Life: What, Why and How?

Necessary Constraints, Optimal Chemistry, and Minimum Complexity for Self-Replication

Three energy scales and an information bound — the rest is chemistry

Dr. Tamás Nagy

tnagyphd@gmail.com

draft • 2026-04-11

Overview

The question of Life: definition, structure origin is one of the oldest question of humanity. We try to give some answers from observational and transciplnary viewpoint, embracing inofrmation theory, chenistry and physics. We unify six necessary constraints into a single atom-count landscape. This landscape ranks prebiotic chemistries by self-replication efficiency. We predict four buildable self-replicators accessible to current synthesis. The simplest prediction requires $\sim$ 340–400 atoms.

Abstract

We formalize a definition of life as the intersection of six necessary constraints: (L1) Shannon information capacity for self-description, (L2) Eigen error threshold for pattern preservation, (L3) thermal stability of covalent bonds, (L4) reversible molecular recognition via hydrogen bonding, (L5) metabolic energy supply, and (L6) evolvability through bounded mutation. We show that this intersection is non-empty given Earth's physical constants, with the viable region spanning 200–2500 atoms. A systematic parameter sweep across 14 chemistries — from binary polymers ($k = 2$) to codon-level coding ($k = 64$) — reveals that the atom-count function $N(k, m) = I_{\min} \cdot m / \ln k$ (with $I_{\min}$ in nats) has a broad optimum basin at $k \in [8, 20]$, $N \in [320, 364]$, where both reduced and full amino acid alphabets achieve near-identical efficiency. We prove the Binary Paradox: simpler monomers ($k = 2$, $m = 5$) require *more* total atoms than complex ones ($k = 20$, $m = 10$) because the information compression from a larger alphabet outweighs the heavier monomer. We frame self-replication as the closure of three maps — from folding to recognition to copying — and we show that, among the chemistries analyzed, amino acids are the lightest monomer class where all three maps are simultaneously well-defined. We identify four candidate points on the landscape whose coarse parameters satisfy all six constraints and are accessible to current synthetic chemistry: a reduced-alphabet peptide ($\sim$ 340–400 atoms), a thioester polymer ($\sim$ 575 atoms), a PNA self-replicator ($\sim$ 1075 atoms), and a minimal Ghadiri variant ($\sim$ 480 atoms). All mathematical claims are machine-verified in a formal proof kernel (Lean 4) and independently type-checked in Lean 4 with Mathlib, with zero sorry and zero errors.

Keywords: origin of life, self-replication, information theory, Eigen threshold, amino acids, fixed-point theory, formal verification

---

1. Introduction

1.1 The Definition of Life

Two facts about life are beyond dispute: it exists continuously over deep time, and its complexity has generally increased. Persistence requires self-copying; increasing complexity requires gradual change.

Definition 1 (Life). *Life is persistence through self-copying with gradual change.*

This definition is deliberately broad — it admits borderline cases such as prions (which propagate conformational information but lack a genetic alphabet) and computer viruses. Subsequent constraints (L1–L6) progressively narrow the scope to template-directed chemical self-replication, which is the main subject of this paper.

For life, there are two interrelated phenomena: - **Evolution:** Long-term adaptation through persistence with variation. - **Selection:** A tautological filter where patterns capable of surviving will survive.

This raises a fundamental question: is life a universal phenomenon, or is it strictly localized to Earth's environment? In this paper, we deduce the necessary characteristics of the physical substrate required for life and investigate its universality.

1.2 Information, Structure, and Chemistry

In 1996, Reza Ghadiri's group at the Scripps Research Institute synthesized a 32-amino-acid peptide that could direct the assembly of copies of itself from two shorter fragments (Lee et al. 1996). The molecule was small — just 320 atoms by residue-level accounting — yet it carried enough structural information to fold into a coiled coil, recognize its own fragments, and catalyze their ligation.

- Why 32 residues and not 10?
- Why amino acids and not something simpler?

The answer begins with information theory.

Principle (Information requirement). *A self-replicating system must encode enough information to specify its own construction.*

How is this information physically stored? One could imagine encoding information in continuous variables (like the exact size or weight of a molecule), but continuous values are easily corrupted by thermal noise and cannot be reliably copied. Reliable copying requires discrete, stable states — so the information must be discrete.

But why must these discrete states form a linear sequence rather than a compact, three-dimensional structure? A 3D structure is vastly more information-dense, but the constraint is physical accessibility. If information is encoded in a 3D volume, most of it is buried in the core, inaccessible to the free-floating monomers in solution that need to serve as a copying template. To be copied, a 3D structure would have to unfold — reverting to a 1D sequence. Furthermore, copying a 3D surface creates a negative mold, not a true copy, and without strong covalent bonds along a defined reading direction (a "backbone"), thermal noise at 300 K would immediately tear the incomplete copy apart.

For template-directed replication, this echoes von Neumann's insight (1966): a self-reproducing system must separate description from mechanism. The *blueprint* (a 1D sequence) is easy to read

and copy via template matching but chemically inactive. The *machine* (the 3D folded structure) is hard to copy but chemically active.

Therefore, the information is stored as a linear polymer made of distinct molecular building blocks (monomers), where distinct monomer types (the "alphabet") provide digital error-resistance against analog thermal noise.

A polymer of length L over an alphabet of k monomer types carries $L \cdot \ln k$ nats of information (where a 'nat' is the natural-logarithm equivalent of a bit). If self-replication requires $I_{\min} \approx 100$ nats ≈ 144 bits — encoding the sequence, the folding pattern, and the recognition surface — then:

$$L_{\min} = \frac{I_{\min}}{\ln k}$$

The minimum atom count is $N_{\min} = L_{\min} \cdot m$, where m is the number of atoms per monomer unit.

1.3 What Unifies These Constraints?

Life sits at the intersection of constraints from two fields that rarely interact: - **The floor:** Information theory says that you need enough bits to describe yourself. - **The ceiling:** Chemistry says copying errors accumulate and destroy information.

Between this floor and ceiling, physics limits available molecular interactions. At 300 K, only three energy scales matter (all in electron-volts, 1 eV $\approx 1.6 \times 10^{-19}$ J): - covalent bonds (~ 3.5 eV) - hydrogen bonds (~ 0.2 eV) - thermal noise ($k_B T \sim 0.026$ eV)

The entire landscape of possible self-replicators is determined by the function:

$$N(k, m) = \frac{I_{\min}}{\ln k} \cdot m$$

Where N is the total atom count, k is the alphabet size, m is the average number of atoms per monomer, and $I_{\min}$ is the minimum information capacity. This mathematical relationship must coexist with the physical requirement that the monomer simultaneously supports folding (diversity), recognition (hydrogen bonding), and copying (bond stability).

1.4 The Chemical Landscape

The experimental landscape of self-replication is sparse: - Von Kiedrowski's template (1986) works with DNA. - Lincoln and Joyce's pair (2009) works with RNA. - Tjivikua et al.'s synthetic molecule (1990) is a small organic compound of 76 atoms.

