).
61. J. F. Lambert, Adsorption and polymerization of amino acids on mineral surfaces: A review. *Origins Life Evol. Biosph.* **38**, 211–242 (2008).
62. W. Takahagi *et al*., Peptide synthesis under the alkaline hydrothermal conditions on Enceladus. *ACS Earth Space Chem.* **3**, 2559–2568 (2019).
63. D. T. Holden, N. M. Morato, R. G. Cooks, Aqueous microdroplets enable abiotic synthesis and chain extension of unique peptide isomers from free amino acids. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2212642119 (2022).
64. E. C. Griffith, V. Vaida, In situ observation of peptide bond formation at the water–air interface. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15697–15701 (2012).
65. A. M. Deal, R. J. Rapf, V. Vaida, Water–air interfaces as environments to address the water paradox in prebiotic chemistry: A physical chemistry perspective. *J. Phys. Chem. A* **125**, 4929–4942 (2021).
66. S. A. Krasnokutski, K. J. Chuang, C. Jäger, N. Ueberschaar, T. Henning, A pathway to peptides in space through the condensation of atomic carbon. *Nat. Astron.* **6**, 381–386 (2022).
67. J. G. Forsythe *et al*., Ester-mediated amide bond formation driven by wet–dry cycles: A possible path to polypeptides on the prebiotic earth. *Angew. Chem. Int. Ed.* **54**, 9871–9875 (2015).
68. M. Rodriguez-Garcia *et al*., Formation of oligopeptides in high yield under simple programmable conditions. *Nat. Commun.* **6**, 8385 (2015).
69. Ei. Imai, H. Honda, K. Hatori, A. Brack, K. Matsuno, Elongation of oligopeptides in a simulated submarine hydrothermal system. *Science* **283**, 831–833 (1999).
70. K. A. Dill *et al*., Principles of protein folding–A perspective from simple exact models. *Prot. Sci.* **4**, 561–602 (1995).
71. R. Koga *et al*., Robust folding of a de novo designed ideal protein even with most of the core mutated to valine. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 31149–31156 (2020).
72. B. Yoo, K. Kirshenbaum, Peptoid architectures: Elaboration, actuation, and application. *Curr. Opin. Chem. Biol.* **12**, 714–721 (2008).
73. A. R. Davidson, R. T. Sauer, Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 2146–2150 (1994).
74. C. Chiarabelli *et al*., Investigation of de novo totally random biosequences. Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display. *Chem. Biodiver.* **3**, 840–859 (2006).
75. A. D. Keefe, J. W. Szostak, Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
76. B. C. Gorske, J. R. Stringer, B. L. Bastian, S. A. Fowler, H. E. Blackwell, New strategies for the design of folded peptoids revealed by a survey of noncovalent interactions in model systems. *J. Am. Chem. Soc.* **131**, 16555–16567 (2009).
77. M. M. Krishna, L. Hoang, Y. Lin, S. W. Englander, Hydrogen exchange methods to study protein folding. *Methods* **34**, 51–64 (2004).
78. R. Li, C. Woodward, The hydrogen exchange core and protein folding. *Prot. Sci.* **8**, 1571–1590 (1999).
79. M. Makarov *et al*., Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. *J. Am. Chem. Soc.* **145**, 5320–5329 (2023).
80. V. Tretyachenko *et al*., Modern and prebiotic amino acids support distinct structural profiles in proteins. *Open Biol.* **12**, 220040 (2022).
81. A. Rathore, T. F. Martinez, Q. Chu, A. Saghatelian, Small, but mighty? Searching for human microproteins and their potential for understanding health and disease. *Expert Rev. Proteom.* **15**, 963–965 (2018).
82. D. M. Anderson *et al*., Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Sci. Sig.* **9**, ra119 (2016).
83. T. F. Martinez *et al*., Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. *Cell Metab.* **35**, 166–183 (2023).
84. X. Cao *et al*., Nascent alt-protein chemoproteomics reveals a pre-60s assembly checkpoint inhibitor. *Nat. Chem. Biol.* **18**, 643–651 (2022).
85. M. G. Durrant, A. S. Bhatt, Automated prediction and annotation of small open reading frames in microbial genomes. *Cell Host Microbe* **29**, 121–131 (2021).
