Original Study • DOI: 10.2478/biocosmos-2025-0013 • BIOCOSMOS • 1 • 2025 • 82-95

BioCosmos

ASSEMBLY THEORY PROVIDES A MEASURE OF SPECIFIED COMPLEXITY THAT QUANTIFIES BUT DOES NOT EXPLAIN SELECTION AND EVOLUTION

Onsi Joe Fakhouri*

Haiku Creations LLC, Denver, CO, USA

Abstract

Categorizing complex objects into 'life' and 'not-life' buckets is hard. Understanding how the 'life' objects came to be in the first place is even harder. The authors of assembly theory (AT) purport to make headway on both issues by providing a measure, dubbed assembly, for categorizing objects and a related mechanism, dubbed selection, underlying their discovery. We show that assembly is a measure of specified complexity, leveraging the assembly index a_i for the complexity of the object and the copy number n_i for its specification. This provides a theoretical foundation for AT's success as an object-categorizer. AT's model for selection yields some helpful parameterizations, but lacks concreteness, and does not grapple with the confounding factors that make the problem of life's origin so challenging. At best, AT's selection model highlights the actual problem that a viable selection process must address: the problem of information.

Keywords

specified complexity • assembly theory • origin of life

Introduction

Arriving at a precise definition for life has proven to be a thorny scientific and philosophical challenge (2). Phenomenologically, however, it is apparent that life is built out of complex objects and that biotic processes – whether cells or third-graders – are, in turn, capable of constructing additional complex objects – whether proteins or poems.

Yet complex objects, by any meaningful measure of *syntactic* complexity (3), abound throughout the universe. What distinguishes non-living complex objects from living ones? And how might the living objects have come about? In particular, if living matter emerges from non-living matter then the causal chain of complex living objects must emanate (whether abruptly or gradually) from non-living objects. What sort of process could have accomplished this?

Assembly theory (AT) – as recently laid out in Sharma et al. (1) and popularized in Walker (4) – aims to make progress on these questions. AT claims to provide a mechanism for distinguishing biological complex objects from abiotic ones and to lay a foundational framework for understanding how these objects came to be via a process of object discovery dubbed **selection**. AT has received substantial engagement and critique [e.g. Uthamacumaran et al. (5) expanded upon in Abrahão et al. (6)] and much of this critique has centered around AT's measure of complexity: the **assembly index**, a_i, which counts the shortest number of steps necessary to recursively construct an object from its constituent parts. Critics have noted that a_i is essentially implementing a simple form of compression and that superior implementations commonly used in the field of Algorithmic Information Theory (AIT) exist.

To this critique, AT proponents have replied (7) arguing that AT is distinct from AIT for two reasons. First, they argue that AT is modeled on a physically-motivated process (object construction via recursive concatenation) thereby allowing it to constrain the probabilities of undirected physical processes. Second, AT does not just measure complexity; it identifies complex objects that exhibit evidence of a directed

* *Corresponding author e-mail: onsijoe@gmail.com*

process of construction (i.e. selection). This is done by augmenting an object's a_i complexity measure with its **copy number** – the number of copies of the object observed. In short, Sharma et al. (1) claim that AT provides a novel method to distinguish complex functional information from random, and equivalently improbable, non-functional information.

But is this contribution novel? As we shall see: no, not exactly.

Moreover, Sharma et al. (1) and Walker (4) push further, claiming that AT provides a framework with the potential to build a 'new physics' of life; one that sheds light on the operation of selection over deep time 'thus unifying key features of life with physics'. Only a few authors (6) have engaged with this particular aspect of AT and the nascent models explored in Sharma et al. (1).

Out present goal is two-fold. We first provide a pedagogical overview of AT and situate it in the broader context of the study of specified complexity (SC). We then explore what AT does, and does not say, about physical processes capable of selection. We argue that while AT *points* to the presence of an information-laden process like selection and can provide some interesting quantitative parameterizations, it does not (indeed, cannot) explain the origin or modes of operation of such a process in any concrete detail.

Specified complexity

In 1998, Bill Dembski published The Design Inference (8) to establish how and when the observation of low probability events can be leveraged to rule out undirected causes. Subsequent refinements of this work [most recently the 2nd edition of The Design Inference; Dembski and Ewert (9)] have placed Dembski's original ideas on a firmer mathematical footing. The core notion underlying The Design Inference is that of SC.

SC is observed when a low probability event (complexity) that conforms to some independently specified pattern (specification) occurs. Dembski argues that it is precisely SC at work when we intuit low-probability events that carry an outsize significance.

Consider, for example, a hand in a game of spades in which 13 cards have been dealt to a player. Any given hand is rare as the odds of getting that *particular* hand are low: if we assume an undirected physical process in which cards are drawn at random with uniform distribution, then the odds of any given hand are ~1 in 6×10^{11}. However, the majority of hands are inconsequential in the context of spades and we are unlikely to attribute anything other than the contingencies of chance and necessity to a given hand. But some hands are *special* and raise suspicion. If the dealer has a hand with 13 spades – a hand guaranteed to win 13 tricks – we might rightly grow suspicious.

Why? The 13-spade hand has exactly the same probability of occurring as any other 13-card hand so it is not *simply* the improbability of the hand that raises our suspicion. Rather, it is the independent pattern 'always wins all 13 tricks' *coupled* with this improbability that leads us to suspect that there may be more at play than merely chance and necessity. The presence of the specification marks the event as *special* and its combination with a low-probability estimate motivates the rejection of the null hypothesis. Specifically: we have cause to doubt that the undirected physical process underlying the probability estimate is a plausible explanation for the event.

Two questions should be addressed immediately. First, what determines whether a probability is *small enough* to rule out chance? To answer this we must evaluate the relevant available probabilistic resources (9) and use them to set a conservative bound. In our spades example the chances of randomly drawing a 13-spade hand is roughly one in a trillion – is this small? If we imagine every human on earth playing a game of spades (about 10 hands) every day for a year we would see ~10^{13} hands of spades. With this many trials, some dealer drawing a 13-spade hand eventually is much more likely than not. However, should we observe multiple 13-spade hands from the same dealer in a given year the available probabilistic resources would be insufficient to quell our suspicions.

Second, how do we rigorously define specification? Different contexts can prompt different definitions for specification; however, Dembski and Ewert (9) takes an AIT approach to generically define specification as the description of the object with minimum length. This allows for the use of English phrases to describe objects (e.g. 'always wins all 13 tricks') and the authors apply this model to numerous problems such as analyzing texts, identifying fraud, and even evaluating claims about the extra-terrestrial origin of the asteroid 'Oumuamua.

Indeed, several examples of SC measures exist, and many have been proposed for biological systems (10). These different models may adopt different measures for complexity and specificity; however, Montañz (10) proposed a unified

schema and showed that all such measures can be expressed in terms of a common form Canonical Specified Complexity (CSC) model:

$$SC(x) = -\log_2 r \frac{p(x)}{\nu(x)} \quad (1)$$

where x denotes an observation taking place in some space χ according to a probability distribution $p(x)$, and $\nu(x)$ is a specification function $\nu : \chi \to \mathbb{R} \geq 0$ that measures the specification of x (higher values of ν denote a higher degree of specification). r is a normalization constant that can be freely chosen – however, for an SC model to be a *canonical* SC model, it must be true that $\nu(\chi) \leq r$ where $\nu(\chi) = \sum_{x \in \chi} \nu(x)$

Given a CSC measure Montañez (10) shows that the probability of observing an object x with $SC(x) \geq b$ satisfies:

$$\Pr(SC(x) \geq b) \leq 2^{-b} \quad (2)$$

This result is referred to as the Conservation of CSC and allows the use of $SC(x)$ as a hypothesis test to rule out the chance explanations underlying $p(x)$.