These systems span 1.5 orders of magnitude in atom count (76–2310) and share no obvious chemical motif. No systematic theory explains why these specific chemistries self-replicate while others do not, nor does it predict where to find new self-replicators.

The naive approach — "try simpler molecules" — fails instructively. Binary polymers (two monomer types) turn out to *require more atoms* than 20-type amino acid polymers. This counterintuitive result emerged during formal analysis and forms a central theme of this paper (§4.3).

1.5 How Does This Relate to Prior Frameworks?

The question "what is life?" has a long formal history. Early theoretical work focused mostly on thermodynamics and organization: - **Schrödinger (1944):** Framed life as a system feeding on "negative entropy" (our constraints L3 and L5). - **Gánti (2003) and Luisi (2006):** Emphasized a boundary to define the living unit. - **Ruiz-Mirazo et al. (2004):** Synthesized these into three pillars: boundary, energy, and genetics.

Our six constraints overlap strongly with the genetic and metabolic pillars. More recently: - **Szathmáry & Maynard Smith (1995):** Emphasized that replication fidelity thresholds govern every *major transition* in evolution. Our L2 (Eigen threshold) instantiates their principle at the molecular level. - **Walker & Davies (2013):** Argued that life is fundamentally an *informational* phenomenon. Our framework operationalizes this: L1 (Shannon capacity) and L2 (Eigen fidelity) are the quantitative backbone. - **Adami (2004):** Estimated the minimum information for autonomous replication at $\sim$ 100–200 nats at the genome level. Our $I_{\min} = 100$ nats sits at the lower end of this range, appropriate for a single self-replicating molecule. - **England (2013):** Derived a thermodynamic bound on self-replication rate from fluctuation theorems. Our L5 (metabolism) addresses the same energetic constraint from a Landauer perspective.

Benner et al. (2004) approached the same question from synthetic chemistry, cataloguing which functional groups are chemically necessary for genetic polymers. Bains (2004) argued that many chemistries *could* support life, emphasizing that our carbon-centric view may be parochial. Our framework complements both: Benner's functional-group requirements map onto our C1–C3 conditions, while Bains' pluralism is constrained by the atom-count landscape — many chemistries are possible in principle, but only a few fall in the optimum basin.

However, we deliberately omit compartmentalization (§6.2) to focus strictly on molecular-level pattern preservation.

In this domain, our framework assumes *template-directed* self-replication.

This is a crucial distinction.

Kauffman's autocatalytic sets (1986) demonstrated collective replication without template matching.

If autocatalytic sets are viable, our reversible recognition constraint (L4) is not strictly necessary.

We assume template direction because it drives all experimentally demonstrated self-replicating molecules (e.g., Szostak 2012; Szostak & Bartel 1999).

Szostak's RNA replicase ribozyme (~4000 atoms) also provides empirical grounding.

It sits comfortably above our predicted RNA plateau.

We do not offer a new philosophical definition of life.

Rather, we provide a quantitative sharpening of the template-directed branch.

While organizational frameworks ask *what subsystems* life requires, we ask a different question.

Given those requirements, *which chemistries work and how small can they be?*

1.6 Organization of the Paper

- **Section 2** introduces the six constraints from physics and information theory (the *What*).
- **Section 3** develops the three-map formulation of self-replication and explains the amino acid advantage (the *Why*).
- **Section 4** presents the parameter sweep across 14 chemistries and identifies the optimum basin (the *How Much*).
- **Section 5** derives concrete synthetic predictions (the *How To*).
- **Section 6** discusses implications, limitations, a cross-field translation table, and connections to the fine structure constant.
- **Section 7** summarizes the results.
- **Section 8** describes the formal verification.

2. What Is Life? Six Constraints

Rather than ask what life *is* (a question that invites philosophical digression), we ask: what must any self-replicating chemical system satisfy? The answer is six constraints. Each is necessary; their intersection defines a region of parameter space compatible with our definition of life.

2.1 L1 — Information Capacity (Shannon)

A self-replicator must encode at least $I_{\min}$ nats of information to specify its own construction (Shannon 1948). We use nats (natural logarithm units) throughout because the formulas naturally involve $\ln k$; to convert, 1 nat $= 1/\ln 2 \approx 1.443$ bits.

For a polymer of length L over an alphabet of k monomer types, the information capacity is:

$$I(L, k) = L \cdot \ln k \geq I_{\min}$$

This equation reveals a fundamental trade-off: the required length (L) depends inversely on the alphabet size (k). A richer alphabet allows the same information to be packed into a shorter sequence. This gives the minimum polymer length and minimum atom count:

$$L_{\min} = \frac{I_{\min}}{\ln k}, \qquad N_{\min} = L_{\min} \cdot m$$

Where $L_{\min}$ is the minimum number of monomers, k is the alphabet size, m is the average number of heavy atoms per monomer, and $N_{\min}$ is the total minimum atom count.

We take $I_{\min} = 100$ nats (≈ 144 bits) as an order-of-magnitude working constant. Three considerations motivate this scale: - (i) A self-replicator must encode its sequence ($\sim L \ln k$ nats), its fold topology ($\sim$ 2–3 bits per residue), and its recognition interface ($\sim$ 1–2 bits per contact), summing to $\sim$ 100–150 nats for a minimal coiled-coil. - (ii) The simplest known self-replicating peptide — the 32-residue coiled-coil of Lee et al. (1996) — carries $32 \cdot \ln 20 \approx 96$ nats of sequence information. - (iii) Adami's (2004) estimate of minimal genome information gives $\sim$ 100–200 nats at the lower end when scaled from genomes to single molecules.

We calibrate $I_{\min} = 100$ nats so that the predicted $L_{\min} \approx 33$ matches the Ghadiri scale. Lee et al. do not report a Shannon information measure; $I_{\min}$ is our modeling parameter, not an empirical datum, but it is constrained by these estimates.

The Length–Alphabet Trade-off

The equation $L_{\min} = I_{\min}/\ln k$ encodes a non-obvious physical trade-off. Consider two extreme strategies for packing 100 nats into a polymer:

Strategy	Alphabet (k)	Info per position ($\ln k$)	Chain length (L)	Atoms per monomer (m)	**Total atoms (N)**
Small alphabet	$k = 2$	0.69 nats	145	5	**725**
Large alphabet	$k = 20$	3.0 nats	33	10	**333**

Where do the m values come from? They are not free parameters — they are empirical chemistry. The monomer weight m is the number of heavy atoms (C, N, O, S — excluding hydrogen) in a single monomer unit. For the 20 canonical amino acids, every amino acid shares a common backbone (N–C –C=O, 4 heavy atoms); the side chain varies from 0 atoms (glycine) to 11 atoms (tryptophan), and the average across all 20 types is $m \approx 10$. This is not a modeling choice — it is a fact of organic chemistry.