86. N. G. D'Lima *et al*., A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **13**, 174–180 (2017).
87. J. Z. Huang *et al*., A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. *Mol. Cell* **68**, 171–184 (2017).
88. J. L. Nieto-Torres, C. Verdiá-Báguena, C. Castaño-Rodriguez, V. M. Aguilella, L. Enjuanes, Relevance of viroporin ion channel activity on viral replication and pathogenesis. *Viruses* **7**, 3552–3573 (2015).
89. S. Ray *et al*., The mlpt/Ubr3/Svb module comprises an ancient developmental switch for embryonic patterning. *eLife* **8**, e39748 (2019).
90. S. A. Slavoff, J. Heo, B. A. Budnik, L. A. Hanakahi, A. Saghatelian, A human short open reading frame (SORF)-encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* **289**, 10950–10957 (2014).
91. R. Shibue *et al*., Comprehensive reduction of amino acid set in a protein suggests the importance of prebiotic amino acids for stable proteins. *Sci. Rep.* **8**, 1227 (2018).
92. M. Makarov *et al*., Enzyme catalysis prior to aromatic residues: Reverse engineering of a dephospho-CoA kinase. *Prot. Sci.* **30**, 1022–1034 (2021).
93. D. S. Riddle *et al*., Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805–809 (1997).
94. M. Kimura, S. Akanuma, Reconstruction and characterization of thermally stable and catalytically active proteins comprising an alphabet of $\sim$ 13 amino acids. *J. Mol. Evol.* **88**, 372–381 (2020).
95. L. M. Longo, C. A. Tenorio, O. S. Kumru, C. R. Middaugh, M. Blaber, A single aromatic core mutation converts a designed "primitive" protein from halophile to mesophile folding. *Prot. Sci.* **24**, 27–37 (2015).
96. S. Yagi *et al*., Seven amino acid types suffice to create the core fold of RNA polymerase. *J. Am. Chem. Soc.* **143**, 15998–16006 (2021).
97. D. Despotović *et al*., Polyamines mediate folding of primordial hyperacidic helical proteins. *Biochemistry* **59**, 4456–4462 (2020).
98. C. S. Foden *et al*., Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water. *Science* **370**, 865–869 (2020).
99. Y. Takahashi, H. Mihara, Construction of a chemically and conformationally self-replicating system of amyloid-like fibrils. *Bioorg. Med. Chem.* **12**, 693–699 (2004).
100. S. K. Rout, M. P. Friedmann, R. Riek, J. Greenwald, A prebiotic template-directed peptide synthesis based on amyloids. *Nat. Commun.* **9**, 234 (2018).
101. F. M. Menger, F. Nome, Interaction vs preorganization in enzyme catalysis. A dispute that calls for resolution. *ACS Chem. Biol.* **14**, 1386–1392 (2019).
102. M. I. Page, W. P. Jencks, Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proc. Natl. Acad. Sci. U.S.A.* **68**, 1678–1683 (1971).
103. E. Y. Lau, K. Kahn, P. A. Bash, T. C. Bruice, The importance of reactant positioning in enzyme catalysis: A hybrid quantum mechanics/molecular mechanics study of a haloalkane dehalogenase. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 9937–9942 (2000).
104. K. A. Dill, Theory for the folding and stability of globular proteins. *Biochemistry* **24**, 1501–1509 (1985).
105. K. Ghosh, K. A. Dill, Computing protein stabilities from their chain lengths. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10649–10654 (2009).
106. S. Bromberg, K. A. Dill, Side-chain entropy and packing in proteins. *Prot. Sci.* **3**, 997–1009 (1994).
107. C. D. Kocher, K. A. Dill, The prebiotic emergence of biological evolution. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2311.13650 (Accessed 22 November 2023).
108. F. J. Dyson, A model for the origin of life. *J. Mol. Evol.* **18**, 344–350 (1982).
109. F. Dyson, *Origins of Life* (Cambridge University Press, 1999).
110. M. Wu, P. G. Higgs, The origin of life is a spatially localized stochastic transition. *Biol. Dir.* **7**, 42 (2012).
111. J. A. Shay, C. Huynh, P. G. Higgs, The origin and spread of a cooperative replicase in a prebiotic chemical system. *J. Theor. Biol.* **364**, 249–259 (2015).
112. P. G. Higgs, R. E. Pudritz, A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **9**, 483–490 (2009).
113. T. Bose *et al*., Origin of life: Protoribosome forms peptide bonds and links RNA and protein dominated worlds. *Nucl. Acids Res.* **50**, 1815–1828 (2022).
114. A. Pross, The evolutionary origin of biological function and complexity. *J. Mol. Evol.* **76**, 185–191 (2013).