We shall show that AT can be expressed as a CSC measure. First we introduce and explore AT.

Assembly theory

AT (1) explores the idea that the discovery and production of complex objects can be modeled as a directed assembly process. Complex objects are recursively constructed from a pool of smaller constituent building blocks. At each step of the construction process two elements in the pool are selected and joined together. The resulting object is then added to the pool and made available for future steps; the pool is presumed to be large with ample provision of each member. Thus a given object, O_i, in AT is the result of some contingent construction-path (sometimes referred to as a lineage); in fact many such paths can exist for O_i. The *complexity* of O_i is measured via its **assembly index**, a_i – defined to be the number of steps in the *shortest* possible construction-path.

We can build a quick intuition for a_i by thinking of the assembly of strings in the English alphabet, see Figure 1. Here the basic building blocks are the 26 letters and a_i can be computed by looking for repeating blocks. The shortest path for a word of length ℓ with no repeating subunits is simply $\ell - 1$ (e.g. $a_i = 4$ for 'short'). However, words with repeating subunits will have a smaller assembly index (e.g. the 12-letter word 'hubbubbubboo' can reuse 'ubb' twice and has assembly index $a_i = 7$).

For ordered linear chains like words we can see that a_i is bounded from above by the length of the chain, ℓ, and from below by $\log_2 \ell$ which represents a chain consisting entirely of a single repeated subunit. Since it tracks object reuse a_i is related to the compressibility of the object; however, AT is not particularly interested in identifying an optimal compression scheme.

Rather, the assembly index was first introduced in Marshall et al. (11) and subsequently formalized in Marshall et al. (12), in order to develop a model-free biosignature to detect the activity of life in the context of astrobiology. By 'model-free' we mean a measure of life that does not assume a particular (i.e. terrestrial)

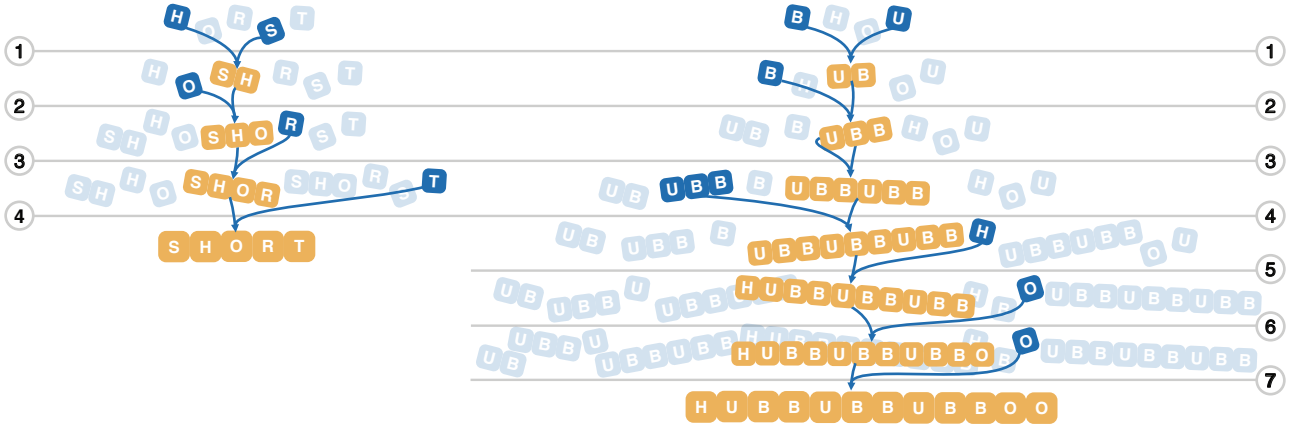

Figure 1. The assembly of English strings. On the left we assemble the word 'short' ($a_i = 4$), on the right the word 'hubbubbubboo' ($a_i = 7$). Shortest pathways are shown, as is the growing pool of building blocks. Note that at each step the resulting product is inserted back in the pool. This allows the reuse of 'ubb' to construct 'hubbubbubboo' in fewer steps.

model of life and can therefore be used to agnostically identify signs of alien life. Indeed, the authors of AT refer to a_i as an *intrinsic* property of the object – though we note that the numerical value of a_i will depend on the choice of constituent building blocks and the set of allowed joining operations.

The assembly index is also a *physically-motivated* measure of complexity because a_i operates as a *lower bound* on the number of physical operations necessary to construct the first instance of O_i. The reasoning here is subtle: in AT each intermediate object in a construction-path is required to be a physically plausible object according to the rules of physics governing the object. However, the *specific* smallest a_i construction-path need not represent physically plausible construction trajectories.

An example will help clarify this distinction. Consider the case of molecular assembly in which atoms are the building blocks and bonds are the joining operation. Each object along a construction-path must correspond to a physically plausible molecule (e.g. the rules of chemical bonds apply, so hydrogen atoms can have at most one bond). However, the *combination* of molecules from prior steps to produce the product at a subsequent step need not correspond to a physically plausible reaction pathway. A pictorial combination of atoms and bonds is sufficient, even if there is no known chemical pathway for the transformation. Despite this, a_i nonetheless represents a valid (if potentially weak) lower bound on the construction history of the object in question as the constraints of a physically plausible path would necessarily require more steps and/or more intermediates.

AT leverages this lower bound to identify evidence for the action of a *directed process* behind the construction of complex objects. To do this AT needs an additional ingredient. AT argues that:

1. The number of objects with high a_i explodes combinatorially as a_i grows. Therefore a given O_i with high a_i has a low probability of being discovered by an undirected process.
2. Observing multiple copies of O_i signifies that some sort of directed process must be at work to discover and produce them.

and this second clause is crucial to AT. To capture both elements, Sharma et al. (1) define a quantity A, called **assembly**, that can be computed for an *ensemble* of objects:

$$A = \sum_{i=0}^{N} A_i = \sum_{i=0}^{N} e^{a_i} \frac{n_i - 1}{N_T} \tag{3}$$

here N represents the number of unique objects, O_i, in the ensemble. Each O_i has a **copy number** n_i representing the number of copies of O_i and the ensemble contains a total of $N_T = \sum_{i=0}^{N} n_i$ objects. The assembly A_i of each O_i is composed of the object's exponentiated assembly index e^{a_i} and its normalized copy number $(n_i - 1)/N_T$. The assembly index is exponentiated to capture the combinatorial explosion implied by a high a_i, while the normalized copy number acts as a filter screening out O_i that do not appear in substantial quantities: $(n_i - 1)/N_T$ is small or even 0 for such objects.