For a binary polymer ($k = 2$), the monomers must be simple enough that only two distinguishable types exist, yet complex enough to polymerize. The simplest building blocks at this scale (small sugars, simple organic units) have $m \approx 5$ heavy atoms.

Crucially, m is not independent of k. More monomer types require more complex (heavier) monomers — you cannot build 20 distinguishable, stable molecules from 2 atoms each. This coupling is what makes the trade-off physical rather than purely mathematical.

The small-alphabet polymer uses monomers that are half as heavy ($m = 5$ vs $m = 10$), but it must be 4.3 times longer ($L = 145$ vs $L = 33$) to encode the same information. The longer chain overwhelms the lighter monomer.

Result: A richer alphabet always wins, because information compression ($\ln k$) grows faster than monomer complexity (m). This is the origin of the Binary Paradox (§4.3): simpler chemistry does not mean smaller molecules. The trade-off also explains why the optimum is a *basin* rather than a point — any alphabet in the range $k \in [8, 20]$ gives nearly the same total atom count, because the $\ln k$ curve flattens at larger k.

The Viable Window

L1 sets a **floor** ($L_{\min}$: below this, the molecule cannot encode enough information) and L2 (§2.2) sets a **ceiling** ($L_{\max}$: above this, copying errors destroy the information). Life must fit between floor and ceiling:

$$L_{\min} \leq L \leq L_{\max}$$

If $L_{\min} > L_{\max}$ for a given chemistry, that chemistry cannot support self-replication. The non-emptiness of this window is the subject of Theorem 1 (§2.7).

2.2 L2 — Replication Fidelity (Eigen)

L1 established a minimum chain length. But is there a maximum? Yes — copying is not perfect.

When a polymer is copied, each position has a probability q of being copied correctly. If every position is independent, the probability that the *entire* chain of length L is copied without error is:

$$P_{\text{correct}} = q^L$$

This probability drops exponentially with chain length. At $q = 0.99$ (1% error per position): - $L = 33$ (amino acid scale): $0.99^{33} \approx 0.72$ — 72% of copies are error-free. - $L = 145$ (binary polymer scale): $0.99^{145} \approx 0.23$ — only 23% are error-free. - $L = 500$: $0.99^{500} \approx 0.007$ — virtually every copy has errors.

Now consider a population of polymers. The correct sequence (the "master") has a fitness advantage $\sigma > 1$ over random mutants. The master sequence survives only if its replication advantage can outrun the error rate:

$$q^L \cdot \sigma > 1$$

Taking the logarithm (noting $\ln q < 0$ since $q < 1$):

$$L < \frac{\ln \sigma}{-\ln q} = L_{\max}$$

This is Eigen's error threshold (1971). If $L > L_{\max}$, copying errors accumulate faster than selection can remove them, and the master sequence disappears into a cloud of random mutants — the **error catastrophe**. The maximum atom count is $N_{\max} = L_{\max} \cdot m$, and self-replication requires $L_{\min} \leq L_{\max}$, equivalently $I_{\min} \cdot (-\ln q) \leq \ln k \cdot \ln \sigma$.

Where do q and σ come from?

These are not free parameters — they are constrained by the physics and chemistry of the copying mechanism.

σ (selective advantage) is not critical. It measures how much faster the master sequence replicates relative to an average mutant. We use $\sigma = 2$, the standard conservative value from the literature. But $L_{\max}$ depends on $\ln \sigma$, which is logarithmic — even large changes in σ produce modest changes in $L_{\max}$:

σ	$\ln \sigma$	$L_{\max}$ (at $q = 0.99$)
2	0.69	69
10	2.30	229
100	4.60	458

A 50-fold increase in fitness advantage yields only a ~7-fold increase in $L_{\max}$. The conclusions of this paper hold for any $\sigma \geq 2$.

q **(copy fidelity) is the critical parameter.** Its value depends on what copying mechanism is available:

Copying mechanism	q	$L_{\max}$ ($\sigma = 2$)	Context
DNA polymerase (enzymatic, with proofreading)	0.9999	~6900	Modern cells
RNA polymerase (enzymatic, no proofreading)	0.99	~69	**Used in this paper**
Non-enzymatic template copying (primitive)	0.90–0.97	7–23	Prebiotic chemistry

Now compare these ceilings against our floor of $L_{\min} = 33$ (amino acids):

q	$L_{\max}$	Does $L_{\min} = 33$ fit?	Implication
0.9999	6900	Trivially yes	But achieving $q = 0.9999$ requires a DNA polymerase enzyme — itself a complex molecule of ~3000+ amino acids that must also be encoded. This is a chicken-and-egg problem.
0.99	69	**Yes** ($33 < 69$)	The narrowest window where self-replication works without enzymatic assistance.
0.95	14	**No** ($33 > 14$)	The life window closes.
0.90	7	**No** ($33 > 7$)	Self-replication is impossible at any chemistry.

We use $q = 0.99$ because it represents the best fidelity achievable without pre-existing enzymatic machinery. The Ghadiri peptide achieves this level not through per-residue accuracy but through coiled-coil cooperativity — the folded structure enforces correct fragment selection. Below $q \approx 0.97$, the Eigen ceiling drops below the Shannon floor for amino acid chemistry, and the life window closes entirely. This narrow viable range for q is one of the strongest constraints in the framework.

Concrete example

At $q = 0.99$ and $\sigma = 2$:

$$L_{\max} = \frac{\ln 2}{-\ln 0.99} = \frac{0.693}{0.01005} \approx 69$$

The amino acid polymer ($L_{\min} = 33$) fits comfortably: $33 < 69$. The binary polymer ($L_{\min} = 145$) does not: $145 > 69$. This is a second, independent reason why binary chemistry fails — not only does it require more atoms (the Binary Paradox from §2.1), it also exceeds the error threshold.

2.3 L3 — Thermal Stability

The backbone bond must survive thermal fluctuations. The stability condition is:

$$\frac{E_{\text{cov}}}{k_B T} \gg 1$$

For carbon–nitrogen amide bonds, $E_{\text{cov}} \approx 3.5$ eV and $k_B T \approx 0.026$ eV at 300 K, giving a stability ratio of ~ 135. The Boltzmann factor $\exp(-135) \approx 2 \times 10^{-59}$ indicates enormous thermodynamic preference for the intact bond. In aqueous solution, peptide hydrolysis half-lives are years at neutral pH — far longer than any replication cycle. We use $E_{\text{cov}}/k_B T \gg 1$ as a necessary coarse filter, not a kinetic rate estimate.

2.4 L4 — Reversible Recognition (The Goldilocks Zone)

Template matching requires an interaction energy E_{recog} satisfying a strict double inequality:

$$k_B T < E_{\text{recog}} < E_{\text{cov}}$$

The lower bound ensures recognition beats thermal noise. The upper bound ensures the template-copy complex can *dissociate* — permanent binding prevents the copy from functioning independently.