Thus, an ensemble containing many identical simple objects (low a_i, high n_i) will have low assembly. As will an ensemble containing many unique individual instances of complex objects (high a_i, but n_i = 1). However, complex objects with high copy numbers will constitute an ensemble with high A. AT argues that a *directed* process, dubbed **selection**, must be postulated to explain the existence of such an ensemble.

For each step in the construction process of an object directed by selection, a specific set of products from earlier steps are *selected* and combined to produce the next iteration of the object. In this way, selection winnows down the combinatorial space of options making the construction of complex objects possible given limited probabilistic resources. Since assembly can be computed for a broad range of contexts, no universal concrete mechanism is proposed for selection. Rather, AT's authors simply argue that given an ensemble of objects with high assembly some sort of selection-like process must have been at work to produce the ensemble.

Assembly as a measure of SC

The parallels between SC and assembly are apparent. In the construction of A, e^{a_i} captures the notion that complex objects inhabit a combinatorially large space and the *probability* of a random undirected process constructing such an object is small: $p(O_i) \sim e^{-a_i}$. But low-probability events occur all the time; recall that drawing any particular hand in a game of spades is a low-probability event, but there is no need to attribute anything other than chance to such a hand. AT, however, seeks to identify objects for which there *is* reason to rule out an undirected construction process. This is done by providing a *specification* in the form of the normalized copy number: $\nu(O_i) \sim \frac{n_i - 1}{N_T}$. Quoting Sharma et al. (1) 'finding more than one identical copy indicates the presence of a non-random process generating the object'.

Such an argument mirrors the formal framing of SC with assembly index acting as a measure of (im)probability,

and copy number acting as an external specification. The combination of complexity (high a_i) *and* specification (high n_i) resists explanation by chance alone – and requires the action of a directed selection process capable of winnowing down combinatorial space.

We explore a concrete application of assembly-as-SC to the problem space of protein discovery below. But first, we further motivate why copy number can serve as a viable specification of interest and explore the properties of assembly index as a measure of complexity.

Copy number as specification

We have proposed that copy number operates as a specification. But *what* does copy number specify? Let's consider two examples from different domains.

First, take the case of written human artifacts. Imagine an archaeologist happening upon a heretofore undiscovered ancient manuscript (given a pool of alphabetical letters any non-trivial manuscript of moderate length will have a high assembly index). At first the copy number of such an artifact will be one – it is unique in the world. But as the manuscript is carefully extracted, preserved, and interpreted, copies will begin to proliferate. There will be copies sent to museums around the world. Copies in academic papers interpreting the text. Copies in anthologies collecting such texts. Over time we would measure $n_i >> 1$ reflecting the *value* ascribed to the manuscript by society.

Contrast this with a random collection of words with assembly index commensurate to the ancient manuscript. These two objects may have similar a_i (and therefore be equivalently improbable); however, the copy number of the random collection of words is unlikely to exceed one. Why? Because such a text carries no value to the society in question. In this context, the long-term copy number acts as a reasonable – if blunt – external specification for 'value ascribed to a written text by a society'.

Now consider a biological case, that of proteins: long chains of sequentially arranged amino acids that can fold into complex functional three-dimensional shapes. Given that cells contain machinery to transcribe sequences of DNA into multiple copies of identical proteins it is common for proteins to exhibit high copy number both within an individual cell and across the biosphere.

But now, imagine a scenario in which random mutation results in a new pair of start and stop codons appearing around a heretofore untranscribed string of DNA. This will result in a new protein. At first the protein will have copy number 1, but as the machinery of the cell repeatedly transcribes the novel DNA strand n_i will increase. If, however, the protein serves no function for the cell (or, worse, is detrimental to the cell) we can expect that future mutations, selection effects, or even error-detection mechanisms will winnow out the production and replication of this strand of random protein. Thus reducing n_i back down to zero.

Here again we see that the long-term behavior of n_i serves as a reasonable proxy for 'value' or 'function'. In the case of proteins in the context of existing cellular infrastructure, the presence and *persistence* of high copy number serves as an external specification that points towards functionality.

We note, however, that there are issues with the use of copy number as a specification. Consider, for example, the presence of stochastic error in the copying process of an object. These errors may prove superficial – the copy still retains functional value commensurate to that of the original – however, without specifying a tolerance range for such variations AT may fail to categorize these superficial variants as members of the same class of object. This would lower the assembly measure for the object (by spreading n_i across multiple O_i) and potentially result in a false negative.

In addition, the normalization of n_i by N_T – the total number of objects in the ensemble – makes normalization of assembly somewhat ambiguous. The actual value of A will depend on what is and is not included in the ensemble making it challenging to speak of a rigorous absolute A_{min} threshold for rejecting an undirected process.

Nonetheless, such limitations are not fatal to the core structure of the theory. As we have seen in both the artifact and biological cases long-term copy number can serve as a workable specification to call out that the random sequence in question – as equally unlikely as any other random sequence of equivalent a_i – plays some sort of externally specified role.

Assembly index as measure of complexity

In the framework of SC the probability under consideration must be computed for some hypothesized chance process. If this probability is low, and the event in question is highly specified, then we have good reason to rule out the chance process as the event's cause. We have proposed that assembly index operates as a probability measure – but what chance process is it associated with?

This is where the physical motivation behind a_i proves valuable. Consider an object with assembly index a_i that is an ordered linear chain of initial building blocks. We assume there are β unique building block elements in the initial pool (for

words in the English language $\beta = 26$, for proteins constructed of amino acids $\beta = 20$). An undirected construction process is one in which the choice of objects to combine at each assembly step is random (we assume a uniform distribution). We can estimate a bound on the probability of a given object being produced by such a process by estimating a bound on the number of possible product objects with index a_i. Doing this is not entirely straightforward, however, as it entails characterizing the minimal-length construction trajectory for objects to determine their a_i. Nonetheless we argue in the Appendix that there are at least β^{a_i} objects with index a_i. Thus the probability of a given object being generated by an undirected process is bounded above by $p(O_i) \leq \beta^{-a_i} = e^{-a_i \ln \beta}$.

Thus, we see how e^{-a_i} operates, heuristically, as a measure of the probability of an undirected process successfully identifying an object of interest in combinatorial construction space. Therefore, when understood as a measure of SC, assembly allows us to conclude that an undirected construction process is unlikely to explain the origin of an object with high a_i and high n_i.

Nonetheless, there are some noteworthy issues with a_i. For example, a highly ordered crystalline structure that occurs naturally by the laws of physics can readily have a high copy number *and* large a_i. If the crystal is simply the repeated presence of a single subunit, then $a_i \sim \log_2 \ell$ where ℓ is a measure of the size of the crystal and can grow quite large. Clearly one need not invoke a directed process to explain the existence of the crystal. Alternative measures of complexity are able to handle such edge cases and this limitation of AT is noted by some of its critics [see Zenil (13)].

Dembski and Ewert (9) emphasize that SC measures must avoid such false positives by carefully considering the space of possible causes when constructing their probability measures. A well-formed SC argument would note that the probability of a repeated crystalline structure is high on the known laws of physics and avoid concluding that a directed process must be at work. Though not articulated in Sharma et al. (1), this nuance is actually described in Figure 4 of Marshall et al. (11), reproduced here in Figure 2, which provides a technique for determining whether to infer a directed process from a given measure of a_i. The authors argue for analyzing the combination of a_i and object size ℓ. Recall that $\log_2 \ell \leq a_i < \ell$. If a_i is 'too close' to its lower bound, $\log_2 \ell$, then objects can be deemed too simple and therefore possibly attainable by an undirected process. Perhaps future iterations of AT will formally include such guards by conjoining all three of (a_i, n_j, ℓ) into A.

Before proceeding, we should note that there are several alternative measures of complexity that predate AT. Perhaps most comparable is the work of Böttcher (14, 15) who develops a local measure of molecular complexity C_m derived from the information content of the degrees of freedom on a per-atom basis. C_m has many advantageous qualities including its computability and its linear nature. Under Böttcher's schema, the molecular complexity of an ensemble of objects is largely given by the sum of the complexities of each member. a_i by comparison can behave in unintuitive ways when we consider ensembles of objects with overlapping assembly histories (see, for example, the complexity in Section 'Computing a bound on N (a_i) for proteins' of the Appendix). Moreover, Böttcher pairs C_m with a measure of the sequence information of biomolecules C_i to develop a model-free schema for understanding biosignatures across a wide spectrum of contexts – much like AT. Indeed, the parallels are striking: both Böttcher and the AT authors apply their measures to the problems of drug discovery, the detection of life in an astrobiological context, the use of experimental methods to estimate complexity [nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS), respectively], and questions of life's origins. A deeper comparison of the two is warranted.

A more detailed example: proteins

To further illustrate how assembly operates as a form of SC we shall work through a concrete example and recast the assembly Eq. (3) in terms of the common form equation for CSC (1).

We consider the case of some protein O with assembly index a observed with copy number $n \geq 1$ in an ensemble of $N \geq n$ total proteins. We explore concrete scenarios and estimate values for these quantities later.

The assembly for this protein, in the context of this ensemble, is:

$$A = e^{a} \frac{n-1}{N} \tag{4}$$

To recast this example in terms of SC we need a probability measure for how (im)probable it is for O to be discovered by an undirected process. As discussed in Section 'Assembly index as measure of complexity' this probability is bounded by $p \leq \beta^{-a}$ where $\beta = 20$ is the size of the pool of amino acid building blocks available for proteins. Since $e^{a \ln \beta} = \beta^{a}$, we can express p in terms of assembly:

$$p < \left(\frac{n-1}{AN} \right)^{\ln \beta} \tag{5}$$

Plugging Eq. (5) into Eq. (1), we find

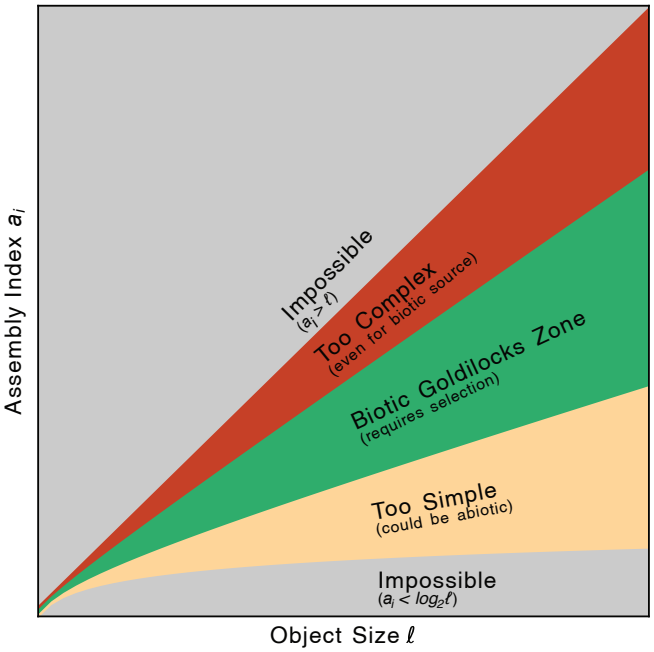

Figure 2. Illustrative rejection regions for a_i and ℓ. Though not referenced in Sharma et al. (1), early work by Marshall et al. (11) (their Figure 4 governed by CC BY 4.0; reproduced here with slight aesthetic modifications) illustrates biosignature interpretations for different observed combinations of a_i and ℓ. The yellow area where a_i is too close to $\log_2 \ell$ is rejected as being too simple – objects in this region can be explained by appeals to random chance and so a directed/biotic selection process cannot be securely inferred. On the other hand, objects in the red region where a_i is too close to ℓ are rejected for being 'so complex that even living systems might have been unlikely to create them' Marshall et al. (11); we discuss this in Section 'Conclusion'. The green Goldilocks zone sits in-between and marks where a directed process can be inferred. Note that while the outer envelope is rigorously defined, these internal subregions are not; they are provided for illustrative purposes only.

$$SC = -\log_2 r\frac{p}{\nu} > -\log_2\left[\left(\frac{n-1}{AN}\right)^{\ln\beta}\frac{r}{\nu}\right]$$

$$SC > \ln\beta\log_2\left[\frac{AN}{n-1}\left(\frac{\nu}{r}\right)^{1/\ln\beta}\right] \quad (6)$$

To ensure this is a measure of CSC we must choose a specification function ν and a normalization r such that $\nu(\chi) = \sum_{x\in\chi}\nu(x) < r$. Recall that χ is the space of observations. Since AT's specification is a function of copy number we shall take $\chi = \{n | 1 \le n \le N\}$ – i.e. we consider the range of outcomes for observing up to N copies of O in the ensemble.

If we set

$$\nu = \left(\frac{n-1}{N}\right)^{\ln\beta} \quad (7)$$

and

$$r = N \quad (8)$$

then it follows that

$$\nu(\chi) = \sum_{n=1}^{N}\left(\frac{n-1}{N}\right)^{\ln\beta} < N = r \quad (9)$$

because $\ln\beta > 0$ and $\frac{n-1}{N} < 1$. Thus these choices of ν and r give us $\nu(\chi) < r$.

Plugging Eqs (7) and (8) into Eq. (6), we find

$$SC > \ln\beta\log_2\frac{A}{N^{1/\ln\beta}} \quad (10)$$

and we see that this normalized form of A serves as a lower bound on a CSC measure. Because of the conservation of CSC it follows that the probability of observing O is:

$$Pr \le 2^{-SC} \le \frac{N}{A^{\ln\beta}} \quad (11)$$

To expose the parameters in our model we can expand A:

$$Pr \le \frac{N^{1+\ln\beta}}{\beta^a(n-1)^{\ln\beta}} \quad (12)$$

These probability equations reflect the features and limitations of assembly we have been discussing. We see that proteins with larger assembly index a and copy number n are, indeed, less probable to attain by undirected processes. However, the assembly (and the resulting probability) is poorly normalized and can be modified substantially by considering different sized ensembles N.

To put some numbers to this we note that proteins routinely have $\ell \gg 100$ amino acids. Since $a < \ell$, we will assume an assembly index of $a = 100$ to probe proteins of moderate complexity. We consider two scenarios.