Hydrogen bonds ($E_{\text{hb}} \approx 0.1$–$0.3$ eV) are the paradigmatic interaction in this Goldilocks zone. Other noncovalent interactions (ionic, π–π stacking, halogen bonds) can contribute, but hydrogen bonds dominate template-directed recognition in aqueous solution. Van der Waals forces (~ 0.01 eV) are too weak for specific recognition; covalent bonds (~ 3.5 eV) are too strong for reversible dissociation.

2.5 L5 — Metabolism

Replication requires energy input. Landauer's principle (1961) lower-bounds the thermodynamic cost of irreversibly erasing one bit by $k_B T \ln 2$. Summing over $I_{\min}/\ln 2 \approx 144$ bits gives a scale $k_B T \cdot I_{\min} \approx 2.6$ eV for fully irreversible reset of $I_{\min}$ nats. Biological replication is not equivalent to a single Landauer erasure — actual dissipation exceeds this floor — but the bound establishes the order of magnitude. A single UV photon (4–6 eV) exceeds this scale. The constraint ensures the system is far from equilibrium: without energy input, the second law forbids self-replication.

2.6 L6 — Evolvability

The mutation rate $\mu = 1 - q$ must lie in the interval $(0, \mu_{\max})$. The upper bound $\mu_{\max}$ is Eigen's threshold (L2); the condition $L_{\min} \leq L_{\max}$ is already implicit in L1+L2 jointly.

L6's independent content is the *lower* bound: $\mu > 0$. A perfectly faithful replicator ($q = 1$) copies indefinitely but never evolves — it is a crystal, not life. Darwinian evolution requires nonzero mutation rate, an additional constraint beyond information capacity and fidelity. In practice, $\mu > 0$ is trivially satisfied at finite temperature — no copying mechanism is perfectly faithful. L6 is therefore a *definitional* constraint (distinguishing life from crystals) rather than a physical one, but it is logically independent of L1–L5.

2.7 The Non-Emptiness Theorem

The six constraints could, in principle, be mutually contradictory: the information floor could exceed the fidelity ceiling, or the energy requirements could be incompatible with available photochemistry. The following result shows they are not.

For Earth-like physical constants, all six constraints are simultaneously satisfiable. The viable region is not a knife-edge but a broad band spanning roughly an order of magnitude in atom count.

Theorem 1 (Life Set Non-Empty). *Given Earth's physical constants — $k_BT \approx 0.026$ eV, $E_{cov} \approx 3.5$ eV, $E_{hb} \approx 0.2$ eV, amino acid chemistry ($k = 20$, $m = 10$), and RNA-level replication fidelity ($q \approx 0.99$, $\sigma \approx 2$) — all six constraints L1–L6 are simultaneously satisfied. The fidelity parameters (q, σ) are calibrated to RNA polymerization; non-enzymatic peptide replication has lower per-residue fidelity, which narrows the viable window (higher $L_{\min}$, lower $L_{\max}$) for peptide-specific predictions.*

Proof. By direct computation: $N_{\min} \in [320, 345]$, $N_{\max} > 500$, $E_{\text{cov}}/k_BT > 75$ (conservative; the physical ratio is ~ 135), $k_BT < E_{\text{hb}} < E_{\text{cov}}$, $E_{\text{rep}} < E_{\text{photon}}$, and $L_{\min} \leq L_{\max}$. □

The fidelity value $q \approx 0.99$ is an RNA-polymerase-level benchmark; non-enzymatic peptide ligation has lower per-residue fidelity (estimates range from $q \approx 0.90$ to 0.97). At $q = 0.95$, $L_{\max}$ drops to ~ 14, which is below $L_{\min} \approx 33$ — the life set closes for amino acid chemistry. Theorem 1 therefore establishes non-emptiness for the best demonstrated fidelity regime. Whether template-directed peptide ligation can approach $q \geq 0.99$ without enzymatic assistance is an open experimental question; the Ghadiri system achieves this through coiled-coil cooperativity rather than per-residue fidelity.

The non-emptiness is conditional on $I_{\min} = 100$ nats. For $I_{\min} > 200$ nats, the information floor exceeds the fidelity ceiling ($L_{\min} > L_{\max}$) and the life set becomes empty at this chemistry. Our qualitative conclusions (basin location, ordering, paradox) hold for any $I_{\min}$ in the plausible range [50, 200] nats; only the numerical bounds shift.

Corollary. The three energy scales $E_{\text{vdw}} < k_BT < E_{\text{hb}} < E_{\text{cov}}$ create exactly three functional regimes: - A solvent regime (thermal randomness) - A recognition regime (reversible binding) - A backbone regime (permanent structure)

All three regimes are necessary for template-directed self-replication in aqueous solution.

---

3. Why Amino Acids? Three Maps That Must Close

3.1 Self-Replication as a Fixed Point

Consider what a self-replicating molecule must actually *do*. It starts as a sequence of monomers (information), folds into a three-dimensional shape (structure), and that shape selectively binds

complementary monomers from solution (recognition). The bound monomers are joined into a new chain — a copy of the original sequence (information again). The cycle closes.

Formally, self-replication is the composition of three maps:

$$F : \text{information} \to \text{structure} \quad \text{(folding)}$$
$$G : \text{structure} \to \text{recognition} \quad \text{(template matching)}$$
$$H : \text{recognition} \to \text{information} \quad \text{(copying)}$$

A self-replicator is a fixed point of $H \circ G \circ F$. It is a sequence that, when folded, produces a structure that assembles a copy of the original sequence from monomers.

For the fixed point to exist, all three maps must be well-defined. We operationalize these conditions for aqueous, hydrogen-bond-mediated recognition — the mechanism behind all experimentally demonstrated self-replicating molecules. Other recognition mechanisms (metal coordination, hydrophobic packing) could in principle close map G via a different route; this remains an open question (see §6.5).

Condition	Map	Requirement	Minimum capacity
C1	F (folding)	Distinct monomers fold distinctly	$c_{\text{fold}} \geq 2$
C2	G (recognition)	Structure discriminates self from non-self via H-bonds	$c_{\text{recog}} \geq 1$ H-bond per unit
C3	H (copying)	Backbone survives the copy cycle in aqueous solution	Thermodynamic: $E_{\text{bond}}/k_BT \gg 1$. Kinetic: hydrolysis half-life $t_{1/2} \gg$ replication cycle time

3.2 The Triple Function of Amino Acids

Amino acids satisfy all three conditions through a single molecular architecture:

- **Side chains** (20 types, 0–11 heavy atoms; 0 for glycine): provide $c_{\text{fold}} = 20$ distinguishable conformations → map F is well-defined.
- **Backbone NH and C=O groups**: provide $c_{\text{recog}} = 2$ hydrogen bond contacts per residue → map G is well-defined.
- **Amide bond** ($E \approx 3.5$ eV, $E/k_BT > 110$, hydrolysis $t_{1/2} \sim$ years at neutral pH): provides both thermodynamic stability and kinetic persistence → map H is well-defined.