First, we consider observing $n \sim 10$ copies of O within the proteome of a single *E. coli* cell. Such a cell typically has $N \sim 10^6$ proteins (16). This yields $A \sim 10^{38}$ and $Pr \le 10^{-109}$. This low probability correctly rules out the chance occurrence of n copies of O within *E. coli*: O is only present in multiple copies because its amino acid sequence is specified (i.e. selected) by the DNA code for O and translated by the machinery in the cell.

Second, we consider a scenario more relevant to evolutionary timescales. Now O is a protein present in the majority of bacteria. There are currently ~10^{30} bacteria on earth (17) so we assume at least $n \sim 10^{30}$ copies of O exist. We draw these n copies from the reservoir of all bacterial proteins that have

existed over the course of evolution. This number is difficult to estimate but we'll assume ~10^{40} bacteria have ever existed, each with ~10^6 proteins giving us $N \sim 10^{46}$. With these numbers $A \sim 10^{26}$ and $Pr \lesssim 10^{-36}$. Again, we see a low probability correctly rule out undirected processes. Some sort of selection mechanism must have been at work to generate O.

Of course these numbers are somewhat contrived – the results are sensitive to the choice of ensemble, N, and the choice of $a = 100$ is somewhat arbitrary.

In particular, if $a = 100$ is too high then our probability bound is too low and we are at risk of incorrectly ruling out the efficacy of an undirected process. However, there is reason to believe that a is not substantially lower than the length of the protein in question and there are many proteins with length $\ell \gg 100$ – several lines of evidence point in this direction, particularly the incompressibility and high entropy content of proteins (18, 19).

In fact, the study of protein compressibility is closely related to the notion of assembly index. Some state-of-the-art approaches have been found that can compress protein sequences substantially: by looking for long-range correlations at the genome-scale of a species Adjeroh and Nan (20) are able to compress protein coding regions by 50%–80% depending on the species. While impressive from a data-storage point of view this does little on its own to paint a convincing picture that proteins are, in fact, broadly constructed recursively from fundamental building blocks at a higher level of abstraction than individual amino acids. Perhaps more sophisticated complexity analyses such as Block Decomposition Method (BDM) (21) will show that proteins are recursively constructed from simpler transformations of building blocks – this would provide a tighter bound on a_i. Or perhaps the presence of minor or synonymous mutations is confounding compression algorithms, and so analyzing three-dimensional protein structure instead of sequences will be necessary to improve compressibility (22). However, absent such an analysis our current knowledge of the field indicates that a_i will likely scale more-or-less linearly with ℓ and that the scaling factor will be modest (i.e. ~$O(0.1)$).

Thus our estimates, though contrived, may nonetheless serve as plausible upper limits. This illustrates how assembly operates as a measure of SC to argue that an undirected process is unlikely to yield high-assembly objects. Indeed, in the case of proteins it is unlikely that random sequences will yield novel complex proteins [see, e.g., Tian and Best (23)]; instead a more complex search process is typically invoked to explain the discovery of novel proteins (24, 25). Assembly correctly points to the need for such a process.

Selection as a material mechanism

Workers in the AT community have applied AT to various domains to explore its utility. Marshall et al. (26) compute the assembly index of molecules and show that these correlate with the number of MS2 peaks observed in tandem MS experiments. They use this correlation to estimate a_i for a series of samples via MS and find that a cutoff value of $a_i \sim 15$ serves as a workable threshold for detecting biosignatures. In the context of MS the copy number specification is implied as only samples containing millions of copies of a molecule will yield measurable MS peaks (4, 7). Uthamacumaran et al. (5) use the same data sets to show that alternative compression algorithms applied to MS results outperform AT. Notably, Uthamacumaran et al. (5) apply compression to a binary representation of the MS distance matrices whereas Marshall et al. (26) simply count the number of peaks (after some filtering and thresholding) to measure a_i.

In more recent work, Jirasek et al. (27) have shown that peak-counting in infrared spectroscopy and NMR spectra also correlates with a_i. These correlations are routinely touted by AT's authors as evidence that assembly is empirically measurable (4). We simply note that the underlying analysis (peak-counting) is relatively straightforward and that these correlations are unsurprising and would be expected with any well-formed measure of complexity (5, 14). They do not, in and of themselves, constitute evidence for the physical interpretation of one model of complexity over another.

Liu et al. (28) use assembly to explore a variety of chemical contexts: for example they illustrate how assembly can inform a search for novel drugs by focusing on candidate molecules that are, themselves, assembled from the common pool of building blocks that underlie existing drugs. They also explore how assembly can help adjudicate how closely related a group of molecules is by exploring the family of basic biomolecules and, also, a set of plasticizers.

Thus Marshall et al. (12) use AT as a biosignature detector and Liu et al. (28) use it as a mechanism to explore molecular groupings and potentially extend them. Importantly, the authors of both papers are clear that assembly index and the underlying assembly pathways are not intended to represent actual physical processes. For example, according to Liu et al. (28) the 'assembly pathway of a molecule does not necessarily correspond to a realistic sequence of chemical reactions that produce this molecule. Instead, the shortest assembly pathway bounds the likelihood of the molecule forming probabilistically… No matter which methods or

synthetic approaches are used, there will be no shorter way than this ideal one, which makes it an intrinsic property of a molecule'. This is consistent with our framing of a_i as a physically motivated measure that can set a bound on physical processes but does not necessarily map on to a known actual physical process.

A shift occurs in Sharma et al. (1); however, according to the paper's title, AT '**explains** and quantifies selection and evolution'. The abstract claims that AT 'enables us to incorporate novelty generation and selection into the physics of complex objects. It **explains** how these objects can be characterized through a forward dynamical process […and…] discloses a new aspect of physics emerging at the chemical scale'. And later, the paper introduces a parameterization scheme (discussed below) that 'suggests that selectivity in an unknown physical process can be **explained** by experimentally detecting the number of objects, their assembly index and copy number as a function of time'. The usage of 'explains' here (emphasis added) is notably vague - does AT provide explanatory mechanisms for selection? Or does it merely explain how the presence and action of selection can be measured and characterized? In contrast, popular-level press releases about AT have been quick to proclaim that AT 'explains both the discovery of new objects and the selection of existing ones' (29). It would seem that AT is no longer simply a measure to rule out undirected processes, but also a framework for quantifying and, somehow, explaining directed mechanisms that can form complex objects.

Modeling selection?

The authors of AT refer to such mechanisms as **selection**: processes that can winnow combinatorial space along an object's construction trajectory. Since their goal is to establish a broad framework for the construction of complex objects, AT's authors choose to focus on the generic properties of selection by exploring abstract models in lieu of concrete physical systems. For example, to explore the dynamics of object discovery under selection, Sharma et al. (1) present a toy model represented by the equation:

$$\frac{dN_{a+1}}{dt} = k_d(N_a)^{\alpha} \qquad (13)$$

where N_a is the number of unique objects at assembly index a, k_d represents the rate of discovery, and α represents the degree of selection. When $\alpha = 1$ there is no selection (all objects at prior assembly index are available to construct subsequent objects) and the combinatorial space of possibilities explodes exponentially. However, if $\alpha < 1$ then a selection process is at work reducing the number of objects from which to generate future objects. In principle, this alleviates the combinatorial explosion at high a and allows a selection process to construct high a_i objects within limited available resources.

But what does this model describe in practice? Solutions to Eq. (13) are provided in the supplementary information of Sharma et al. (1) and take the form $N_a \propto t^x$ where x is a function of a and α and always satisfies $x \geq 1$ (its detailed behavior is unimportant here). This means the number of *unique* objects at assembly index a grows unbounded with time, even for small a. This is an unphysical result: the proposed solutions do not account for the combinatorial reality that even for large numbers of building blocks the possible number of unique objects at a given assembly index must be finite. Thus, it is unclear how to apply these solutions to realistic physical systems. Other exemplar models in Sharma et al. (1) are similarly limited. For example, the authors explore the construction of 'linear chains defined as integers which are equivalent to linear polymers constructed from a single monomeric unit'. For such linear monomers the only distinguishing feature is their length; however, objects of identical length that have different construction histories are treated as distinct objects (e.g. there are two chains of length $4: 1 \xrightarrow{1+1} 2 \xrightarrow{2+2} 4$ and $1 \xrightarrow{1+1} 2 \xrightarrow{2+1} 3 \xrightarrow{3+1} 4$).This is a confusing choice that obscures the relationship between actual objects (e.g. '4') and the various trajectories that could have constructed the objects.

These linear monomers are subsequently used to explore the effects of directed selection vs undirected construction. A simulation is performed in which an undirected process constructs chains by drawing two chains from the assembly pool at random, attaching them together, noting the product, then adding it to the pool. This is compared to a directed process in which, at each step, the **longest chain** in the pool is chosen and combined with a randomly chosen chain from the pool. The authors then note that the products of the directed process are substantially longer than the products of the undirected process. Since assembly index simply scales with the log of length for these monomeric chains, the directed process therefore generates more complex objects. But both the model and this result are *trivial* in light of the sorts of real-life physical problems that selection must be invoked to solve.

These various toy models illustrate a key limitation of selection as construed in AT. Selection is conceived to operate via object concatenation and is parameterized with α as such. However, the real-life construction of objects (whether by the hands of human workers or the blind forces of evolution) does not only operate via concatenation. Consider, again, the example of proteins: it is well known that enzymes exhibiting promiscuous behavior can be optimized to catalyze related

molecules via a process of stochastic exploration of navigable fitness landscapes via mutation (30). Modeling such a change-by-modification process is awkward in AT given its change-by-concatenation perspective of selection.

Finally, Sharma et al. (1) further characterize the selection process by stipulating that the timescale for producing copies, τ_P, and the timescale for discovering new objects, τ_D, must be commensurate. If $\tau_P \gg \tau_D$ then many unique objects with low copy number will be discovered and proliferate (these have low A); and if $\tau_D \gg \tau_P$ then there will be many copies of objects with low assembly index (these will also have low A). Expressing these constraints ($\alpha < 1$, $\tau_D \sim \tau_P$) provides helpful language, however, neither these constraints nor the simple models explored in Sharma et al. (1) describe a concrete set of mechanisms for implementing an actual viable material selection process.

Selection and information

Indeed for a material selection process to *actually* exist, function, and succeed a few things need to be true:

1. There must be a mechanism for *creating copies* of objects such that earlier elements in an assembly trajectory are available for combination into future elements.
2. There must be a mechanism for *generating variations* of objects – envisioned in AT as recombining prior elements in new ways.
3. There must be a mechanism for *filtering* the objects at each step in the trajectory in a way that collapses combinatorial space.

Readers will recognize these as the three Lewontin conditions for evolution to hold (31). These are necessary but not sufficient. A viable selection process must also be shown to actually succeed in discovering high-a_i objects within the available probabilistic resources (time, number of trials, material, etc.) and in spite of any confounding factors (e.g. decay rates, polluting cross-linkages, etc.). AT does not attempt to concretely address any of these thorny issues for any actual physical system of interest – at best, AT narrowly focuses on the degree of challenge posed solely by the information dynamics of the system in question.

Nonetheless, AT's senior authors argue that a *material* selection process *must* exist by dint of the *existence* of objects with high assembly. Lee Croning makes precisely this argument in the context of abiogenesis, arguing that 'existence is the proof' (32). Such a conviction, however, neither demonstrates that a material process capable of the necessary selection exists nor does it shed light on how it might operate. In fact, Marshall et al. (26) explicitly casts doubt, noting that high-a_i objects are 'those so complex they require a vast amount of encoded information to produce their structures, which means that their abiotic formation is very unlikely'.

More pointedly, Sharma et al. (1) note that some known prebiotic synthesis reactions, 'such as the formose reaction, end up producing tar, which is composed of a large number of molecules with too low a copy number to be individually identifiable'. This is one of many paradoxes associated with origin of life research (33) and it is not clear how a prebiotic selection process might emerge that is capable of selecting subcomponents of the combinatorially complex tar and moving the ensemble towards a high-copy number target such as life (34, 35). Indeed, it is increasingly apparent that origin of life experiments that manage to make progress typically do so only because of human intervention (36).

This brings us to the crux of the issue. In pointing to selection, AT draws attention to a core problem underpinning the formation of complex objects: that of information. Consider the toy model in Eq. (13), the parameterization of selection with a simple scalar α is, on its own, somewhat misleading. The challenge to solve is not merely identifying a low-α process capable of selecting *arbitrary* subset, N_a^α of objects. The challenge is to identify a *directed* process that can select *particular* subsets of objects that result in the eventual construction of relevant high-a_i objects. This is what it *means* for a directed selection process to tame the combinatorial explosion of possibilities: there must be some non-random bias that drives the exclusion of some possibilities and the inclusion of others.

Indeed we can estimate the information content of such a process (37). As described in Shannon (3), information can be computed in terms of probability via:

$$I = -\log_2 P \tag{14}$$

so at each step in which a viable subset N_a^α is selected the selection process must utilize

$$I = -\log_2 \frac{N_a^\alpha}{N_a} = (1-\alpha)\log_2 N_a \tag{15}$$

bits of information to propagate the system forward. Where does this information come from? And how is it marshaled by a viable physical process to do the rele-vant work? Assembly theory does not (indeed, cannot) answer these questions; it merely helps us ask them in a more quantitative frame.

It is striking to compare this conclusion with what senior authors Cronin and Walker have argued in the past. In Cronin and Walker (38), both advocate for the need to '[focus] on information' to make headway in the origin-of-life field. And in Walker and Davies (39) and Walker and Davies (40), Walker eloquently argues that the origin-of-life represents a phase shift in which *information* transitions from playing a passive role to taking on a more active role as a top-down causal force [drawing on Ellis (41)]. Walker describe this as 'the hard problem' of life. However, AT, as an adjudicator of undirected processes, can at best point us to *where* this 'hard problem' is at work; it does not make progress explaining *how* it works.