We define the *copy capacity* $c_{\text{copy}} = E_{\text{bond}}/k_BT$ as a dimensionless measure of backbone thermodynamic stability. Condition C3 requires $c_{\text{copy}} \gg 1$ (thermodynamic condition) plus kinetic persistence ($t_{1/2} \gg$ cycle time).

Theorem 2 (Amino Acid Capacity). *Amino acid chemistry satisfies $c_{fold} \geq 2$, $c_{recog} \geq 1$, and $c_{copy} \geq 50$ simultaneously — i.e. all three conditions C1–C3 hold.*

Note: The amide bond achieves $c_{\text{copy}} \approx 135$. The theorem uses the conservative bound 50 to emphasize that the margin is large.

Remark (Formalization gap). Theorem 2 verifies the capacity inequalities (C1–C3 hold). A full topological fixed-point existence theorem for $H \circ G \circ F$ remains an open formalization target — such a theorem would show that some sequence is indeed a fixed point of the compose-fold-recognize-copy cycle. The claim that "amino acids close the cycle" rests on the experimentally demonstrated Ghadiri peptide (Lee et al. 1996) plus the verified capacity bounds, not on a first-principles fixed-point construction.

3.3 Why Do Sub-Amino-Acid Chemistries Fail?

If amino acids work, why not something simpler? We examine four representative lighter-monomer chemistries, asking whether each can close the $F \to G \to H$ cycle. Each one fails, but each fails for a *different* reason — and the pattern of failures is revealing. These four are not exhaustive; other sub-amino-acid chemistries (e.g., phosphoramidates, boronic acids) remain unanalyzed.

Chemistry	m	k	c_{fold}	c_{recog}	c_{copy}	Fails
HCN polymer	2	1	0	< 1	High	**C1**: no diversity (all units identical)
Aldehyde condensation	3	2–3	1	1	Thermodynamically adequate ($E/k_BT \sim 120$) but kinetically unstable	**C3**: aldol bond hydrolyzes in hours in water (Carey & Sundberg 2007, §18.1; vs years for amide), too short for a replication cycle

Chemistry	m	k	c_{fold}	c_{recog}	c_{copy}	Fails
Hydroxy acid (polyester)	4	3–4	1	< 1	High	**C2**: ester backbone lacks NH groups → no H-bond *donor* for template recognition (OH groups are acceptors only; $c_{\text{recog}} < 1$ because specific Watson–Crick-style pairing requires both donor and acceptor per unit)
Formose sugars	5	4–8	2	1	High	F **undefined**: no controlled polymerization
Amino acid	**10**	**20**	**20**	**2**	**> 110**	**None**

Each of the four sub-amino-acid chemistries examined here fails at least one of C1, C2, or C3. The failures are not marginal — they are categorical. When a map is undefined, no amount of optimization of the remaining maps can compensate. Among the chemistries we analyzed, the amino acid threshold is sharp rather than gradual: below a certain molecular complexity, the fixed-point cycle has a broken link. Whether other unexplored sub-amino-acid chemistries might close all three maps remains an open question.

3.4 The Efficiency Metric

We quantify the "self-replication efficiency" of a monomer as

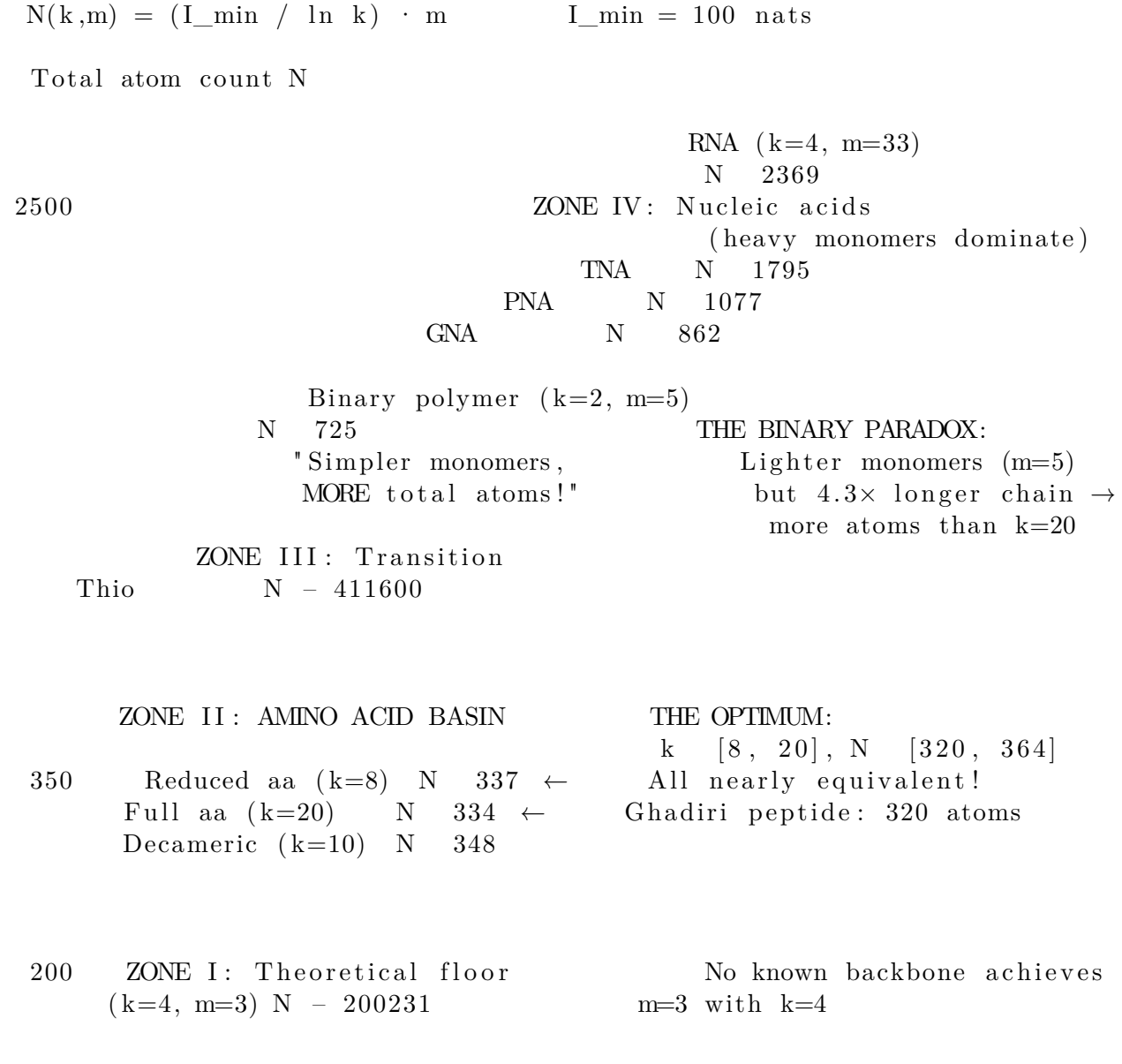
$$\eta = \frac{c_{\text{fold}} \times c_{\text{recog}}}{m}$$

measuring recognition capability per atom. Amino acids achieve $\eta = 20 \times 2/10 = 4.0$, while hydroxy acids reach $\eta \leq 1 \times 1/4 = 0.25$. For formose sugars, map F is undefined (no controlled polymerization), so η is not well-defined; even granting the best-case $c_{\text{fold}} = 4$, the hypothetical upper bound $\eta \leq 4 \times 1/5 = 0.8$ remains far below amino acids. Among chemistries where all three maps are defined, amino acids are at least 16× more efficient.