This can be seen in the shape of Cronin and Walker's latest research program. Walker (4) discusses Cronin's development of a 'chemputer' (42) – an impressive 'grad-student-in-a-box' programmatically driven chemical synthesis machine. According to Walker (4):

> *Lee right now has several chemputers running in his lab with primordial-soup chemistries in them, which he is evolving by programmable iteration over possible environments… Because we have assembly theory, we can quantify the assembly of the starting molecules and track how assembled things become over time to look for evidence of the emergence of evolution and selection within the boundary conditions of the experiment… Because the chemputer is automated, we hope to scale this up to search large volumes of chemical space all at once… [with] a cloud of potentially millions of chemical reactors.*

Notably, while AT can help define success criteria for these experiments, it offers little insight into how to conduct them. Instead, the experimental approach most resembles a brute force search [albeit one in which each search attempt is heavily information laden – see Sˇiaučiulis et al. (43) for a sense of how much information is encoded in the protocols]. Ironically, AT itself casts doubt that such a random-sampling approach will succeed; indeed we can hazard a prediction: To the extent that the chemputer cluster can be directed towards finding a solution (perhaps by an AI-driven algorithm iterating on protocols to eventually generate high assembly objects) it will be because of a target-oriented design – an externally imposed selection – that has no material analogue in the prebiotic context of the early earth.

Conclusion

We have shown that AT includes a measure of SC: the conjoining of assembly index (complexity) and copy number (specification) can rule out undirected processes as explanations for high assembly objects. But while AT helps us identify the action of a directed process such as selection, and provides some helpful vocabulary for parameterizing such a process – it does not provide insight into how a physically viable material mechanism for the selection process might itself arise and operate. Such a selection process would need to overcome a wide range of confounding factors [see, e.g., Benner (33) for the origin-of-life context] that AT does not attempt to address.

What might explain the origin and nature of a viable selection process, particularly in the context of abiogenesis? Walker and Davies (40) argues that the causal agency of information is the 'hard part of the problem' of life, drawing an analogy to Chalmer's 'hard problem of consciousness' (44). What if these two problems are actually one and the same, and the unified 'hard problem' to grapple with is the nature of conscious agents and their capacity to infuse and affect the material world with information? After all, in our uniform and repeated experience (45) we routinely observe that conscious agents are capable of sifting through combinatorial space by providing and acting upon the information necessary to perform selection and accomplish outcomes. The presence of selection may well be evidence of the action of a selector (46).

Competing interests

The author declares no competing interests.

References

1. Sharma A, Czégel D, Lachmann M, Kempes CP, Walker SI, Cronin L. Assembly theory explains and quantifies selection and evolution. *Nature*. 2023;622(7982): 321–328. doi: 10.1038/s41586-023-06600-9
2. Tirard S, Morange M, Lazcano A. The definition of life: a brief history of an elusive scientific endeavor. *Astrobiology*. 2010;10(10): 1003–1009. doi: 10.1089/ast.2010.0535
3. Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal*. 1948;27: 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x
4. Walker SI. *Life as no one knows it*. New York, NY: Riverhead Books; 2024.
5. Uthamacumaran A, Abrahão FS, Kiani NA, Zenil H. On the salient limitations of the methods of assembly theory and their classification of molecular biosignatures. *NPJ Systems Biology and Applications*. 2024;10(1): 82. doi: 10.1038/s41540-024-00403-y
6. Abrahão FS, Hernández-Orozco S, Kiani NA, Tegnér J, Zenil H. Assembly theory is an approximation to algorithmic complexity based on LZ compression that does not explain selection or

evolution. *PLOS Complex Systems*. 2024;1(1): e0000014. doi: 10.1371/journal.pcsy.0000014

7. Mathis C. *On the salient misunderstandings of assembly theory*. 2022. Available from: https://colemathis.github.io/blog/2022-10-25-SalientMisunderstandings
8. Dembski W. The design inference: eliminating chance through small probabilities, Cambridge studies in probability, induction and decision theory. Cambridge, England: Cambridge University Press; 1998. Available from: https://books.google.com/books?id=Zanic8M0PjgC
9. Dembski W, Ewert W. *The design inference: eliminating chance through small probabilities*. Discovery Institute; 2023. Available from: https://books.google.com/books?id=kGsj0AEACAAJ
10. Montañez, George D. A unified model of complex specified information. *BIO-Complexity.* 2018;2018(4). doi: 10.5048/bio-c.2018.4
11. Marshall SM, Murray ARG, Cronin L. A probabilistic framework for identifying biosignatures using pathway complexity. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2017;375(2109): 20160342. doi: 10.1098/rsta.2016.0342
12. Marshall SM, Moore DG, Murray ARG, Walker SI, Cronin L. Formalising the pathways to life using assembly spaces. *Entropy (Basel, Switzerland)*. 2022;24(7): 884. doi: 10.3390/e24070884
13. Zenil H. *The 8 fallacies of assembly theory*. 2022. Available from: https://hectorzenil.medium.com/the-8-fallacies-of-assembly-theory-ba54428b0b45
14. Böttcher T. An additive definition of molecular complexity. *Journal of Chemical Information and Modeling*. 2016;56(3): 462–470. doi: 10.1021/acs.jcim.5b00723
15. Böttcher T. From molecules to life: quantifying the complexity of chemical and biological systems in the universe. *Journal of Molecular Evolution*. 2018;86(1): 1–10. doi: 10.1007/s00239-017-9824-6
16. Alon U. An introduction to systems biology: design principles of biological circuits. Boca Raton, Fla: Chapman and Hall/CRC; 2019. doi: 10.1201/9780429283321
17. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*. 1998;95(12): 6578–6583. doi: 10.1073/pnas.95.12.6578
18. Giancarlo R, Scaturro D, Utro F. Textual data compression in computational biology: a synopsis. *Bioinformatics (Oxford, England)*. 2009;25(13): 1575–1586. doi: 10.1093/bioinformatics/btp117
19. Schmitt AO, Herzel H. Estimating the entropy of DNA sequences. *Journal of Theoretical Biology*. 1997;188(3): 369–377. doi: 10.1006/jtbi.1997.0493
20. Adjeroh D, Nan F. On compressibility of protein sequences. In: *Proceedings of the Data Compression Conference (DCC'06)*, March 2006, Snowbird, Utah, USA; 2006.p.1–10.
21. Zenil H, Hernández-Orozco S, Kiani N, Soler-Toscano F, Rueda-Toicen A, Tegnér J. A decomposition method for global evaluation of Shannon entropy and local estimations of algorithmic complexity. *Entropy (Basel, Switzerland)*. 2018;20(8): 605. doi: 10.3390/e20080605
22. Al-Fatlawi A, Menzel M, Schroeder M. Is protein BLAST a thing of the past? *Nature Communications*. 2023;14(1): 8195. doi: 10.1038/s41467-023-44082-5
23. Tian P, Best RB. How many protein sequences fold to a given structure? A coevolutionary analysis. *Biophysical Journal*. 2017;113(8): 1719–1730. doi: 10.1016/j.bpj.2017.08.039
24. Copley SD. Evolution of new enzymes by gene duplication and divergence. *The FEBS Journal*. 2020;287(7): 1262–1283. doi: 10.1111/febs.15299
25. Glasner ME, Truong DP, Morse BC. How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation. *The FEBS Journal*. 2020;287(7): 1323–1342. doi: 10.1111/febs.15185
26. Marshall SM, Mathis C, Carrick E, Keenan G, Cooper GJT, Graham H, et al. Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nature Communications*. 2021;12(1): 3033. doi: 10.1038/s41467-021-23258-x
27. Jirasek M, Sharma A, Bame JR, Mehr SHM, Bell N, Marshall SM, et al. Investigating and quantifying molecular complexity using assembly theory and spectroscopy. *ACS Central Science*. 2024;10(5): 1054–1064. doi: 10.1021/acscentsci.4c00120
28. Liu Y, Mathis C, Bajczyk MD, Marshall SM, Wilbraham L, Cronin L. Exploring and mapping chemical space with molecular assembly trees. *Science Advances*. 2021;7(39): eabj2465. doi: 10.1126/sciadv.abj2465
29. EurekAlert A. New "assembly theory" unifies physics and biology to explain evolution and complexity. 2023. Available from: https://www.eurekalert.org/news-releases/1003613 [Accessed 14th October 2025].
30. Papkou A, Garcia-Pastor L, Escudero JA, Wagner A. A rugged yet easily navigable fitness landscape. *Science (New York, N.Y.)*. 2023;382(6673): eadh3860. doi: 10.1126/science.adh3860
31. Lewontin RC. The units of selection. *Annual Review of Ecology and Systematics*. 1970;1(1): 1–18. doi: 10.1146/annurev.es.01.110170.000245
32. Cronin L, Tour J. Dr. Lee Cronin vs Dr. James Tour debate the origin of life at Harvard Cambridge Faculty Roundtable. 2023. Available from: https://www.youtube.com/watch?v=6GDv4f2zUus
33. Benner SA. Paradoxes in the origin of life. *Origins of Life and Evolution of Biospheres*. 2014;44(4): 339–343. doi: 10.1007/s11084-014-9379-0
34. Fry I. The role of natural selection in the origin of life. *Origins of Life and Evolution of Biospheres*. 2010;41(1): 3–16. doi: 10.1007/s11084-010-9214-1
35. Freeland S. Undefining life's biochemistry: implications for abiogenesis. *Journal of The Royal Society Interface*. 2022;19(187). doi: 10.1098/rsif.2021.0814
36. Richert, Clemens. Prebiotic chemistry and human intervention. *Nature Communications.* 2018;9(1): ISSN: 2041-1723. doi: 10.1038/s41467-018-07219-5