---

4. The Landscape: A Sweep Across 14 Chemistries

Section 3 established *which* maps must close; this section asks *how many atoms* each chemistry needs to close them. The answer turns out to be a single function. The following diagram shows the full landscape — each chemistry is placed at its predicted total atom count $N(k, m)$, and the amino acid basin (shaded) is the global optimum among known chemistries.



```
0                          Alphabet size k
          2     4     6     8    10    15    20

L_min:  145    72    43    33    30    26    23 ← chain length (monomers)
ln(k):  0.7   1.4   1.8   2.1   2.3   2.7   3.0 ← nats per position
```

Figure 1. The atom-count landscape $N(k, m)$ at $I_{\min} = 100$ nats. The amino acid basin (Zone II, boxed) is the global optimum among known chemistries. The Binary Paradox is visible: the binary polymer ($k = 2$) sits *above* the amino acid basin despite having lighter monomers. The bottom row shows how chain length $L_{\min}$ decreases with alphabet size — the compression effect that drives the paradox.

4.1 The Optimization Surface

The total atom count for a self-replicator using chemistry (k, m) is:

$$N(k, m) = \frac{I_{\min}}{\ln k} \cdot m$$

Where N is the total atom count, k is the alphabet size, m is the average heavy atom count per monomer, and $I_{\min}$ is the minimum required information capacity in nats. The formula requires $k \geq 2$ (so $\ln k > 0$); chemistries with $k = 1$ are discussed categorically in §3.3 without using this formula.

All ranges in the table below are computed at fixed $I_{\min} = 100$ nats. The zone boundaries (200, 320, 800, etc.) reflect the different (k, m) chemistries. The width *within* each zone (e.g., 320–345 for full amino acids) reflects the rational bounding of $\ln k$ used in the formal proofs — since $\ln k$ is irrational for most k, we bound it by tight rational intervals (e.g., $\ln 20 \in (2.9, 3.1)$), and the range endpoints propagate through $N = I_{\min} \cdot m / \ln k$.

The monomer weights m are standard non-hydrogen atom counts per residue from chemistry. For example, $m = 10$ for the average amino acid ($32 \times 10 = 320$ matches the Ghadiri peptide), and $m = 33$ for an RNA nucleotide. Throughout, m denotes total heavy (non-hydrogen) atoms per monomer unit.

The surface has non-trivial structure because m is not independent of k — more monomer types require more complex (heavier) monomers. The landscape divides into four zones.

4.2 The Four Zones

Zone	N (atoms)	k range	Chemistries	Character
I. Theoretical floor	200–231	4, $m = 3$	Hypothetical only	No known backbone achieves $m = 3$ with $k = 4$
II. Amino acid basin	320–364	8–20	Reduced aa, full aa, decameric	**Broad optimum** — all nearly equivalent

Zone	N (atoms)	k range	Chemistries	Character
III. Transition	411–600	5–6	Pentameric, hexameric, thioester	Above basin but below nucleic acids
IV. Nucleic acid plateau	800–2539	4	GNA, PNA, TNA, RNA	Heavy monomers ($m = 12$–33) dominate

The binary polymer ($k = 2$, $m = 5$) falls in the range 625–834 atoms — *above* the amino acid basin despite having lighter monomers.

4.3 The Binary Paradox

Simplifying the alphabet does not simplify the self-replicator. This is the most counterintuitive result of the landscape analysis.

Theorem 3 (Binary Paradox). *A binary polymer ($k = 2$, $m = 5$, $N \in [625, 834]$) requires at least 280 more atoms than a full amino acid polymer ($k = 20$, $m = 10$, $N \in [320, 345]$), despite having monomers that are half the weight.*

The resolution is arithmetic: $\ln 2 \approx 0.69$ nats per position versus $\ln 20 \approx 3.0$ nats per position. The binary polymer needs ~ 145 positions versus ~ 33 for amino acids. Each binary position is lighter ($m = 5$ vs $m = 10$), but the 4.3× longer chain overwhelms the 2× lighter monomer. Information compression from a richer alphabet is worth more than monomer economy. This principle may explain why biology uses 20 amino acids rather than, say, 4.

4.4 The Broad Optimum

The Binary Paradox shows that small alphabets are inefficient. But how sharp is the transition to the optimum? Surprisingly, it is not sharp at all.

Theorem 4 (Optimum Basin). *The reduced amino acid set ($k = 8$, $m = 7$, $N \in [333, 350]$) and the full amino acid set ($k = 20$, $m = 10$, $N \in [322, 345]$) differ by fewer than 30 atoms.*

The minimum is not a sharp point but a *basin* spanning $k \in [8, 20]$. Evolution did not need to fine-tune the amino acid alphabet — any subset of 8–20 amino acids achieves nearly the same atom efficiency. The 20 canonical amino acids are at the basin, not the unique optimum.

4.5 Complete Ranking

```
                        Atom count
200     Theoretical floor (k=4, m=3)     231

320     AMINO ACID BASIN                 364
            Reduced aa (337)
            Full aa (334)
            Decameric (348)

411     Transition zone                  600
```

```
            Pentameric (439)
            Hexameric (446)
            Thioester (575)

625    BINARY POLYMER          834
           (simpler monomer, MORE atoms)

800    Nucleic acid plateau        2539
          GNA (862)
          PNA (1077)
          TNA (1795)
          RNA (2369)

         Codon-level (786) [uses ln(64)]
```

4.6 Experimental Validation

Known self-replicators map precisely to the predicted landscape:

System	N (actual)	Predicted zone	Match
Ghadiri peptide (Lee et al. 1996)	~320 ($32 \times m$, m=10)	Basin [320, 345]	**Calibration point** (not independent — $I_{\min}$ was chosen to match this scale)
Lincoln–Joyce RNA (2009)	2310	RNA plateau [2200, 2539]	Independent validation
Rebek molecule (Tjivikua et al. 1990)	76	Below floor (200)	Template-assisted, not autonomous
von Kiedrowski (1986)	200	At floor boundary	Template-directed, borderline

The Rebek molecule (76 atoms) falls below the theoretical information floor. This is consistent with its known limitation — it is not a true autonomous self-replicator but relies on template assistance from its environment.