37. Bartlett JL. Measuring active information in biological systems. *BIO-Complexity*. 2020;2020(2). doi: 10.5048/BIO-C.2020.2
38. Cronin L, Walker SI. Beyond prebiotic chemistry. *Science (New York, N.Y.)*. 2016;352(6290): 1174–1175. doi: 10.1126/science.aaf6310
39. Walker, Sara Imari and Davies, Paul C. W. The algorithmic origins of life. *Journal of The Royal Society Interface*. 2013;10(79): 20120869. doi: 10.1098/rsif.2012.0869
40. Walker, Sara Imari and Davies, Paul C. W. The “Hard Problem” of Life. *From Matter to Life*, Cambridge University Press. 2017; 19–37. doi 10.1017/9781316584200.002
41. Ellis GFR. Top-down causation and emergence: some comments on mechanisms. *Interface Focus*. 2012;2(1): 126–140. doi: 10.1098/rsfs.2011.0062
42. Steiner S, Wolf J, Glatzel S, Andreou A, Granda JM, Keenan G, et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science (New York, N.Y.)*. 2019;363(6423): eaav2211. doi: 10.1126/science.aav2211
43. Šiaučiulis M, Knittl-Frank C, Mehr SH, Clarke E, Cronin L. Reaction blueprints and logical control flow for parallelized chiral synthesis in the chemputer. *Nature Communications*. 2024;15(1): 10261. doi: 10.1038/s41467-024-54238-6
44. Chalmers D. Facing up to the problem of consciousness. *Journal of Consciousness Studies*. 1995;2(3): 200–219. doi: 10.1093/acprof:oso/9780195311105.003.0001
45. Meyer SC. Signature in the cell: DNA and the evidence for intelligent design. HarperOne; 2010.
46. Voie OA. Biological function and the genetic code are interdependent. *Chaos, Solitons and Fractals*. 2006;28(4): 1000–1004. doi: 10.1016/j.chaos.2005.08.146

Appendix

Computing a bound on $N(a_i)$ for proteins

In Section 'Assembly index as measure of complexity', we estimate the probability bound on a random process generating a linear-chain object composed of β building blocks with assembly index a_i. To do this we assume that $N(a_i) \geq \beta^{a_i}$. We motivate this assumption here.

Let S_a be the set of objects with assembly index a and T_ℓ be the set of objects with length ℓ. By definition $|T_\ell| = \beta^\ell$. Our goal will be to show that the size of $|S_a|$ is at least as large as the size of $T_{\ell=a}$. This will imply that $N_a = |S_a| \geq T_{\ell=a} \geq \beta^a$.

First we characterize the objects in S_a. Recall from Section 'Assembly theory' that S_a contains objects with length ℓ satisfying $\log_2 \ell \leq a < \ell$. This means that the smallest object in S_a has length $\ell = a + 1$.

Thus $T_{\ell=a+1}$, the set of all objects with length $\ell = a + 1$, must be distributed across the sets S_x where $x \leq a$. This is simply the statement that objects with $\ell = a + 1$ must have assembly index $\leq a$.

Some of these objects will have assembly index a and be in S_a itself, but some will have smaller assembly indices $< a$. Assume there is an object $X \in T_{\ell=a+1}$ that is not in S_a. Such an object would have to have an assembly index $x < a$ and must, therefore, be in some S_x with $x < a$.

X would be available for subsequent assembly. Which means there are objects with assembly index $>x$ with construction trajectories that incorporate X. In particular, if X is combined with an object Y with assembly index $y = (a-x-1)$ then it will form an object $Z = X + Y$ with $Z \in S_a$ since $y + x + 1 = a$.

Unfortunately, not just any partner object Y with $y = a - x - 1$ will do. It is possible – because of the awkward definition of the assembly index as the size of the *minimal* length construction trajectory – that a more optimal construction of the resulting object, Z, exists. This more optimal construction would have assembly index $z < a$. However, we can then combine Z with yet another object Q with assembly $q = a - z - 1$ to find an object that *could* be in S_a. We can repeat this process until we find an object in S_a.

Thus we see that every object in $T_{\ell=a+1}$ is either in S_a or is part of a construction trajectory that results in some other larger sized object in S_a. Thus $|S_a| \geq |T_{\ell=a+1}| = \beta^{a+1} > \beta^a$ (since $\beta > 1$). In practice, $|S_a| \gg |T_{\ell=a+1}|$ for large a as many objects with $\log_2 \ell \leq a < \ell$ are mapped into S_a.

In Section 'Assembly index as measure of complexity', we use this size bound to estimate the probability bound to be $P \leq 1/\beta^a$. Implicit in this assumption is the sense that only one of the objects in S_a corresponds to the object of interest. However, to rule out an undirected process generating this object, we need to use a probability bound that accounts for what an undirected process is actually capable of. In reality, there could be several different construction paths to construct the object. Indeed we expect there are more paths to assemble highly compressible objects with many repeating blocks than there are paths to assemble less compressible objects. This complexity will tend to inflate both the numerator and denominator for our probability measure in different ways depending on the object in question. In the interest of obtaining an order-of-magnitude estimate, however, we conjecture that $P \leq 1/\beta^a$ is a reasonably conservative upper bound on the probabilities in question.