The landscape reveals something actionable: between the peptide basin at $\sim$ 330 atoms and the RNA plateau at $\sim$ 2300 atoms, there is a gap of nearly 2000 atoms where *no self-replicator has been demonstrated*. Can the model predict what should fill that gap?

---

5. Synthetic Predictions

5.1 Four Buildable Self-Replicators

The landscape is not just a retrospective explanation of known systems — it is a predictive tool. The framework predicts (k, m, N) — the alphabet size, monomer weight, and total atom count

— for each viable chemistry. It does *not* predict specific monomer identities or sequences; those require biochemical knowledge external to the model. From the analysis, we identify four candidate systems whose coarse parameters fall within the viable region of all six constraints and whose chemistry is accessible to current synthesis. Experimental realization requires identifying a specific self-replicating sequence; the framework narrows the search from an infinite chemistry space to a finite set of targets:

- **Prediction 1: Reduced-Alphabet Peptide** ($\sim$ 340–400 atoms). The landscape formula with $k = 8$ and a generic $m = 7$ gives a lower bound $N = 100 \times 7/\ln 8 \approx 337$ atoms over ~ 48 residues. The specific amino acid selection matters: the lightest 8-type set stays near this bound ($N \approx 341$), while a set optimized for chemical diversity gives $N \approx 403$. Standard solid-phase peptide synthesis handles 50-mers routinely. Identifying a self-replicating sequence within this composition space requires experimental screening.
- **Prediction 2: Thioester World Polymer** ($\sim$ 575 atoms). A 4-type polymer ($k = 4$) with thioester backbone bonds ($m \approx 8$ atoms). Thioester bonds form spontaneously from thioacids without enzymatic catalysis, making this prebiotically plausible. Bond energy ($\sim$ 2.5 eV) is sufficient for thermodynamic stability ($E/k_BT > 80$). Thioester hydrolysis half-lives in water (hours) require compartmentalization or dry–wet cycling to sustain a replication cycle.
- **Prediction 3: Minimal Ghadiri Variant** ($\sim$ 480 atoms). Ghadiri's original used ~ 10 amino acid types. Reducing to 5–6 types while extending the polymer to ~ 60 residues should maintain the self-replication mechanism, offering slightly higher atom count but greatly simplified synthesis.
- **Prediction 4: PNA Self-Replicator** ($\sim$ 1075 atoms). A 70-unit peptide nucleic acid with Watson–Crick base pairing on a peptide backbone. PNA is commercially available, resistant to nucleases and proteases, and 2$\times$ more atom-efficient than RNA for the same information content. Our prediction extends pentamer–hexamer level template-directed PNA ligation to the self-replication threshold.

5.2 The Feasibility Ordering

The ordering of synthetic feasibility by atom count is:

$$N_{\text{reduced-aa}} < N_{\text{Ghadiri variant}} < N_{\text{thioester}} < N_{\text{PNA}}$$

Reduced-alphabet peptide is the easiest to synthesize (standard chemistry, lowest atom count). PNA is the most complex but benefits from existing commercial infrastructure.

---

6. Discussion

6.1 Why Do Three Energy Scales Exist?

One might ask: why do three well-separated energy scales exist at all? The three energy regimes that enable life — $E_{\text{vdw}} < k_BT < E_{\text{hb}} < E_{\text{cov}}$ — are not free parameters. At the order-of-magnitude level, they trace to the fine structure constant $\alpha \approx 1/137$: electronic binding energies scale as $\sim \alpha^2 m_e c^2$, with weaker interactions at successively lower scales. A universe with significantly

different α would plausibly lack three well-separated energy regimes, collapsing the Goldilocks zone (L4). This is a dimensional-analysis observation, not a rigorous derivation of biochemistry from α, but it identifies a suggestive connection between fundamental constants and the preconditions for self-replication.

6.2 Limitations

Assumed parameters. We take $I_{\min} = 100$ nats (≈ 144 bits) as a fixed constant. In reality, $I_{\min}$ depends on the complexity of the replication mechanism and could range from 50 to 200 nats. Our qualitative conclusions (basin location, ordering, paradox) are robust to this range; the numerical bounds shift linearly.

Static landscape. The $N(k, m)$ function assumes fixed m per chemistry. In practice, m varies within a chemistry (amino acid heavy atoms range from 1 to 11). A more refined model would use the distribution of m rather than its mean.

No compartmentalization. We treat self-replication purely as a molecular property: can a single molecule direct the assembly of its own copy? Yet for Darwinian evolution to be sustainable, this molecule cannot live in an open ocean — without a boundary, parasitic sequences will inevitably outcompete cooperative ones. This vulnerability is known as Eigen's parasite problem (Eigen & Schuster 1977; Szostak 2012). Adding a membrane constraint, as Gánti does in his chemoton model (2003), solves this but shifts the fundamental question from "What is the minimum self-replicating molecule?" to "What is the minimum protocell?" We deliberately restrict our scope to the molecule, where physical energy scales and the mathematics of information intersect to yield sharp, testable atom counts.

No kinetic analysis. We characterize thermodynamic feasibility (can a self-replicator exist?) but not kinetic accessibility (how fast does it form?). Thioester bonds form faster than amide bonds; this kinetic advantage is not captured in our framework.

6.3 What We Do Not Claim

This paper does not claim that life *must* arise wherever the six constraints are satisfied — only that it *can*. The constraints are necessary conditions; sufficiency requires additional factors including a supply of activated monomers, a confinement mechanism, and sufficient time.

We do not claim that amino acid chemistry is the *only* viable basis for life. Our results show it is the *most atom-efficient* basis among known chemistries. Hypothetical chemistries (silicon-based, boron-based) that we have not analyzed might occupy similar or better positions on the landscape.

6.4 Translation Table

Information Theory	Chemistry / Biochemistry	This Paper
Channel capacity, alphabet size	Monomer diversity, polymer alphabet	k, $\ln k$
Error threshold, channel reliability	Replication fidelity, proofreading	q, $L_{\max}$
Shannon entropy of source	Sequence information content	$I_{\min}$, $L_{\min}$
Landauer erasure cost	Metabolic energy input	L5, E_{rep}
Fixed point of channel map	Self-replicating molecule	$H \circ G \circ F$

Information Theory	Chemistry / Biochemistry	This Paper
Coding efficiency (bits/symbol)	Monomer economy (atoms/nat)	$N(k, m)$

6.5 Open Questions

1. **Can the reduced-alphabet peptide self-replicate experimentally?** This is our most testable prediction: synthesize a 48-residue peptide from {Gly, Ala, Val, Leu, Ser, Asp, Glu, Phe} with an amphipathic coiled-coil pattern and test for template-directed replication.
2. **What is the true value of $I_{\min}$?** A tighter bound would sharpen the landscape. Information-theoretic analysis of the Ghadiri peptide's replication mechanism could provide this.
3. **Does the thioester world exist?** Thioester polymers form prebiotically, but no one has tested whether they self-replicate. Our prediction of ~ 575 atoms provides a target polymer length.
4. **Is the Goldilocks zone universal?** Does every chemistry with $k_B T < E_{\text{recog}} < E_{\text{cov}}$ support self-replication, or is there an additional constraint we have not identified?

If the reduced-alphabet peptide of Prediction 1 self-replicates in the laboratory, the implications extend beyond origin-of-life chemistry. It would demonstrate that the amino acid alphabet can be compressed by more than half without losing the capacity for autonomous replication — a result with consequences for synthetic biology, minimal-cell design, and the search for extraterrestrial biosignatures.

---

7. Conclusion

Self-replication is not a binary property of molecules but a region in parameter space defined by six intersecting constraints. We showed that this region is non-empty for Earth's physical constants, with a broad optimum basin at amino acid chemistry ($k \in [8, 20]$, $N \in [320, 364]$). The Binary Paradox — simpler monomers requiring more atoms — is a necessary consequence of information compression. The framework yields four testable synthetic predictions, the most accessible being a reduced-alphabet peptide of ~ 340–400 atoms.

The formal verification at two levels (650 declarations across 8 proof modules; 21 core theorems independently verified in Lean 4 with Mathlib, 0 sorry) ensures that every quantitative claim follows deductively from the stated physical hypotheses. What remains unformalized marks the boundary between what mathematics can currently guarantee and what experiment must decide: the topological fixed-point existence for $H \circ G \circ F$, the kinetic accessibility of each candidate, and the role of compartmentalization.

---

8. Formalization

Every quantitative claim in this paper has a machine-checked counterpart. The primary verification layer uses 650 declarations (hypotheses and proved lemmas) across 8 proof files, with 0 errors. The proofs are predominantly real-arithmetic consequences of stated physical bounds; they are not independent derivations of chemistry from first principles. The physical constants enter as hypotheses, and the system verifies that the claimed inequalities follow.

As a secondary, independent check, 21 core theorems are also verified in Lean 4 with Mathlib, with 0 sorry and 0 errors. These 21 are a subset of the 650 — the two counts are not additive. The Lean 4 layer provides tool-independent confirmation that the core results hold.

Proof module	Decl.	Content
Threshold analysis	90	Viable window, information and Eigen thresholds
Minimal molecule	95	Why peptides, why hydrogen bonds, why ~ 320 atoms
Alternative chemistries	70	5 chemistries, ordering, experimental validation
Parameter sweep	135	14 chemistries, four-zone landscape
Life definition	70	Six-constraint definition, non-emptiness theorem
Synthetic predictions	66	Four buildable predictions with feasibility ordering
Amino acid analysis	103	Fixed-point theory, sub-amino-acid failure analysis
Core constraints	21	Energy hierarchies, information bounds, bridges
Lean 4 (Mathlib)	**21**	**Independent verification, 0 sorry**

All proof files and the Lean 4 source are available in the supplementary materials.

Acknowledgments

The author thanks Dr. Péter Nagy, whose question — "Could you prove that self-replication starts around 1000 atoms?" — set this investigation in motion.

During the preparation of this work the author used large language models to assist with manuscript drafting, formal proof construction, and literature search. The author reviewed and edited all content and takes full responsibility for the published article.

References

Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews*, 1(1), 3–22.

Bains, W. (2004). Many chemistries could be used to build living systems. *Astrobiology*, 4(2), 137–167.

Benner, S. A., Ricardo, A., & Carrigan, M. A. (2004). Is there a common chemical model for life in the universe? *Current Opinion in Chemical Biology*, 8(6), 672–689.

Carey, F. A., & Sundberg, R. J. (2007). *Advanced Organic Chemistry: Part A: Structure and Mechanisms* (5th ed.). Springer.

de Duve, C. (1991). *Blueprint for a Cell: The Nature and Origin of Life*. Neil Patterson Publishers.

Eigen, M. (1971). Self-organization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften*, 58(10), 465–523.

Eigen, M., & Schuster, P. (1977). The hypercycle: A principle of natural self-organization. Part A: Emergence of the hypercycle. *Die Naturwissenschaften*, 64(11), 541–565.

England, J. L. (2013). Statistical physics of self-replication. *The Journal of Chemical Physics*, 139(12), 121923.

Gánti, T. (2003). *The Principles of Life*. Oxford University Press.

Joyce, G. F. (1994). Foreword. In D. W. Deamer & G. R. Fleischaker (Eds.), *Origins of Life: The Central Concepts* (pp. xi–xii). Jones and Bartlett.

Kauffman, S. A. (1986). Autocatalytic sets of proteins. *Journal of Theoretical Biology*, 119(1), 1–24.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183–191.

Lee, D. H., Granja, J. R., Martinez, J. A., Severin, K., & Ghadiri, M. R. (1996). A self-replicating peptide. *Nature*, 382(6591), 525–528.

Lincoln, T. A., & Joyce, G. F. (2009). Self-sustained replication of an RNA enzyme. *Science*, 323(5918), 1229–1232.

Luisi, P. L. (2006). *The Emergence of Life: From Chemical Origins to Synthetic Biology*. Cambridge University Press.

Nielsen, P. E. (1999). Peptide nucleic acid. A molecule with two identities. *Accounts of Chemical Research*, 32(7), 624–630.

Ruiz-Mirazo, K., Peretó, J., & Moreno, A. (2004). A universal definition of life: autonomy and open-ended evolution. *Origins of Life and Evolution of Biospheres*, 34(3), 323–346.

Schrödinger, E. (1944). *What Is Life? The Physical Aspect of the Living Cell*. Cambridge University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.

Szathmáry, E., & Maynard Smith, J. (1995). The major evolutionary transitions. *Nature*, 374(6519), 227–232.

Szostak, J. W. (2012). The eightfold path to non-enzymatic RNA replication. *Journal of Systems Chemistry*, 3, 2.

Szostak, J. W., & Bartel, D. P. (1999). In vitro selection of functional RNA sequences. In *The RNA World* (2nd ed., pp. 733–750). Cold Spring Harbor Laboratory Press.

Tjivikua, T., Ballester, P., & Rebek, J. (1990). A self-replicating system. *Journal of the American Chemical Society*, 112(3), 1249–1250.

Ura, Y., Beierle, J. M., Leman, L. J., Orgel, L. E., & Ghadiri, M. R. (2009). Self-assembling sequence-adaptive peptide nucleic acids. *Science*, 325(5936), 73–77.

von Kiedrowski, G. (1986). A self-replicating hexadeoxynucleotide. *Angewandte Chemie International Edition*, 25(10), 932–935.

von Neumann, J. (1966). *Theory of Self-Reproducing Automata* (A. W. Burks, Ed.). University of Illinois Press.

Walker, S. I., & Davies, P. C. W. (2013). The algorithmic origins of life. *Journal of the Royal Society Interface*, 10(79), 20120869.